Graph Kernels in Gaussian Process Surrogates for Bayesian Optimization in AutoML

Generated Literature Survey

December 29, 2025

Contents

1 Introduction

Graph kernels play a pivotal role in extending Gaussian process (GP) surrogates for Bayesian optimization (BO) to graph-structured inputs, enabling the modeling of similarities in combinatorial search spaces that underpin AutoML tasks like neural architecture search (NAS). By integrating these advanced kernels with GP regression, traditional limitations of Euclidean kernels in non-vector spaces are overcome, facilitating efficient surrogate modeling over discrete and structured domains. This section defines core concepts, elucidates their significance in AutoML, and surveys recent advances, setting the stage for detailed discussions on kernel design, theoretical foundations, and empirical applications.

1.1 Graph Kernels for GP Surrogates in Graph BO

Graph kernels play a pivotal role in extending Gaussian process (GP) surrogates for Bayesian optimization (BO) to graph-structured inputs, enabling the modeling of similarities in combinatorial search spaces prevalent in AutoML tasks such as neural architecture search (NAS) [1, 2]. Traditional GP-based BO methods, which rely on kernels for scalar or vector inputs, face challenges with graph data due to their discrete topology and variable structure; however, graph kernels address this by quantifying pairwise similarities between graphs, supporting uncertainty-aware surrogates essential for sample-efficient optimization [1]. For instance, existing approaches predominantly employ GPs paired with graph kernels to approximate black-box functions over graphs, facilitating applications like NAS where architectures are inherently represented as graphs [1].

In the context of AutoML, NAS leverages BO with GP surrogates to predict architecture performance from validation results, guiding the search for optimal configurations [2]. Yet, GP inference scales cubically with observations and struggles with variable-length neural architectures, prompting the use of specialized kernels—such as phenotypic distance kernels or those for conditional parameter spaces—to better capture architectural similarities [2]. Recent advancements, including shortest-path graph kernels, formulate GP acquisition functions as mixed-integer programs, enabling feasible exploration of graph domains while accommodating mixed features and constraints, thus enhancing BO efficacy for NAS and related tasks [1].

Pre-trained GP priors further bolster these frameworks by incorporating domain knowledge from prior tasks, potentially integrating structured kernels like graph variants to improve transfer in high-dimensional AutoML settings [3]. This foundation underscores the scope of subsequent reviews on graph kernel innovations in GP-driven BO.

1.2 GP Surrogates with Advanced Kernels for Non-Vector AutoML

The integration of Gaussian process (GP) surrogates with advanced kernels, particularly graph kernels, effectively addresses the limitations of Euclidean kernels in modeling non-vector spaces, such as graph-structured inputs prevalent in AutoML tasks like neural architecture search (NAS) [4]. Traditional GP models rely on kernels suited for scalar or vector inputs, which fail to capture the discrete topology and variable structures of graphs; in contrast, graph kernels quantify pairwise similarities directly over graph domains, enabling accurate prediction and uncertainty quantification in Bayesian optimization (BO) [4]. For instance, these kernels facilitate generalization from non-structural spaces—implicitly Euclidean vector representations—to complex graph spaces, as demonstrated by implementations distinguishing vectorized embeddings (e.g., adjacency matrices) from direct graph modeling [4].

Advanced graph kernels, such as linear and exponential shortest-path (SP and ESP) forms alongside Weisfeiler-Lehman (WL) variants, further enhance GP surrogates by embedding graph properties like shortest paths and reachability into kernel computations, implemented via libraries like GPflow [4]. Experimental comparisons on NAS benchmarks reveal strong predictive performance for these kernels: ESP achieves the lowest RMSE (0.11) and highest Spearman correlation (0.93) on NAS-Bench-101, outperforming simpler random walk (RW) and base SP kernels, while WL also excels on converted node-labeled graphs [4].

This approach not only supports sample-efficient BO in AutoML but also paves the way for global optimization of graph acquisition functions, mitigating scalability issues in non-Euclidean domains [4].

2 Thematic or Methodological Landscape

This section maps the thematic and methodological landscape of recent literature on graph kernels integrated into Gaussian process (GP) surrogates for Bayesian optimization (BO) over discrete graph spaces, categorizing prominent types such as path-based, Weisfeiler-Lehman (WL), and diffusion/heat kernels. These kernels enable effective encoding of graph similarities, uncertainty quantification, and optimization in domains like combinatorial problems and neural architecture search (NAS). The following subsections examine shortest-path and WL kernels, diffusion-based extensions, and their applications in AutoML.

