Surprisingly Strong Performance Prediction with Neural Graph Features

Gabriela Kadlecová [1 2] **Jovita Lukasik** [3] **Martin Pilát** [1] **Petra Vidnerová** [2] **Mahmoud Safari** [4] **Roman Neruda** [2] **Frank Hutter** [4 5]

Abstract

Performance prediction has been a key part of the neural architecture search (NAS) process, allowing to speed up NAS algorithms by avoiding resource-consuming network training. Although many performance predictors correlate well with ground truth performance, they require training data in the form of trained networks. Recently, zero-cost proxies have been proposed as an efficient method to estimate network performance without any training. However, they are still poorly understood, exhibit biases with network properties, and their performance is limited. Inspired by the drawbacks of zero-cost proxies, we propose neural graph features (GRAF), simple to compute properties of architectural graphs. GRAF offers fast and interpretable performance prediction while outperforming zero-cost proxies and other common encodings. In combination with other zero-cost proxies, GRAF outperforms most existing performance predictors at a fraction of the cost.

1. Introduction

With the rising popularity of deep learning with applications across different domains, finding a well-performing neural network architecture for a given application has become a key problem. The field of Neural Architecture Search (NAS) automatizes the process and has gained noticeable attention in the past years (Elsken et al., 2019; White et al., 2023). Due to the high costs of neural network training, speedup techniques are a key component of the NAS process. One of these techniques is based on performance predictors – models that help us to reduce the number of trained architectures by predicting their performance based on a small number of sampled and trained networks (White et al., 2021b). Since these prediction models still require some networks to be trained and in some cases even include an overhead when fitting on the sampled train set, other, even more efficient techniques were needed. As an answer, *zero-cost proxies* have been recently proposed as a variant that does not require any network training (Abdelfattah et al., 2021; Tanaka et al., 2020; Mellor et al., 2021). For every network, these proxies need mostly a single mini-batch of data to compute a score that correlates with the true performance of the network after full training on the downstream task of interest.

However, the reason for the correlation is still underexplored. Furthermore, Krishnakumar et al. (2022) even discovered biases between these proxies and network properties like the number of skip-connections. On the large number of various tasks that a NAS method should be able to solve nowadays (Duan et al., 2021), zero-cost proxies show inconsistent performances and most of these proxies even have a lower average correlation with the network performance than simple metrics, such as `flops` (White et al., 2022; Krishnakumar et al., 2022). In addition, most performance prediction methods lack interpretability. This is in particular an important property that would eventually allow us to determine what architectural properties drive the performance for a given task.

To address these limitations of zero-cost proxies, we first examine in an in-depth analysis the reason behind the zero-cost proxy biases, showing that it is in some cases a direct dependency rather than an actual correlation. Inspired by that, we take advantage of these biases and propose in this paper *neural graph features* – simple to compute network properties like operation counts or node degrees. We show that using the proposed neural graph features as an input to several (simple) prediction models provides strong and interpretable performance prediction on several tasks, ranging from the common accuracy prediction to hardware metrics prediction and lastly to (multi-objective) robustness prediction.

Our main contributions are summarized as follows:

1. We examine biases in zero-cost proxies (ZCP) and

[1]Charles University, Faculty of Mathematics and Physics [2]The Czech Academy of Sciences, Institute of Computer Science [3]University of Siegen [4]University of Freiburg [5]ELLIS Institute Tübingen. Correspondence to: Gabriela Kadlecová <gabi.kadlecova@outlook.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

show that some of them directly depend on the number of convolutions. We also demonstrate that they are poor at distinguishing structurally similar networks.

2. We propose *neural graph features* (GRAF), interpretable features that outperform ZCP in accuracy prediction tasks using a random forest, and yield an even better performance when combined with ZCP.

3. Using GRAF's interpretability, we demonstrate that different tasks favor diverse network properties.

4. We evaluate GRAF on tasks beyond accuracy prediction, and compare with different encodings and predictors. The combination of using ZCP and GRAF as prediction input outperforms most existing methods at a fraction of the compute.

We make the code of our contributions available publicly[1]

2. Related work

Recent works have proposed different kinds of performance predictors, with the most popular being model-based methods, learning curve extrapolation methods, zero-cost methods, and weight-sharing methods. Using the library NASLib, a wide variety of predictors were evaluated with unified fit and query time settings across common benchmarks, as well as in search settings (White et al., 2021b). In this work, we focus on model-based predictors and zero-cost proxies, as they are directly comparable to our method.

Model-based predictors include tree-based models like random forest (Liaw & Wiener, 2002), or graph neural networks like BRP-NAS (Dudziak et al., 2020). An extended overview of model-based predictors is provided in the appendix (Section B).

Zero-cost proxies can be used as direct predictors of the target metric, or also as input to model-based predictors (Abdelfattah et al., 2021; Krishnakumar et al., 2022). Notably, Multi-Predict has used zero-cost proxies with an MLP accuracy predictor that enables transfer to other search spaces (Akhauri & Abdelfattah, 2023). Another work used a random forest predictor and zero-cost proxies to estimate robust accuracy of NAS-Bench201 (Jung et al., 2023; Lukasik et al., 2023).

As for interpretable predictors, NASBOWL enables the extraction of so-called motifs – network subgraphs that are found in high or low-performing networks (Ru et al., 2021). Other non-predictor works have derived a dependence of network performance on depth (Chen et al., 2023), or that limiting search spaces to a specific subset of operations entails high-performing networks (Lopes et al., 2023).

[1] https://github.com/gabikadlecova/zc_combine

Many of the well-performing predictors take several minutes to train for larger sample sizes – a limitation in budget-restricted search settings, as it introduces a computational overhead. Another drawback is, with the notable exception of NASBOWL, a lack of interpretability. Even NASBOWL does not consider properties like network depth or operation count. In our work, we discover that these properties are highly influential on the prediction quality.

The drawback of zero-cost proxies is that they often need access to training data, and in some settings, this might be impossible, e.g. due to privacy reasons. Most proxies are computed by providing a mini-batch to networks, thus requiring us to compile the whole model. For very large architectures, or for networks with proprietary training routines, it might be advantageous to work only with encodings without any overhead. Additionally, although they are labeled zero-cost, some of them can take minutes per architecture to compute (Krishnakumar et al., 2022). The advantage of our proposed graph features is that they are completely data-independent.

3. Introducing Neural Graph Features

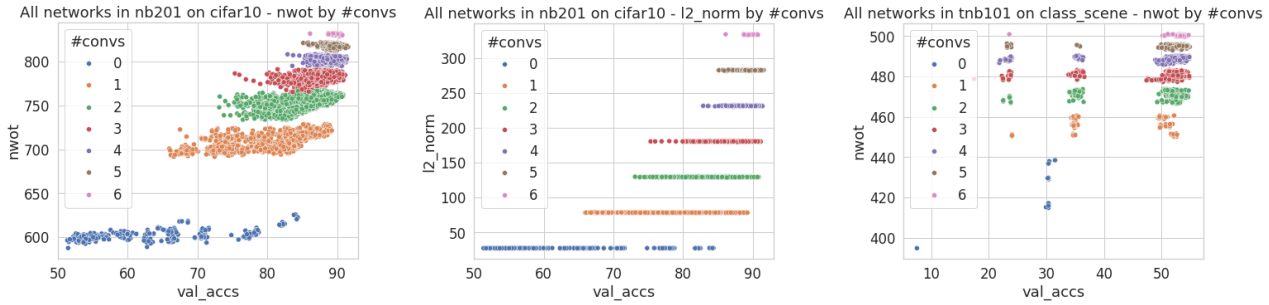
In our work, we will use precomputed zero-cost proxy scores from NAS-Bench-Suite-Zero (Krishnakumar et al., 2022). Refer to Section C.3 in the appendix for more details about the provided zero-cost proxies, and Section C.1 for the used benchmarks, datasets, and abbreviations. For NB101 and NB301, zero-cost proxies were computed only for a fraction of the search space, and all subsequent experiments are evaluated on these samples. For NB201 and TNB101-micro, we also exclude networks with unreachable branches due to zero operations. Detailed information about the process is in Section C.2 in the appendix.

To motivate our approach, we first examine the limitation of many zero-cost proxies, namely their correlation with the layer count. Based on these insights, we then propose our simple-to-compute graph properties which we name neural graph features (GRAF).

3.1. Limitations of zero-cost proxies

As shown in NB-Suite-Zero, many existing zero-cost proxies have a good correlation with validation accuracy on NB201 and `cifar-10`. In this section, we investigate the reason for the good results.

From all ZCP available in NB-Suite-Zero, `nwot` has the best spearman correlation with validation accuracy. However, there is a hidden property of the proxy that brings about the good correlation. Figure 1a shows the `nwot` score of all NB201 networks plotted against validation accuracy, where the color of a point indicates the number of `conv1x1` plus `conv3x3` (i.e. all convolutions) in the architecture. We can



(a) `nwot` against `cifar-10` val. acc. (b) `l2_norm` against `cifar-10` val. acc. (c) `nwot` against `class_scene` val. acc.

Figure 1. ZCP score of all networks from NB201 against the validation accuracy - colors indicate the number of conv3x3 and 1x1

see that each cluster corresponds to a specific number of convolutions, and the `nwot` score increases with the number of convolutions. Figure 1b depicts the same behavior for `l2_norm`, where the score is constant for every cluster. Similar results were shown with NB-Suite-Zero, however, they discovered only the correlation with the number of convolutions, not direct dependence.

In Figure 2a, we can see that although most proxies have a good correlation over the whole search space, many have trouble distinguishing networks with the same number of convolutions, and most have trouble distinguishing networks with a high number of convolutions. For example, the row with `jacov` has a correlation equal to 0.8 for the cluster with 1 convolution, but in the cluster with 5 convolutions, the correlation is 0.

The same property applies for `cifar-100` and `ImageNet16-120` (Figure 12 in the appendix). Contrary to NB201, the proxies had a fairly low correlation with TNB101-micro tasks. Figure 1c explains why – unlike `cifar-10`, the task performance does not correlate with the number of convolutions, and `nwot` does not capture some property that would be important for the task. Figure 2 also shows, that for `class_scene`, the proxies can distinguish networks with the same number of convolutions. However, for clusters with the same number of `conv3x3`, the proxies have a negative correlation with validation accuracy. On NB101 and NB301, the dependence on the number of convolutions is not direct, but there are other dependencies (Section D in the appendix).

To summarize, zero-cost proxies capture the number of convolutions, an important property for NB201 tasks like `cifar-10` classification, but most of them cannot distinguish structurally similar networks – they lack some other important properties. These findings have inspired us to examine the potential of using operation counts and other graph properties as input for performance predictors.

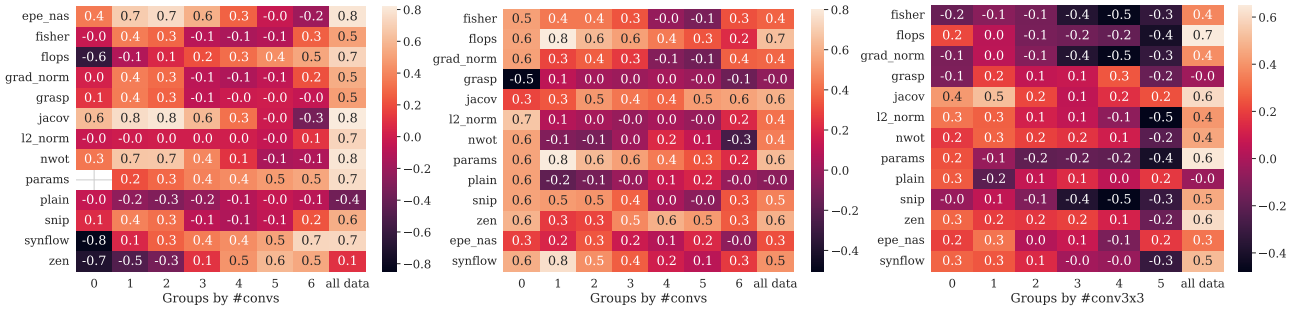
3.2. Neural graph features

In this section, we describe the *neural graph features* (GRAF). Given operation set O and an architecture graph $G = (L, V, E)$, where V is the vertex set, E is the edge set, and L is the set of labels (associated with edges or vertices based on the search space type, see Section C.1), we define the following features:

- Number of times the operation $o \in O$ is used in L
- Minimum path length from the input node to output node going only over operations $O' \subseteq O$
- Maximum path length from the input node to the output node going only over operations $O' \subseteq O$
- Output degree of the input node counting only operations $O' \subseteq O$
- Input degree of the output node counting only operations $O' \subseteq O$
- Mean input/output degree of intermediate nodes counting only operations $O' \subseteq O$

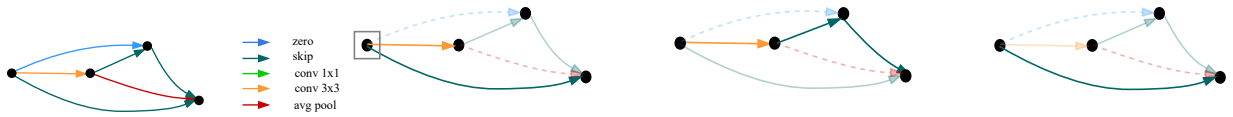
These features are computed for all possible subsets $O' \subseteq O$. They are the same for all cell-based benchmarks with one exception – for NB301, min path length is not computed, since all nodes are directly connected to the output. Figure 3 visualizes three different example features computed for an architecture. For node degree (3b), the computed value is 2, since we count the conv3x3 and skip edges going from the input node, but not the zero edge. For the first max path length (3c), the computed value is 3, as we can take a long way over 1 skip edge and 2 conv3x3 edges. However, for the second max path length (3d), the value is only 1, since we use only the skip edge.

We designed GRAF to reflect findings in recent work as well as zero-cost proxy properties. Operation counts stem from the analysis from the previous section. Maximum path



(a) NB201 `cifar10` (b) TNB101-micro `class_scene` (c) TNB101-micro `class_scene`

Figure 2. Spearman correlation of zero-cost proxies with validation accuracy by clusters of architectures with the same number of 1x1 and 3x3 convolutions (a,b), and **only** 3x3 convolutions (c). *0-6* represents clusters with corresponding convolution counts, *all data* is the correlation over all available networks (same as in NB-Suite-Zero (Krishnakumar et al., 2022)).



(a) Exemplary cell (b) Input node degree in allowed [conv 3x3, skip] = 2 (c) Maximal path length over [conv 3x3, skip] = 3 (d) Maximal path length over [skip]=1

Figure 3. Exemplary overview of neural graph features in the NAS-Bench-201 search space.

length represents the depth of the network, and minimum path length tells us whether there are shortcuts from the input to the output – for example, whether there is a skip connection. Node degree features can be compared to motifs from NASBOWL – some best-performing motifs included the input node with a specific operation pattern (Ru et al., 2021).

In the appendix, we include Table 4 that shows GRAF counts with computation time per benchmark (always well below one second per network). We also include macro features in Section E.1; these are used for TNB101-macro and are based on channel and stride counts.

3.3. Using Graph Features in Prediction

In this section, we use GRAF as input data for validation accuracy prediction with a random forest predictor. We also include zero-cost proxies (ZCP), one-hot encoding (onehot) and combinations of these encodings. We examine variants where we include only `flops` and `params` (FP) instead of all ZCP, since they can be calculated without any batches passed through the networks. We evaluate the different settings on all available benchmarks and datasets, for 3 train sample sizes (32, 128 and 1024) and across 50 seeds. We report Kendall tau for every run. Full results are available in Section F in the appendix.

Figure 4 shows results for 1024 sampled train networks on `cifar-10` and NB101, NB201, and NB301. GRAF performs better than ZCP and onehot in all cases, and ZCP + GRAF yields the overall best results. Figure 5 shows results for the `autoencoder` task and TNB101-micro. We can see that using GRAF leads to better results than ZCP or OH across all sample sizes. For other TNB-micro tasks (Figures 16, 17) and TNB-macro (Figures 18, 19), the best configurations always include GRAF. Overall, GRAF performs the same or better than ZCP even on the smaller sample sizes (32, 128) across all benchmarks and tasks (Section F in the appendix) except for NB201, where it is slightly worse (Table 7).

These results show that while zero-cost proxies have some flaws, they still capture important network properties that help the prediction. The strong performance of ZCP + GRAF suggests that GRAF does not capture all network properties, and ZCP and GRAF complement each other well for prediction. It is important to note that for some tasks beyond image classification, using GRAF on its own is sufficient.

3.4. Feature importance

We analyze the feature importance of random forest fit on ZCP + GRAF using SHAP (Lundberg & Lee, 2017). For each of the 50 runs, we compute the mean absolute Shapley value for all features. Then, we sort the values, obtain ranks, compute mean rank across the 50 runs and list the 10

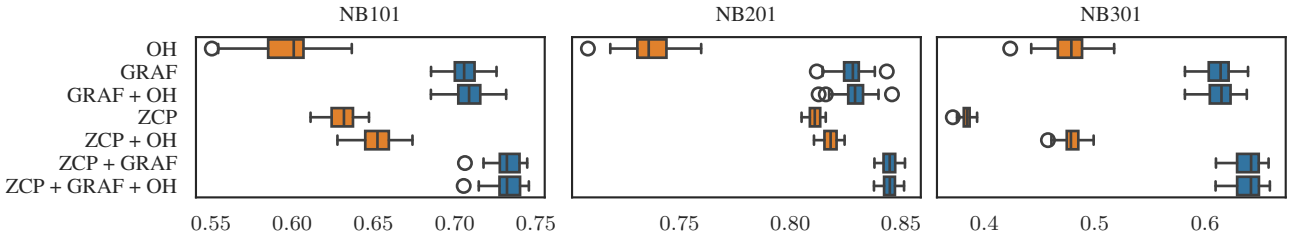


Figure 4. Validation accuracy prediction (`cifar-10`) across 3 benchmarks with 1024 sampled networks. Comparison of 3 encodings (ZCP, onehot, and GRAF) and their combinations. Blue boxes denote runs including GRAF.

Table 1. The ten most important features on NB201 `cifar-10` and TNB101-micro - `autoencoder`. Ranking is based on Shapley values computed from random forest trained on 1024 training samples with ZCP+GRAF.

NB201 - `cifar-10`		TNB101-micro - `autoencoder`	
Feature name	Mean rank	Feature name	Mean rank
jacov	0.00	min path over skip	0.00
nwot	1.12	jacov	1.00
flops	3.62	fisher	2.00
synflow	4.08	min path over [skip,C3x3]	5.50
min path over [skip,C3x3,C1x1]	4.78	snip	5.58
params	5.04	min path over [skip,C1x1]	5.64
epe_nas	6.04	grad_norm	6.64
zen	6.36	zen	8.08
min path over [skip,C3x3]	11.08	grasp	9.34
min path over skip	11.88	l2_norm	9.74

features with the lowest (best) ranks.

Table 1 shows the most important features for NB201 and `cifar-10`, and TNB101-micro and `autoencoder`. For `cifar-10`, many ZCP have low ranks, and from GRAF, skip-connection and convolution min path length (shortcuts) are the most important. However, on TNB101-micro, the situation is different. For the `autoencoder` task, shortcuts with skip-connection and average pooling are important features (confirming findings of Lopes et al. (2023) – they found that `autoencoder` had skip-connections in well-performing networks). In the appendix, Table 9 shows that while `ImageNet16-120` important features overlap with `cifar-10`, for `class_scene`, node degree features are important.

In Section G in the appendix, we list feature importances for the other benchmarks and tasks. Although NB101 and NB301 also evaluate architectures on `cifar-10`, the important features are different – notably, `jacov` was highly ranked on NB201 but is not in the top 10 features for these benchmarks. Similarly, node degree features are important for NB101 and NB301, but not present in NB201. This indicates that search space design has a great impact on performance prediction – models may need to learn orthogonal architectural properties.

3.5. Analysing feature redundancy

Group level First, we analyze the contribution of the following feature groups: max path length (MAX), min path length (MIN), operation count (OPS), and node degree (DEG). We look at performance of ZCP + one feature group, and ZCP + all but one feature group. We compare two cases: with all ZCP, and only with `flops` and `params` (FP). Figure 6 shows random forest prediction across 50 seeds on 1024 networks (NB201, `cifar-10`). We see that DEG has the highest individual contribution. However, when using all but one group, leaving out MIN leads to the greatest drop in performance. It also seems that OP do not have a large influence when other feature groups are present, indicating they might possibly be computed from the other features or from FP. However, we still use them in our models for better interpretability. Figures 20 and 21 in the appendix show similar results for NB101 and NB201 + `ImageNet16-120` — all groups except OP are needed, but the most influential groups differ.

Feature level Since GRAF is designed for all subsets of the operation set, some features might be redundant. To identify redundancies, we examine the linear dependence of features. Using linear regression on the GRAF feature set, we predict one feature using all other features. Then, if the prediction for a feature is perfect, we remove it from the dataset. Doing so, we iteratively eliminate all redundant features until we get a linearly independent set of features.

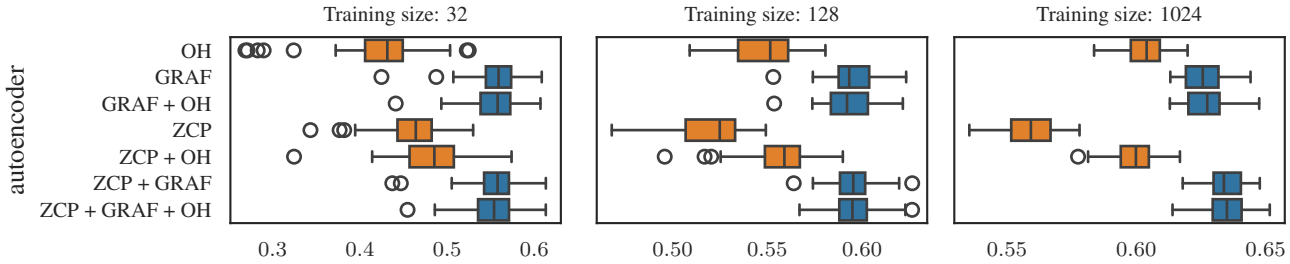


Figure 5. TNB101_micro `autoencoder` task across 3 train sample sizes. Comparison of 3 encodings (ZCP, onehot, and GRAF) and their combinations. Blue boxes denote runs including GRAF.

This process resulted in 30 features for NB101, 72 for NB201, and 546 for NB301. However, when we compare the results of a random forest on the smaller feature set versus the original set, the results with the smaller set are worse if using only the GRAF subset (refer to Table 18 in the appendix). When combined with ZCP, the performance is the same as the full GRAF + ZCP on NB101 and NB201, yet still worse on NB301. A possible explanation could be that the models are not able to construct the removed features from the subset of features during learning. ZCP cover some of the performance loss, but fail for a more complex search space like NB301.

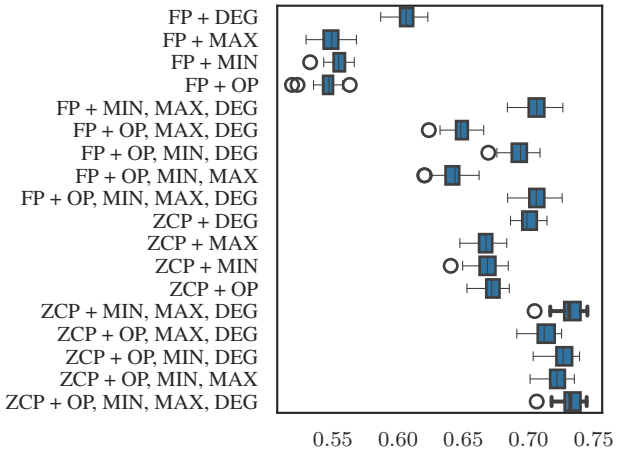


Figure 6. Ablation study on GRAF groups – analyzing cases when only 1 group is present vs when keeping all but 1 group.

4. Empirical Evaluation on Diverse Tasks

In this section, we evaluate GRAF on tasks beyond validation accuracy prediction – namely hardware metrics and robustness tasks. Then, we examine how ZCP and GRAF can be included into BRP-NAS. We also compare GRAF with existing encodings and performance predictors. Lastly, we use the ZCP + GRAF predictor in a search setting on NB201. Hardware, robustness and BRP-NAS experiments have the same experiment settings as in Section 3.3.

4.1. Other Prediction Tasks

Hardware tasks Our proposed GRAF features were also evaluated on hardware metrics from HW-NAS-Bench (Li et al., 2021). This includes mainly energy and latency on different devices, in total 10 metrics for 3 datasets. The prediction tasks are of varying difficulty, ranging from easy tasks to difficult ones. Onehot is generally a good predictor in HW-NAS-Bench tasks as shown by Laube et al. (2022). Figure 7 shows results on the `edgegpu_energy` task. GRAF is better than ZCP on all sample sizes, and better than onehot on smaller sample sizes. GRAF + onehot yields the overall best results. Additional results can be found in the appendix J, they show that the best setting contains GRAF among predictors on the vast majority of tasks.

