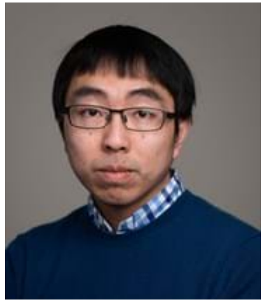
UNIVERSITY OF OXFORD

Interpretable Neural Architecture Search
Using Bayesian Optimisation with Weisfeiler-Lehman Kernels
(NAS-BOWL)






Robin Ru* Xingchen Wan* Xiaowen Dong Michael Osborne

Interpretable NAS

- **Current NAS:**

Task + Objective ⇨ **NAS** ⇨ **Architecture**

- **Interpretable NAS:**

Task + Objective ⇨ **NAS** ⇨ **Architecture**

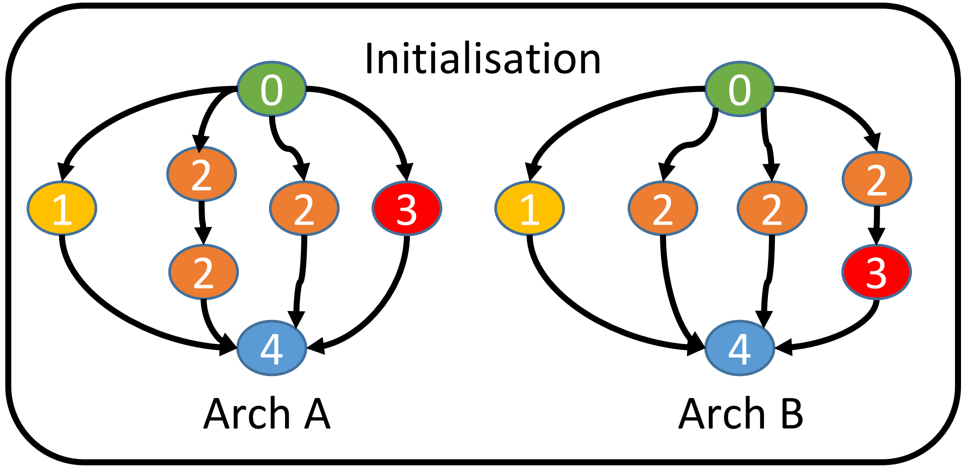
⇨ **Interpretable features that explaining arch. performance**

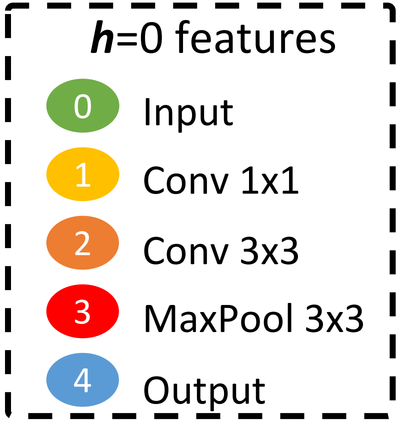
Contributions

NAS-BOWL:

- **Query-efficient** BO-based NAS strategy
  - GP surrogate with the Weisfeiler-Lehman (WL) graph kernel achieves good predictive performance
  - Enable BO to handle graph inputs directly
- **Interpretability**
  - Interpretable graph features extracted by WL kernel
  - Help explain the performance of architecture design
  - Example: motif-based transfer learning to warm start a related task
- **Superior empirical performance**
  - GP surrogate on various search spaces
  - BO strategy on various NAS datasets

