AFRL-AFOSR-UK-TR-2022-0016

Error quantification and complexity limits of deep learning models

Cosse, Augustin
ECOLE NORMALE SUPERIEURE
45 RUE D'ULM
PARIS, , 75005
FRA

01/13/2022
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
European Office of Aerospace Research and Development
Unit 4515 Box 14, APO AE 09421

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
20220113	Final	START DATE 20180930	END DATE 20210929

4. TITLE AND SUBTITLE
Error quantification and complexity limits of deep learning models

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
	FA9550-18-1-7007	61102F
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Augustin Cosse

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
ECOLE NORMALE SUPERIEURE
45 RUE D'ULM
PARIS 75005
FRA

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)
EOARD UNIT 4515 APO AE 09421-4515	AFRL/AFOSR IOE	AFRL-AFOSR-UK-TR-2022-0016

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Only $5 000 of the total grant funds of $50,000 were spent on this grant. The remainder was returned by the university Ecole Normale Superieure. in July 2021. The PI left the school in Aug of 2020 and the school could not support transferring his grant to John Hopkins University. No further work was performed since the last report that summarized work performed up to Sept 2020. Attached is the last report filed for this grant.
The grant although aborted did result in a publication in publication in NeurIPS 2020, the top venue for publication in machine learning.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	SAR	103

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
MARK FRIEND	314-235-6292

EOARD FA9550-18-1-7007
Error quantification and complexity limits in deep learning

PI: Augustin Cosse, co-PI: Soledad Villar

Report September 2019 - September 2020

1 Summary

The project's goal is to produce mathematical theory to shed light on the theoretical limits of deep learning.

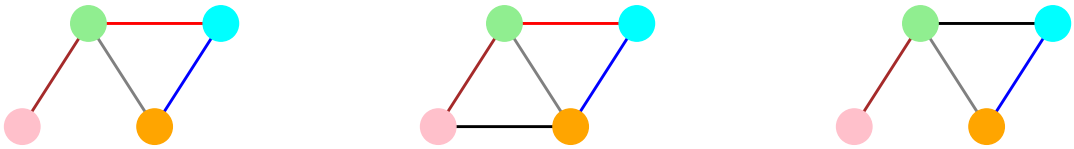
Our efforts from September 2019 to September 2020 focused on the following research directions:

- We continue our research program on the study of the theoretical limits of learning of graph neural networks (GNNs). Building on the theoretical foundations from our last year's result [1], we produce a new research article where we characterize graph neural networks (GNNs) by their ability to count substructures. This research article recently got accepted for publication in NeurIPS 2020, the top venue for publication in machine learning.
- We collaborate with sonar researcher Zoi-Heleni Michalopoulou to study the feasibility of machine learning techniques for seabed classification from underwater acoustics signals. Our work was accepted for publication in The Journal of the Acoustical Society of America, which is a top journal in acoustics related research.
- Finally, we continue the investigation of semidefinite programming relaxations for the resolution of the compressed and multidimensional super resolution problems as a step towards the use of such relaxation in the understanding of the training of neural networks.

2 Introduction

Deep learning has proven to be the state of the art method in many domain problems. Computer vision and natural language processing were completely revolutionized by deep learning in the past years.

Both natural language and computer vision are very different fields, but they do have something in common, there are incredibly large amounts of data are available in the internet, for anyone to use, and train their models (in form of images and text). In other domains, where data is a scarce resource, deep learning techniques have proven their worth with different amounts of success. There exist skepticism among practitioners regarding the applicability of these techniques and their advantages. In practice, the use of deep learning effectively requires a large amount of training data and computational resources, and often the quality of results is highly dependent on the choice of not well-understood parameters and architectural design.



$G^{[P]}$ $G^{[1]}$ $G^{[2]}$

Figure 1: Illustration of the two types of counts of the pattern $G^{[P]}$ in graphs $G^{[1]}$ and $G^{[2]}$. The edge and node features are represented by colors.
An induced subgraph of a graph is another graph, formed from a subset of the vertices of the graph and all of the edges connecting pairs of vertices in that subset. Therefore the induced-subgraph-count of $G^{[P]}$ in $G^{[1]}$ is 0. A subgraph of a graph is another graph where the vertices and edges are a subset of the former's. No additional constraints, therefore the subgraph-count of $G^{[P]}$ in $G^{[1]}$ is 1.
Both induce-subgraph-count and subgraph-count for $G^{[P]}$ in $G^{[2]}$ are 0 since the edge features do not match.