Robustness tasks In the following section, we evaluate how GRAF improves the results on the task of predicting the robustness of the architectures. We use the evaluated robustness of the NAS-Bench-201 (Dong & Yang, 2020b) from the robustness dataset (Jung et al., 2023). This dataset evaluated the architectures from NAS-Bench-201 on different adversarial attacks and different perturbation strength; we consider in this paper two white-box attacks (FGSM (Goodfellow et al., 2015), and PGD (Kurakin et al., 2017). Important to note that the adversarial robustness is highly correlated with the validation accuracy since the former can never improve over the latter. Here, we will differentiate the robustness evaluation into two different tasks: (i) predicting only the robustness accuracy, and (ii) predicting both, the validation accuracy **and** the robustness accuracy jointly (*Multi_Obj*). In more detail, we evaluate the prediction ability on both attacks for three different perturbation strengths ($\epsilon = \{0.1/255, 0.5/255, 1.0/255\}$) on `cifar-10` using three different training sizes for learning the random forest prediction model. For the first prediction type (i), the combination of zero-cost proxies, the proposed graph features, and the additional onehot encoding shows the highest Kendall tau in all considered tasks (see Section K in the appendix for a detailed overview). Furthermore, with increasing perturbation strength, using only the graph features shows a better

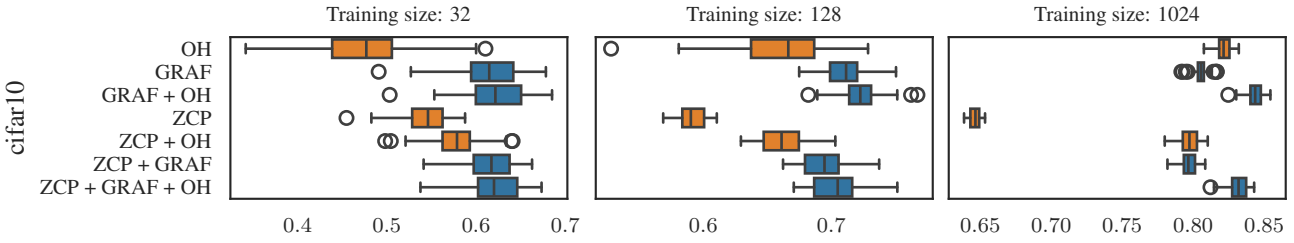


Figure 7. Random forest `edgegpu_energy` prediction on NB201 `cifar-10` (different training sample sizes).

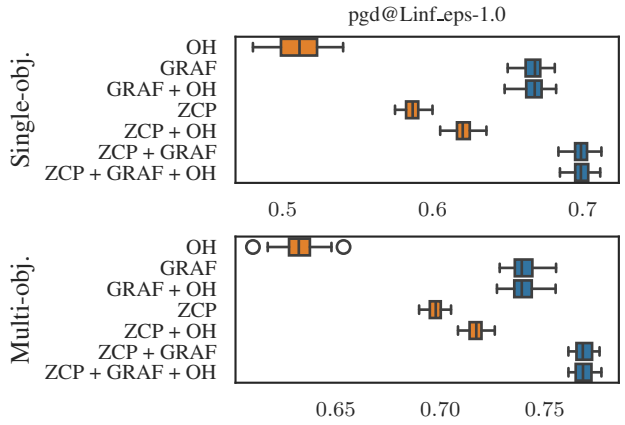


Figure 8. Robust accuracy prediction for the PGD adversarial attack with 1024 training data on the attack perturbations $\epsilon = 1.0/255$ for the objectives, only robust accuracy (top), and both, validation and robust accuracy (bottom). NB201 and `cifar-10`.

prediction ability than using only the zero-cost proxies from the literature. The same behavior is visible for the joint objective prediction task (ii). The latter task of predicting both accuracies (validation and robust) at once shows also a higher Kendall tau. Thus, this seems to be the easier task for the prediction model, which was also shown in Lukasik et al. (2023). We present the boxplots for the PGD attack for perturbation strength $\epsilon = 1.0/255$ for both evaluation tasks in 8 and more results on both adversarial attacks in both settings in Section K in the appendix.

4.2. Other Types of Encoding

Apart from the GRAF features and the zero-cost proxies, we also experimented with other types of network architecture encodings – path encoding, arch2vec (Dong & Yang, 2020a), and NASBOWL's Weisfeiler-Lehman features (WL) for the number of iterations equal to 1. (Ru et al., 2021). Detailed results of these experiments are included in the appendix (Tables 20 and 21). Here, we give just a brief overview.

Adding path encoding to the GRAF features does not improve the results, with the exception of `cifar-10` (NB101, NB201 and NB301) and `cifar-100` (NB201) on the largest sample size. However, the increase from ZCP + GRAF is small and is not observable on `ImageNet16-120` or TNB101-micro tasks. When GRAF is not used, ZCP + path encoding outperforms ZCP. This indicates that the GRAF features already contain most of the useful information from path encoding. The same behavior can be observed for the WL features and arch2vec, as they improve the results without GRAF, but when GRAF is present, the improvement is only small.

4.3. Existing performance predictors

BRP-NAS We compare our results with BRP-NAS (Dudziak et al., 2020), also including variants when BRP-NAS takes additional feedback from ZCP, GRAF, or onehot. This additional feedback is concatenated with the output from the BRP-NAS GCN and used in the output layer. A detailed description of this process and of the results of the experiments can be found in Section L in the appendix.

On most tasks, the best combination of BRP-NAS with the ZCP + GRAF features typically does not have better results than the best random forest combination. However, both ZCP + GRAF improve the BRP-NAS performance over the pure BRP-NAS in most cases.

On the accuracy task (cf. Table 28) on NB201 BRP-NAS is slightly better than the RF-based models only with the largest training size for the ImageNet16-120 target. On the NB301 benchmark, however, BRP-NAS has the same or better results for all training sizes, which may indicate that the GCN may be able to extract additional important information from the graph structure. On the hardware tasks (Table 29) , the original BRP-NAS has a similar performance to the best combination of features with RF-based predictors. Including the other information improves performance in all cases, with BRP-NAS + GRAF + OH being among the best configurations. Interestingly, very often the performance of BRP-NAS + ZCP is not much better than the performance of BRP-NAS alone. On the other hand, on the robustness tasks (Tables 30,31), BRP-NAS has slightly worse results than RF-based models with GRAF and onehot features. Additionally, unlike the other tasks, additional

information does not improve the results, except for the smallest perturbation and the smallest training set. An explanation could be that the properties relevant for prediction are captured both by BRP-NAS and ZCP + GRAF.

NASLib In this section, we compare the graph features with other predictors using the same setting as in the performance predictor survey by White et al. (2021b). We chose ZCP + GRAF for evaluation, as it performed the best across various settings. We use subsets of NB301 and NB101, as ZCP from NB-Suite-Zero are available only for a limited number of architectures. Figure 9 shows results for NB101, and detailed results are in Section N in the appendix, along with more information about the different predictors and ZCP + GRAF model variants.

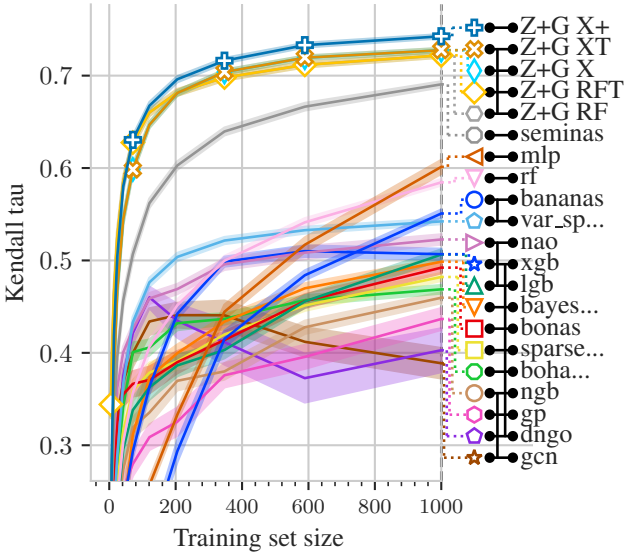


Figure 9. Evaluation of NASLib performance predictors on a subset of NB101, `cifar-10`.

In all cases, ZCP + GRAF models are the best-performing predictors. The predictor results match the original results from the survey except for NB101 (Figure 9), where BANANAS (White et al., 2021a) and GCN (Wen et al., 2020) are much worse in our case. An explanation could be that the sampled NB101 set is too diverse, and when sampling a train set from it, all remaining networks are too dissimilar for both predictors. This claim is supported by the encoding study by White et al. (2020), where path encoding had poor results on networks outside of a train set limited to a subspace of the search space.

TA-GATES Next, we compared GRAF with TA-GATES, a well-performing graph neural network predictor. We used the same models as in the previous section (without tuning). On small sample sizes, ZCP + GRAF performs better than TA-GATES. On larger sample sizes, only XGB+ has a similar performance. Although predictors with ZCP + GRAF are faster and more interpretable, using well-performing graph neural networks might be a promising direction in achieving better performance prediction. Full details are provided in the Section M.

4.4. NASLib search run

Lastly, we evaluate ZCP + GRAF in a search setting. As in Section 4.3, we repeat the experiments from the predictor survey and use the predictor as a surrogate in Bayesian optimization. We again use the same models for ZCP + GRAF except for the tuning. More details are included in Section O in the appendix.

Figure 10 shows the results of the search on `ImageNet16-120`. ZCP + GRAF is on average more sample efficient than the other predictors. On `cifar-10` (Figure 30), SemiNAS (Luo et al., 2020) has on average the same performance as GRAF in later search stages. Thus, a more extensive search evaluation would be needed as a part of future work, ideally including various tasks and optimizers.

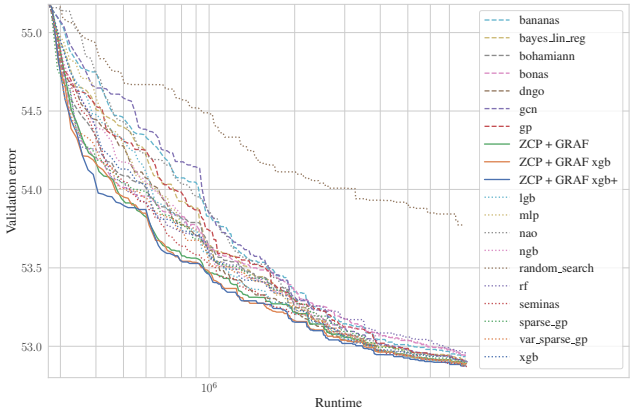


Figure 10. Bayesian optimization – search on `ImageNet16-120` and NB201.

5. Conclusion

To summarize, we introduced GRAF, simple-to-compute graph features inspired by the shortcomings of zero-cost proxies. When used as the input to a random forest predictor, they outperform ZCP and other common encodings. Their interpretability enabled us to highlight that different tasks favor different network properties. We evaluated GRAF extensively across a variety of tasks (including metrics beyond validation accuracy), showing very strong results overall. We demonstrated that they can also improve existing models such as the graph neural network BRP-NAS. Lastly, we evaluated them in prediction and search settings, where they outperformed all available predictors.

6. Discussion

The potential of GRAF is to bring a better understanding of existing search spaces, tasks, and existing performance predictors. In fact, due to the simple design, GRAF could be a good baseline for more complex predictors – especially since many predictors for NASLib had worse results while consuming more resources. GRAF could also inspire new zero-cost proxies that capture specific network properties and correlate better with tasks beyond `cifar-10` classification. Lastly, since GRAF macro had better results than proxies and the onehot encoding, it could be easily extended to new macro search spaces and transformer search spaces. An interesting question is whether we could scale up the networks according to the most important features while keeping a good performance.

When compared with other predictors, TA-GATES in performance prediction and SemiNAS in `cifar-10` search matched XGBoost (Chen & Guestrin, 2016) trained on ZCP + GRAF for the respective tasks. Also, BRP-NAS combined with GRAF had the best results on NB301 compared to the random forest results. This suggests that more complex models are still promising and combination with additional information like ZCP or GRAF may be the key to better results across search spaces.

A disadvantage of GRAF is that if a search space differs significantly from the cell-based search spaces and macro search space used in this work, new graph features need to be designed. Also, even for similar search spaces, transferability is limited due to different operation sets. However, due to their simplicity, designing new graph features should not be hard. Transfer between search spaces might even be an ill-posed problem due to different architecture training pipelines – also supported by the fact, that important features for `cifar-10` were different across search spaces.

A major drawback is that for the best results, zero-cost proxies are still needed. One possible explanation could be that there might be some graph properties that we have not used. Another hypothesis could be that while network performance depends on the neural graph, some part of it also depends on some properties of the information flow or training dynamics that cannot be captured by analyzing the architecture only. This problem opens an interesting direction for future research.

Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under grant number 417962828. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. This research was partially supported by the SVV project number 260 698. G. Kadlecová was supported by Charles University Grant Agency project no. 246322. P. Vidnerová was supported by the project 22-02067S of the Czech Science Foundation. J. Lukasik acknowledges support by the DFG research unit 5336 - Learning to Sense. Frank Hutter is a Hector Endowed Fellow at the ELLIS Institute Tübingen. We would like to thank Arjun Krishnakumar for the valuable discussions that helped to shape the initial ideas of this paper.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. Specifically, our work is in the field of performance prediction, which aims to reduce the costs of network evaluation. We believe we contributed to this goal, as our predictors outperform existing methods while being cheap to fit. This enables energy savings in future NAS applications.

Since our method is interpretable and has a good performance across various tasks, we could gain insight into how network components influence other important metrics, namely fairness. This could get us closer to designing more fair deep learning applications. The interpretability could also make NAS more attractive for fields where it is important – for example medical or decision-making fields.

References

Abdelfattah, M. S., Mehrotra, A., Dudziak, Ł., and Lane, N. D. Zero-Cost Proxies for Lightweight NAS. In *International Conference on Learning Representations (ICLR)*, 2021.

Akhauri, Y. and Abdelfattah, M. S. Multi-predict: Few shot predictors for efficient neural architecture search. In *AutoML Conference 2023*, 2023. URL https://openreview.net/forum?id=14U6uzrh-wr.

Bauer, M., van der Wilk, M., and Rasmussen, C. E. Understanding probabilistic sparse gaussian process approximations. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 1533–1541, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

Chen, W., Huang, W., and Wang, Z. "no free lunch" in neural architectures? a joint analysis of expressivity, convergence, and generalization. In *AutoML Conference 2023*, 2023. URL https://openreview.net/forum?id=EMys3eIDJ2.

Dong, X. and Yang, Y. Nas-bench-201: Extending the scope of reproducible neural architecture search. In *International Conference on Learning Representations (ICLR)*, 2020a. URL https://openreview.net/forum?id=HJxyZkBKDr.

Dong, X. and Yang, Y. Nas-bench-201: Extending the scope of reproducible neural architecture search. In *International Conference on Learning Representations (ICLR)*, 2020b. URL https://openreview.net/forum?id=HJxyZkBKDr.

Dreczkowski, K., Grosnit, A., and Ammar, H. B. Framework and benchmarks for combinatorial and mixed-variable bayesian optimization, 2023.

Duan, T., Avati, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A., and Schuler, A. Ngboost: natural gradient boosting for probabilistic prediction. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Duan, Y., Chen, X., Xu, H., Chen, Z., Liang, X., Zhang, T., and Li, Z. Transnas-bench-101: Improving transferability and generalizability of cross-task neural architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5251–5260, 2021.

Dudziak, L., Chau, T., Abdelfattah, M. S., Lee, R., Kim, H., and Lane, N. D. Brp-nas: prediction-based nas using gcns. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Elsken, T., Metzen, J. H., and Hutter, F. Neural architecture search: A survey. *J. Mach. Learn. Res.*, 20:55:1–55:21, 2019.

Erickson, N., Mueller, J., Shirkov, A., Zhang, H., Larroy, P., Li, M., and Smola, A. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2015.

Jung, S., Lukasik, J., and Keuper, M. Neural architecture design and robustness: A dataset. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=p8coElqiSDw.

Kandasamy, K., Neiswanger, W., Schneider, J., Poczos, B., and Xing, E. P. Neural architecture search with bayesian optimisation and optimal transport. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/f33ba15effa5c10e873bf3842afb46a6-Paper.pdf.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Krishnakumar, A., White, C., Zela, A., Tu, R., Safari, M., and Hutter, F. NAS-bench-suite-zero: Accelerating research on zero cost proxies. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=yWhuIjIjH8k.

Kurakin, A., Goodfellow, I. J., and Bengio, S. Adversarial machine learning at scale. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2017.

Laube, K. A., Mutschler, M., and Zell, A. What to expect of hardware metric predictors in NAS. In *First Conference on Automated Machine Learning (Main Track)*, 2022. URL https://openreview.net/forum?id=HHrzAgpHUgq.

Lee, N., Ajanthan, T., and Torr, P. H. S. Snip: single-shot network pruning based on connection sensitivity. In *International Conference on Learning Representations ICLR*. OpenReview.net, 2019.

Li, C., Yu, Z., Fu, Y., Zhang, Y., Zhao, Y., You, H., Yu, Q., Wang, Y., and Lin, Y. C. HW-NAS-bench: Hardware-aware neural architecture search benchmark. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=_0kaDkv3dVf.

Liaw, A. and Wiener, M. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002. URL https://CRAN.R-project.org/doc/Rnews/.

Lin, M., Wang, P., Sun, Z., Chen, H., Sun, X., Qian, Q., Li, H., and Jin, R. Zen-nas: A zero-shot NAS for high-performance image recognition. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV*. IEEE, 2021.

Lindauer, M. and Hutter, F. Best practices for scientific research on neural architecture search, 2020.

Liu, H., Simonyan, K., and Yang, Y. DARTS: Differentiable architecture search. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1eYHoC5FX.

Lopes, V., Alirezazadeh, S., and Alexandre, L. A. EPE-NAS: efficient performance estimation without training for neural architecture search. In *Artificial Neural Networks and Machine Learning - ICANN*, volume 12895 of *Lecture Notes in Computer Science*, pp. 552–563. Springer, 2021.

Lopes, V., Degardin, B., and Alexandre, L. A. Are neural architecture search benchmarks well designed? a deeper look into operation importance. *Information Sciences*, 650:119695, 2023. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2023.119695. URL https://www.sciencedirect.com/science/article/pii/S002002552301280X.

Lukasik, J., Moeller, M., and Keuper, M. An evaluation of zero-cost proxies – from neural architecture performance to model robustness. In *DAGM German Conference on Pattern Recognition*, 2023.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4765–4774. Curran Associates, Inc., 2017.

Luo, R., Tian, F., Qin, T., Chen, E.-H., and Liu, T.-Y. Neural architecture optimization. In *Advances in neural information processing systems*, 2018.

Luo, R., Tan, X., Wang, R., Qin, T., Chen, E., and Liu, T.-Y. Semi-supervised neural architecture search. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 10547–10557. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/77305c2f862ad1d353f55bf38e5a5183-Paper.pdf.

Ma, L., Cui, J., and Yang, B. Deep neural architecture search with deep graph bayesian optimization. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 500–507, 2019.

Mehrotra, A., Ramos, A. G. C. P., Bhattacharya, S., Dudziak, Ł., Vipperla, R., Chau, T., Abdelfattah, M. S., Ishtiaq, S., and Lane, N. D. {NAS}-bench-{asr}: Reproducible neural architecture search for speech recognition. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=CU0APx9LMaL.

Mellor, J., Turner, J., Storkey, A. J., and Crowley, E. J. Neural architecture search without training. In *Proceedings of the 38th International Conference on Machine Learning, ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7588–7598. PMLR, 2021.

Ning, X., Tang, C., Li, W., Zhou, Z., Liang, S., Yang, H., and Wang, Y. Evaluating efficient performance estimators of neural architectures. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Ning, X., Zhou, Z., Zhao, J., Zhao, T., Deng, Y., Tang, C., Liang, S., Yang, H., and Wang, Y. TA-GATES: An encoding scheme for neural network architectures. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=74fJwNrBlPI.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Pham, H., Guan, M., Zoph, B., Le, Q., and Dean, J. Efficient neural architecture search via parameters sharing. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4095–4104. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/pham18a.html.

Rasmussen, C. E. *Gaussian Processes in Machine Learning*, pp. 63–71. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. doi: 10.1007/978-3-540-28650-9_4. URL https://doi.org/10.1007/978-3-540-28650-9_4.

Ru, B., Wan, X., Dong, X., and Osborne, M. Interpretable neural architecture search via bayesian optimisation with weisfeiler-lehman kernels. In *International Conference*

on Learning Representations, 2021. URL https://openreview.net/forum?id=j9Rv7qdXjd.

Shervashidze, N., Schweitzer, P., van Leeuwen, E. J., Mehlhorn, K., and Borgwardt, K. M. Weisfeiler-lehman graph kernels. *Journal of Machine Learning Research*, 12(77):2539–2561, 2011. URL http://jmlr.org/papers/v12/shervashidze11a.html.

Shi, H., Pi, R., Xu, H., Li, Z., Kwok, J., and Zhang, T. Bridging the gap between sample-based and one-shot neural architecture search with bonas. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1808–1819. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/13d4635deccc230c944e4ff6e03404b5-Paper.pdf.

Shu, Y., Wang, W., and Cai, S. Understanding architectures learnt by cell-based neural architecture search. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJxH22EKPS.

Snoek, J., Rippel, O., Swersky, K., Kiros, R., Satish, N., Sundaram, N., Patwary, M. M. A., Prabhat, P., and Adams, R. P. Scalable bayesian optimization using deep neural networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 2171–2180. JMLR.org, 2015.

Springenberg, J. T., Klein, A., Falkner, S., and Hutter, F. Bayesian optimization with robust bayesian neural networks. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/a96d3afec184766bfeca7a9f989fc7e7-Paper.pdf.

Tanaka, H., Kunin, D., Yamins, D. L. K., and Ganguli, S. Pruning neural networks without any data by iteratively conserving synaptic flow. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Titsias, M. Variational learning of inducing variables in sparse gaussian processes. In van Dyk, D. and Welling, M. (eds.), *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pp. 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL https://proceedings.mlr.press/v5/titsias09a.html.

Turner, J., Crowley, E. J., O'Boyle, M. F. P., Storkey, A. J., and Gray, G. Blockswap: Fisher-guided block substitution for network compression on a budget. In *International Conference on Learning Representations ICLR*. OpenReview.net, 2020.

Wang, C., Zhang, G., and Grosse, R. B. Picking winning tickets before training by preserving gradient flow. In *International Conference on Learning Representations ICLR*. OpenReview.net, 2020.

Wen, W., Liu, H., Chen, Y., Li, H., Bender, G., and Kindermans, P.-J. Neural predictor for neural architecture search. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX*, pp. 660–676, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58525-9. doi: 10.1007/978-3-030-58526-6_39. URL https://doi.org/10.1007/978-3-030-58526-6_39.

White, C., Neiswanger, W., Nolen, S., and Savani, Y. A study on encodings for neural architecture search. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20309–20319. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/ea4eb49329550caaa1d2044105223721-Paper.pdf.

White, C., Neiswanger, W., and Savani, Y. Bananas: Bayesian optimization with neural architectures for neural architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021a.

White, C., Zela, A., Ru, R., Liu, Y., and Hutter, F. How powerful are performance predictors in neural architecture search? *Advances in Neural Information Processing Systems*, 34, 2021b.

White, C., Khodak, M., Tu, R., Shah, S., Bubeck, S., and Dey, D. A deeper look at zero-cost proxies for lightweight nas. In *ICLR Blog Track*, 2022. URL https://iclr-blog-track.github.io/2022/03/25/zero-cost-proxies/. https://iclr-blog-track.github.io/2022/03/25/zero-cost-proxies/.

White, C., Safari, M., Sukthanker, R., Ru, B., Elsken, T., Zela, A., Dey, D., and Hutter, F. Neural architecture search: Insights from 1000 papers, 2023.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *International Conference*

on Learning Representations, 2019. URL https://openreview.net/forum?id=ryGs6iA5Km.

Ying, C., Klein, A., Christiansen, E., Real, E., Murphy, K., and Hutter, F. NAS-bench-101: Towards reproducible neural architecture search. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7105–7114, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/ying19a.html.

Zela, A., Siems, J. N., Zimmer, L., Lukasik, J., Keuper, M., and Hutter, F. Surrogate NAS benchmarks: Going beyond the limited search spaces of tabular NAS benchmarks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=OnpFa95RVqs.

A. NAS Best Practice Checklist

We now describe how we addressed the individual points of the NAS best practice checklist (Lindauer & Hutter, 2020).

1. **Best Practices for Releasing Code**

   For all experiments you report:

   (a) Did you release code for the training pipeline used to evaluate the final architectures? [N/A] We query architectures from NAS benchmarks (via NASLib)

   (b) Did you release code for the search space [N/A] We use search spaces from NASLib

   (c) Did you release the hyperparameters used for the final evaluation pipeline, as well as random seeds? [N/A]

   (d) Did you release code for your NAS method? [Yes] Link to a public github repository will be provided upon acceptance. Code is a part of the submission.