Weisfeiler-Lehman(WL) Graph Kernel

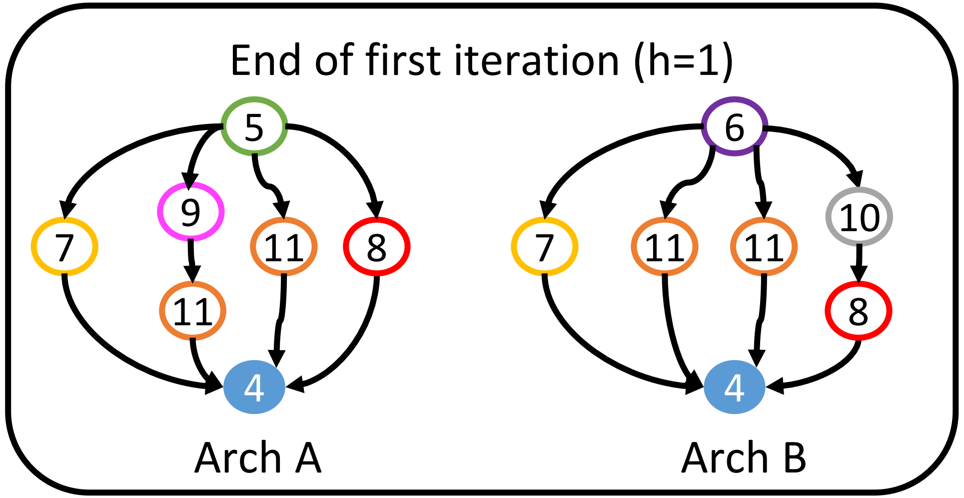


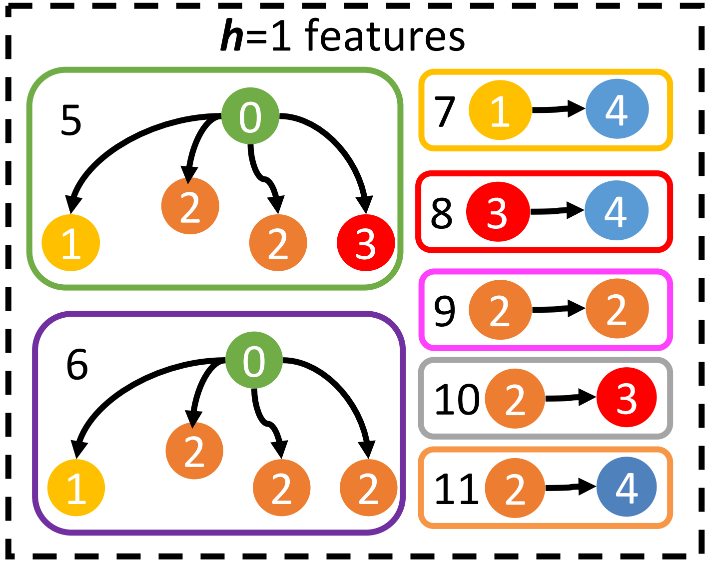


$$k_{\mathrm{WL}}^{H}(G_A, G_B) = \sum_{h=0}^{H} k_{\mathrm{b}}\big(\phi_h(G_A), \phi_h(G_B)\big)$$

h: the WL iteration
= depth of the subtree features

As ***h*** increases, WL captures higher-order features corresponding to larger neighborhoods

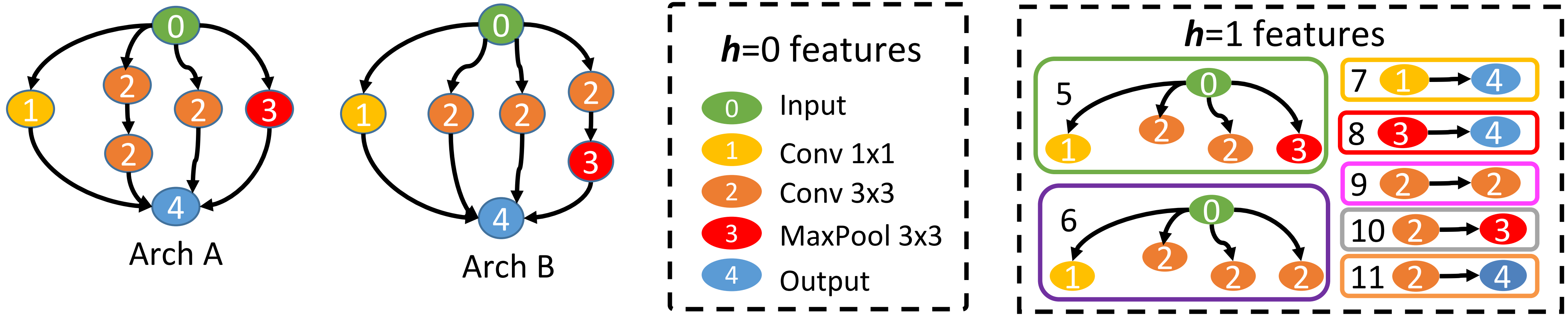




(Shervashidze et al., 2011)