2.1 Shortest-Path and Weisfeiler-Lehman Kernels for Graph BO

Shortest-path (SP) and Weisfeiler-Lehman (WL) kernels stand out as prominent mechanisms for encoding graph similarities within Gaussian process (GP) surrogates for Bayesian optimization (BO) over discrete graph spaces, owing to their ability to capture structural properties like paths and neighborhood features [1, 4]. These kernels enable direct similarity computation between graphs, addressing the limitations of vector-based methods in combinatorial domains such as molecular design and neural architecture search (NAS). In BoGrape, four SP kernel variants—simplified SP (SSP), SP, exponential SSP (ESSP), and exponential SP (ESP)—are developed for attributed graphs, with SP and ESP demonstrating superior predictive performance on larger graphs by imposing stricter criteria on path comparisons, as evidenced by lower mean negative log likelihoods (MNLLs) on QM7 and QM9 datasets [1]. Implemented in GPflow, these kernels support mixed-integer programming (MIP) formulations for global acquisition optimization, facilitating feasible exploration under constraints [1].

NAS-GOAT further advances SP kernels for NAS, formulating linear and exponential variants as MIPs to optimize acquisition functions like lower confidence bound (LCB) over graph-encoded spaces, handling node/edge labels and variable topologies [4]. Experimental benchmarks on NAS-Bench-101 and NAS-Bench-201 reveal ESP's strong generalization, achieving the lowest RMSE (0.11) and highest Spearman correlation (0.93) on NAS-Bench-101, outperforming random walk (RW) kernels [4]. WL kernels exhibit robust performance when applied to node-labeled graph conversions of NAS architectures (e.g., RMSE 0.23, Spearman 0.81 on NAS-Bench-201), significantly outperforming WL on original edge-labeled graphs (WL-e: RMSE 0.37, Spearman 0.11) [4].

These kernels' prominence stems from their balance of expressiveness and tractability in GP-based BO, with SP variants enabling precise path-based encoding and WL providing effective stabilization on adapted graph representations [1, 4].

2.2 Graph Kernels Extend GPs for Combinatorial BO

Diffusion and heat kernels, often termed diffusion kernels, alongside Matérn variants on graphs, extend Gaussian process (GP) surrogates to combinatorial Bayesian optimization (BO) over graph-structured domains, providing principled covariance measures that enhance uncertainty quantification [5]. Heat kernels, derived from the heat equation on graphs, serve as discrete analogs to radial basis function (RBF) kernels in Euclidean spaces, enabling direct modeling of similarities within graph topologies rather than between separate graphs [5]. A unifying framework demonstrates that prominent combinatorial kernels, such as those in CASMOPOLITAN and COMBO, are equivalent to heat kernels on Hamming graphs, simplifying computation via analytical eigendecompositions of graph Laplacians and yielding automatic relevance determination (ARD) forms proportional to RBF after one-hot encoding [5].

Matérn Gaussian processes on graphs further advance this paradigm by incorporating smoothness priors tailored to discrete graph structures, as evidenced by their repeated invocation in analyses of heat kernel generalizations [5]. These constructions address challenges in combinatorial spaces like neural architecture search (NAS), where graphs exhibit variable node counts and labels; adaptations involve padding to uniform sizes and grouping vertices by type, facilitating permutation-invariant heat kernel applications [5].

Empirically, heat and Matérn kernels improve surrogate accuracy in combinatorial BO, yielding faster implementations (e.g., $O(n)$ versus $O(\sum |X_i|^3)$ for COMBO) and competitive performance on graph benchmarks, including NAS, through superior covariance estimation that bolsters uncertainty-aware acquisition [5]. This enhances sample efficiency and exploration in non-Euclidean graph domains compared to prior path-based or Weisfeiler-Lehman approaches.

2.3 Graph Kernels for GP BO in AutoML/NAS

Applications in AutoML, particularly neural architecture search (NAS), increasingly leverage graph kernels within Gaussian process (GP) surrogates for Bayesian optimization (BO) to navigate architecture spaces represented as directed acyclic graphs (DAGs) [2, 4]. NAS, a prominent AutoML subfield, automates the discovery of high-performing neural architectures by selecting operations from predefined search spaces, where GP-based BO models validation performance to guide optimization; however, standard GPs struggle with cubic scaling and variable-length architectures [2]. Graph kernels address these by directly measuring similarities over graph-structured inputs, enabling GP BO to handle the discrete topology of NAS search spaces [4].

