Graph perceiver network for lung tumor and premalignant lesion stratification from histopathology

Rushin H. Gindra[1,2,3], Yi Zheng[1,4], Emily J. Green[1], Mary E. Reid[5], Sarah A. Mazzilli[1], Daniel T. Merrick[6], Eric J. Burks[7], Vijaya B. Kolachalama[1,4,8,*], Jennifer E. Beane[1,*]

[1]Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, 72 E. Concord Street, Boston, 02118, MA, USA.
[2]Institute of AI for Health, Helmholtz Zentrum München—German Research Center for Environmental Health, Ingoldstädter Landstr. 1, München, 85764, Germany.
[3]Center for Translational Cancer Research (TranslaTUM), School of Medicine, Technische Universität München, Einsteinstraße 25, München, 81675, Germany.
[4]Department of Computer Science, Boston University, 665 Commonwealth Avenue, Boston, 02215, MA, USA.
[5]Roswell Park Comprehensive Cancer Center, 665 Elm Street, Buffalo, 14203, NY, USA.
[6]Department of Pathology, University of Colorado School of Medicine, 12631 East 17th Avenue, Aurora, 80045, CO, USA.
[7]Department of Pathology & Laboratory Medicine, Boston University Chobanian & Avedisian School of Medicine, 72 E. Concord Street, Boston, 02118, MA, USA.
[8]Faculty of Computing & Data Sciences, Boston University, 665 Commonwealth Avenue, Boston, 02215, MA, USA.

*Corresponding author(s). E-mail(s): vkola@bu.edu; jbeane@bu.edu;

Abstract

Bronchial premalignant lesions (PMLs) precede the development of invasive lung squamous carcinoma (LUSC); however, it is challenging to distinguish between

PMLs that will progress to LUSC from those that will regress without intervention. We developed a computational pipeline using H&E whole slide images (WSIs) to stratify PMLs on the spectrum from normal to tumor tissue. Implementing a study across four data cohorts, including lung tumor resection tissue and endobronchial biopsies, our network distinguished carcinoma in situ biopsies by progression status, indicating that WSIs contain important information that could be used in combination with molecular markers to predict progression to cancer.

Keywords: Lung cancer, digital pathology, premalignant lesions, deep learning

Lung squamous cell carcinoma (LUSC), the second most common type of non-small cell lung cancer, is preceded by the development of bronchial premalignant lesions (PMLs) that can progress toward invasive carcinoma or regress without intervention. Recently, genomic and proteomic profiling of PMLs has demonstrated that PML progression is associated with impaired immunosurveillance [1–5]. Currently, there are ongoing efforts to build a Lung Pre-Cancer Atlas [6] to understand molecular alterations in their spatial context associated with disease severity and progression. To date, H&E stained tissue slides of PMLs have not been fully utilized as a data source to augment our insights into PML biology. Endobronchial biopsies of PMLs are heterogeneous and contain a range of histologic grades, a variety of structures including cartilage and submucosal glands, and varying degrees of inflammatory cell infiltrate. We hypothesized that computational methods that utilize digitized H&E whole slide images (WSIs) may be able to capture PML heterogeneity and stratify PMLs by histologic severity or their ability to progress to invasive carcinoma, providing a standardized and informative WSI-level assessment of PMLs.

We developed a Graph Perceiver Network (GPN) to characterize PMLs using lung tissue WSIs from public cohorts, consisting of Normal (non-tumor adjacent), lung adenocarcinoma (LUAD) and LUSC tissue samples. Training with TCGA (534 Normal, 740 LUSC, 808 LUAD)[7] and Testing on CPTAC (719 Normal, 685 LUSC, 667 LUAD)[8], the GPN learned WSI-level features and classes of endobronchial biopsy-based WSIs from two datasets: 1) University College London (UCL) (112 CIS samples)[4], defining progression as moving to LUSC; and 2) Roswell Park Comprehensive Cancer Institute (Roswell) (346 samples, normal to CIS spectrum)[1], where progression was persistence of dysplasia or advancement to at least mild dysplasia. The GPN identified signatures that differentiated LUSC and LUAD patterns, and CIS samples progressing to LUSC (Fig. 1a).