Interpretability

- WL kernel extracts **interpretable features**:



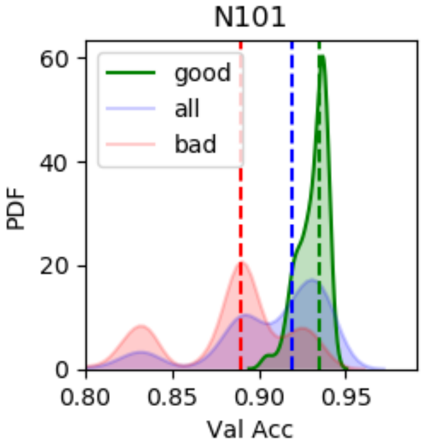
- Average the **derivatives of the posterior mean** w.r.t these features

$$\mathrm{AG}(\phi^j) = \int_{\phi^j(G)>0} \frac{\partial \mu}{\partial \phi^j(G)} p\left(\phi^j(G)\right) d\phi^j(G)$$

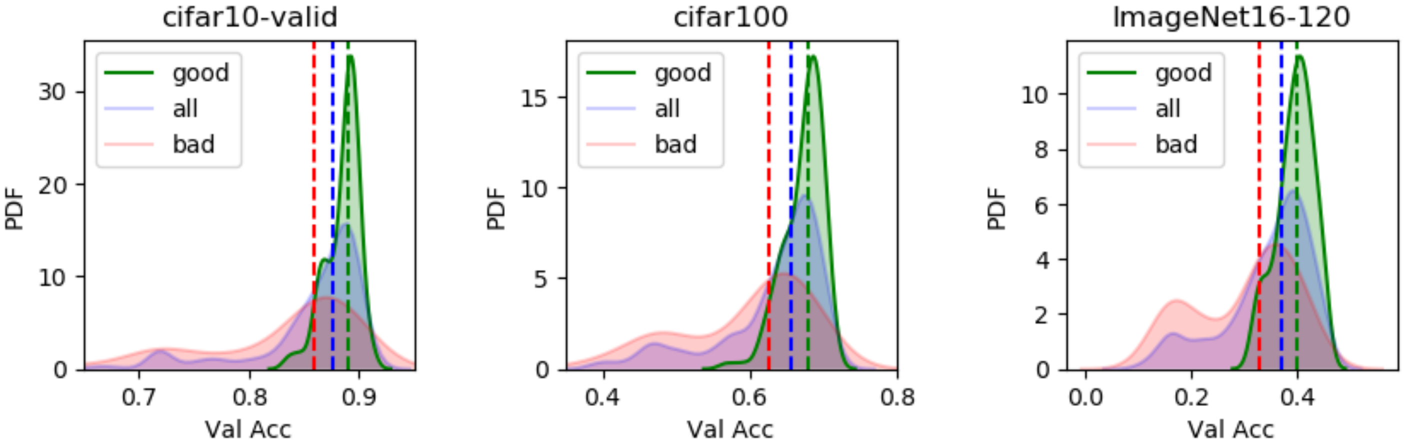
→ the **global**/overall sensitivity of the objective due to presence of certain motifs.