   (e) Did you release hyperparameters for your NAS method, as well as random seeds? [Yes] We provide hyperparameters to the XGB+ model and the BRP-NAS model. Other models have default hyperparameters or are the same as in previous work (NASLib, TA-GATES).

2. **Best practices for comparing NAS methods**

   (a) For all NAS methods you compare, did you use exactly the same NAS benchmark, including the same dataset (with the same training-test split), search space and code for training the architectures and hyperparameters for that code? [Yes] For BRP-NAS, we used different splits than in RF experiments, but we compute the average over 50 seeds

   (b) Did you control for confounding factors (different hardware, versions of DL libraries, different runtimes for the different methods)? [Yes] For experiments where predictors are compared in terms of runtime, we ran the predictors on the same hardware.

   (c) Did you run ablation studies? [Yes]

   (d) Did you use the same evaluation protocol for the methods being compared? [Yes]

   (e) Did you compare performance over time? [Yes]

   (f) Did you compare to random search? [Yes]

   (g) Did you perform multiple runs of your experiments and report seeds? [Yes] Seeds are provided in the codebase, default values were used.

   (h) Did you use tabular or surrogate benchmarks for in-depth evaluations? [Yes]

3. **Best practices for reporting important details**

   (a) Did you report how you tuned hyperparameters, and what time and resources this required? [Yes]

   (b) Did you report the time for the entire end-to-end NAS method (rather than, e.g., only for the search phase)? [Yes] Applies mostly to NASLib experiments

   (c) Did you report all the details of your experimental setup? [Yes]

B. Related Work (extended)

We extend Section 2 and provide more information and examples of model-based predictors, as well as works on interpretability. Given a train set of architectures and their performance $\{a, f(a)\}$, a model-based predictor f' solves a regression task by learning to estimate $f(a)$ from the train set. Many model-based predictors are based on graph neural networks that work directly with the graph structure of architectures. One example is BRP-NAS, a graph neural network latency and accuracy predictor. Its authors also demonstrated that in search, predicting binary relations between networks proved to be more sample efficient for accuracy prediction (Dudziak et al., 2020).

Other predictors use models like Gaussian processes, tree-based methods, or neural networks that take as input architectures encoded as vectors (White et al., 2021b). A study on neural encodings showed that the success of different encodings (one-hot encoding, path-encoding, and their variants) depends on the context in which they are used – good encodings for search may fare worse in performance prediction (White et al., 2020). Another possibility how to encode the architectures is through architectural representation learning. Dong & Yang (2020a) have demonstrated that unsupervised representation learning (named arch2vec) leads to better embedding quality compared to the embedding extracted in the supervised accuracy prediction task (Dong & Yang, 2020a).

In terms of interpretability, recent work analyzed the "no free lunch" theorem for architectures, where given a fixed budget, it is impossible to maximize the expressivity, convergence and generalization of an architecture (Chen et al., 2023). The authors discovered that while expressive networks tend to be deep and narrow, convergence and generalization are biased toward wide and shallow topologies. Interestingly, previous work showed that NAS optimizers like DARTS (Liu et al., 2019) or ENAS (Pham et al., 2018) favor wide and shallow architectures (Shu et al., 2020).

A recent analysis of cell-based search spaces has shown that only a subset of the operation set is needed to generate high-performing architectures (Lopes et al., 2023). In fact, for `cifar10`, `cifar100`, and `ImageNet16-120` – the most popular datasets for NAS evaluation – the number of `conv3x3` is a crucial factor in network performance. The authors demonstrated more variability on TransNAS-Bench-101 tasks, encouraging the evaluation on various datasets and tasks.

C. Benchmarks, Zero-cost Proxies and Encodings

C.1. NAS Benchmarks

In this work, we used the same benchmarks as in NAS-Bench-Suite-Zero (NB-Suite-Zero) (Krishnakumar et al., 2022) – NAS-Bench-101 (Ying et al., 2019), NAS-Bench-201 (Dong & Yang, 2020b), NAS-Bench-301 (Zela et al., 2022) and TransNAS-Bench micro and macro (Duan et al., 2021). Out of these benchmarks, only NB101 has operation labels on vertices, the other have operation labels on edges. It is important to note that TNB101-micro is a subset of NB201, and contains all networks without max pooling operations. All except for the macro search space TNB101-macro are cell-based search spaces. Table 2 lists all benchmarks with abbreviations used throughout the paper, the number of sampled architectures to be used in experiments, and the total number of architectures in the search space. We used only the subsets of the search spaces for which pre-evaluated zero-cost proxies are available. For NB201 and TNB101-micro, we also decreased the number of networks due to a problem described in Section C.2.

Table 3 lists the datasets used for each benchmark, with non-classification tasks marked.

Table 2. List of benchmarks used in our experiments with abbreviations and the number of sampled architectures. Total search space size is included for reference.

	NAS-Bench-101	NAS-Bench-201	NAS-Bench-301	TransNAS-Bench-101 -micro	TransNAS-Bench-101 - macro
abbreviation	NB101	NB201	NB301	TNB101-micro	TNB101-macro
# arch (sampled)	10 000	9 445	11 221	2 128	3 256
# arch (total)	423 624	9 445	10^{18}*	4 096	3 256

*surrogate benchmark

Table 3. List of datasets used in our experiments. Task that are *not* classification tasks are marked (*).

benchmark	dataset
NB101	`cifar-10`
NB201	`cifar-10`, `cifar-100`, `ImageNet16-120`
NB301	`cifar-10`
TNB101 (micro and macro)	`class_scene`, `class_object`, `autoencoder` (*), `jigsaw`, `normal` (*), `room_layout` (*), `segmentsemantic` (*)

C.2. Unreachable Branches in NB201 and TNB101-micro

For NB201, due to the presence of zero operation, some networks have edges (non-zero operations) that do not receive any non-zero input, and in other networks, some operations are not connected to the output. Figure 11 illustrates the problem, where multiple operations do not contribute to the output.

In our work, we keep only networks without unreachable branches (Figure 11) for all NB201 and TNB101-micro experiments. The main reason is that some zero-cost proxies have scores influenced by the unreachable operations – for example, `params` includes parameters of these operations. For validation accuracy prediction, this effect would be harmful, since `params` does not correspond to the true information flow. However, for hardware tasks, the unreachable operations might contribute to higher energies and latencies, and including unreachable parameters makes sense.

All in all, we believe removing unreachable operations in NAS experiments is good practice – these networks should not be used in practice (due to higher energy costs), and removing the operations requires just a DFS run, i.e. $\mathcal{O}(\mathcal{V}+\mathcal{E})$ for V nodes and E edges. In related work, NAS-Bench-ASR includes removing unreachable branches in the same manner as we do for NB201 (Mehrotra et al., 2021).

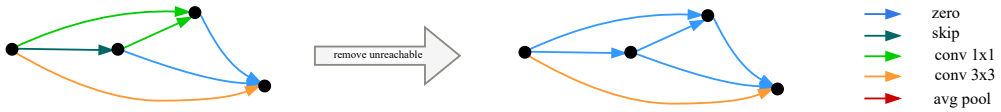


Figure 11. Visualization of unreachable nodes in the NAS-Bench-201 search space and our adaptation.

C.3. Zero-cost proxies

Here, we will provide an overview of the used zero-cost proxies. We use 13 zero-cost proxies from NB-Suite-Zero in our work: `epe_nas`, `fisher`, `flops`, `grad_norm`, `grasp`, `jacov`, `l2_norm`, `nwot`, `params`, `plain`, `snip`, `synflow`, `zen`. In general, these proxies can be differentiated into two distinct groups: data-independent, and data-dependent. Within these groups, there are different types of proxies, e.g., jacobian-based zero-cost proxies.

The data-independent group contains proxies that ignore the downstream dataset (for example `CIFAR-10`) entirely. Especially, so-called *baseline* proxies, fall into that group. These proxies are based on basic network information, such as the number of parameters (`params`) (Abdelfattah et al., 2021), or the sum of the weight's L2-norm `l2-norm` (Ning et al., 2021). In addition to these two baseline proxies, `synflow` (Tanaka et al., 2020), a pruning-at-initialization based score multiplying all weights in the network, was also successfully used as a zero-cost proxy (Abdelfattah et al., 2021). The last data-independent zero-cost proxy is the `zen-score` (Lin et al., 2021), which approximates the neural network by piecewise linear functions conditioned on activation patterns.

The majority of zero-cost proxies however is data-dependent; note, the data are not used to update the weights but only for score calculation using mostly one mini-batch. Mellor et al. (2021) used heuristics based on the Jacobian of the network to calculate zero-cost proxies, resulting in `jacov` and `nwot`. The former measures the covariance of the Jacobian, whereas the latter calculates the number of active linear regions in the network. Based on that, `epe-nas` was introduced (Lopes et al., 2021), which measures ability to distinguish different classes based on the correlation matrix of the Jacobian. Alternatively, `grad-norm` (Abdelfattah et al., 2021) sums the Euclidean norm of the gradients. In addition to these Jacobian-based scores, there exists several different techniques based on the pruning-at-initialization literature using data. Lee et al. (2019) introduced `snip`, which approximates the change in loss. Wang et al. (2020) propose a technique. `grasp`, which approximates the change in gradient norm. The last pruning-at-initialization based technique `fisher` (Turner et al.,

2020), which is defined by the sum of the gradients of the network's activation. These techniques were used in (Abdelfattah et al., 2021) as zero-cost proxies. Also natural network baselines can be data-dependent; `plain`, which is the multiplication of the weights and its gradients, and `flops` are both used as zero-cost proxies in (Abdelfattah et al., 2021).

C.4. Encodings

Throughout the paper, we work with these additional encodings: one-hot encoding (onehot), path encoding, arch2vec, NASBOWL's Weisfeiler-Lehman features, and ZCP.

One-hot and path encoding One-hot encoding is composed from the one-hot encoding of operations and a flattened adjacency matrix. Path encoding has the size of all possible paths in a cell from the search space – if the path is present, the corresponding index is 1, otherwise 0. For example, the cell on the right in Figure 11 would have 1 at positions corresponding to a) conv3x3, b) zero, and c) zero-zero. Both encodings were studied in an encoding study by White et al. (2020). The path-encoding combined with BANANAS was shown to have a very good performance on NB101 (White et al., 2021a).

arch2vec Arch2vec is a graph neural network autoencoder model (Dong & Yang, 2020a). It contains a GIN encoder (Xu et al., 2019) and inner product matrix and MLP operations decoder. It enables us to extract latent representations of the neural graphs.

NASBOWL We use NASBOWL's Weisfeiler-Lehman kernel features – features based on isomorphism test of graphs (Ru et al., 2021; Shervashidze et al., 2011). NASBOWL uses them as a part of a Gaussian process predictor, which we do not use in this work.

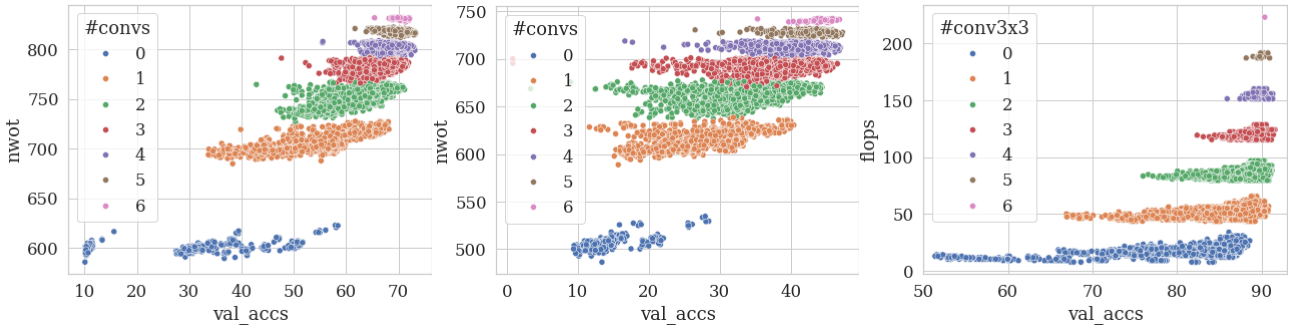
D. ZCP Biases – More Results

We present additional results on ZCP dependence on the number of convolutions

Figure 12 shows more results for NB201 – for `cifar-100` and `ImageNet16-120`, we see a similar dependence of `nwot` on the number of convolutions. For `flops`, we instead see a dependence on the number of conv3x3.

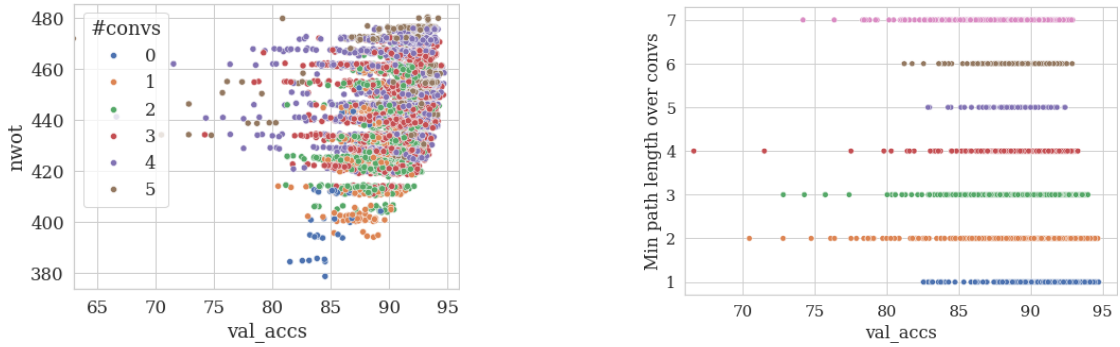
For NB101, the dependence on the number of convolution for NB101 is not simply observable, although we could possibly find some dependence based on other features and the networks size (Figure 13a). We can observe a different interesting property – NB101 has all the best-performing networks with min path length over conv1x1 and conv3x3 equal to 1 or 2 (Figure 13b) – and in fact, Table 10 shows that it is the most performing feature on `cifar-10`.

NB301 does not have a direct dependence on the number of convolutions, but #conv (sum of all sep_convs and dil_convs) is a stronger proxy than all ZCP except `nwot`, which is slightly better (Figure 14).



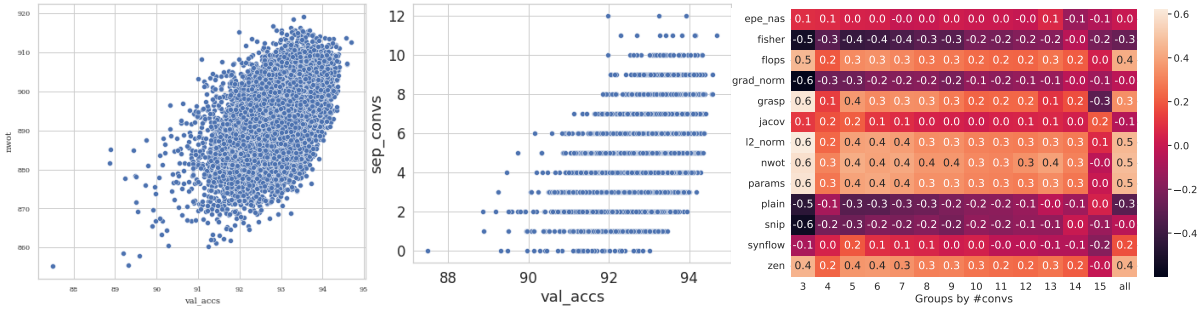
(a) NB201 `cifar-100` – nwot (b) NB201 `ImageNet16-120` – nwot (c) NB201 `cifar-10` – `flops` by #conv3x3

Figure 12. Additional results of proxy dependence on the number of layers (left, middle) – # convs, right – #conv3x3.



(a) NB101 `cifar-10` – nwot (b) Min path over convs

Figure 13. For `NB101`, the dependence on the number of convolutions is not so simple as in NB201 (left). From the features, min path length over convolutions has all best-performing networks with lengths 1 or 2.



(a) NB301 `cifar-10` – nwot (b) NB301 #sep_conv (c) NB201 `cifar-10` – `flops` by #conv3x3

Figure 14. NB301 does not show direct dependence on #convs, however, `nwot` (left) has only a slightly better correlation than #sep_conv (middle), and it is the only proxy with a better correlation (right)

E. Details about GRAF and Models

We list the total feature count and GRAF computation time for each of the benchmarks in Table 4.

Table 4. Numbers of GRAF features and time needed for GRAF computation (average of 10 evaluations, across all available networks from the benchmark samples) on AMD Ryzen 7 3800X.

	NB101	NB201	NB301	TNB101 -micro	TNB101 - macro
count	47	191	1540	95	16
time	6.1 s	27.3 s	427.1 s	3.6 s	0.65 s

For min and max path lengths, it can happen that no path exists, e.g. when none of the network operations are in the set of allowed operations for a feature. Then, the value would be defined as infinity. To avoid too large numbers in feature columns, we set the value to $|V'| + 1$, where $|V'|$ is the maximum number of nodes in a cell from the search space. The other features are defined in all cases (when the set of allowed nodes is restrictive, the feature value is 0).

For most tasks, we use a random forest model with default hyperparameters from scikit-learn (Pedregosa et al., 2011). When comparing with other predictors, we also include an XGBoost (Chen & Guestrin, 2016) with non-default hyperparameters (AutoGluon (Erickson et al., 2020) default parameters with slight tuning), denoted XGB+, see Table 5.

Table 5. XGB+ hyperparameters

tree_method	hist
subsample	0.9
n_estimators	10000
learning_rate	0.01

E.1. Macro features

TNB101-macro is a linear search space of 4-6 modules with four different module types – normal, downsampling (strided), channels increasing, and strided + channel increasing. We introduce two classes of features:

- Total number of channel increases/strided convolutions until position i.
- Total number of a module type in the architecture.

The motivation is that since the four types are only one-hot encoded, the model would need to learn which index is strided and which index increases channels. Including the information removes this need. Similarly, the number of increases/strides until a specific position reflects the behavior of input image processing in the network.

F. GRAF Validation Accuracy – Full Results

For the experiments, we have chosen the random forest predictor (with default scikit-learn hyperparameters) due to its fast fitting time. For all benchmarks and sample sizes, we used 18 CPU hours on Intel Xeon CPU E5-2620.

Figures 15–19 and Tables 6–8 show full results of the validation accuracy prediction task.