The GPN demonstrated high classification performance on TCGA (mean AUROC=0.98 $\pm$ 0.01 & mean AUPRC=0.98 $\pm$ 0.01) and CPTAC data (mean AUROC=0.93 $\pm$ 0.01 & mean AUPRC=0.95 $\pm$ 0.02)(Fig. 1b), utilizing significantly fewer resources compared to two published frameworks (Supplementary Table 1, $GPN \sim$ 202K vs $GTP \sim$ 664K trainable parameters). It generated class-specific heatmaps, consistent with expert pathologist annotations from 20 CPTAC samples

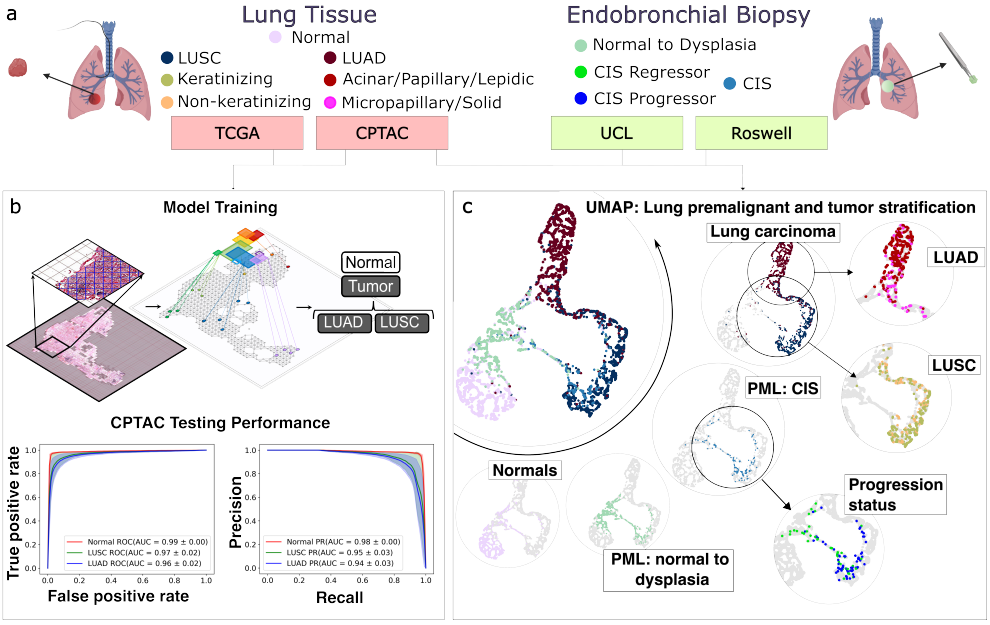

Fig. 1 "GPN stratification of lung tumors and PMLs relates to histologic features and outcome: (a) GPN, trained on TCGA, tested with CPTAC lung resection WSIs, identifies WSI-level signatures in UCL and Roswell PMLs. **(b)** Accepting patched WSIs as graph input, GPN clusters patches with similar embeddings into overlapping sets, with ROC & PR curves displaying 3-label classification on the CPTAC cohort. **(c)** UMAP plots of CPTAC, UCL, and Roswell WSI features stratified samples from normal to invasive carcinoma. The model separated LUAD and LUSC WSIs by tumor histologic patterns and PMLs by histologic grade and outcome."

(selected at random), accurately pinpointing areas corresponding to same-label annotations (Fig. 2a and supplementary Fig. 6), and reliably identified pathologically significant regions across varied samples/tumor types, showing robust model learning and identification.

We hypothesized that our trained model could differentiate between predominant histologic patterns of LUSC and LUAD tumors and lung squamous PML by discerning subtle morphological features. To test this, we used the WSI latent features from the CPTAC, UCL, and Roswell samples to perform UMAP-based clustering (Fig. 1c). Clustering segregated normal, LUAD, and LUSC with a high adjusted Rand score of 0.69 and adjusted mutual information score of 0.64. Principal component analysis (PCA) [9] on the latent features (Supplementary Fig 5) revealed significant PC1 differences between keratinizing and non-keratinizing LUSC tumors (PC1: $p = 0.01$), with the latter PC2 values similar to those of LUAD tumors. The LUAD tumors showed significant PC1 differences between aggressive (solid/micropapillary) and non-aggressive (lepidic/acinar/papillary) histologic patterns (PC1: $p = 5.4e-14$). Notably, LUAD solid/micropapillary tumors exhibited PC2 values similar to LUSC tumors,