NAS-GOAT exemplifies this approach, employing shortest-path (SP) graph kernels—linear and exponential variants—for GP surrogates in graph BO tailored to NAS [4]. Neural architectures are encoded as weakly connected DAGs, with kernels capturing properties like node/edge labels and paths; the lower confidence bound (LCB) acquisition function is globally optimized via mixed-integer programming (MIP) formulations, restricting to valid architectures [4]. This contrasts with baselines relying on mutation or sampling, such as NASBOT (GP over graphs), by enabling exact optimization over graph domains.

Empirical results on NAS-Bench-101 and NAS-Bench-201 demonstrate the efficacy of SP kernels in GP BO, with exponential SP (ESP) achieving competitive or superior performance against state-of-the-art methods like NAS-BOWL, even under noisy validation settings [4]. By overcoming limitations of vectorized embeddings and scalability issues noted in early GP applications to NAS, graph kernel-enabled BO enhances sample efficiency for architecture optimization [2, 4].

3 Synthesis & Critical Discussion

In this section, we synthesize findings across graph kernels and their applications in Bayesian optimization (BO), critically comparing empirical performance, theoretical guarantees, and key challenges such as scalability and kernel expressiveness. Path-based kernels (e.g., shortest-path and exponential shortest-path) and Weisfeiler-Lehman kernels outperform traditional random walk kernels in graph BO for neural architecture search, yet struggle with computational demands on large graphs. Meanwhile, heat and Matérn kernels bolster uncertainty quantification in combinatorial BO through tailored covariance structures, though they necessitate careful hyperparameter tuning akin to general Gaussian process challenges.

3.1 Path-based and WL Kernels: Gains vs. Scalability Issues

Path-based kernels, including shortest-path (SP) and exponential SP (ESP), alongside Weisfeiler-Lehman (WL) kernels, demonstrate superior predictive performance over traditional random walk (RW) kernels in graph Bayesian optimization (BO) for neural architecture search (NAS) [4]. On NAS-Bench-101, ESP achieves the lowest root mean squared error (RMSE: 0.11) and highest Spearman correlation (0.93), outperforming RW (RMSE: 0.29, Spearman: 0.81) and linear SP (RMSE: 0.21, Spearman: 0.83), while WL yields competitive results (RMSE: 0.15, Spearman: 0.87) [4]. Similarly, on NAS-Bench-201, WL on converted node-labeled graphs excels (RMSE: 0.23, Spearman: 0.81) relative to RW (RMSE: 0.32, Spearman: 0.78) and far surpasses WL on original edge-labeled graphs (RMSE: 0.37, Spearman: 0.11), with ESP also strong (RMSE: 0.30, Spearman: 0.64) [4]. These gains stem from SP kernels' stricter path comparisons and WL's effective neighborhood encoding, enhancing uncertainty quantification via lower mean negative log likelihoods (MNLLs) [4].

This outperformance extends to general graph BO tasks, where SP and ESP kernels excel on larger graphs by better capturing structural properties, as evidenced by lower MNLLs on QM7/QM9 datasets with node counts up to 30 [1]. However, these advanced kernels introduce computational scalability challenges, particularly for large graphs, due to their complex formulations [1, 4]. Exponential variants like ESP yield more intricate mixed-integer programs (MIPs) for acquisition optimization, demanding greater resources and longer runtimes, though they offer improved representation at the cost of query efficiency [1]. Simpler path kernels (e.g., SSP) sometimes outperform in BO under time limits by reducing MIP complexity, highlighting trade-offs in expressive kernels for expansive graph domains [1].

3.2 Heat and Matérn Kernels: Uncertainty Gains, Tuning Challenges

Heat and Matérn graph kernels enhance uncertainty quantification in combinatorial Bayesian optimization (BO) by providing principled covariance structures tailored to discrete graph domains, outperforming certain baselines on benchmarks including neural architecture search [5]. These kernels, including heat kernels derived from graph Laplacians and Matérn constructions on graphs, unify prominent combinatorial kernels (e.g., those in CASMOPOLITAN, COMBO, and Bounce) as equivalents to RBF or Matérn after one-hot encoding, yielding near-identical empirical performance across variants [5]. Notably, heat kernels demonstrate robustness to relocated optima—unlike spectrum-based kernels (e.g., SSK in BOSS)—maintaining state-of-the-art results on tasks like Pest Control, LABS, and Cluster Expansion, where superior covariance estimation supports effective exploration-exploitation trade-offs [5]. Graph Matérn variants further extend this by incorporating smoothness priors on graph topologies, contributing to competitive BO outcomes in permutation-invariant settings [5].