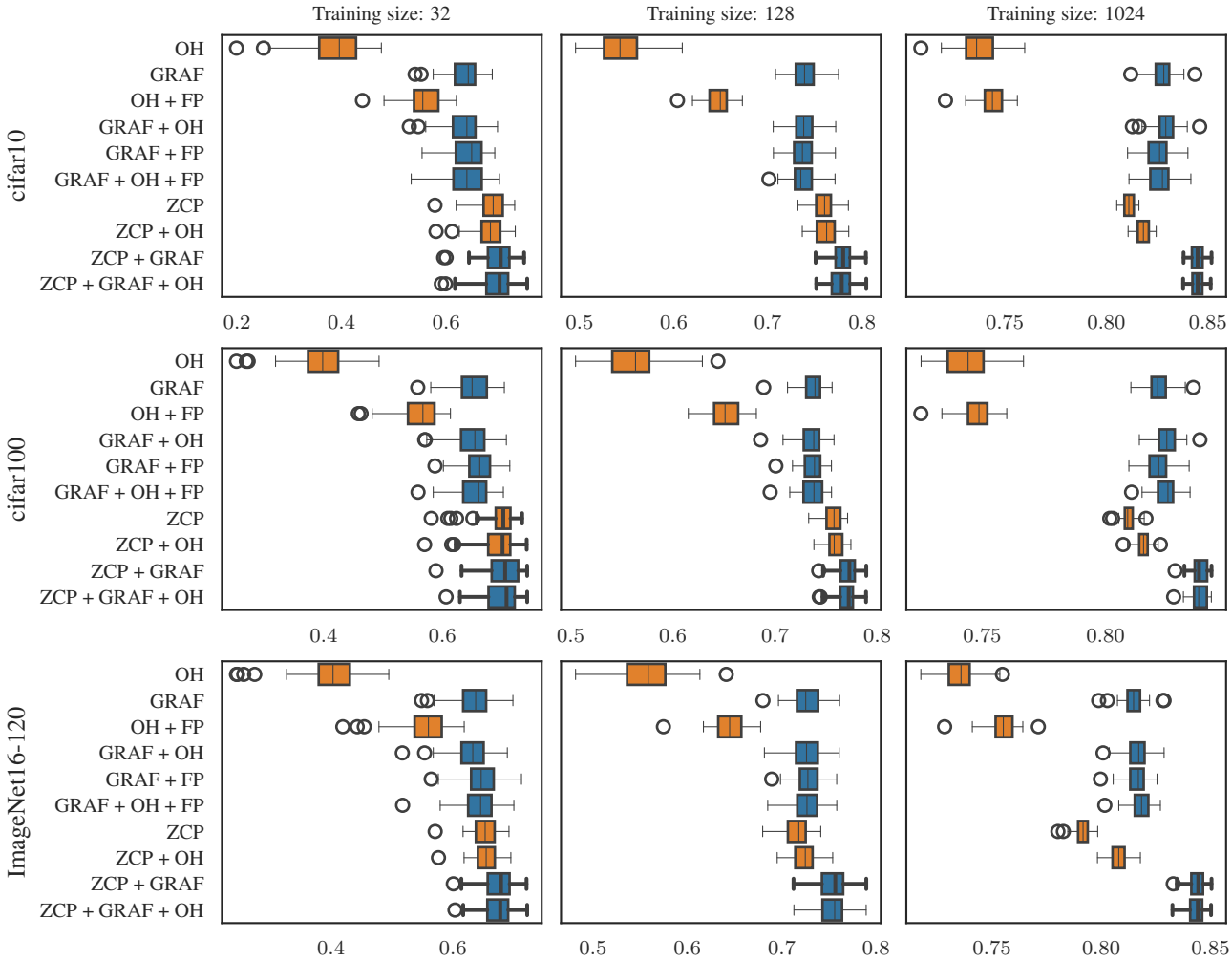


Figure 15. NAS-Bench201, accuracy

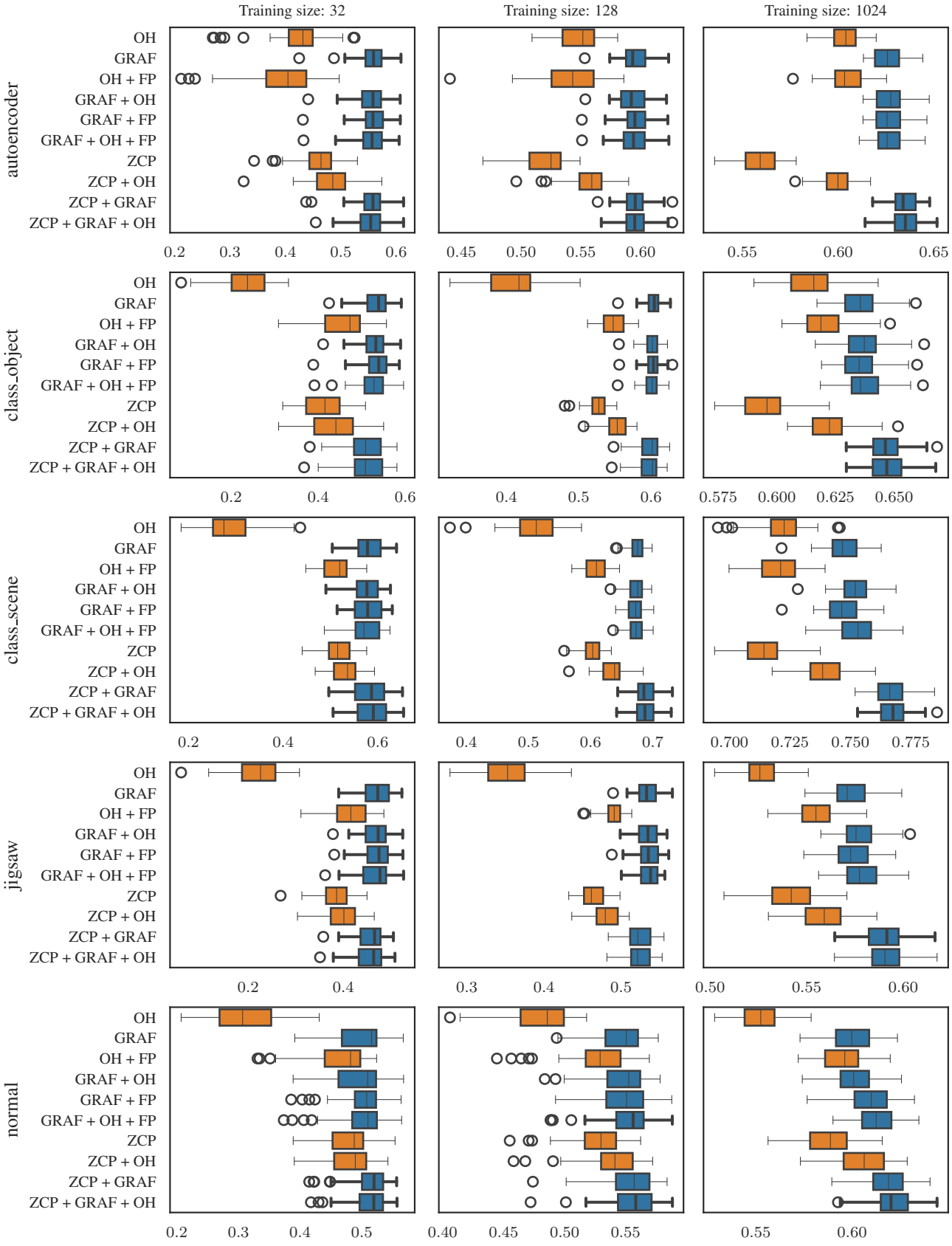


Figure 16. Trans-NAS-Bench101-micro, different targets

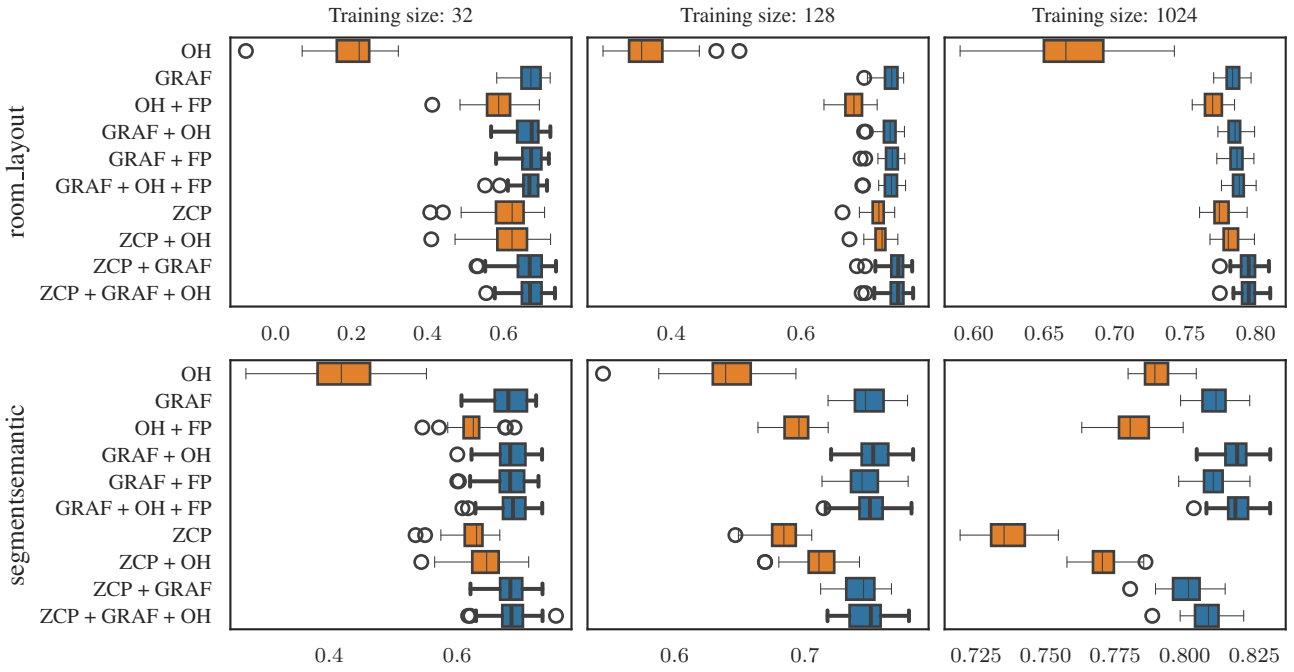


Figure 17. Trans-NAS-Bench101-micro, different targets

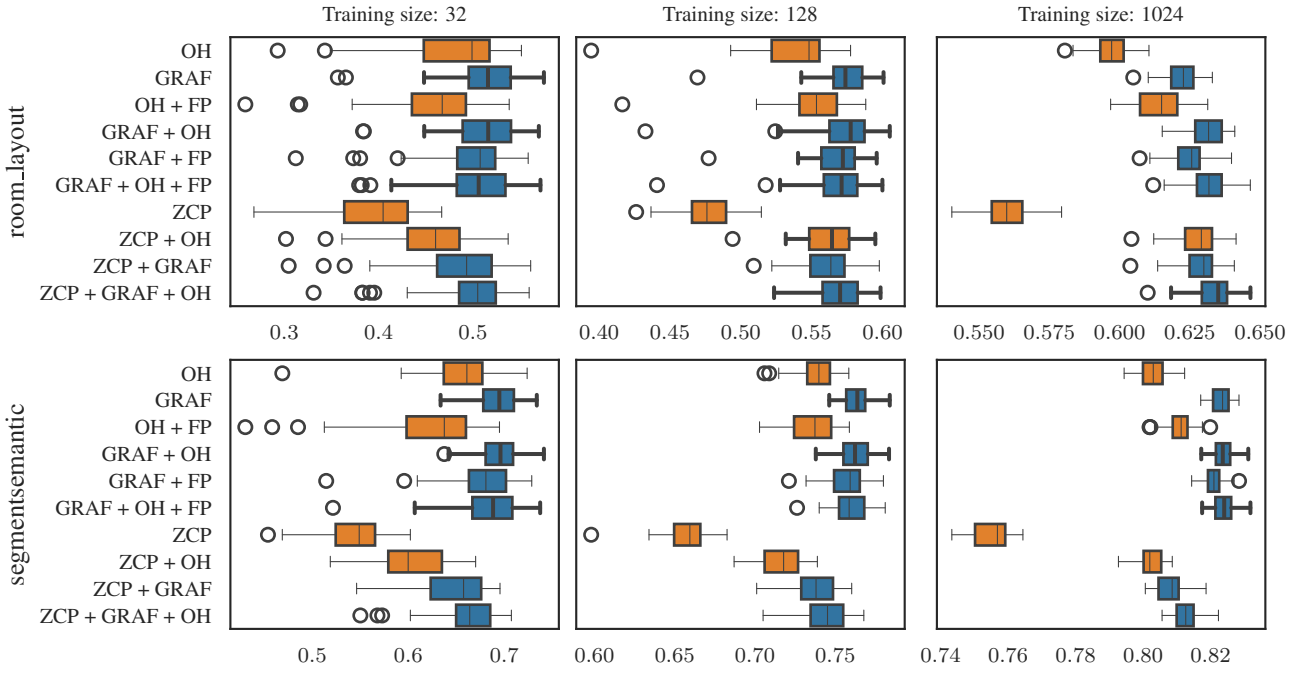


Figure 18. Trans-NAS-Bench101-macro, different targets

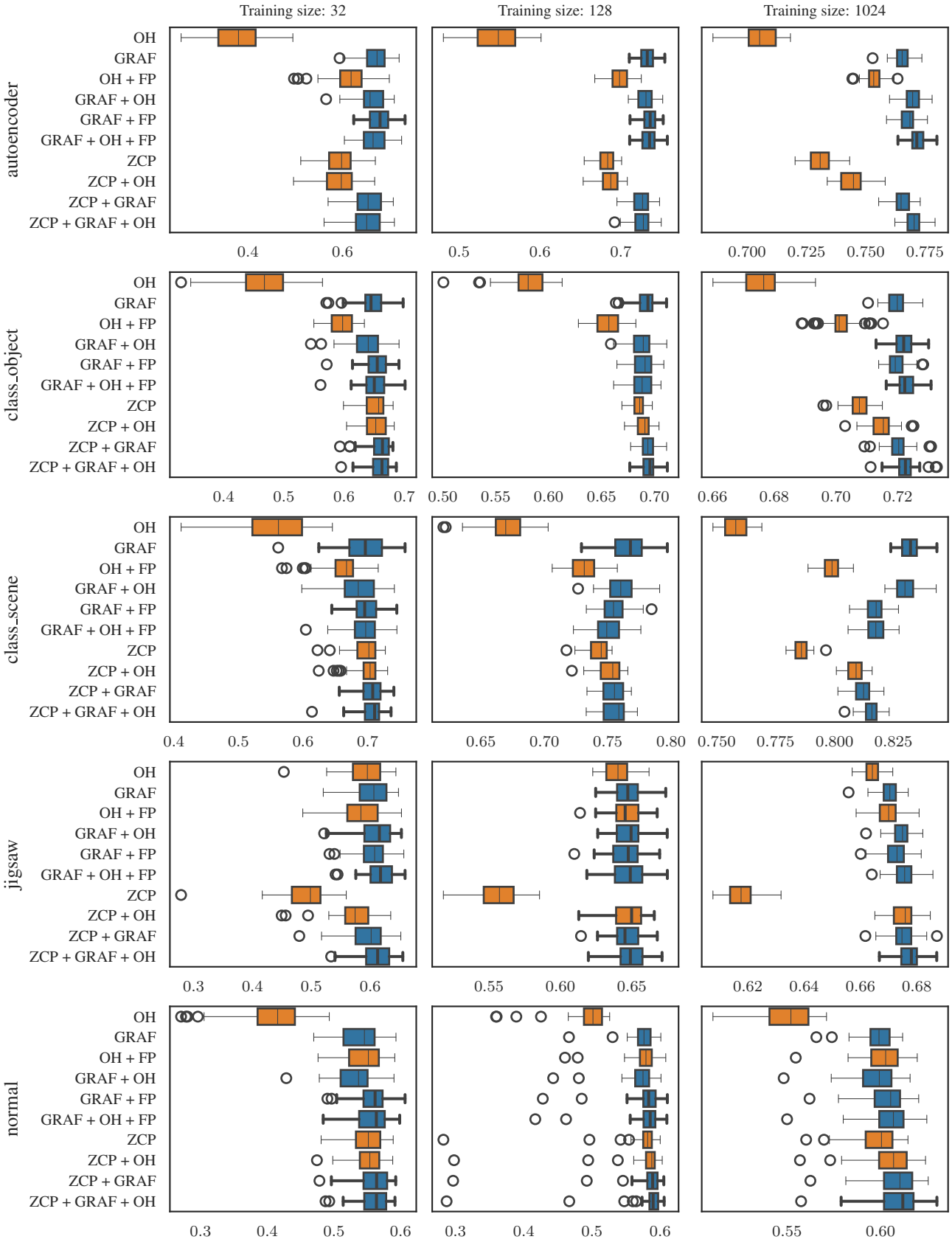


Figure 19. Trans-NAS-Bench101-macro, different targets

Table 6. Results for predicting validation accuracy on cifar10 on the NB101 benchmark. Average and standard deviation of Kendall tau over 50 independent runs. Bold values indicate the best results and those that are not statistically different from them.

dataset	cifar10		
train_size	32	128	1024
OH	$0.24^{0.09}$	$0.40^{0.06}$	$0.60^{0.02}$
OH + FP	$0.32^{0.09}$	$0.45^{0.06}$	$0.60^{0.02}$
GRAF	$0.54^{0.08}$	$0.63^{0.04}$	$0.71^{0.01}$
GRAF + FP	$0.53^{0.08}$	$0.63^{0.04}$	$0.71^{0.01}$
GRAF + OH	$0.52^{0.08}$	$0.63^{0.04}$	$0.71^{0.01}$
GRAF + OH + FP	$0.53^{0.08}$	$0.63^{0.04}$	$0.71^{0.01}$
ZCP	$0.48^{0.05}$	$0.56^{0.02}$	$0.63^{0.01}$
ZCP + OH	$0.47^{0.05}$	$0.57^{0.02}$	$0.65^{0.01}$
ZCP + GRAF	$\mathbf{0.56^{0.07}}$	$\mathbf{0.66^{0.04}}$	$0.73^{0.01}$
ZCP + GRAF + OH	$\mathbf{0.55^{0.07}}$	$\mathbf{0.66^{0.04}}$	$0.73^{0.01}$
OH + PE	$0.32^{0.07}$	$0.50^{0.04}$	$0.66^{0.01}$
OH + FP + PE	$0.39^{0.06}$	$0.52^{0.04}$	$0.66^{0.01}$
GRAF + PE	$0.54^{0.07}$	$0.64^{0.03}$	$0.71^{0.01}$
GRAF + FP + PE	$0.54^{0.07}$	$0.64^{0.03}$	$0.71^{0.01}$
GRAF + OH + PE	$0.52^{0.07}$	$0.64^{0.03}$	$0.72^{0.01}$
GRAF + OH + FP + PE	$0.53^{0.07}$	$0.63^{0.03}$	$0.72^{0.01}$
ZCP + PE	$0.49^{0.05}$	$0.58^{0.03}$	$0.67^{0.01}$
ZCP + OH + PE	$0.48^{0.05}$	$0.59^{0.03}$	$0.67^{0.01}$
ZCP + GRAF + PE	$\mathbf{0.56^{0.07}}$	$\mathbf{0.67^{0.03}}$	$\mathbf{0.74^{0.01}}$
ZCP + GRAF + OH + PE	$\mathbf{0.55^{0.07}}$	$\mathbf{0.67^{0.03}}$	$\mathbf{0.74^{0.01}}$

Table 7. Results for predicting validation accuracy on different datasets on the NB201 benchmark. Average and standard deviation of Kendall tau over 50 independent runs. Bold values indicate the best results and those that are not statistically different from them.

dataset	cifar10			cifar100			ImageNet16-120		
train_size	32	128	1024	32	128	1024	32	128	1024
OH	$0.39^{0.06}$	$0.54^{0.03}$	$0.74^{0.01}$	$0.39^{0.05}$	$0.56^{0.03}$	$0.74^{0.01}$	$0.40^{0.06}$	$0.56^{0.03}$	$0.74^{0.01}$
OH + FP	$0.56^{0.04}$	$0.65^{0.02}$	$0.74^{0.01}$	$0.56^{0.04}$	$0.65^{0.02}$	$0.75^{0.01}$	$0.55^{0.04}$	$0.64^{0.02}$	$0.75^{0.01}$
GRAF	$0.63^{0.03}$	$0.74^{0.01}$	$0.83^{0.01}$	$0.65^{0.03}$	$0.74^{0.01}$	$0.82^{0.01}$	$0.63^{0.03}$	$0.73^{0.02}$	$0.81^{0.01}$
GRAF + FP	$0.64^{0.04}$	$0.74^{0.01}$	$0.83^{0.01}$	$0.66^{0.03}$	$0.74^{0.01}$	$0.82^{0.01}$	$0.64^{0.03}$	$0.73^{0.02}$	$0.82^{0.01}$
GRAF + OH	$0.63^{0.04}$	$0.74^{0.01}$	$0.83^{0.01}$	$0.65^{0.03}$	$0.74^{0.01}$	$0.82^{0.01}$	$0.63^{0.03}$	$0.73^{0.02}$	$0.82^{0.01}$
GRAF + OH + FP	$0.63^{0.04}$	$0.74^{0.02}$	$0.83^{0.01}$	$0.65^{0.03}$	$0.74^{0.01}$	$0.82^{0.00}$	$0.64^{0.03}$	$0.73^{0.02}$	$0.82^{0.01}$
ZCP	$0.69^{0.03}$	$0.76^{0.01}$	$0.81^{0.00}$	$\mathbf{0.69^{0.03}}$	$0.76^{0.01}$	$0.81^{0.00}$	$0.65^{0.02}$	$0.72^{0.01}$	$0.79^{0.00}$
ZCP + OH	$0.68^{0.03}$	$0.76^{0.01}$	$0.82^{0.00}$	$\mathbf{0.69^{0.04}}$	$0.76^{0.01}$	$0.82^{0.00}$	$0.65^{0.02}$	$0.72^{0.01}$	$0.81^{0.00}$
ZCP + GRAF	$\mathbf{0.70^{0.03}}$	$\mathbf{0.78^{0.01}}$	$0.84^{0.00}$	$\mathbf{0.70^{0.03}}$	$\mathbf{0.77^{0.01}}$	$0.84^{0.00}$	$\mathbf{0.67^{0.03}}$	$\mathbf{0.75^{0.02}}$	$\mathbf{0.84^{0.00}}$
ZCP + GRAF + OH	$\mathbf{0.69^{0.04}}$	$\mathbf{0.78^{0.01}}$	$0.84^{0.00}$	$\mathbf{0.70^{0.03}}$	$\mathbf{0.77^{0.01}}$	$0.84^{0.00}$	$\mathbf{0.67^{0.02}}$	$0.75^{0.02}$	$0.84^{0.00}$
OH + PE	$0.42^{0.06}$	$0.57^{0.03}$	$0.75^{0.01}$	$0.42^{0.05}$	$0.58^{0.03}$	$0.75^{0.01}$	$0.40^{0.05}$	$0.56^{0.03}$	$0.74^{0.01}$
OH + FP + PE	$0.56^{0.04}$	$0.66^{0.01}$	$0.76^{0.01}$	$0.57^{0.04}$	$0.66^{0.02}$	$0.76^{0.01}$	$0.55^{0.04}$	$0.65^{0.02}$	$0.76^{0.01}$
GRAF + PE	$0.63^{0.03}$	$0.74^{0.01}$	$0.83^{0.01}$	$0.65^{0.03}$	$0.74^{0.01}$	$0.82^{0.01}$	$0.63^{0.04}$	$0.73^{0.02}$	$0.82^{0.01}$
GRAF + FP + PE	$0.64^{0.03}$	$0.74^{0.01}$	$0.83^{0.01}$	$0.66^{0.03}$	$0.74^{0.01}$	$0.82^{0.01}$	$0.64^{0.03}$	$0.73^{0.02}$	$0.82^{0.01}$
GRAF + OH + PE	$0.63^{0.04}$	$0.74^{0.02}$	$0.83^{0.01}$	$0.65^{0.03}$	$0.74^{0.01}$	$0.83^{0.01}$	$0.63^{0.04}$	$0.72^{0.02}$	$0.82^{0.01}$
GRAF + OH + FP + PE	$0.64^{0.04}$	$0.74^{0.02}$	$0.83^{0.01}$	$0.66^{0.03}$	$0.74^{0.01}$	$0.82^{0.01}$	$0.64^{0.03}$	$0.73^{0.02}$	$0.82^{0.01}$
ZCP + PE	$0.68^{0.03}$	$0.76^{0.01}$	$0.82^{0.00}$	$\mathbf{0.69^{0.03}}$	$0.76^{0.01}$	$0.82^{0.00}$	$0.65^{0.02}$	$0.72^{0.01}$	$0.80^{0.00}$
ZCP + OH + PE	$0.68^{0.03}$	$0.76^{0.01}$	$0.82^{0.00}$	$\mathbf{0.69^{0.04}}$	$0.76^{0.01}$	$0.82^{0.00}$	$0.65^{0.02}$	$0.72^{0.01}$	$0.81^{0.00}$
ZCP + GRAF + PE	$\mathbf{0.69^{0.04}}$	$\mathbf{0.78^{0.01}}$	$\mathbf{0.85^{0.00}}$	$\mathbf{0.70^{0.03}}$	$\mathbf{0.77^{0.01}}$	$\mathbf{0.84^{0.00}}$	$\mathbf{0.67^{0.03}}$	$\mathbf{0.75^{0.02}}$	$\mathbf{0.84^{0.00}}$
ZCP + GRAF + OH + PE	$\mathbf{0.69^{0.03}}$	$\mathbf{0.78^{0.01}}$	$\mathbf{0.85^{0.00}}$	$\mathbf{0.70^{0.03}}$	$\mathbf{0.77^{0.01}}$	$\mathbf{0.84^{0.00}}$	$\mathbf{0.67^{0.02}}$	$0.75^{0.02}$	$\mathbf{0.84^{0.00}}$

Table 8. Results for predicting validation accuracy on cifar10 on the NB301 benchmark. Average and standard deviation of Kendall tau over 50 independent runs. Bold values indicate the best results and those that are not statistically different from them.

dataset	cifar10		
train_size	32	128	1024
OH	$0.19^{0.05}$	$0.32^{0.03}$	$0.48^{0.02}$
OH + FP	$0.27^{0.06}$	$0.38^{0.02}$	$0.48^{0.01}$
GRAF	$0.38^{0.05}$	$0.52^{0.03}$	$0.61^{0.01}$
GRAF + FP	$0.39^{0.05}$	$0.53^{0.03}$	$0.63^{0.01}$
GRAF + OH	$0.39^{0.05}$	$0.52^{0.03}$	$0.61^{0.01}$
GRAF + OH + FP	$0.39^{0.05}$	$0.53^{0.03}$	$0.63^{0.01}$
ZCP	$0.28^{0.06}$	$0.34^{0.02}$	$0.38^{0.00}$
ZCP + OH	$0.31^{0.06}$	$0.40^{0.02}$	$0.48^{0.01}$
ZCP + GRAF	$\mathbf{0.40^{0.05}}$	$0.53^{0.03}$	$0.64^{0.01}$
ZCP + GRAF + OH	$\mathbf{0.40^{0.05}}$	$\mathbf{0.53^{0.03}}$	$0.64^{0.01}$
OH + PE	$0.23^{0.05}$	$0.40^{0.04}$	$0.55^{0.02}$
OH + FP + PE	$0.31^{0.07}$	$0.47^{0.04}$	$0.59^{0.01}$
GRAF + PE	$0.38^{0.05}$	$0.52^{0.03}$	$0.61^{0.01}$
GRAF + FP + PE	$0.39^{0.05}$	$0.53^{0.03}$	$0.64^{0.01}$
GRAF + OH + PE	$0.39^{0.05}$	$0.52^{0.03}$	$0.61^{0.01}$
GRAF + OH + FP + PE	$0.39^{0.05}$	$0.53^{0.03}$	$0.64^{0.01}$
ZCP + PE	$0.33^{0.08}$	$0.46^{0.04}$	$0.58^{0.00}$
ZCP + OH + PE	$0.33^{0.07}$	$0.47^{0.04}$	$0.59^{0.00}$
ZCP + GRAF + PE	$\mathbf{0.40^{0.05}}$	$\mathbf{0.53^{0.03}}$	$\mathbf{0.64^{0.01}}$
ZCP + GRAF + OH + PE	$\mathbf{0.40^{0.05}}$	$\mathbf{0.53^{0.03}}$	$\mathbf{0.64^{0.01}}$

G. Feature Importances

Now we list additional results for Section 3.4. The Tables 10–17 list the top ten important features for a given task and benchmark. We can observe a great variability across tasks and search space.

For the macro search spaces (Tables 14–17), an interesting observation is that the number of strides until pos. 0 (effectively *at* position 0) is often the most important feature. This indicates that the predictor benefits from the information about which operation is strided and which increases channels.

Table 9. The ten most important features on NB201 `ImageNet16-120` and TNB101-micro `class_scene` from ranking based on Shapley values computed from random forest trained on 1024 training samples with ZCP+GRAF.

NB201 - `ImageNet16-120`		TNB101-micro - `class_scene`	
Feature name	Mean rank	Feature name	Mean rank
jacov	0.00	params	0.28
min path over [skip,C3x3,C1x1]	2.30	flops	0.72
nwot	2.30	Average out deg. - C3x3	2.48
params	3.56	Average out deg. - [zero,skip,C1x1]	2.52
synflow	4.08	Average in deg. - [zero,skip,C1x1]	4.32
flops	4.30	Average in deg. - C3x3	4.68
min path over [skip,C1x1]	5.86	number of C3x3	6.00
zen	6.56	synflow	7.30
fisher	9.40	Input node degree - [zero,C1x1]	9.34
grad_norm	11.56	Input node degree - [zero,C1x1,C3x3]	9.42

Table 10. The ten most important features on NB101 and NB301 (`cifar-10` dataset) from ranking based on Shapley values computed from random forest trained on 1024 training samples with ZCP+GRAF.

NB101 - `cifar-10`		NB301 - `cifar-10`	
Feature name	Mean rank	Feature name	Mean rank
min path over [C1x1,C3x3]	2.16	nwot	2.02
l2_norm	2.16	params	5.16
zen	2.38	max path over [MP3x3,AP3x3]from input 2 (normal)	7.68
fisher	3.72	Input node 2 degree - skip (normal)	9.00
Average input node degreeC3x3	5.42	Average in deg. - [skip,SC3x3,SC5x5] (normal)	9.70
Average output node degreeC3x3	5.48	fisher	10.76
number of C3x3	6.06	Input 1 degree - [MP3x3,skip,SC3x3] (normal)	11.58
grasp	8.26	l2_norm	11.58
Output node degree - MP3x3	9.72	Average out deg. - [skip,SC3x3,SC5x5] (normal)	11.60
plain	9.88	Average out deg. - [MP3x3,AP3x3,DC3x3,DC5x5] (normal)	12.10

Table 11. The ten most important features on TNB101-micro (`class_object` and `normal` datasets) from ranking based on Shapley values computed from random forest trained on 1024 training samples with ZCP+GRAF.