Interpretable WL features + tractable GP derivatives = Interpretable NAS

Interpretability: Validation

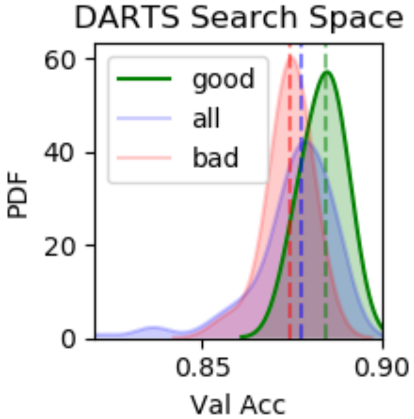


NAS-Bench-101



NAS-Bench-201 (CIFAR-10)

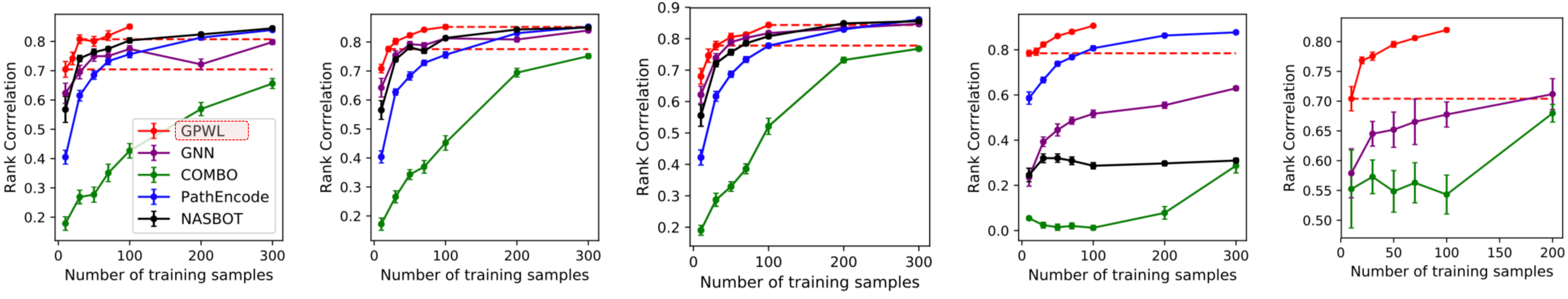
Features are indicative of architecture performance;
Features learnt on CIFAR-10 also transfers well to the other tasks!



DARTS Search Space

Experiments: Surrogate Regression

- Comparison with alternate surrogates
  - Increasing training data, 400 validation data, 20 repetitions
  - GPWL outperforms all competing methods with ***much less*** training data esp. on datasets with ***larger*** search spaces (N101 and Flowers102)



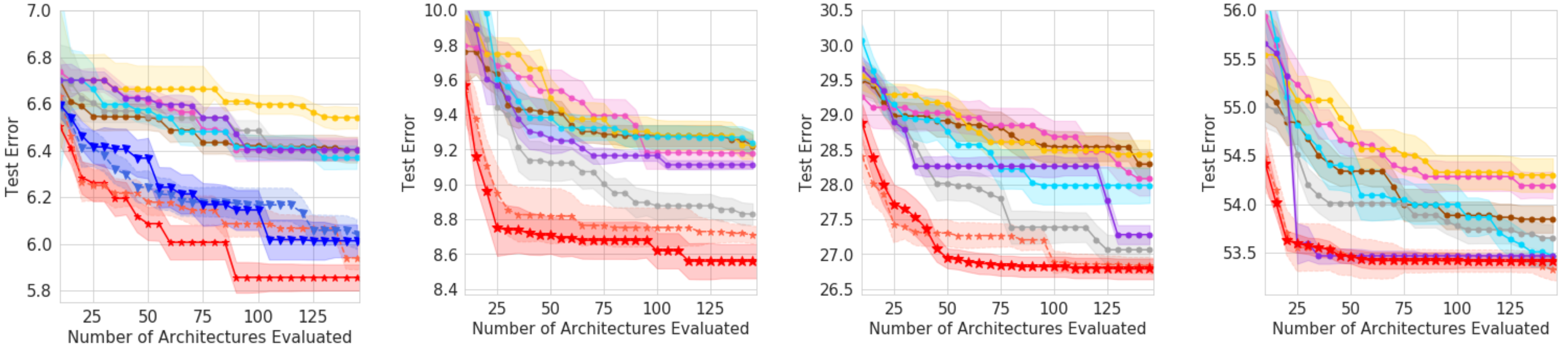
(a) N201(C-10) (b) N201(C-100) (c) N201(ImageNet) (d) N101 (e) Flowers102