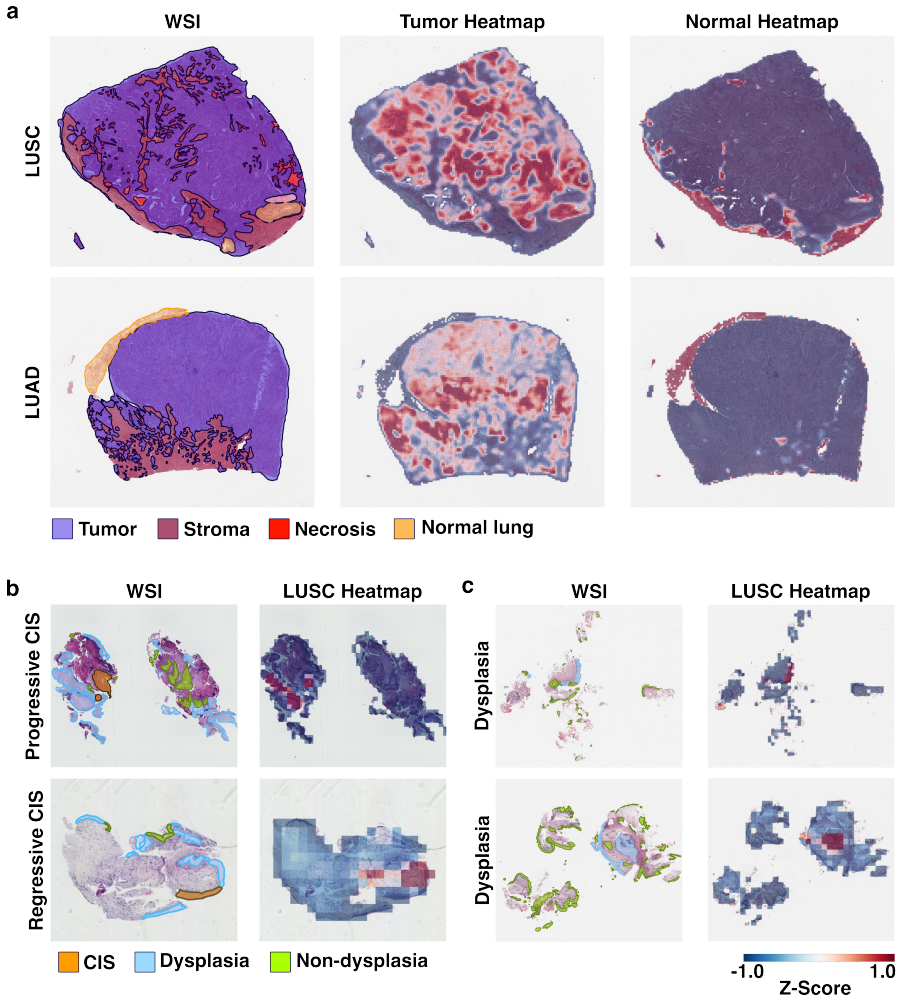

Fig. 2 The model identifies tissue regions that correspond with pathologic annotations. **(a)** The GPN identified regions annotated as tumor and normal on LUAD and LUSC WSIs from CPTAC. **(b, c)** The network was also used to classify PMLs as one of the three labels. The GPN LUSC-specific heatmaps for PMLs predicted to be LUSC identified regions of dysplasia and CIS as well as other regions that are likely contributing to its ability to separate CIS progressors from CIS regressors.

as solid LUAD tumors require immunohistochemical analysis for diagnosis of adenocarcinomatous differentiation as they can be difficult to distinguish from LUSC tumors.[10].