TNB101-micro - `class_object`		TNB301-micro - `normal`	
Feature name	Mean rank	Feature name	Mean rank
params	0.36	jacov	0.02
flops	0.78	flops	0.98
Average out deg. - [zero,skip,C1x1]	2.68	Average in deg. - [zero,skip,C1x1]	4.22
Average out deg. - C3x3	2.72	Average out deg. - [zero,skip,C1x1]	4.64
Average in deg. - [zero,skip,C1x1]	4.58	Average out deg. - C3x3	4.72
Average in deg. - C3x3	5.00	params	5.18
min path over skip	5.02	number of C3x3	5.34
number of C3x3	6.88	Average in deg. - C3x3	5.44
Input node degree - skip	9.08	Input node degree - C3x3	8.68
Input node degree - [zero,C1x1,C3x3]	9.32	Input node degree - [zero,skip,C1x1]	8.92

Table 12. The ten most important features on TNB101-micro (`jigsaw` and `room_layout` datasets) from ranking based on Shapley values computed from random forest trained on 1024 training samples with ZCP+GRAF.

TNB101-micro - `jigsaw`		TNB301-micro - `room_layout`	
Feature name	Mean rank	Feature name	Mean rank
Input node degree - [zero,C1x1]	1.84	params	1.10
flops	2.34	fisher	1.70
params	2.48	flops	1.82
Input node degree - [skip,C3x3]	3.74	Average out deg. - [zero,skip,C1x1]	3.74
Average out deg. - C3x3	6.56	Average out deg. - C3x3	3.86
jacov	6.86	jacov	5.84
Average out deg. - [zero,skip,C1x1]	7.22	Average in deg. - [zero,skip,C1x1]	6.08
Average in deg. - [zero,skip,C1x1]	7.46	Average in deg. - C3x3	6.38
Input node degree - skip	7.48	number of C3x3	8.52
Average in deg. - C3x3	7.96	min path over skip	8.56

Table 13. The ten most important features on TNB101-micro (`segmentsemantic` dataset) from ranking based on Shapley values computed from random forest trained on 1024 training samples with ZCP+GRAF.

TNB101-micro - `segmentsemantic`	
Feature name	Mean rank
flops	0.60
Average out deg. - [zero,skip,C1x1]	0.60
Average in deg. - C3x3	1.96
Average in deg. - [zero,skip,C1x1]	3.12
Average out deg. - C3x3	4.36
params	4.60
number of C3x3	5.76
jacov	7.68
snip	9.54
Input node degree - [skip,C3x3]	9.86

Table 14. The ten most important features on TNB101-macro (`autoencoder` and `class_scene` datasets) from ranking based on Shapley values computed from random forest trained on 1024 training samples with ZCP+GRAF.

TNB101-macro - `autoencoder`		TNB101-macro - `class_scene`	
Feature name	Mean rank	Feature name	Mean rank
Number of strides until pos. 4	0.04	nwot	0.00
Number of strides until pos. 5	1.04	Number of strides until pos. 0	2.10
flops	1.92	grad_norm	2.36
params	3.00	flops	3.14
nwot	4.72	grasp	4.04
number of convs - channel increased	5.86	jacov	5.52
Number of strides until pos. 3	5.92	snip	5.58
number of simple convs	8.96	Number of channel increases until pos. 1	9.32
zen	9.86	l2_norm	9.34
l2_norm	10.68	fisher	10.38

Table 15. The ten most important features on TNB101-macro (`class_object` and `normal` datasets) from ranking based on Shapley values computed from random forest trained on 1024 training samples with ZCP+GRAF.

TNB101-macro - `class_object`		TNB101-macro - `normal`	
Feature name	Mean rank	Feature name	Mean rank
nwot	0.00	nwot	0.46
Number of strides until pos. 5	1.50	flops	0.54
grasp	1.62	params	3.44
Number of strides until pos. 4	4.08	l2_norm	3.72
flops	4.22	jacov	4.76
grad_norm	4.22	Number of strides until pos. 5	6.58
jacov	7.38	plain	8.74
number of convs - strided	7.70	fisher	9.14
Number of strides until pos. 0	8.62	zen	9.56
epe_nas	11.22	grad_norm	9.78

Table 16. The ten most important features on TNB101-macro (`jigsaw` and `room_layout` datasets) from ranking based on Shapley values computed from random forest trained on 1024 training samples with ZCP+GRAF.

TNB101-macro - `jigsaw`		TNB101-macro - `room_layout`	
Feature name	Mean rank	Feature name	Mean rank
Number of strides until pos. 0	0.00	Number of strides until pos. 0	0.00
Number of channel increases until pos. 0	1.00	Number of channel increases until pos. 0	1.12
nwot	2.26	nwot	2.98
flops	4.06	zen	5.66
Number of channel increases until pos. 1	5.48	fisher	5.90
grasp	5.78	Number of strides until pos. 1	6.14
jacov	6.86	grad_norm	6.50
fisher	8.18	synflow	8.20
plain	8.52	jacov	10.02
number of convs - strided	9.54	Number of strides until pos. 4	10.18

Table 17. The ten most important features on TNB101-macro (`segmentsemantic` dataset) from ranking based on Shapley values computed from random forest trained on 1024 training samples with ZCP+GRAF.

TNB101-macro - `segmentsemantic`	
Feature name	Mean rank
Number of strides until pos. 0	0.00
nwot	1.02
Number of strides until pos. 1	2.00
Number of channel increases until pos. 0	2.98
number of convs - strided	4.00
flops	5.52
Number of strides until pos. 2	5.64
params	7.30
Number of channel increases until pos. 1	8.58
Number of strides until pos. 5	9.58

H. Feature Redundancy

For the group redundancy, we used 2 CPU hours on Intel Xeon CPU E5-2620.

Table 18 lists average Kendall tau values comparing accuracy prediction using Random Forest with all GRAF and selected linearly independent GRAF. We compare cases with only GRAF and GRAF with ZCP. We denote the independent set of GRAF as sel. GRAF.

When using only GRAF, GRAF outperforms sel. GRAF in all cases except for NB201 and sample size 1024. When ZCP are also included, GRAF and sel. GRAF perform on par on NB101 and NB201 while being worse on NB301. This suggests that ZCP capture the same information as some of the redundant features, but cannot capture all important properties on NB301.

Table 18. Comparison of prediction with all GRAF features and selected (linearly independent) GRAF features. Average of Kendall tau over 50 runs. All benchmarks evaluated for `CIFAR-10`

benchmark	NB101			NB201			NB301		
train size	32	128	1024	32	128	1024	32	128	1024
features									
GRAF	$0.54^{0.08}$	$0.63^{0.04}$	$0.71^{0.01}$	$0.63^{0.03}$	$0.74^{0.01}$	$0.83^{0.01}$	$0.38^{0.05}$	$0.52^{0.03}$	$0.61^{0.01}$
sel. GRAF	$0.53^{0.07}$	$0.62^{0.04}$	$0.7^{0.01}$	$0.61^{0.05}$	$0.72^{0.02}$	$0.83^{0.0}$	$0.31^{0.06}$	$0.44^{0.02}$	$0.57^{0.01}$
GRAF + ZCP	$0.56^{0.07}$	$0.66^{0.04}$	$0.73^{0.01}$	$0.7^{0.03}$	$0.78^{0.01}$	$0.84^{0.0}$	$0.4^{0.05}$	$0.53^{0.03}$	$0.64^{0.01}$
sel. GRAF + ZCP	$0.56^{0.06}$	$0.66^{0.03}$	$0.73^{0.01}$	$0.7^{0.04}$	$0.78^{0.01}$	$0.84^{0.0}$	$0.35^{0.06}$	$0.48^{0.03}$	$0.61^{0.01}$

We include an additional experiment in Table 19 – we study how the results differ when GRAF is used only with data-dependent proxies (as per (Krishnakumar et al., 2022)). They seem to capture most of the performance gain, and including `params` leads to the same performance as with all ZCP.

Table 19. Results of GRAF used along with data-dependent ZCP (DD-ZCP), params (P) and all ZCP (ZCP).

benchmark	nb101			nb201			nb301		
train size	32	128	1024	32	128	1024	32	128	1024
features									
GRAF	$0.54^{0.08}$	$0.63^{0.04}$	$0.71^{0.01}$	$0.63^{0.03}$	$0.74^{0.01}$	$0.83^{0.01}$	$0.38^{0.05}$	$0.52^{0.03}$	$0.61^{0.01}$
GRAF + DD_ZCP	$0.56^{0.07}$	$0.66^{0.04}$	$0.73^{0.01}$	$0.69^{0.03}$	$0.78^{0.01}$	$0.84^{0.0}$	$0.4^{0.05}$	$0.53^{0.03}$	$0.63^{0.01}$
GRAF + DD_ZCP + P	$0.55^{0.07}$	$0.66^{0.04}$	$0.73^{0.01}$	$0.7^{0.04}$	$0.78^{0.01}$	$0.84^{0.0}$	$0.4^{0.05}$	$0.53^{0.03}$	$0.64^{0.01}$
GRAF + ZCP	$0.56^{0.07}$	$0.66^{0.04}$	$0.73^{0.01}$	$0.7^{0.03}$	$0.78^{0.01}$	$0.84^{0.0}$	$0.4^{0.05}$	$0.53^{0.03}$	$0.64^{0.01}$

We list additional plots of the feature group ablation in Figures 20 and 21.

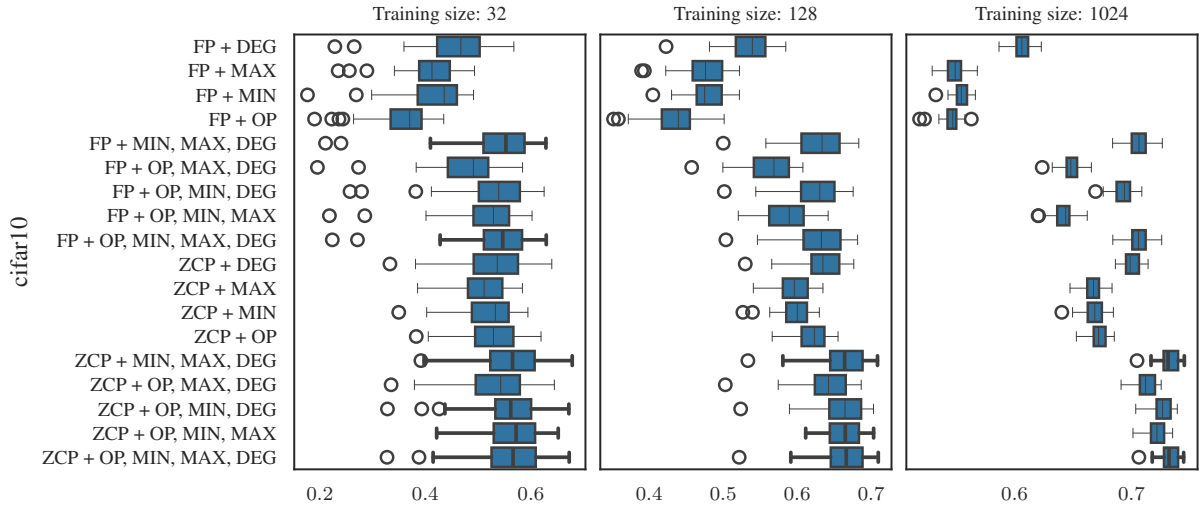


Figure 20. NB101 feature group ablation for `cifar-10`.

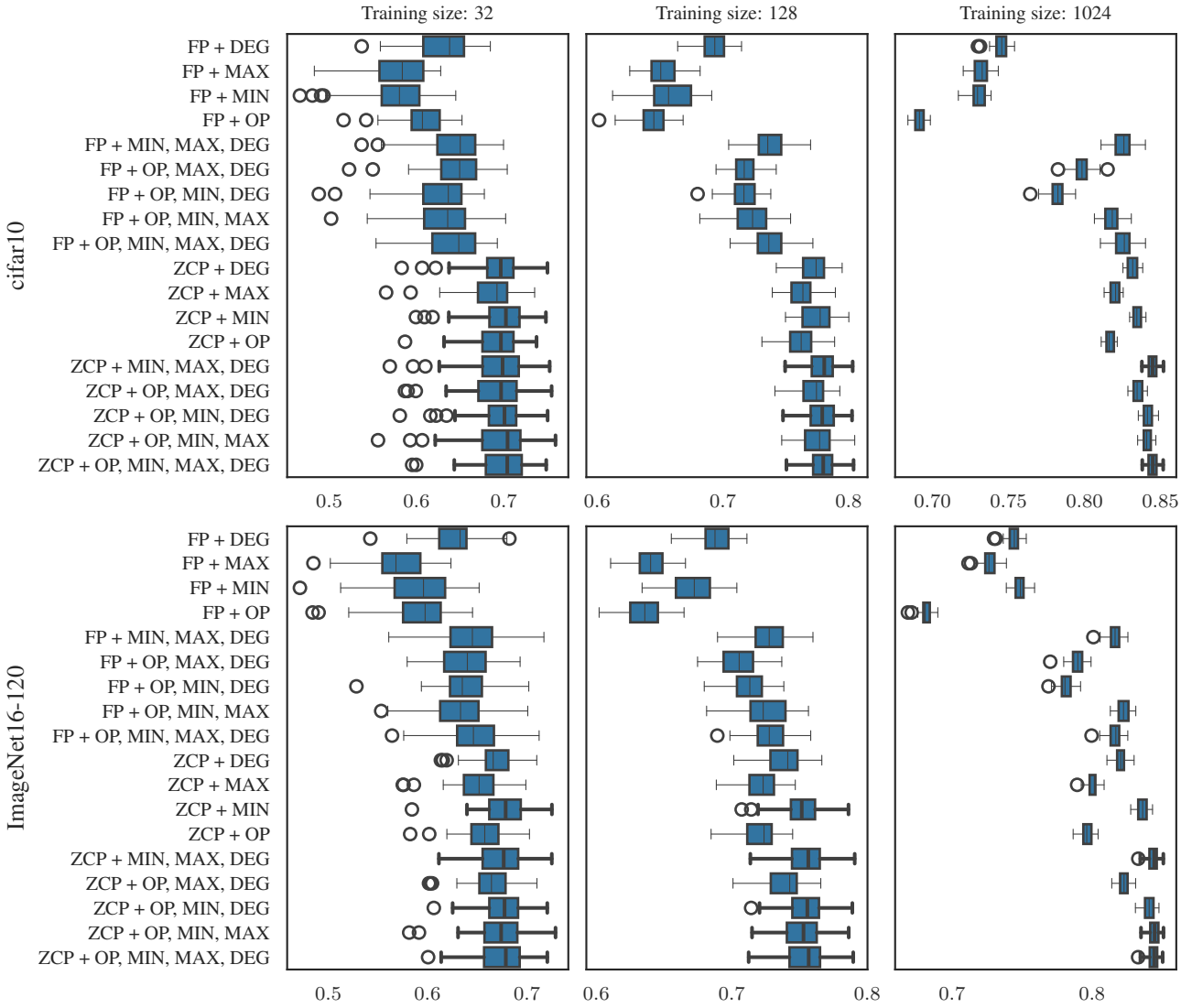


Figure 21. NB201 feature group ablation for `cifar-10` and `ImageNet16-120`.

I. Other Encoding Types

In this section, we list the full results of other encoding types (listed in Section C.4) in Tables 20 and 21. It is important to note that NASBOWL uses the WL features in a Gaussian process predictor, and their performance in a random forest might be limited.

Table 20. Results on the validation accuracy prediction tasks for different datasets on NB201 using Arch2vec and WL embedding. Average and standard deviation of Kendall tau over 50 independent runs. Bold values indicate the best results and those that are not statistically different from them.

dataset	cifar10			cifar100			ImageNet16-120		
train_size	32	128	1024	32	128	1024	32	128	1024
OH + WL	$0.54^{0.04}$	$0.67^{0.02}$	$0.78^{0.01}$	$0.55^{0.04}$	$0.67^{0.02}$	$0.78^{0.00}$	$0.55^{0.05}$	$0.66^{0.02}$	$0.77^{0.01}$
OH + FP + WL	$0.60^{0.04}$	$0.68^{0.01}$	$0.77^{0.01}$	$0.60^{0.03}$	$0.68^{0.01}$	$0.78^{0.01}$	$0.59^{0.04}$	$0.67^{0.02}$	$0.77^{0.01}$
GRAF + WL	$0.63^{0.04}$	$0.74^{0.01}$	$0.83^{0.01}$	$0.65^{0.03}$	$0.73^{0.01}$	$0.82^{0.01}$	$0.63^{0.03}$	$0.73^{0.02}$	$0.81^{0.01}$
GRAF + FP + WL	$0.64^{0.03}$	$0.74^{0.01}$	$0.82^{0.01}$	$0.66^{0.03}$	$0.73^{0.01}$	$0.82^{0.01}$	$0.64^{0.03}$	$0.73^{0.02}$	$0.82^{0.01}$
GRAF + OH + WL	$0.63^{0.04}$	$0.74^{0.01}$	$0.83^{0.01}$	$0.65^{0.03}$	$0.73^{0.01}$	$0.82^{0.01}$	$0.63^{0.03}$	$0.72^{0.02}$	$0.82^{0.01}$
GRAF + OH + FP + WL	$0.63^{0.04}$	$0.73^{0.01}$	$0.83^{0.01}$	$0.65^{0.03}$	$0.73^{0.01}$	$0.82^{0.01}$	$0.64^{0.03}$	$0.73^{0.02}$	$0.82^{0.01}$
ZCP + WL	$\mathbf{0.69^{0.03}}$	$0.76^{0.01}$	$0.82^{0.00}$	$\mathbf{0.70^{0.03}}$	$0.76^{0.01}$	$0.82^{0.00}$	$0.66^{0.02}$	$0.72^{0.01}$	$0.81^{0.00}$
ZCP + OH + WL	$0.69^{0.03}$	$0.76^{0.01}$	$0.83^{0.00}$	$\mathbf{0.69^{0.03}}$	$0.76^{0.01}$	$0.82^{0.00}$	$0.66^{0.02}$	$0.73^{0.01}$	$0.82^{0.00}$
ZCP + GRAF + WL	$\mathbf{0.69^{0.03}}$	$\mathbf{0.78^{0.01}}$	$\mathbf{0.85^{0.00}}$	$\mathbf{0.70^{0.03}}$	$\mathbf{0.77^{0.01}}$	$\mathbf{0.84^{0.00}}$	$\mathbf{0.67^{0.02}}$	$\mathbf{0.75^{0.02}}$	$\mathbf{0.84^{0.00}}$
ZCP + GRAF + OH + WL	$\mathbf{0.69^{0.04}}$	$0.78^{0.01}$	$\mathbf{0.84^{0.00}}$	$\mathbf{0.70^{0.04}}$	$0.77^{0.01}$	$\mathbf{0.84^{0.00}}$	$\mathbf{0.67^{0.03}}$	$\mathbf{0.75^{0.02}}$	$0.84^{0.00}$
OH + A2V	$0.56^{0.05}$	$0.65^{0.02}$	$0.72^{0.00}$	$0.57^{0.04}$	$0.66^{0.02}$	$0.73^{0.00}$	$0.56^{0.04}$	$0.64^{0.02}$	$0.72^{0.01}$
OH + FP + A2V	$0.58^{0.05}$	$0.66^{0.01}$	$0.73^{0.00}$	$0.60^{0.04}$	$0.67^{0.01}$	$0.74^{0.00}$	$0.59^{0.04}$	$0.66^{0.01}$	$0.73^{0.01}$
GRAF + A2V	$0.62^{0.05}$	$0.71^{0.02}$	$0.81^{0.01}$	$0.64^{0.04}$	$0.72^{0.02}$	$0.80^{0.01}$	$0.62^{0.03}$	$0.71^{0.02}$	$0.80^{0.01}$
GRAF + FP + A2V	$0.62^{0.04}$	$0.72^{0.02}$	$0.81^{0.01}$	$0.65^{0.03}$	$0.72^{0.01}$	$0.80^{0.01}$	$0.63^{0.03}$	$0.71^{0.02}$	$0.80^{0.01}$
GRAF + OH + A2V	$0.62^{0.04}$	$0.71^{0.02}$	$0.81^{0.01}$	$0.64^{0.04}$	$0.72^{0.02}$	$0.81^{0.01}$	$0.62^{0.03}$	$0.71^{0.02}$	$0.80^{0.01}$
GRAF + OH + FP + A2V	$0.62^{0.04}$	$0.72^{0.02}$	$0.81^{0.01}$	$0.64^{0.03}$	$0.72^{0.01}$	$0.81^{0.01}$	$0.63^{0.03}$	$0.71^{0.02}$	$0.80^{0.01}$
ZCP + A2V	$0.67^{0.04}$	$0.76^{0.01}$	$0.82^{0.00}$	$\mathbf{0.69^{0.03}}$	$0.75^{0.01}$	$0.81^{0.00}$	$0.65^{0.03}$	$0.72^{0.01}$	$0.80^{0.00}$
ZCP + OH + A2V	$0.67^{0.04}$	$0.76^{0.01}$	$0.82^{0.00}$	$0.68^{0.03}$	$0.76^{0.01}$	$0.82^{0.00}$	$0.65^{0.03}$	$0.72^{0.01}$	$0.81^{0.00}$
ZCP + GRAF + A2V	$0.68^{0.04}$	$0.77^{0.01}$	$0.84^{0.00}$	$\mathbf{0.69^{0.03}}$	$0.76^{0.01}$	$0.83^{0.00}$	$\mathbf{0.67^{0.03}}$	$0.75^{0.02}$	$0.84^{0.00}$
ZCP + GRAF + OH + A2V	$0.68^{0.04}$	$0.77^{0.01}$	$0.84^{0.00}$	$\mathbf{0.69^{0.03}}$	$0.76^{0.01}$	$0.83^{0.00}$	$\mathbf{0.67^{0.03}}$	$0.75^{0.02}$	$0.84^{0.00}$

Table 21. Results on the validation accuracy prediction tasks for cifar10 on NB101 using Arch2vec and WL embedding. Average and standard deviation of Kendall tau over 50 independent runs. Bold values indicate the best results and those that are not statistically different from them.

dataset	cifar10		
train_size	32	128	1024
GRAF + WL	$0.54^{0.06}$	$0.63^{0.03}$	$0.71^{0.01}$
GRAF + FP + WL	$0.53^{0.07}$	$0.63^{0.03}$	$0.71^{0.01}$
ZCP + WL	$0.51^{0.05}$	$0.62^{0.02}$	$0.68^{0.01}$
ZCP + GRAF + WL	$\mathbf{0.55^{0.07}}$	$\mathbf{0.67^{0.03}}$	$\mathbf{0.73^{0.01}}$
GRAF + A2V	$0.49^{0.09}$	$0.60^{0.04}$	$0.68^{0.01}$
GRAF + FP + A2V	$0.49^{0.09}$	$0.60^{0.04}$	$0.68^{0.01}$
ZCP + A2V	$0.45^{0.07}$	$0.56^{0.04}$	$0.63^{0.01}$
ZCP + GRAF + A2V	$0.52^{0.09}$	$0.64^{0.04}$	$0.72^{0.01}$

J. Hardware Metrics Results

In this section, we present the complete results from hardware metrics predictions. The experiment was realized using HW-NAS-Bench (Li et al., 2021). This benchmark contains networks from NB201 and provides ten hardware statistics on `cifar-10`, `cifar-100` and `ImageNet16-120`. The prediction tasks are of varying difficulty.

Figure 22 shows the results for `edgegpu_energy` prediction. The following Tables 22, 23 and 24 list average Kendall tau for selected nontrivial tasks and various predictors. In all cases, the best results are produced with settings containing GRAF among predictors. The similar results were obtained also on the rest of the tasks from HW-NAS-Bench, the only exceptions are 5 tasks, where onehot + FP produced the best result (from total 90 tasks; three datasets, three different training sizes, 10 hardware metrics).

In general, onehot is a good predictor for HW tasks (as shown in (Laube et al., 2022)), but on the majority of tasks, the prediction can be improved by adding GRAF to the predictors.

The computational cost of the experiment (50 evaluations for ten prediction tasks, three datasets and 10 settings, resulting in 15000 runs) was 1577 s for 32 training samples, 3087 s for 128 training samples, and 16 348 s for 1024 training samples on AMD Ryzen 7 3800X.