Despite these advances, heat and Matérn kernels necessitate careful hyperparameter tuning, such as lengthscales in automatic relevance determination (ARD) forms or diffusion parameters $_i$ in heat kernel expressions, to adapt to varying graph cardinalities and structures [5]. ARD-enabled variants, proportional to RBF post-encoding, require per-dimension optimization to assess input relevance, with $_i$ parameters bounded by graph sizes $g_i = |X_i|$ [5].

This tuning sensitivity mirrors longstanding challenges in general GP-based BO, where maximum likelihood estimation of kernel hyperparameters often underperforms due to data scarcity and non-representative function assumptions, prompting alternatives like Wasserstein barycenters of multiple fixed-hyperparameter GPs [6]. Such issues underscore the need for robust hyperparameter strategies to fully leverage kernel improvements in combinatorial graph BO.

4 Conclusion

In this concluding section, we synthesize the key advancements in graph kernel-enhanced Gaussian process (GP) surrogates for Bayesian optimization (BO), emphasizing their critical role in enabling effective surrogates over non-Euclidean domains. Recent works have solidified graph kernels as indispensable tools, particularly for advancing neural architecture search (NAS) in AutoML applications, demonstrating the field's emerging maturity. The following subsections recap these contributions and outline promising directions ahead.

4.1 Graph Kernels for Non-Euclidean BO in NAS

Recent advancements in graph kernel design have solidified their role as essential components for Gaussian process (GP) surrogates in Bayesian optimization (BO) over non-Euclidean domains, particularly graphs, enabling effective navigation of combinatorial spaces in AutoML applications like neural architecture search (NAS) [1, 2]. BoGrape exemplifies this progress by introducing shortest-path graph kernels—such as simplified SP (SSP), SP, exponential SSP (ESSP), and exponential SP (ESP)—tailored for GP surrogates on attributed graphs, which quantify structural similarities via path comparisons and support mixed-integer programming (MIP) formulations for global acquisition optimization [1]. These kernels address prior limitations in graph BO, where traditional methods relied on evolutionary algorithms or sampling, by facilitating feasible exploration under constraints and demonstrating superior predictive performance on larger graphs, as measured by lower mean negative log likelihoods (MNLLs) [1].

In NAS, an AutoML cornerstone, GP-based BO struggles with cubic inference scaling and variable-length architectures modeled as graphs; graph kernels mitigate this by directly encoding topological similarities, as noted in comprehensive surveys, with specialized variants like phenotypic distance or conditional parameter space kernels enhancing surrogate efficacy [2]. BoGrape's framework aligns with these needs, explicitly positioning shortest-path kernels for adaptation to NAS alongside molecular design, building on established BO applications in architecture optimization [1, 2].

Pre-trained GP priors further amplify graph kernel potential by incorporating domain knowledge from prior tasks, potentially integrating structured kernels to boost transfer learning in high-dimensional AutoML settings [3]. Collectively, these developments mark the field's maturity, transforming graph BO into a sample-efficient tool for non-Euclidean AutoML challenges.

5 Future Directions

While significant advances have been made in graph kernels for Gaussian process surrogates in Bayesian optimization (BO) within AutoML—particularly for neural architecture search (NAS)—key challenges persist in scalability, expressiveness, and theoretical guarantees. This section identifies critical gaps, including scalable approximations for expressive kernels like shortest-path, exponential shortest-path, and Weisfeiler-Lehman variants on massive graphs, hybrid deep-graph kernels integrated with neural processes, and regret bounds under kernel misspecification. We further explore promising extensions to domains beyond NAS, such as molecular design, outlining avenues for future research.

5.1 Scalable Graph Kernels and Neural Processes for AutoML

While expressive graph kernels such as shortest-path (SP), exponential SP (ESP), and Weisfeiler-Lehman (WL) variants enable Gaussian process surrogates for Bayesian optimization (BO) in neural architecture search (NAS)—a core AutoML task—optimizing acquisition functions over combinatorial graph spaces remains challenging, prompting reliance on sample-based or evolutionary algorithms that require only function evaluations [4]. These methods, including mixed-integer programming (MIP) formulations solved with time limits (e.g., 1800s per MIP in Gurobi), facilitate global optimization but highlight efficiency limitations, particularly for variable graph sizes that complicate kernel normalizations [4].