32-node architecture cell

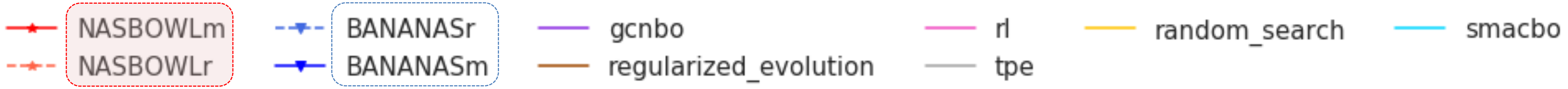
(Shi et al., 2019; Oh et al., 2019; White et al., 2019; Kandasamy et al., 2018)

Experiments: NAS Performance

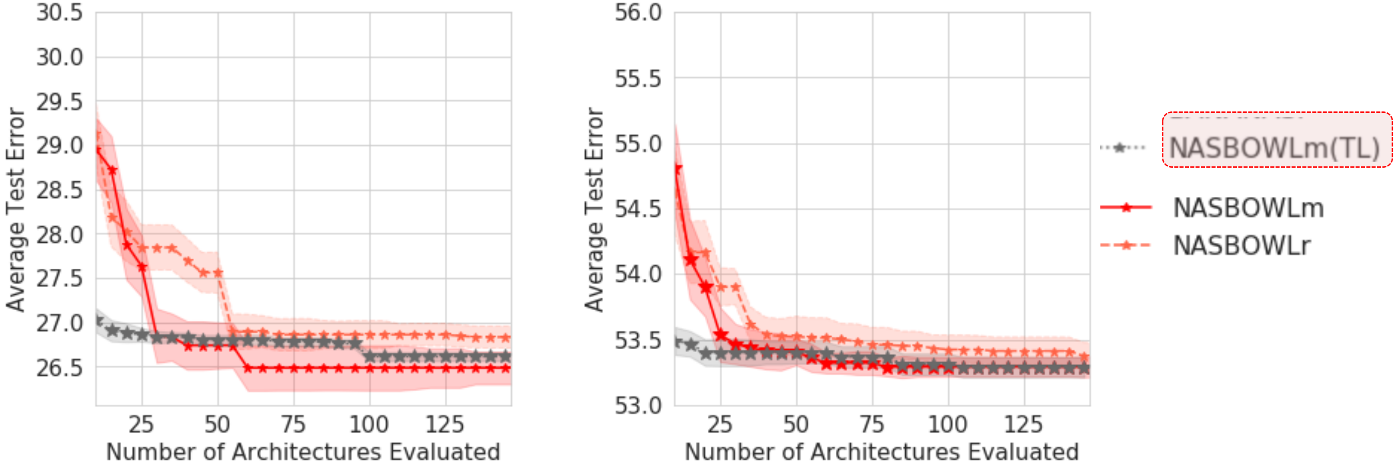
For NASBOWL and BANANAS, ***r/m*** indicates *random sampling/mutation* for optimising the acq. func.



(e) N101 (f) N201(C-10) (g) N201(C-100) (h) N201(ImageNet)



Transferring from features learnt on CIFAR-10 to CIFAR-100 and ImageNet tasks in NAS-Bench-201 ;
Grey line is NAS-BOWL with warm-starting



(a) N201(CIFAR100) (b) N201(ImageNet)

Summary and Discussion

- **Our NAS-BOWL**:
  - Query-efficient NAS strategy and data-efficient surrogate model;
  - Learn interpretable motifs responsible for architecture performance along with the search; First step towards ***interpretable NAS***
  - An example use of interpretability: motif-based transfer learning;

- **Future directions**:
  - Alternative ways to extract interpretable insights
  - Better ways to use interpretability for improving NAS (e.g. robustness)
  - Use interpretable motifs for pruning the architectures

- **Poster ID 1775:** Session 7, 1 am and 3am (PDT), May 5, 2021
- **Code**: https://github.com/xingchenwan/nasbowl
- **Email**: robin@robots.ox.ac.uk