UMAP and PCA revealed most PMLs were proximal to the normal cluster, with some high-grade dysplasia samples clustering with LUAD and LUSC cases ($p < 0.01$ for PML vs. LUAD, PML vs. LUSC, on both PC1 and PC2). Roswell PMLs did not show significant separation by molecular subtype or progression status (Supplementary Fig 5). All CIS samples from Roswell and some from UCL grouped with LUSC tumor

cases (52% of CIS samples were classified as LUSC tumors with $p < 0.01$ for CIS vs. LUAD, CIS vs. LUSC on both PC1 and PC2). The model robustly stratified regressive versus progressive CIS and was associated with lymphocyte count [4]($r = 0.3$, $p = 0.001$ between PC2 and lymphocyte count), highlighting high-grade histology regions in PMLs predicted to be LUSC (Fig. 2b and c).

The GPN generated latent features observable across various sample types (resected lung tissue versus endobronchial biopsies), histologic patterns, and multiple cohorts, aiding in tumor and PML stratification. Notably, it identified CIS endobronchial biopsy samples at risk of progressing to LUSC, potentially assisting in treatment planning. The GPN also produced class-specific heatmaps, highlighting tissue areas associated with specific labels and facilitating morphological feature understanding linked to lung tumor and PML histologic patterns. Future research should assess the GPN across broader cohorts and explore additional genomic data integration to enhance PML stratification.

Methods

Ethics statement

The institutional review boards at Boston University Medical Center and Roswell Park Comprehensive Cancer Center approved the study. All subjects provided written informed consent. Data from TCGA, CPTAC, and UCL are publicly available.

Data acquisition and design

We obtained lung tumor type from the TCGA and CPTAC samples and histologic grade and progression status data from the UCL and Roswell samples. In the CPTAC samples, we obtained the predominant histologic pattern of the tumors. The UCL CIS samples included supplementary details about the samples' lymphocyte counts in CIS and stromal regions. The Roswell samples included molecular phenotypes identified using RNA sequencing data such as PML molecular subtype that included 4 categories – proliferative, inflammatory, secretory, and normal-like [1].

To evaluate our model's explanations, we selected a few WSIs from CPTAC, Roswell, and UCL and uploaded them to a web-based software (PixelView, deepPath, Boston, MA). Several regions on the WSIs were annotated by board certified thoracic pathologists (D.M. and E.B.) using the software. In the LUAD specimens, tumor areas were annotated by their histologic patterns (solid, micropapillary, cribiform, papillary, acinar, and lepidic). For both LUAD and LUSC tumors, histologic features, including necrosis, lymphatic and vascular invasions, were also annotated. Airways present in the tissues were graded on the spectrum of normal to CIS and non-tumor tissue areas were labeled as healthy or inflamed lung tissue, stroma, or necrosis. The LUAD and LUSC annotations were combined into tumor, stroma, normal lung, and necrosis for analysis. For the endobronchial biopsy samples, epithelial regions were annotated based on histologic grade (normal, hyperplasia, immature squamous metaplasia, squamous metaplasia, mild dysplasia, moderate dysplasia, severe dysplasia, and CIS. These annotations were combined into CIS, dyplasia, and non-dyplasia for analysis. The overlap between the model-produced heatmaps and the pathologist annotations was analyzed by exporting them as binary images.

Graph perceiver network

Our architectural design consists of three main components (Fig 3). (1) A graph module that retains position information in the input features via edge connections between neighboring nodes, enabling the model to capture the relationship between tokens. (2) A cross-attention module, inspired from perceiver [11], maps variable length patch embeddings to predefined latent clusters (or sets). (3) A self-attention module comprising of multiple dot-product attention layers to learn relevant interactions between the sets. The number of sets used in cross-attention is a hyperparameter, typically much smaller than the expected number of visual tokens. Additionally, to address the variable scale of visual tokens, we utilized the CTransPath [12], a Swin-Transformer, as our feature extractor. CTransPath was pre-trained on the TCGA pan-cancer dataset in a self-supervised manner. Our model introduces novelty by integrating the graph

module with the perceiver architecture, enabling sparse graph computations on the visual tokens and computationally efficient modeling of the perceiver.