Emerging approximations provide initial steps toward scalability, as seen in Monte Carlo sampling for permutation-invariant heat kernels, which approximates sums over graph automorphisms using 200 random permutations to match or exceed WL performance on NAS-Bench-101 and NAS-Bench-201 benchmarks [5]. Such techniques support faster implementations in combinatorial BO, yet scalable approximations for more discriminative kernels like SP or WL—crucial for nuanced graph similarities—remain underdeveloped [5, 4].

These gaps are amplified in full-pipeline AutoML, where inter-step dependencies (e.g., dataset choices influencing preprocessing and architecture optimization) enlarge search spaces across diverse tasks [7]. Developing scalable approximations for expressive graph kernels will thus be essential for enabling large-scale graph BO in comprehensive AutoML frameworks [4, 5, 7].

5.2 Misspecified Graph Kernels: Regret and Beyond-NAS Applications

Theoretical analyses of regret bounds under kernel misspecification remain limited in Gaussian process (GP) Bayesian optimization (BO), particularly for specialized kernels like those adapted to graph-structured inputs, underscoring a key gap for applications in neural architecture search (NAS) and beyond, such as molecular design. Existing theoretical results, including improved cumulative regret bounds of $\mathcal{O}(\sqrt{T \log^d T/2})$ for GP-UCB with squared exponential kernels and near-optimal simple regret rates $\mathcal{O}(t^{-\nu/(d+\epsilon)})$ for Matérn kernels under noise-free observations, explicitly assume a well-specified setting where the objective function resides in the reproducing kernel Hilbert space (RKHS) defined by the chosen kernel [8]. These bounds rely on precise posterior uncertainty quantification and RKHS norm constraints, yet they do not extend to misspecified scenarios prevalent in complex domains.

In practice, kernel misspecification—where the true function falls outside the assumed RKHS due to unknown hyperparameters—is commonplace, causing GP confidence bounds to converge to suboptimal local optima via maximum likelihood estimation [6]. Heuristic remedies, such as iteratively decreasing the kernel length-scale to expand the RKHS and boost exploration, incur higher regret despite outperforming standard GP lower confidence bound (GP-LCB) methods in misspecified real-world settings [6]. Such challenges amplify for graph kernels, which encode non-Euclidean similarities in discrete spaces like NAS architectures or molecular graphs, yet lack tailored regret analyses under misspecification.

Extending regret bounds to graph kernel misspecification would clarify convergence guarantees in these settings, informing robust acquisition strategies for BO over graphs. Given the strict well-specified assumptions in current theory [8] and the ubiquity of misspecification in practical applications [6], rigorous analyses for graph-structured domains warrant further study.

References

[1] Yilin Xie et al. “BoGrape: Bayesian optimization over graphs with shortest-path encoded”. In: (2025). arXiv: 2503.05642 [cs]. URL: https://arxiv.org/abs/2503.05642.

[2] Xin He, Kaiyong Zhao, and Xiaowen Chu. “AutoML: A Survey of the State-of-the-Art”. In: (2019). arXiv: 1908.00709 [cs]. URL: https://arxiv.org/abs/1908.00709.

[3] Zi Wang et al. “Pre-trained Gaussian Processes for Bayesian Optimization”. In: (2021). arXiv: 2109.08215 [cs]. URL: https://arxiv.org/abs/2109.08215.

[4] Yilin Xie et al. “Global optimization of graph acquisition functions for neural architecture search”. In: (2025). arXiv: 2505.23640 [cs]. URL: https://arxiv.org/abs/2505.23640.

[5] Colin Doumont et al. “Heat Kernels in Combinatorial Bayesian Optimization”. In: (2025). arXiv: 2510.26633 [cs]. URL: https://arxiv.org/abs/2510.26633.

[6] Antonio Candelieri, Andrea Ponti, and Francesco Archetti. “Wasserstein Barycenter Gaussian Process based Bayesian Optimization”. In: (2025). arXiv: 2505.12471 [cs]. URL: https://arxiv.org/abs/2505.12471.

[7] Patara Trirat, Wonyong Jeong, and Sung Ju Hwang. “AutoML-Agent: A Multi-Agent LLM Framework for Full-Pipeline AutoML”. In: (2024). arXiv: 2410.02958 [cs]. URL: https://arxiv.org/abs/2410.02958.

[8] Hwanwoo Kim and Daniel Sanz-Alonso. “Enhancing Gaussian Process Surrogates for Optimization and Posterior Approximation”. In: (2024). arXiv: 2401.17037 [cs]. URL: https://arxiv.org/abs/2401.17037.