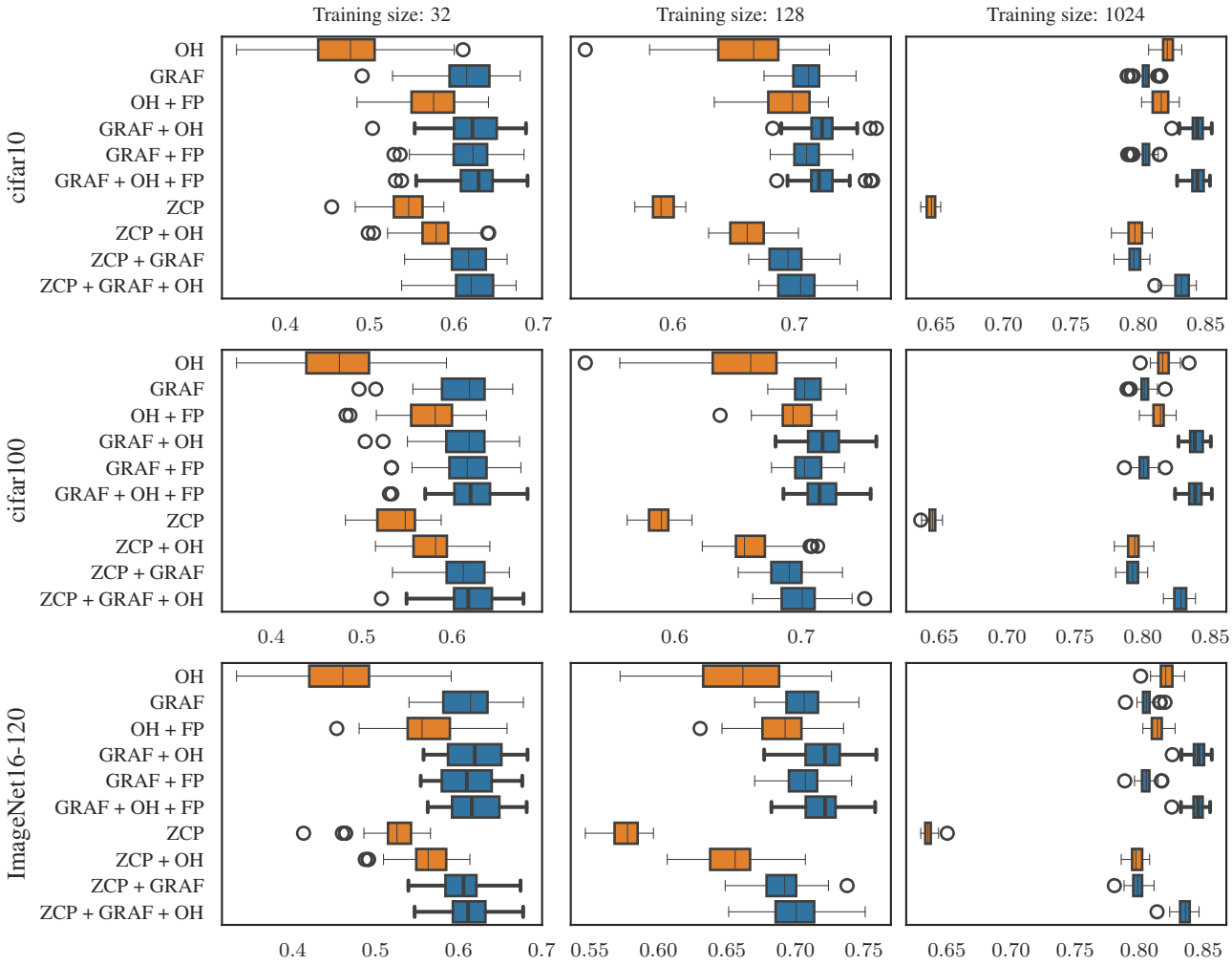


Figure 22. Random forest `edgegpu_energy` prediction on NB201 `cifar-10`, `cifar-100` and `ImageNet16-120` (different training sample sizes).

Table 22. Results on the `eyeriss_arithmetic_intensity` task for various datasets. Average and standard deviation of Kendall tau over 50 independent runs. Bold values indicate the best results and those that are not statistically different from them.

dataset	cifar10			cifar100			ImageNet16-120		
train_size	32	128	1024	32	128	1024	32	128	1024
OH	$0.53^{0.08}$	$0.73^{0.04}$	$0.88^{0.01}$	$0.53^{0.08}$	$0.73^{0.04}$	$0.88^{0.01}$	$0.53^{0.08}$	$0.73^{0.04}$	$0.88^{0.01}$
OH + FP	$0.65^{0.04}$	$0.77^{0.02}$	$0.94^{0.01}$	$0.65^{0.04}$	$0.77^{0.02}$	$0.94^{0.01}$	$0.64^{0.04}$	$0.76^{0.03}$	$0.92^{0.01}$
GRAF	$\mathbf{0.78^{0.05}}$	$\mathbf{0.92^{0.01}}$	$\mathbf{0.99^{0.00}}$	$\mathbf{0.78^{0.05}}$	$0.92^{0.01}$	$\mathbf{0.99^{0.00}}$	$0.78^{0.05}$	$\mathbf{0.93^{0.01}}$	$\mathbf{1.00^{0.00}}$
GRAF + FP	$\mathbf{0.78^{0.05}}$	$0.91^{0.02}$	$0.99^{0.00}$	$\mathbf{0.78^{0.05}}$	$0.91^{0.02}$	$0.99^{0.00}$	$\mathbf{0.79^{0.05}}$	$0.93^{0.02}$	$\mathbf{1.00^{0.00}}$
GRAF + OH	$\mathbf{0.78^{0.05}}$	$\mathbf{0.92^{0.02}}$	$\mathbf{0.99^{0.00}}$	$\mathbf{0.78^{0.05}}$	$\mathbf{0.92^{0.02}}$	$0.99^{0.00}$	$0.78^{0.05}$	$\mathbf{0.93^{0.01}}$	$\mathbf{1.00^{0.00}}$
GRAF + OH + FP	$\mathbf{0.78^{0.05}}$	$0.91^{0.02}$	$0.99^{0.00}$	$\mathbf{0.78^{0.05}}$	$0.91^{0.02}$	$0.99^{0.00}$	$\mathbf{0.79^{0.05}}$	$0.93^{0.02}$	$1.00^{0.00}$
ZCP	$0.61^{0.04}$	$0.72^{0.02}$	$0.88^{0.00}$	$0.61^{0.04}$	$0.72^{0.02}$	$0.88^{0.01}$	$0.62^{0.03}$	$0.70^{0.01}$	$0.85^{0.01}$
ZCP + OH	$0.63^{0.04}$	$0.74^{0.02}$	$0.89^{0.00}$	$0.63^{0.04}$	$0.74^{0.02}$	$0.89^{0.00}$	$0.64^{0.03}$	$0.73^{0.02}$	$0.88^{0.01}$
ZCP + GRAF	$0.75^{0.05}$	$0.89^{0.02}$	$0.99^{0.00}$	$0.75^{0.05}$	$0.89^{0.02}$	$0.99^{0.00}$	$\mathbf{0.78^{0.04}}$	$0.91^{0.02}$	$0.99^{0.00}$
ZCP + GRAF + OH	$0.75^{0.05}$	$0.89^{0.02}$	$0.99^{0.00}$	$0.75^{0.05}$	$0.89^{0.02}$	$0.99^{0.00}$	$\mathbf{0.78^{0.04}}$	$0.91^{0.02}$	$0.99^{0.00}$

Table 23. Results on the `pixel3_latency` task for various datasets. Average and standard deviation of Kendall tau over 50 independent runs. Bold values indicate the best results and those that are not statistically different from them.

dataset	cifar10			cifar100			ImageNet16-120		
train_size	32	128	1024	32	128	1024	32	128	1024
OH	$0.56^{0.07}$	$0.72^{0.03}$	$0.83^{0.00}$	$0.56^{0.08}$	$0.72^{0.03}$	$0.83^{0.00}$	$0.57^{0.08}$	$0.72^{0.03}$	$0.83^{0.00}$
OH + FP	$0.76^{0.02}$	$0.81^{0.01}$	$0.86^{0.00}$	$0.76^{0.02}$	$0.81^{0.01}$	$0.86^{0.00}$	$0.78^{0.02}$	$0.82^{0.01}$	$0.86^{0.00}$
GRAF	$\mathbf{0.80^{0.03}}$	$\mathbf{0.85^{0.01}}$	$0.87^{0.00}$	$\mathbf{0.79^{0.03}}$	$\mathbf{0.85^{0.01}}$	$0.86^{0.00}$	$\mathbf{0.81^{0.03}}$	$\mathbf{0.85^{0.01}}$	$0.87^{0.00}$
GRAF + FP	$\mathbf{0.80^{0.02}}$	$0.85^{0.01}$	$0.87^{0.00}$	$\mathbf{0.80^{0.02}}$	$0.84^{0.01}$	$0.86^{0.00}$	$\mathbf{0.81^{0.02}}$	$0.85^{0.01}$	$0.87^{0.00}$
GRAF + OH	$0.79^{0.03}$	$\mathbf{0.85^{0.01}}$	$0.87^{0.00}$	$\mathbf{0.79^{0.03}}$	$\mathbf{0.85^{0.01}}$	$0.87^{0.00}$	$\mathbf{0.80^{0.03}}$	$\mathbf{0.85^{0.01}}$	$0.87^{0.00}$
GRAF + OH + FP	$\mathbf{0.80^{0.02}}$	$0.85^{0.01}$	$0.87^{0.00}$	$\mathbf{0.80^{0.03}}$	$0.84^{0.01}$	$0.87^{0.00}$	$\mathbf{0.81^{0.03}}$	$0.85^{0.01}$	$0.87^{0.00}$
ZCP	$0.75^{0.02}$	$0.80^{0.01}$	$0.86^{0.00}$	$0.75^{0.02}$	$0.80^{0.01}$	$0.85^{0.00}$	$0.76^{0.02}$	$0.80^{0.01}$	$0.86^{0.00}$
ZCP + OH	$0.75^{0.02}$	$0.81^{0.01}$	$0.86^{0.00}$	$0.75^{0.02}$	$0.81^{0.01}$	$0.85^{0.00}$	$0.77^{0.02}$	$0.81^{0.01}$	$0.86^{0.00}$
ZCP + GRAF	$0.79^{0.02}$	$0.84^{0.01}$	$0.87^{0.00}$	$\mathbf{0.79^{0.02}}$	$0.84^{0.01}$	$\mathbf{0.87^{0.00}}$	$\mathbf{0.80^{0.03}}$	$0.85^{0.01}$	$0.87^{0.00}$
ZCP + GRAF + OH	$0.79^{0.02}$	$0.84^{0.01}$	$\mathbf{0.87^{0.00}}$	$\mathbf{0.79^{0.03}}$	$0.84^{0.01}$	$\mathbf{0.87^{0.00}}$	$0.80^{0.03}$	$0.85^{0.01}$	$\mathbf{0.87^{0.00}}$

Table 24. Results on the `raspi4_latency` task for various datasets. Average and standard deviation of Kendall tau over 50 independent runs. Bold values indicate the best results and those that are not statistically different from them.

dataset	cifar10			cifar100			ImageNet16-120		
train_size	32	128	1024	32	128	1024	32	128	1024
OH	$0.58^{0.08}$	$0.73^{0.03}$	$0.84^{0.00}$	$0.56^{0.08}$	$0.72^{0.03}$	$0.83^{0.00}$	$0.46^{0.09}$	$0.61^{0.02}$	$0.69^{0.00}$
OH + FP	$0.80^{0.01}$	$0.84^{0.01}$	$0.87^{0.00}$	$0.80^{0.02}$	$0.84^{0.01}$	$0.87^{0.00}$	$0.65^{0.03}$	$0.69^{0.01}$	$0.72^{0.00}$
GRAF	$0.80^{0.04}$	$\mathbf{0.86^{0.01}}$	$0.88^{0.00}$	$\mathbf{0.81^{0.04}}$	$\mathbf{0.86^{0.01}}$	$0.88^{0.00}$	$0.63^{0.04}$	$\mathbf{0.70^{0.01}}$	$0.73^{0.00}$
GRAF + FP	$\mathbf{0.81^{0.02}}$	$\mathbf{0.86^{0.01}}$	$0.88^{0.00}$	$\mathbf{0.81^{0.02}}$	$\mathbf{0.86^{0.00}}$	$0.88^{0.00}$	$\mathbf{0.66^{0.03}}$	$0.70^{0.01}$	$0.73^{0.00}$
GRAF + OH	$0.80^{0.03}$	$\mathbf{0.86^{0.01}}$	$\mathbf{0.88^{0.00}}$	$0.80^{0.03}$	$\mathbf{0.86^{0.01}}$	$0.88^{0.00}$	$0.63^{0.04}$	$\mathbf{0.70^{0.01}}$	$0.73^{0.00}$
GRAF + OH + FP	$\mathbf{0.81^{0.03}}$	$\mathbf{0.86^{0.01}}$	$0.88^{0.00}$	$\mathbf{0.81^{0.02}}$	$\mathbf{0.86^{0.00}}$	$0.88^{0.00}$	$\mathbf{0.66^{0.02}}$	$\mathbf{0.70^{0.01}}$	$0.73^{0.00}$
ZCP	$0.79^{0.02}$	$0.84^{0.01}$	$0.87^{0.00}$	$0.79^{0.02}$	$0.83^{0.01}$	$0.87^{0.00}$	$0.64^{0.03}$	$0.68^{0.01}$	$0.72^{0.00}$
ZCP + OH	$0.79^{0.02}$	$0.84^{0.01}$	$0.87^{0.00}$	$0.79^{0.02}$	$0.84^{0.01}$	$0.87^{0.00}$	$0.64^{0.03}$	$0.68^{0.01}$	$0.72^{0.00}$
ZCP + GRAF	$\mathbf{0.81^{0.02}}$	$\mathbf{0.86^{0.01}}$	$0.88^{0.00}$	$\mathbf{0.81^{0.02}}$	$0.86^{0.00}$	$\mathbf{0.88^{0.00}}$	$\mathbf{0.66^{0.03}}$	$\mathbf{0.70^{0.01}}$	$0.73^{0.00}$
ZCP + GRAF + OH	$\mathbf{0.81^{0.02}}$	$\mathbf{0.86^{0.01}}$	$\mathbf{0.88^{0.00}}$	$\mathbf{0.81^{0.02}}$	$0.86^{0.00}$	$\mathbf{0.88^{0.00}}$	$\mathbf{0.66^{0.02}}$	$\mathbf{0.70^{0.01}}$	$\mathbf{0.73^{0.00}}$

K. Robustness tasks

Here we present the evaluations on the robustness task on `cifar-10` on a subset of results for clarity. The Figure 23 presents the results on the FGSM and PGD attacks for the perturbation strengths $\epsilon = \{0.1/255, 0.5/255, 1.0/255\}$ and the training sizes $32, 128, 1024$ for the first task, predicting only the robust accuracy. The combination of all ZCP, the proposed graph features, and additionally onehot encoding leads to the highest Kendall tau in almost all cases. We present the results on the PGD attack also in Tab. 25.

If we turn to the multi-objective case, in which the prediction targets are both the validation **and** robust accuracy, we see the same conclusions (see Figs. 24 for visualizations of fgsm and pgd, and Tab. 26.)

Furthermore, we can see that the multi-objective task has in general a higher Kendall tau than the single robust accuracy prediction task. Interestingly, the prediction task for the attack strength $\epsilon = 0.5/255$, shows in most cases the lowest correlation, though $\epsilon = 1.0/255$ is the stronger perturbation attack.

The computational cost of the experiment (50 evaluations for two adversarial prediction tasks (single and multi-objectives), two adversarial attacks, three perturbations strengths and three training sizes resulting in 18 000 runs) was 677 s for 32 training samples, 1 537 s for 128 training samples, and 11 404 s for 1024 training samples.

Table 25. Results on predicting the `pgd` adversarial accuracy on cifar10 for various values of ϵ on the NB201 benchmark. Average and standard deviation of Kendall tau over 50 independent runs. Bold values indicate the best results and those that are not statistically different from them.

dataset	pgd@Linf_eps-0.1			pgd@Linf_eps-0.5			pgd@Linf_eps-1.0		
train_size	32	128	1024	32	128	1024	32	128	1024
OH	$0.38^{0.06}$	$0.55^{0.03}$	$0.74^{0.01}$	$0.21^{0.04}$	$0.32^{0.02}$	$0.48^{0.01}$	$0.19^{0.05}$	$0.31^{0.03}$	$0.51^{0.02}$
OH + FP	$0.57^{0.03}$	$0.65^{0.01}$	$0.74^{0.01}$	$0.27^{0.04}$	$0.37^{0.02}$	$0.52^{0.01}$	$0.28^{0.05}$	$0.41^{0.03}$	$0.57^{0.01}$
GRAF	$0.64^{0.03}$	$0.73^{0.01}$	$0.82^{0.01}$	$0.32^{0.04}$	$0.46^{0.03}$	$0.63^{0.01}$	$0.32^{0.06}$	$0.52^{0.03}$	$0.67^{0.01}$
GRAF + FP	$0.65^{0.03}$	$0.73^{0.01}$	$0.82^{0.01}$	$0.32^{0.05}$	$0.47^{0.02}$	$0.63^{0.01}$	$0.33^{0.06}$	$0.52^{0.03}$	$0.66^{0.01}$
GRAF + OH	$0.64^{0.03}$	$0.73^{0.01}$	$0.82^{0.01}$	$0.32^{0.04}$	$0.46^{0.02}$	$0.63^{0.01}$	$0.31^{0.06}$	$0.51^{0.03}$	$0.67^{0.01}$
GRAF + OH + FP	$0.65^{0.03}$	$0.73^{0.01}$	$0.82^{0.01}$	$0.32^{0.05}$	$0.46^{0.02}$	$0.62^{0.01}$	$0.33^{0.06}$	$0.52^{0.03}$	$0.66^{0.01}$
ZCP	$\mathbf{0.69^{0.02}}$	$0.75^{0.01}$	$0.80^{0.00}$	$\mathbf{0.35^{0.05}}$	$0.44^{0.02}$	$0.55^{0.00}$	$\mathbf{0.34^{0.06}}$	$0.47^{0.03}$	$0.59^{0.01}$
ZCP + OH	$0.68^{0.03}$	$0.75^{0.01}$	$0.81^{0.00}$	$\mathbf{0.35^{0.03}}$	$0.45^{0.02}$	$0.58^{0.01}$	$0.33^{0.05}$	$0.48^{0.02}$	$0.62^{0.01}$
ZCP + GRAF	$\mathbf{0.69^{0.02}}$	$\mathbf{0.77^{0.01}}$	$0.83^{0.00}$	$\mathbf{0.36^{0.04}}$	$\mathbf{0.50^{0.02}}$	$\mathbf{0.64^{0.01}}$	$\mathbf{0.36^{0.06}}$	$\mathbf{0.56^{0.03}}$	$\mathbf{0.70^{0.01}}$
ZCP + GRAF + OH	$\mathbf{0.69^{0.03}}$	$0.77^{0.01}$	$\mathbf{0.83^{0.00}}$	$\mathbf{0.36^{0.04}}$	$\mathbf{0.50^{0.02}}$	$\mathbf{0.64^{0.01}}$	$\mathbf{0.35^{0.06}}$	$\mathbf{0.56^{0.03}}$	$0.70^{0.01}$

Table 26. Results on predicting both the `pgd` adversarial accuracy and clean validation accuracy on cifar10 for various values of ϵ on the NB201 benchmark. Average and standard deviation of Kendall tau over 50 independent runs. Bold values indicate the best results and those that are not statistically different from them.

dataset	pgd@Linf_eps-0.1			pgd@Linf_eps-0.5			pgd@Linf_eps-1.0		
train_size	32	128	1024	32	128	1024	32	128	1024
OH (MO)	$0.39^{0.06}$	$0.55^{0.03}$	$0.74^{0.01}$	$0.29^{0.04}$	$0.44^{0.02}$	$0.63^{0.01}$	$0.28^{0.04}$	$0.43^{0.02}$	$0.63^{0.01}$
OH + FP (MO)	$0.56^{0.04}$	$0.65^{0.01}$	$0.74^{0.01}$	$0.41^{0.04}$	$0.50^{0.01}$	$0.63^{0.01}$	$0.41^{0.04}$	$0.52^{0.02}$	$0.65^{0.01}$
GRAF (MO)	$0.64^{0.03}$	$0.73^{0.01}$	$0.82^{0.01}$	$0.49^{0.03}$	$0.61^{0.02}$	$0.73^{0.00}$	$0.49^{0.04}$	$0.63^{0.02}$	$0.74^{0.01}$
GRAF + FP (MO)	$0.64^{0.03}$	$0.73^{0.01}$	$0.82^{0.01}$	$0.50^{0.03}$	$0.61^{0.02}$	$0.72^{0.00}$	$0.50^{0.04}$	$0.63^{0.02}$	$0.74^{0.00}$
GRAF + OH (MO)	$0.63^{0.03}$	$0.73^{0.01}$	$0.83^{0.01}$	$0.49^{0.03}$	$0.61^{0.02}$	$0.73^{0.01}$	$0.49^{0.04}$	$0.62^{0.02}$	$0.74^{0.01}$
GRAF + OH + FP (MO)	$0.64^{0.04}$	$0.73^{0.01}$	$0.82^{0.01}$	$0.49^{0.03}$	$0.61^{0.02}$	$0.72^{0.00}$	$0.50^{0.04}$	$0.62^{0.02}$	$0.74^{0.01}$
ZCP (MO)	$\mathbf{0.69^{0.02}}$	$0.76^{0.01}$	$0.81^{0.00}$	$0.51^{0.03}$	$0.60^{0.01}$	$0.68^{0.00}$	$0.51^{0.03}$	$0.61^{0.02}$	$0.70^{0.00}$
ZCP + OH (MO)	$0.68^{0.03}$	$0.76^{0.01}$	$0.82^{0.00}$	$0.52^{0.03}$	$0.61^{0.01}$	$0.70^{0.00}$	$0.51^{0.03}$	$0.61^{0.02}$	$0.72^{0.00}$
ZCP + GRAF (MO)	$\mathbf{0.69^{0.03}}$	$\mathbf{0.77^{0.01}}$	$0.84^{0.00}$	$\mathbf{0.54^{0.03}}$	$\mathbf{0.64^{0.02}}$	$\mathbf{0.75^{0.00}}$	$\mathbf{0.53^{0.04}}$	$\mathbf{0.66^{0.02}}$	$\mathbf{0.77^{0.00}}$
ZCP + GRAF + OH (MO)	$\mathbf{0.69^{0.03}}$	$\mathbf{0.77^{0.01}}$	$\mathbf{0.84^{0.00}}$	$\mathbf{0.53^{0.04}}$	$\mathbf{0.64^{0.02}}$	$\mathbf{0.75^{0.00}}$	$0.53^{0.04}$	$\mathbf{0.66^{0.02}}$	$\mathbf{0.77^{0.00}}$

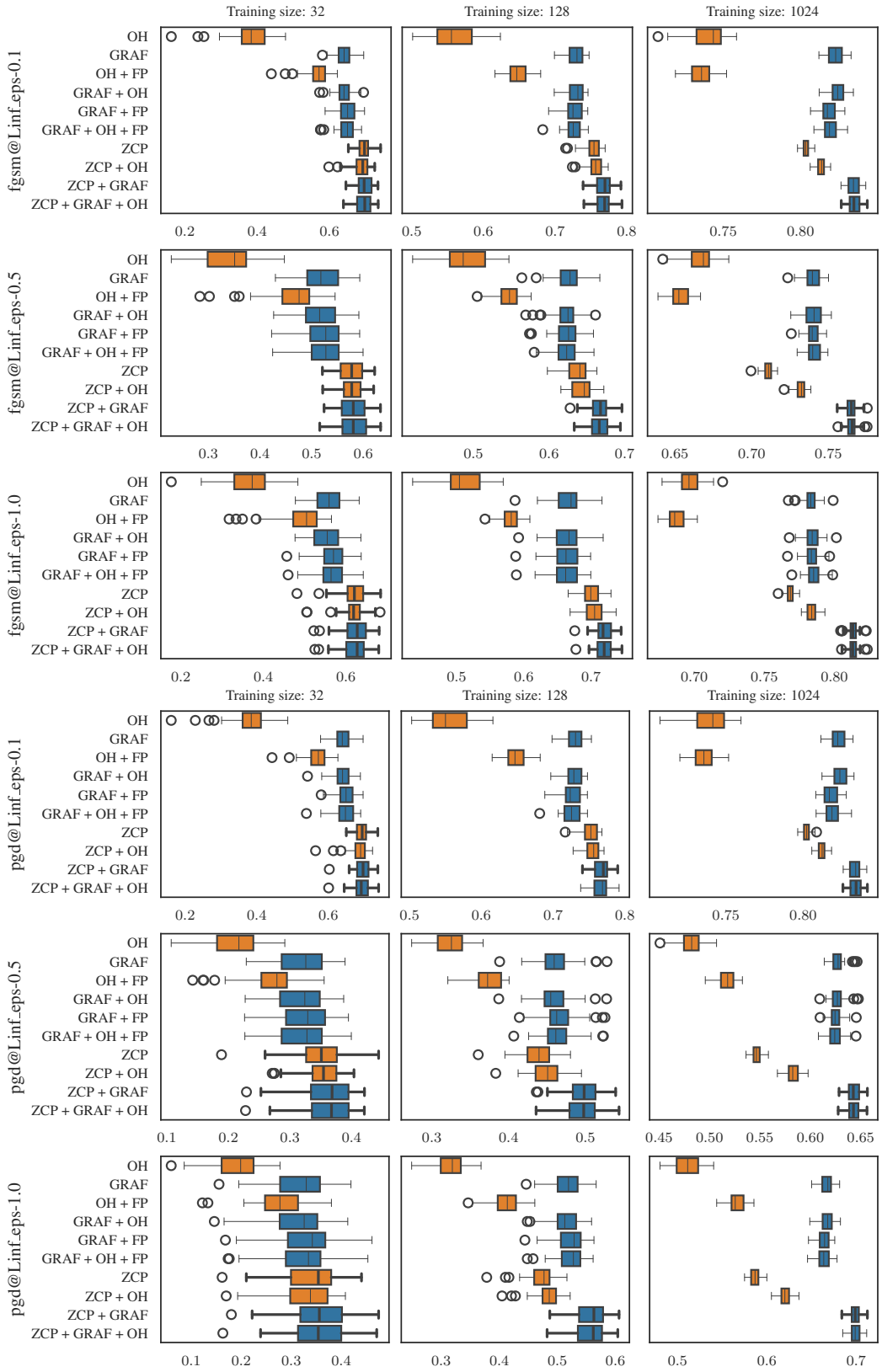


Figure 23. Results on predicting the robust accuracy on cifar-10 for the two different adversarial attacks (**FGSM, PGD**) and different strengths of perturbations from Jung et al. (2023).