Inspired by the recent work on using min-cut pooling on a graph-transformer network (GTP) for WSI-level classification [13], and adaptive aggregation functions used in recent graph classification problems [14], we adapted a cross-attention module (PMA) from perceivers. This module further simplifies GTP [13], by replacing the min-cut pooling mechanism with cross-attention layers, which cluster the graph nodes with similar embeddings and spatial proximity to a fixed size latent embedding. We denote these embeddings as 'sets'. These cross-attention pooling strategies paired with sparsity in computations have shown comparable results [11], while minimizing the computational footprint, thus allowing for extremely large WSIs to be processed efficiently.

The input sets, accompanied by a classification token CLS, were provided to the self-attention module (SAB), mirroring the approach employed in vision transformers [15]. SAB assigns weights to each set embedding for classification. The PMA and SAB modules were alternatively stacked, allowing the set embeddings to extract information from the input WSI graph based on previous layer's learnings. This strategy preserves structural information. We followed the weight sharing strategy of the perciever [11] to optimize computations. The block was followed by an MLP, which facilitated multi-class prediction. The CLS token embedding represents the slide-level aggregation vector of the weighted sets. The utilization of the CLS token instead of averaging the embeddings of c sets facilitated explanability in the network. Essentially, our model serves as an efficient, end-to-end, amortized clustering algorithm, which clusters distinct regions of the tissue into overlapping sets, preserving vital spatial interactions that enhance our comprehension of the biological processes within tumor tissues.

Experimental design

Each WSI was passed through a fast patching pipeline at 20x magnification, creating non-overlapping tissue region patches of 256×256 pixels while filtering out background slide information. Key epithelial regions were retained via Otsu thresholding [16]. GPN accepts graphs as inputs, where each WSI is represented as an undirected, unweighted graph with nodes representing the tissue image patches and set of edges connecting the nodes. The graph follows an 8-connectivity neighborhood structure. The node embeddings in the graph are represented as a matrix of n vectors; where n is the total number of nodes in the graph. Each vector is of size 768-dimensions obtained from the CTransPath feature extractor for the corresponding image patch. These WSI graphs feed into the GPN, which was trained as a three-label classifier (LUAD, LUSC tumor, or Normal) using the TCGA dataset, with stratified patient-level sampling and 5-fold cross-validation.

Optimal hyperparameters for our network configuration involved a hidden dimension (D) of 64, 3 GINConv layers, cross-attention pooling configuration of 200 sets and a 3 self-attention layers per self-attention block (SAB). We used 8 heads for multi-head attention training. To avoid overfitting, we used dropouts with a rate of 0.2 in feed-forward blocks, and binary cross-entropy loss to enhance robustness. Training was completed over 30 epochs in 8-sample mini-batches, using early stopping, a 0.0003

initial learning rate, a step scheduler, and the Adam optimizer [17] for expedited convergence. The experiments were performed on a single GeForce RTX 2080Ti 11Gb workstation.

For post-hoc analysis, the WSI-level features for each sample were extracted by giving the respective graphs as input to the trained 3-label GPN classifier. The 64-dimensional features from the final layer of the network were considered as the representations for each WSI and used for further analysis. The performance of the classifier was evaluated on the CPTAC cohort. For the 3-label classification performance, 2071 samples were used. The receiver operating characteristic (ROC) and precision-recall (PR) curves for each class were computed along with the accuracy, precision, sensitivity and specificity. UMAP manifold representations [18] and principal component analysis (PCA) representations were plotted to illustrate the relationship between image features and sample phenotypes. Sample phenotypes were grouped as follows: normal adjacent lung tissue, LUSC, and LUAD for the lung tissue data (CPTAC). Within CPTAC, we examined differences in the principal components associated with LUSC predominant histologic patterns (keratinizing or non-keratinizing) and LUAD predominant histologic patterns grouped as follows: (solid/micropapillary/cribiform or lepidic/acinar/papillary). Within the CIS samples, we examined differences in the principal components by progression status. Post-hoc analysis results suggest that the model is robust to cohort-based batch effects between Roswell and UCL cohorts.

Explainability analysis

GPN learns the classification label-relevant contributions of the tissue regions during training. We show these explanations as heatmaps using a relevance propagation mechanism [19]. This mechanism is a generic attention-based model explanations for bimodal transformers. It uses attentions and gradients from the SAB and PMA blocks to produce relevancy maps for each interaction between the sets C and the assigned nodes N. The relevancy map from PMA provides the individual contributions of each node (patch) for the final prediction instead of distributing the contribution of each set as an average to all the nodes within the set. These relevancy maps are constructed on the WSI using the adjacency matrix and coordinates of all the patches on the WSI.

Statistical analysis

We compared the performance of GPN with other approaches for the tumor classification task in lung cancer (train on TCGA-Lung, test on CPTAC-Lung) highlighting it's efficiency while performing similar to the state-of-the-art methods. UMAP and PCA clustering analyses were conducted using Scanpy [9], with statistical significance between principal components and sample characteristics assessed through two-tailed t-tests. Statistical analyses were performed at a significance level of 0.05.

Data availability

Data from TCGA and CPTAC can be downloaded freely from the public domain. UCL data can be obtained from: https://idr.openmicroscopy.org with IDR0082. Roswell images can be provided upon request.

Code availability

Computer scripts and manuals will be made available at the time of manuscript acceptance.

Acknowledgments

This work was supported by grants from the National Institutes of Health (R21-CA253498, U2C-CA233238, R01-HL159620, R43-DK134273, RF1-AG062109, and P30-AG073104), Johnson & Johnson Enterprise Innovation, Inc., the American Heart Association (20SFRN35460031), and the Karen Toffler Charitable Trust.

References

[1] Beane, J. E. *et al.* Molecular subtyping reveals immune alterations associated with progression of bronchial premalignant lesions. *Nature Communications* **10**, 1856 (2019).

[2] Merrick, D. T. *et al.* Altered cell-cycle control, inflammation, and adhesion in high-risk persistent bronchial dysplasia. *Cancer research* **78**, 4971–4983 (2018).

[3] Mascaux, C. *et al.* Immune evasion before tumour invasion in early lung squamous carcinogenesis. *Nature* **571**, 570–575 (2019).

[4] Pennycuick, A. *et al.* Immune Surveillance in Clinical Regression of Preinvasive Squamous Cell Lung Cancer. *Cancer Discovery* **10**, 1489–1499 (2020). URL https://doi.org/10.1158/2159-8290.CD-19-1366.

[5] Ning, B. *et al.* Convergence of yap/taz, tead and tp63 activity is associated with bronchial premalignant severity and progression. *Journal of Experimental & Clinical Cancer Research* **42**, 1–18 (2023).

[6] Rozenblatt-Rosen, O. *et al.* The human tumor atlas network: charting tumor transitions across space and time at single-cell resolution. *Cell* **181**, 236–249 (2020).

[7] TCGA Research Network. The cancer genome atlas program URL https://portal.gdc.cancer.gov/. Available: https://portal.gdc.cancer.gov/.

[8] Edwards, N. J. *et al.* The CPTAC data portal: A resource for cancer proteomics research. *Journal of Proteome Research* **14**, 2707–2713 (2015). URL https://doi.org/10.1021/pr501254j. PMID: 25873244.

[9] Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology* **19**, 15 (2018).

[10] Kadota, K. *et al.* Reevaluation and reclassification of resected lung carcinomas originally diagnosed as squamous cell carcinoma using immunohistochemical analysis. *The American journal of surgical pathology* **39**, 1170 (2015).

[11] Jaegle, A. *et al.* Perceiver: General perception with iterative attention (2021). 2103.03206.

[12] Wang, X. *et al.* Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis* (2022).

[13] Zheng, Y. *et al.* A graph-transformer for whole slide image classification. *IEEE Transactions on Medical Imaging* **41**, 3003–3015 (2022).

[14] Buterez, D., Janet, J. P., Kiddle, S. J., Oglic, D. & Liò, P. Graph neural networks with adaptive readouts (2022). 2211.04952.

[15] Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale (2021). 2010.11929.

[16] Otsu, N. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics* **9**, 62–66 (1979).

[17] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] McInnes, L., Healy, J., Saul, N. & Grossberger, L. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software* **3**, 861 (2018).

[19] Chefer, H., Gur, S. & Wolf, L. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers (2021). 2103.15679.

[20] Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How powerful are graph neural networks? (2019). 1810.00826.