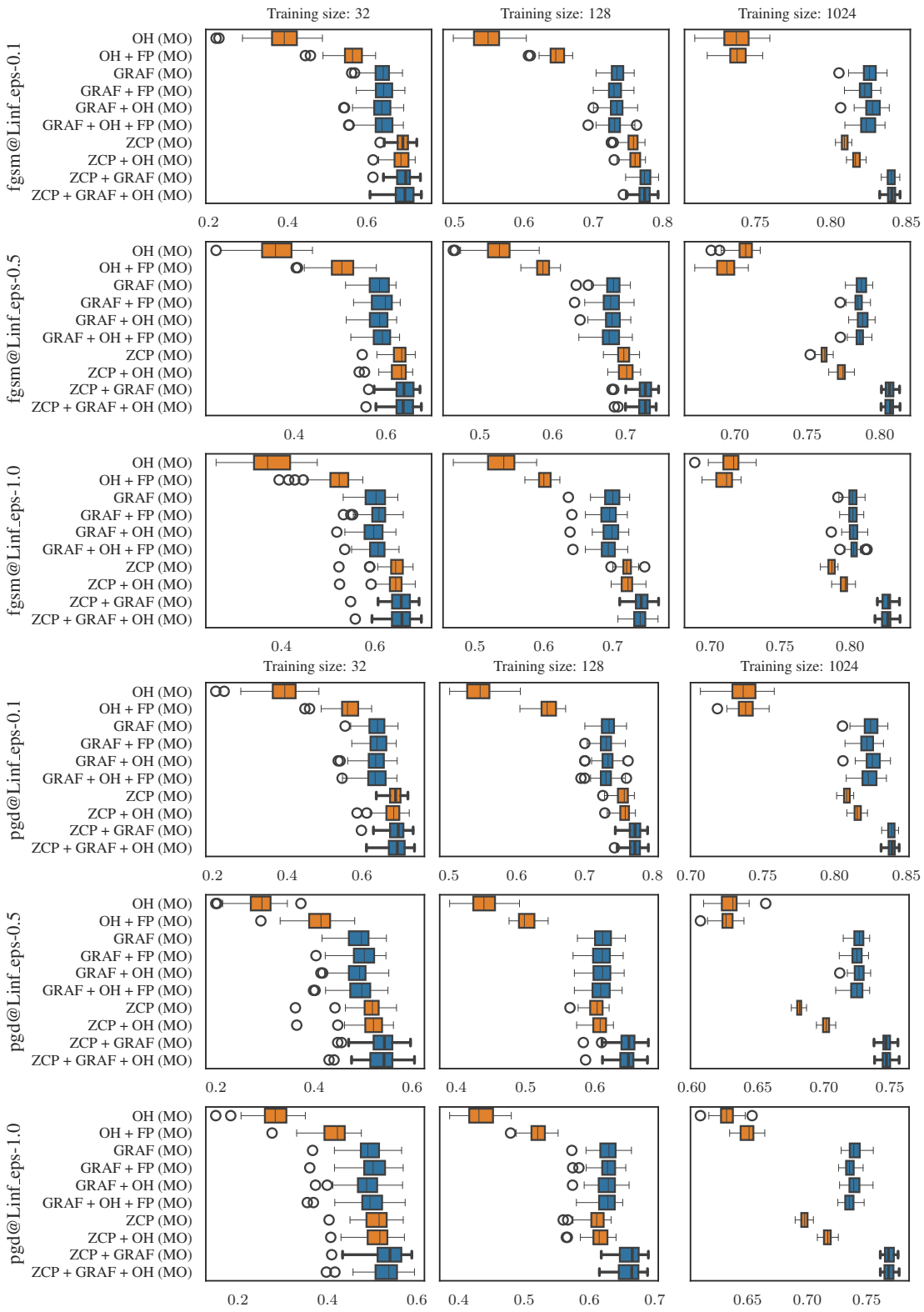


Figure 24. Results on the multi-objective prediction tasks for the validation **and** robust accuracy on cifar-10 for the two different adversarial attacks (**FGSM, PGD**) and different strengths of perturbations from Jung et al. (2023).

L. BRP-NAS full results

The BRP-NAS model (Dudziak et al., 2020) is a graph convolutional network connected to a single-layer MLP, i.e. a linear regressor on the embedding of the architecture. The additional features are added using so-called *augments*, also supported by the original BRP-NAS model. These get appended to the embedding produced by the GCN. We standardize these features before appending them.

The BRP-NAS model can be used in two modes – either as a binary classifier on two inputs or as a regressor that predicts the given target. We use the latter approach for better comparison with the random forest models used in the other experiments.

Table 27. BRP-NAS hyper-parameters

Model Parameters	
Parameter	Value
num_layers	4
num_hidden	600
dropout_ratio	0.002
weight_init	thomas
bias_init	thomas
Training Parameters	
Parameter	Value
epochs	128
learning_rate	4×10^{-4}
weight_decay	5×10^{-5}
lr_patience	10
batch_size	1 (train = 32) or 16 (otherwise)
optim_name	adamw
lr_scheduler	plateau

The hyper-parameters of the BRP-NAS model are in Table 27. These are based on the settings published together with the original model, but to make the experiments more comparable to the other experiments in this paper, we do not use any validation set and early stopping, and always run the experiments for 128 epochs. The number of epochs was set during preliminary experiments, when we noticed that the early stopping typically stops the training after roughly this number of epochs. We use different batch size for the experiments with only 32 training networks, as for this setting we observed much worse performance with the larger bach size used in the other experiments. Experiments with 128 and 1024 training networks use batch size of 16 – we did not observe any difference in the performance between the batch size 1 and 16 in these experiments, and the experiments with larger step size run faster.

Every setting of BRP-NAS was trained 50 times with different, independently sampled training sets. All the results were first evaluated using the Kruskal test to assess statistical significance, then the Mann-Whitney rank sum test was used as a posthoc test with significance level $\alpha = 0.05$ and Holm correction for multiple testing. The statistical treatment is different for BRP-NAS compared to the other experiments, as with BRP-NAS, we do not use the same training sets for runs with different sets of features. Therefore, we use unpaired statistical tests instead of paired ones.

Training the BRP-NAS models is much slower than training the tree-based models. We run the experiments in a heterogeneous cluster environment. All the experiments always run on a single CPU and a single GPU. A single training and evaluation run on the NB201 benchmark with 32 training networks takes roughly between 35s on an NVIDIA A100-SXM4-40GB and 70s on an NVIDIA GeForce GTX 1080 Ti. With 128 training networks, it takes similar time between 40s and 70s thanks to the larger batch size. With 1024 training networks, the runs take between 3 minutes on an NVIDIA A100-SXM4-40GB and around 6.5 minutes on an NVIDIA Tesla T4.

The experiments on the NB301 dataset are even slower, mainly due to the large number of features. While without the additional features the experiment times are similar to those on NB201, with the full set of features, they take between 4 and 6 minutes on an NVIDIA GeForce GTX 1080 Ti even for the smallest training set. For 128 training networks with all features each run takes between 6 and 7 minutes on an NVIDIA Tesla T4 and for the largest training set, they take between 20 minutes on an NVIDIA A40 and 37 minutes on an NVIDIA GeForce GTX 1080 Ti.

Overall, we needed around 100 GPU days to finish all these experiments (including initial testing, development, and results not included in the paper for brevity).

Table 28. BRP-NAS model results for validation accuracy prediction on cifar10 and ImageNet16-120 on NB201. Average and standard deviation of Kendall tau over 50 independent runs. Bold values indicate the best results and those that are not statistically different from them.

dataset	NB201: cifar10			NB201: ImageNet16-120			NB301: cifar10		
train_size	32	128	1024	32	128	1024	32	128	1024
None (BRP-NAS)	$0.48^{0.05}$	$0.58^{0.04}$	$0.80^{0.02}$	$0.48^{0.06}$	$0.61^{0.03}$	$0.84^{0.01}$	$0.37^{0.05}$	$0.48^{0.02}$	$0.58^{0.01}$
OH	$0.53^{0.05}$	$0.63^{0.02}$	$0.77^{0.02}$	$0.54^{0.04}$	$0.63^{0.02}$	$0.83^{0.01}$	$0.21^{0.08}$	$0.46^{0.03}$	$0.59^{0.01}$
OH + FP	$0.57^{0.04}$	$0.64^{0.02}$	$0.77^{0.02}$	$0.54^{0.04}$	$0.64^{0.02}$	$0.83^{0.02}$	$0.28^{0.06}$	$0.47^{0.02}$	$0.58^{0.01}$
GRAF	$0.62^{0.04}$	$0.69^{0.01}$	$0.79^{0.01}$	$0.62^{0.04}$	$0.69^{0.02}$	$0.83^{0.01}$	$\mathbf{0.38^{0.05}}$	$\mathbf{0.54^{0.03}}$	$0.67^{0.01}$
GRAF + FP	$0.62^{0.03}$	$0.69^{0.02}$	$0.78^{0.01}$	$0.62^{0.03}$	$0.69^{0.02}$	$0.83^{0.01}$	$\mathbf{0.37^{0.07}}$	$\mathbf{0.55^{0.02}}$	$\mathbf{0.67^{0.01}}$
GRAF + OH	$0.59^{0.05}$	$0.69^{0.02}$	$0.79^{0.01}$	$0.61^{0.04}$	$0.69^{0.02}$	$0.83^{0.01}$	$\mathbf{0.35^{0.09}}$	$\mathbf{0.54^{0.02}}$	$0.67^{0.01}$
GRAF + OH + FP	$0.61^{0.04}$	$0.69^{0.02}$	$0.79^{0.01}$	$0.61^{0.04}$	$0.68^{0.02}$	$0.83^{0.02}$	$0.35^{0.07}$	$\mathbf{0.54^{0.03}}$	$0.66^{0.01}$
ZCP	$\mathbf{0.63^{0.03}}$	$0.66^{0.02}$	$0.80^{0.02}$	$0.56^{0.04}$	$0.63^{0.03}$	$\mathbf{0.85^{0.01}}$	$\mathbf{0.40^{0.06}}$	$0.51^{0.02}$	$0.61^{0.00}$
ZCP + OH	$\mathbf{0.64^{0.04}}$	$0.68^{0.02}$	$0.79^{0.01}$	$0.58^{0.03}$	$0.66^{0.02}$	$0.83^{0.02}$	$0.32^{0.06}$	$0.50^{0.02}$	$0.61^{0.01}$
ZCP + GRAF	$\mathbf{0.65^{0.05}}$	$\mathbf{0.72^{0.02}}$	$\mathbf{0.81^{0.01}}$	$\mathbf{0.64^{0.03}}$	$\mathbf{0.71^{0.02}}$	$0.84^{0.01}$	$0.35^{0.06}$	$\mathbf{0.55^{0.02}}$	$\mathbf{0.67^{0.01}}$
ZCP + GRAF + OH	$\mathbf{0.64^{0.04}}$	$\mathbf{0.72^{0.02}}$	$\mathbf{0.81^{0.01}}$	$\mathbf{0.63^{0.03}}$	$\mathbf{0.71^{0.02}}$	$0.84^{0.01}$	$\mathbf{0.37^{0.07}}$	$\mathbf{0.55^{0.02}}$	$0.67^{0.01}$

Table 29. BRP-NAS model results for various hardware tasks on cifar10 and NB201. Average and standard deviation of Kendall tau over 50 independent runs. Bold values indicate the best results and those that are not statistically different from them.

dataset	edgegpu_energy			edgegpu_latency			edgetpu_latency		
train_size	32	128	1024	32	128	1024	32	128	1024
None (BRP-NAS)	$0.59^{0.05}$	$0.71^{0.01}$	$0.80^{0.00}$	$0.59^{0.05}$	$0.72^{0.01}$	$0.80^{0.00}$	$\mathbf{0.59^{0.08}}$	$0.69^{0.04}$	$\mathbf{0.84^{0.02}}$
OH	$0.67^{0.05}$	$0.77^{0.01}$	$0.87^{0.00}$	$\mathbf{0.73^{0.04}}$	$\mathbf{0.83^{0.01}}$	$\mathbf{0.87^{0.00}}$	$\mathbf{0.55^{0.13}}$	$\mathbf{0.73^{0.04}}$	$\mathbf{0.83^{0.02}}$
OH + FP	$0.69^{0.04}$	$0.78^{0.01}$	$\mathbf{0.87^{0.00}}$	$\mathbf{0.74^{0.03}}$	$\mathbf{0.83^{0.01}}$	$\mathbf{0.87^{0.00}}$	$\mathbf{0.56^{0.12}}$	$\mathbf{0.73^{0.04}}$	$\mathbf{0.83^{0.02}}$
GRAF	$0.75^{0.03}$	$0.81^{0.01}$	$0.86^{0.00}$	$\mathbf{0.73^{0.04}}$	$0.82^{0.01}$	$0.87^{0.00}$	$0.49^{0.16}$	$\mathbf{0.73^{0.04}}$	$0.81^{0.02}$
GRAF + FP	$0.75^{0.02}$	$0.82^{0.01}$	$0.86^{0.00}$	$0.73^{0.03}$	$0.82^{0.01}$	$0.87^{0.00}$	$0.44^{0.16}$	$\mathbf{0.73^{0.03}}$	$0.81^{0.02}$
GRAF + OH	$\mathbf{0.77^{0.03}}$	$\mathbf{0.83^{0.01}}$	$\mathbf{0.87^{0.00}}$	$\mathbf{0.75^{0.03}}$	$\mathbf{0.83^{0.01}}$	$\mathbf{0.87^{0.00}}$	$0.41^{0.15}$	$\mathbf{0.72^{0.04}}$	$0.81^{0.02}$
GRAF + OH + FP	$\mathbf{0.77^{0.02}}$	$\mathbf{0.83^{0.01}}$	$\mathbf{0.87^{0.00}}$	$\mathbf{0.75^{0.04}}$	$\mathbf{0.83^{0.01}}$	$\mathbf{0.87^{0.00}}$	$0.42^{0.14}$	$\mathbf{0.72^{0.04}}$	$0.81^{0.02}$
ZCP	$0.63^{0.05}$	$0.72^{0.01}$	$0.81^{0.00}$	$0.62^{0.05}$	$0.72^{0.01}$	$0.81^{0.00}$	$\mathbf{0.60^{0.09}}$	$0.70^{0.04}$	$\mathbf{0.83^{0.02}}$
ZCP + OH	$0.71^{0.04}$	$0.78^{0.01}$	$0.87^{0.00}$	$0.72^{0.05}$	$0.83^{0.01}$	$\mathbf{0.87^{0.00}}$	$0.47^{0.13}$	$\mathbf{0.72^{0.04}}$	$0.82^{0.02}$
ZCP + GRAF	$0.74^{0.03}$	$0.81^{0.01}$	$0.86^{0.00}$	$0.69^{0.05}$	$0.82^{0.01}$	$0.87^{0.00}$	$0.45^{0.12}$	$\mathbf{0.72^{0.03}}$	$0.81^{0.02}$
ZCP + GRAF + OH	$0.75^{0.04}$	$0.83^{0.01}$	$\mathbf{0.87^{0.00}}$	$0.72^{0.05}$	$0.82^{0.01}$	$0.87^{0.00}$	$0.41^{0.16}$	$0.71^{0.04}$	$0.81^{0.02}$

Table 30. BRP-NAS model results on the `fgsm` adversarial accuracy on cifar10 for various values of ϵ on the NB201 benchmark. Average and standard deviation of Kendall tau over 50 independent runs. Bold values indicate the best results and those that are not statistically different from them.

dataset	fgsm@Linf_eps-0.1			fgsm@Linf_eps-0.5			fgsm@Linf_eps-1.0		
train_size	32	128	1024	32	128	1024	32	128	1024
None (BRP-NAS)	$0.48^{0.05}$	$0.58^{0.04}$	$\mathbf{0.80^{0.01}}$	$0.42^{0.07}$	$0.55^{0.03}$	$0.74^{0.01}$	$0.44^{0.06}$	$0.60^{0.03}$	$0.78^{0.01}$
OH	$0.53^{0.05}$	$0.64^{0.03}$	$0.78^{0.02}$	$0.44^{0.05}$	$0.55^{0.04}$	$\mathbf{0.74^{0.01}}$	$0.47^{0.06}$	$0.60^{0.02}$	$0.79^{0.01}$
OH + FP	$0.56^{0.04}$	$0.65^{0.02}$	$0.78^{0.02}$	$0.45^{0.05}$	$0.57^{0.02}$	$0.74^{0.01}$	$0.48^{0.05}$	$\mathbf{0.61^{0.02}}$	$0.79^{0.01}$
GRAF	$0.60^{0.04}$	$0.69^{0.02}$	$0.78^{0.02}$	$\mathbf{0.50^{0.04}}$	$\mathbf{0.58^{0.02}}$	$0.69^{0.03}$	$\mathbf{0.54^{0.05}}$	$\mathbf{0.62^{0.02}}$	$0.77^{0.03}$
GRAF + FP	$\mathbf{0.61^{0.05}}$	$0.69^{0.02}$	$0.78^{0.01}$	$\mathbf{0.48^{0.06}}$	$0.58^{0.01}$	$0.69^{0.02}$	$\mathbf{0.55^{0.05}}$	$\mathbf{0.62^{0.02}}$	$0.77^{0.03}$
GRAF + OH	$0.58^{0.04}$	$0.68^{0.02}$	$0.79^{0.01}$	$\mathbf{0.47^{0.06}}$	$0.58^{0.02}$	$0.70^{0.02}$	$\mathbf{0.54^{0.04}}$	$\mathbf{0.62^{0.02}}$	$0.77^{0.03}$
GRAF + OH + FP	$0.59^{0.05}$	$0.68^{0.02}$	$0.78^{0.02}$	$\mathbf{0.49^{0.05}}$	$0.58^{0.02}$	$0.69^{0.02}$	$\mathbf{0.53^{0.05}}$	$\mathbf{0.62^{0.02}}$	$0.76^{0.04}$
ZCP	$\mathbf{0.62^{0.04}}$	$0.67^{0.02}$	$0.80^{0.01}$	$\mathbf{0.48^{0.06}}$	$0.56^{0.03}$	$\mathbf{0.74^{0.01}}$	$0.50^{0.04}$	$\mathbf{0.61^{0.03}}$	$\mathbf{0.79^{0.01}}$
ZCP + OH	$\mathbf{0.62^{0.04}}$	$0.68^{0.02}$	$0.79^{0.01}$	$\mathbf{0.48^{0.04}}$	$0.58^{0.02}$	$\mathbf{0.74^{0.01}}$	$0.51^{0.06}$	$\mathbf{0.62^{0.02}}$	$\mathbf{0.79^{0.01}}$
ZCP + GRAF	$\mathbf{0.63^{0.04}}$	$\mathbf{0.71^{0.02}}$	$\mathbf{0.80^{0.02}}$	$\mathbf{0.50^{0.05}}$	$\mathbf{0.59^{0.01}}$	$0.70^{0.02}$	$\mathbf{0.55^{0.06}}$	$\mathbf{0.62^{0.02}}$	$0.77^{0.02}$
ZCP + GRAF + OH	$\mathbf{0.62^{0.03}}$	$\mathbf{0.71^{0.02}}$	$\mathbf{0.81^{0.01}}$	$\mathbf{0.49^{0.05}}$	$\mathbf{0.59^{0.01}}$	$0.70^{0.02}$	$\mathbf{0.54^{0.04}}$	$\mathbf{0.62^{0.02}}$	$0.77^{0.03}$

Table 31. BRP-NAS model results on the `pgd` adversarial accuracy on cifar10 for various values of ϵ on the NB201 benchmark. Average and standard deviation of Kendall tau over 50 independent runs. Bold values indicate the best results and those that are not statistically different from them.

dataset	pgd@Linf_eps-0.1			pgd@Linf_eps-0.5			pgd@Linf_eps-1.0		
train_size	32	128	1024	32	128	1024	32	128	1024
None (BRP-NAS)	$0.48^{0.05}$	$0.58^{0.05}$	$0.79^{0.02}$	$0.24^{0.07}$	$0.36^{0.04}$	$\mathbf{0.61^{0.01}}$	$0.21^{0.06}$	$\mathbf{0.37^{0.06}}$	$0.68^{0.01}$
OH	$0.53^{0.04}$	$0.64^{0.03}$	$0.78^{0.02}$	$0.24^{0.06}$	$0.37^{0.03}$	$\mathbf{0.61^{0.02}}$	$0.22^{0.07}$	$0.36^{0.05}$	$\mathbf{0.68^{0.01}}$
OH + FP	$0.56^{0.03}$	$0.65^{0.02}$	$0.78^{0.02}$	$0.25^{0.07}$	$0.36^{0.03}$	$\mathbf{0.60^{0.02}}$	$0.24^{0.05}$	$\mathbf{0.38^{0.05}}$	$\mathbf{0.68^{0.01}}$
GRAF	$0.61^{0.04}$	$0.69^{0.02}$	$0.78^{0.01}$	$\mathbf{0.28^{0.07}}$	$\mathbf{0.39^{0.02}}$	$0.55^{0.03}$	$\mathbf{0.28^{0.05}}$	$\mathbf{0.39^{0.04}}$	$0.66^{0.04}$
GRAF + FP	$0.60^{0.05}$	$0.69^{0.01}$	$0.79^{0.01}$	$\mathbf{0.28^{0.08}}$	$\mathbf{0.39^{0.02}}$	$0.53^{0.03}$	$\mathbf{0.27^{0.07}}$	$\mathbf{0.39^{0.03}}$	$0.66^{0.05}$
GRAF + OH	$0.59^{0.04}$	$0.68^{0.02}$	$0.79^{0.01}$	$\mathbf{0.29^{0.05}}$	$\mathbf{0.39^{0.02}}$	$0.54^{0.04}$	$\mathbf{0.27^{0.06}}$	$\mathbf{0.39^{0.04}}$	$0.66^{0.04}$
GRAF + OH + FP	$0.60^{0.05}$	$0.68^{0.02}$	$0.79^{0.01}$	$\mathbf{0.31^{0.04}}$	$\mathbf{0.39^{0.02}}$	$0.54^{0.03}$	$0.25^{0.09}$	$\mathbf{0.38^{0.03}}$	$0.65^{0.05}$
ZCP	$\mathbf{0.64^{0.04}}$	$0.67^{0.02}$	$\mathbf{0.80^{0.02}}$	$0.25^{0.07}$	$0.37^{0.03}$	$\mathbf{0.61^{0.02}}$	$0.24^{0.06}$	$\mathbf{0.38^{0.04}}$	$\mathbf{0.68^{0.01}}$
ZCP + OH	$\mathbf{0.63^{0.04}}$	$0.68^{0.02}$	$0.79^{0.01}$	$0.26^{0.06}$	$0.37^{0.03}$	$\mathbf{0.61^{0.02}}$	$0.25^{0.06}$	$\mathbf{0.38^{0.05}}$	$\mathbf{0.69^{0.01}}$
ZCP + GRAF	$\mathbf{0.64^{0.04}}$	$\mathbf{0.72^{0.02}}$	$\mathbf{0.80^{0.01}}$	$\mathbf{0.30^{0.06}}$	$\mathbf{0.40^{0.03}}$	$0.55^{0.03}$	$\mathbf{0.30^{0.07}}$	$\mathbf{0.40^{0.04}}$	$0.66^{0.04}$
ZCP + GRAF + OH	$\mathbf{0.62^{0.04}}$	$\mathbf{0.71^{0.02}}$	$\mathbf{0.81^{0.01}}$	$\mathbf{0.28^{0.06}}$	$\mathbf{0.39^{0.03}}$	$0.55^{0.03}$	$\mathbf{0.27^{0.07}}$	$\mathbf{0.40^{0.04}}$	$0.66^{0.05}$

M. Comparison with TA-GATES

TA-GATES demonstrates how zero-cost proxies can improve graph neural network predictors. It is a graph neural network that mimics the forward and backward pass of information, and uses zero-cost proxies to distinguish operations that would otherwise have the same encoding due to similar contexts in the network (Ning et al., 2022). The model outperforms some strong existing predictors, such as XGBoost (Chen & Guestrin, 2016) or GCN (Wen et al., 2020).

For the TA-GATES experiments, we use the same evaluation setting as in their original paper, as they provide a fixed train and test set that cannot be mixed due to different data formats (the test set is missing some information). We evaluate only on NB201, since NB101 and NB301 are different samples than those from NB-Suite-Zero. After removing networks with unreachable branches (see Section C.2), we obtain 4 675 train and 4 770 test networks. For every evaluation, we use a subset of the train set for fitting, and evaluate on the whole held-out test set. We use the following fractions of the train set – 0.01, 0.05, 0.1, 0.2 (corresponding to 46, 233, 467 and 935 networks), and evaluate each model on 50 different seeds.

TA-GATES requires over 10 minutes to train on a single seed (for all sample sizes), but ZCP + GRAF + XGB+ needs only 12 seconds on the largest sample size. The total time required to run all 50 repetitions and 4 sample sizes is listed in Table 32 (on Intel Xeon CPU E5-2620, NVIDIA GeForce GTX 1080 Ti, no networks from the search space were trained since we use precomputed search space data). ZCP computation time was not included for both models, as TA-GATES provides only precomputed scores. Figure 25 shows the results of all models and shows that ZCP + GRAF is competitive with TA-GATES at a fraction of the train time.

Table 32. Runtimes of different models across all experiments. ZCP + GRAF experiments take less then 1 hour on a CPU, while TA-GATES needs over 37 GPU hours.. Time for computing ZCP is not included.

predictor	runtime
ZCP + GRAF + RF	4.9 CPU min
ZCP + GRAF + XGB	0.86 CPU min
ZCP + GRAF + XGB+	42.7 CPU min
TA-GATES	37.65 GPU hours

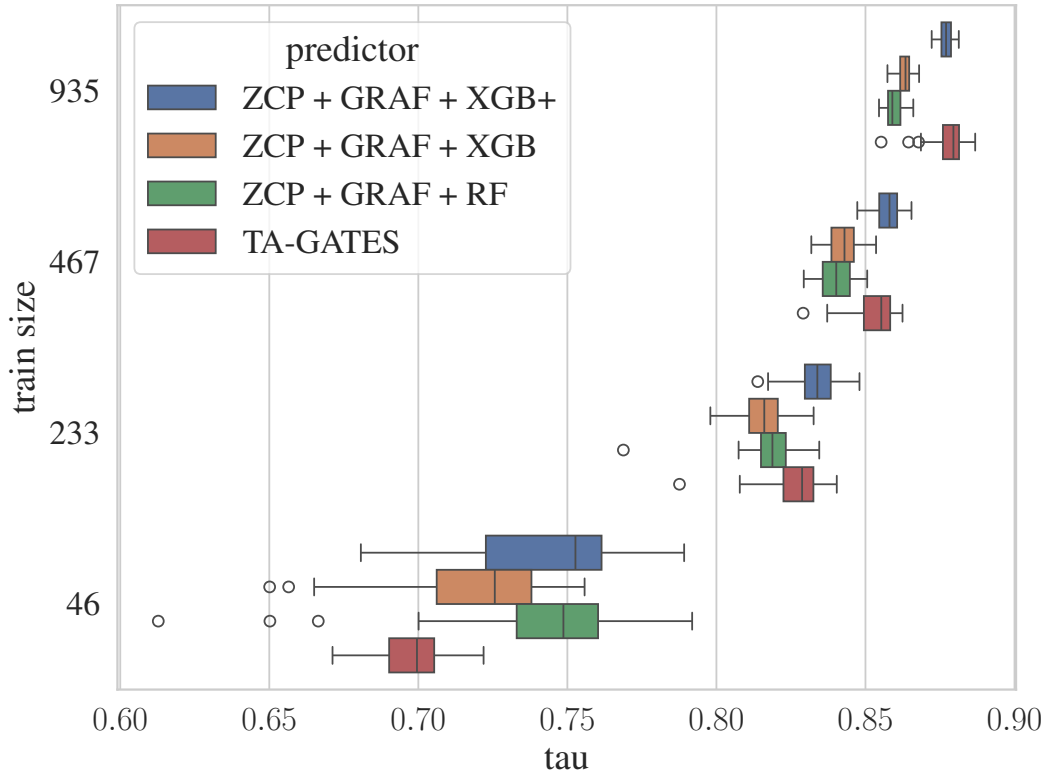


Figure 25. Results on NB201 `cifar-10` – comparison of ZCP + GRAF models and TA-GATES.

N. NASLib Predictor Evaluation

Unlike the original study, we evaluate our approach on the NB101 and NB301 sampled networks, and on NB201 without unreachable networks. For predictor evaluation, we use the model-based predictors from NASLib (White et al., 2021b). These are as follows:

- Neural predictors – BANANAS (White et al., 2021a), BONAS (Shi et al., 2020), GCN (Wen et al., 2020), MLP (White et al., 2021a), NAO (Luo et al., 2018), SemiNAS (Luo et al., 2020)
- Based on Bayesian optimization – Bayesian Linear Regression (White et al., 2021b), BOHAMIANN (Springenberg et al., 2016), DNGO (Snoek et al., 2015), BONAS
- Tree-based methods – LGBoost (Ke et al., 2017), NGBoost (Duan et al., 2020), random forest (Liaw & Wiener, 2002), XGBoost (XGB) (Chen & Guestrin, 2016)
- Gaussian process methods – GP (Rasmussen, 2004), Sparse GP (Bauer et al., 2016), Variational Sparse GP (Titsias, 2009)

The methods have up to 15 minutes of hyperparameter tuning available. For more details, refer to the original study (White

et al., 2021b). For ZCP + GRAF evaluation, we use 5 different models – random forest (no tuning), XGB (no tuning), random forest and XGB with hyperparameter tuning, and the XGB+ model (see Section E for details).

We evaluate the quality of the predictors based on the size of the training sets. The plots show the mean Kendall tau over 100 runs with the shaded areas showing the standard deviation of the results. All the methods use the same seeds for generating their sets in the 100 runs. The methods connected by the vertical lines in the critical difference plot on the left-hand side are not statistically significantly different for the largest training size. The script to create these plots is a part of the MCBO library (Dreczkowski et al., 2023).

All the results were first evaluated using the Friedmann test to assess statistical significance, then the Wilcoxon signed-rank test was used as a posthoc test with significance level $\alpha = 0.05$ and Holm correction for multiple testing.

In total, we used 21 CPU days per each of the 4 tasks, with 32GB RAM and 8 cores allocated per run, on Intel Xeon Gold 6242 CPU. No networks from the search space were trained, as we queried the accuracies.

Figures 26–26 show results on NB101, NB201, and NB301 for `cifar-10`, and NB201 for `ImageNet16-120`. ZCP + GRAF + XGB+ is the overall best predictor, and all our ZCP + GRAF predictors outperform existing models.

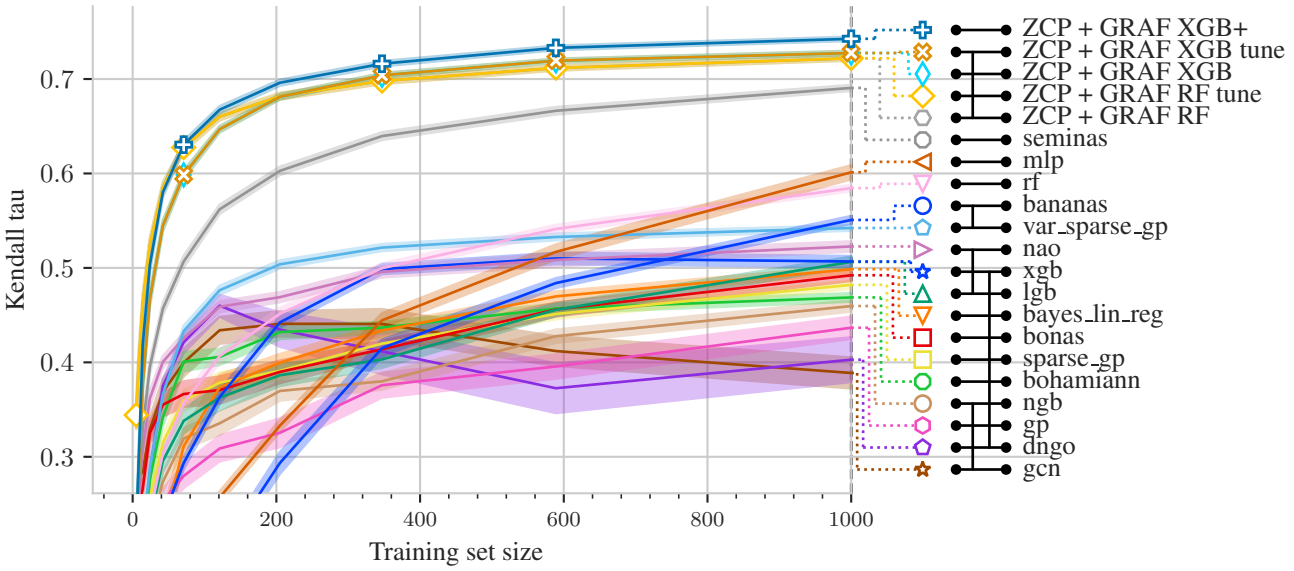


Figure 26. NASLib predictor evaluation, NAS-Bench-101 and `cifar-10`.

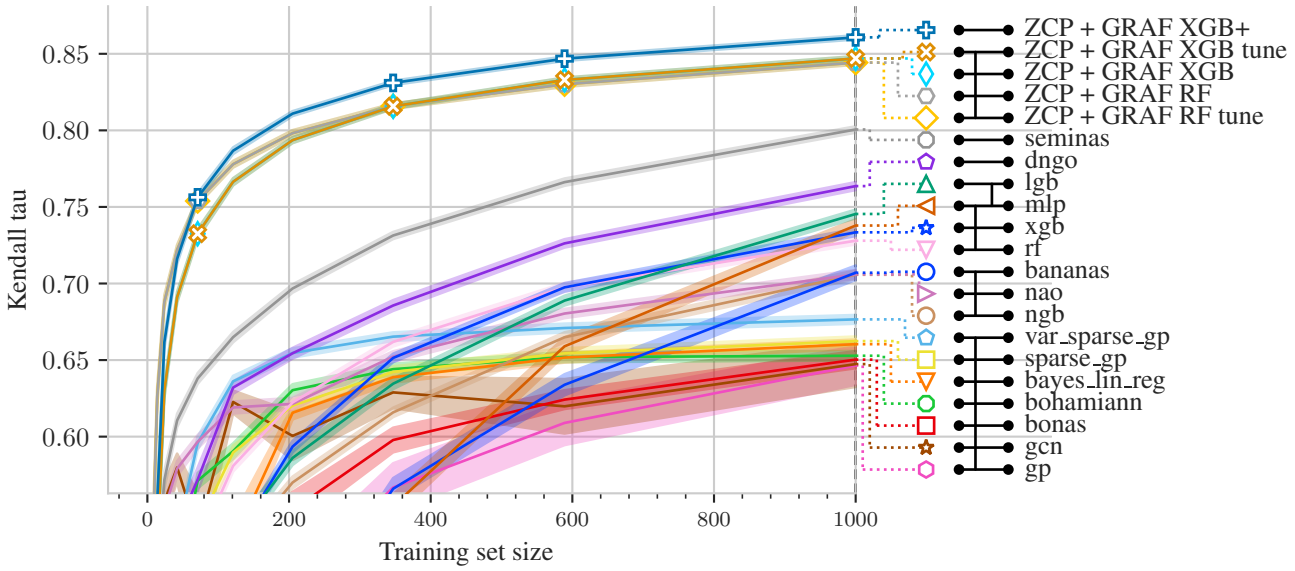


Figure 27. NASLib predictor evaluation, NAS-Bench-201 and `cifar-10`.

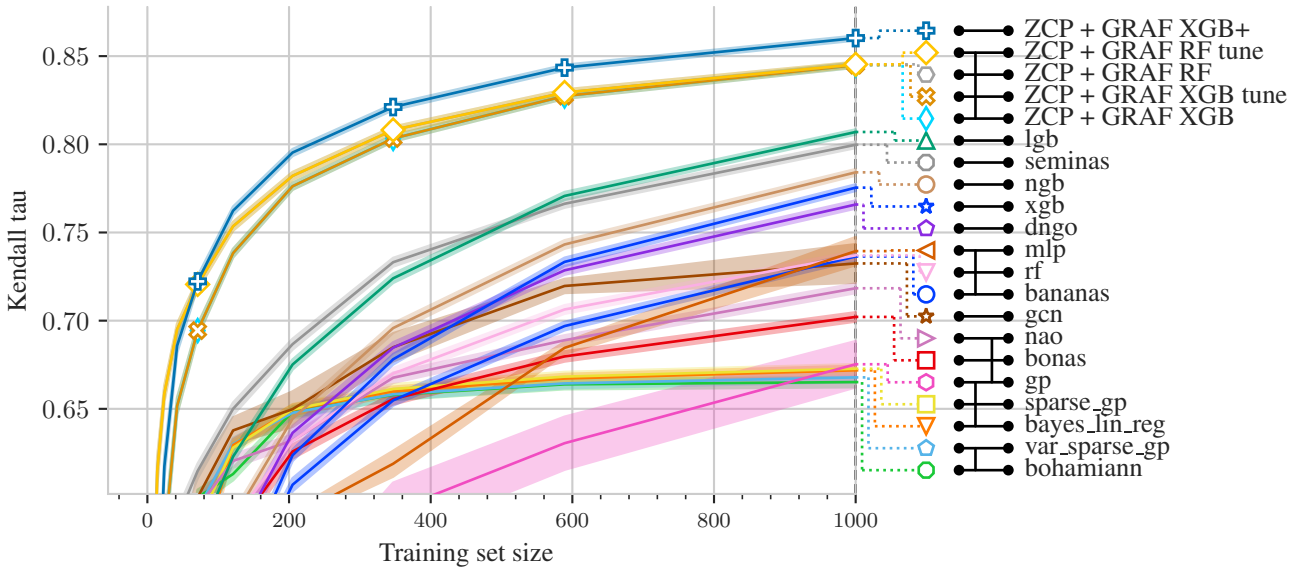


Figure 28. NASLib predictor evaluation, NAS-Bench-201 and `ImageNet16-120`

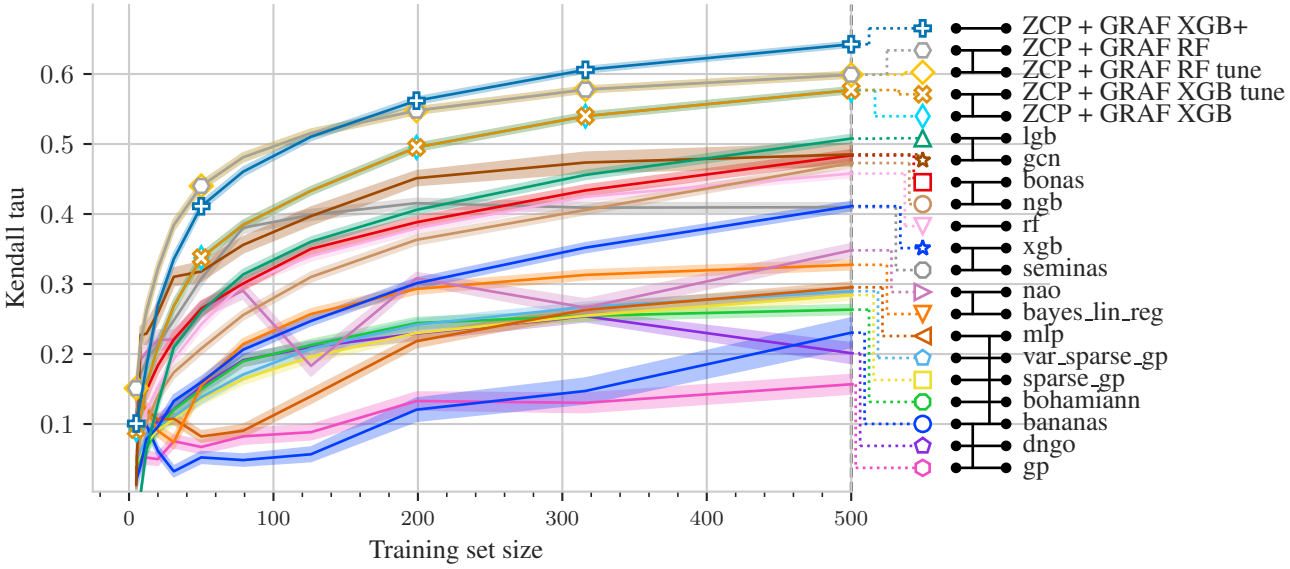


Figure 29. NASLib predictor evaluation, NAS-Bench-301 and `cifar-10`.

O. NAS Search Evaluation

For NAS search experiments, we use the same Bayesian optimization + predictor framework (Kandasamy et al., 2018; Ma et al., 2019; Shi et al., 2020) as in the NASLib study (White et al., 2021b). We compare the runs with a random search run. In total, we used 21 CPU days for each of the 2 tasks, with 32GB RAM and 8 cores allocated per run, on Intel Xeon Gold 6242 CPU. No networks from the search space were trained, as we queried the accuracies.

In the BO + predictor framework, an ensemble of three performance predictors is used for uncertainty estimates for each prediction. In each of the 25 iterations, 200 architectures are randomly sampled and out of these, 20 candidate architectures whose predictions maximize an acquisition function are evaluated. White et al. (2021b) use independent Thompson sampling (White et al., 2021a) as the acquisition function.

Figures 30 and 31 show results of the search on `cifar-10` and `ImageNet16-120` respectively.

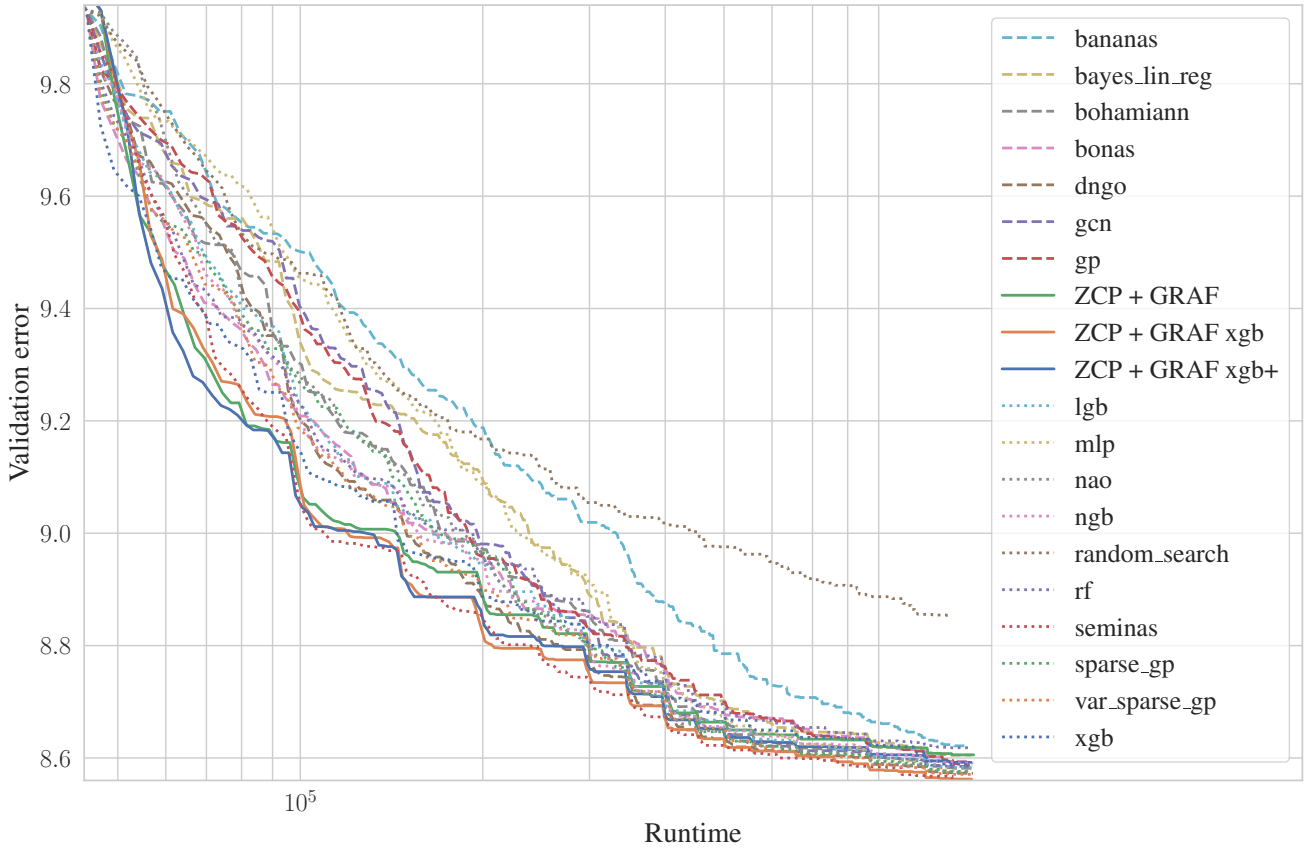


Figure 30. NASLib search experiment, NAS-Bench-201 and `cifar-10`.

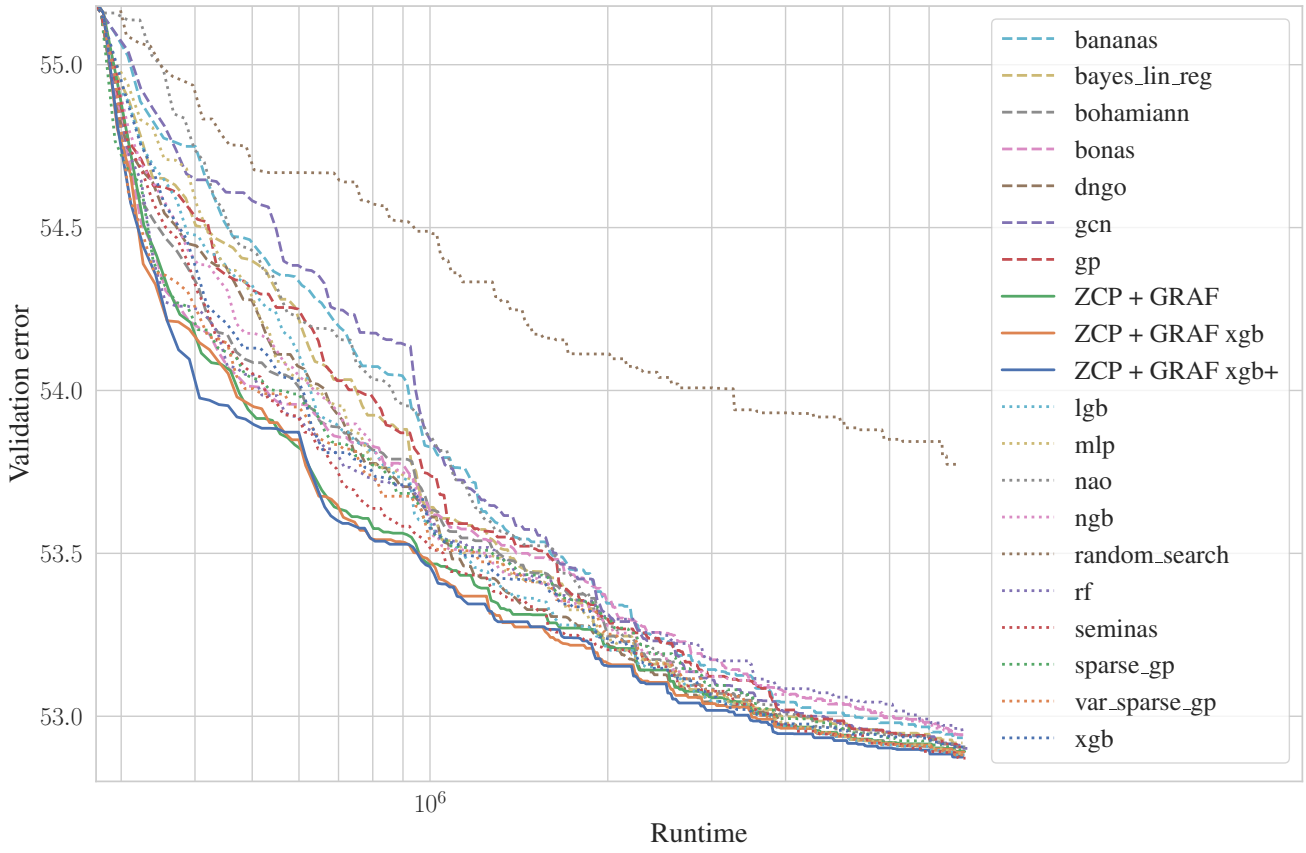


Figure 31. NASLib search experiment, NAS-Bench-201 and `ImageNet16-120`.