Another important challenge in deep learning is the non convexity inherent to the training of neural networks and the multiplicity of local minima that can be returned during that step. Several approaches have been proposed to improve the intuition on the impressive generalization properties of the networks despite the highly non convex (and seemingly arbitrary) learning step. One of them relies on the relaxation of the traning step as a minimization problem on the space of measures, similarly to what had been previously applied to the super-resolution problem (see for example [2]).

In order to improve our understanding of the above questions, we focus on three independent research directions. In the first we study the applicability of deep learning to graph data, in a continuation of last year work. In the second we study applicability of machine learning techniques to sonar data from an experimental point of view. Finally, the last direction continues the work that was initiated in 2018-2019 regarding the improvement of existing relaxation approaches for the super resolution problem. By improving our understanding of relaxations over measures and their implementation via semidefinite programming on the simpler super-resolution problem, we hope to be able to develop a better understanding of the use of those relaxations for the training of neural networks.

3 Methods, Assumptions and Procedures

3.1 Procedures regarding funded expenses

Due to difficulties that were amplified by the COVID-19 pandemic we could not hire students to work in this project in this period. Since our issues with the PI's Institution administration did not improve, we are considering transferring the funds to the co-PI institution so we can hire students to work as interns in the final year of the grant.

3.2 Methodological approach to studying graph neural networks

We believe that in order for deep learning techniques to be useful in general data-science problems, and general domain sciences, they should be able to work with graph-structured data. In this vein, recently many models for deep learning in graphs had been proposed [3, 4, 5, 6]. We continue our work in the direction of understanding the computational limits of graph neural networks. In this direction we coauthored the

research article *Can graph neural networks count substructures?*, accepted for publication in arguably the top machine learning publication venue which is NeurIPS [7].

Graph neural networks are parametrized functions F_θ (implemented by neural networks with learnable parameters θ) that take a graph as an input and outputs either an embedding of the graph in Euclidean space, or, in some cases, a real number. Graphs are usually given by their adjacency matrix, and sometimes they can be endowed with node features or edge features (extra information that can be represented as a vector). Graphs as objects are independent of their representation. In particular the if the nodes are ordered in a different way the graph doesn't change but the adjacency matrix does. In this sense we can think as a graph to be represented by all adjacency matrices $\Pi A \Pi^T$ where A is its adjacency matrix and Π is a permutation matrix. Since graph neural networks aim to learn a function of the *graph* that does not depend on any specific representation, graph neural networks need to satisfy the following property:

$$F_\theta(\Pi A \Pi^T) = \Pi F_\theta(A).$$

This notion, known as equivariance with respect to permutations, can be written in more general terms, for instance, by thinking of Π as acting on each of the dimensions of a k-tensor.

Having these symmetries reduces the number of functions that can be expressed by graph neural networks, but it also introduces the technical difficulty of how to parametrize equivariant functions that can be implemented efficiently and are *expressive* enough. There is no consensus in the field for a gold-standard of graph neural networks, as opposed to image processing, where convolutional neural networks are without a doubt the winner of the field. This is a reason why graph neural networks has been such a fruitful area of research recently.

In the work supported by this grant last year we produced the following results:

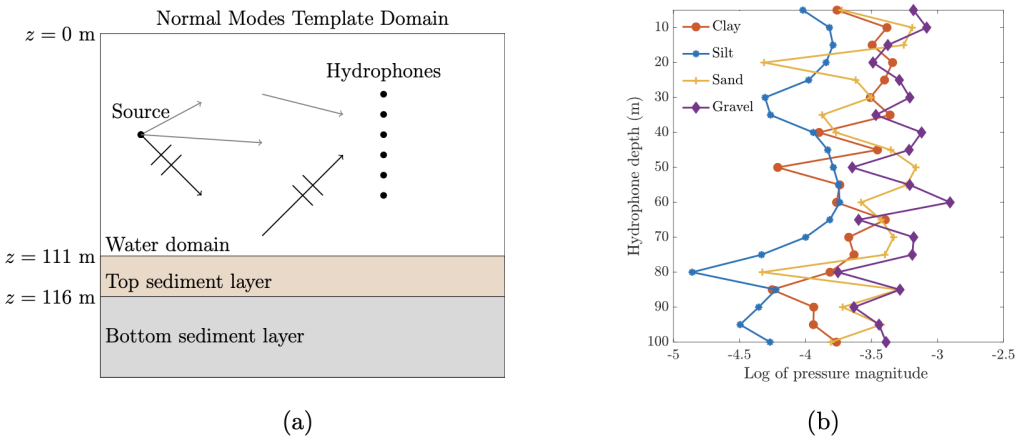
- We mathematically characterize what functions can and cannot be expressed by different types of graph neural networks in terms of the *graph isomorphism problem* [1],
- we use the insights from our mathematical theory to propose an architecture that overcome the shortcomings that other architectures have [1],
- we provide extensive numerical experiments to characterize the power of graph neural networks for certain problems.[8]

Our work this year uses our previous theory to answer the following question. Given a graph G and a pattern, substructure or motif a (for instance, one can think a to be a triangle, or a clique of size k) we consider the function f_a such that $f_a(G)$ is the number of times the substructure a is present in the graph G. We consider two ways of counting the substructure a, as a subgraph, or as an induced subgraph (see Figure 1 for a definition). The question is, can we express f_a as a graph neural network?

This is a very natural problem. Counting substructures is a natural functions on graphs, not only from a theoretical point of view, but also useful in computational biology applications [9].

3.3 Use of machine learning for sonar data

We collaborate with researchers in sonar and numerical methods and study acoustic models along with machine learning techniques to classify sediments in oceanic environments based on the geoacoustic properties of a two-layer seabed [11]. Two different scenarios are investigated. First, a simple low-frequency case is set up, where the acoustic field is modeled with normal modes. Four different hypotheses are made for seafloor



(a) (b)

Figure 2: (a) Sound propagation in an oceanic waveguide. (b) Simulated pressure field measurements for four top sediments (pressure is in μPa, simulations are produced with the standard simulation tool KRAKEN [10]).

sediment possibilities (sand, silt, clay and gravel) and these are explored using both various machine learning techniques and a simple matched-field approach. For most noise levels, the latter has an inferior performance to the machine learning methods. This setting is depicted in Figure 2.

Second, the high-frequency model of the scattering from a rough, two-layer seafloor is considered. Again, four different sediment possibilities are classified with machine learning. We illustrate the model in Figure 3 For higher accuracy, 1D Convolutional Neural Networks (CNNs) are employed. In both cases we see that the machine learning methods, both in simple and more complex formulations, lead to effective sediment characterization. Our results assess the robustness to noise and model misspecification of different classifiers.

3.4 Relaxations of non convex problems

In order to improve our understanding of relaxations over measures and their efficient implementation via semidefinite programming, we investigated two questions related to super resolution. The first question is the use of compressed semidefinite programs to solve the one dimensional super resolution problem (see for example [2]). The second is the study of the multidimensional problem and its application to radar. During the year 2018-2019, we generated numerical experiments which indicated that high level of compression (typically $< \frac{1}{\sqrt{n}}$) prevented the recovery of the measure through the resulting semidefinite programs.

During this second year, we focused on proving recovery of the measure through compressed semidefinite programs for compression rates matching the numerical experiments conducted in 2018-2019. As indicated in the previous report, proving recovery of the measure via compressed semidefinite programming requires constructing an interpolating polynomial p such that $1 - |p(\theta)|^2$ admits a decomposition as a sum of squares. In order to construct such a certificate in the compressed framework, we proposed to find such polynomial and its associated SoS decomposition in the non decimated framework and then sample those two components in order to obtain a corresponding certificate for the compressed formulation.

As a first step, we focused on proving that the ansatz based on the projector $\mathcal{P}_U^{\perp}$ (see 2018-

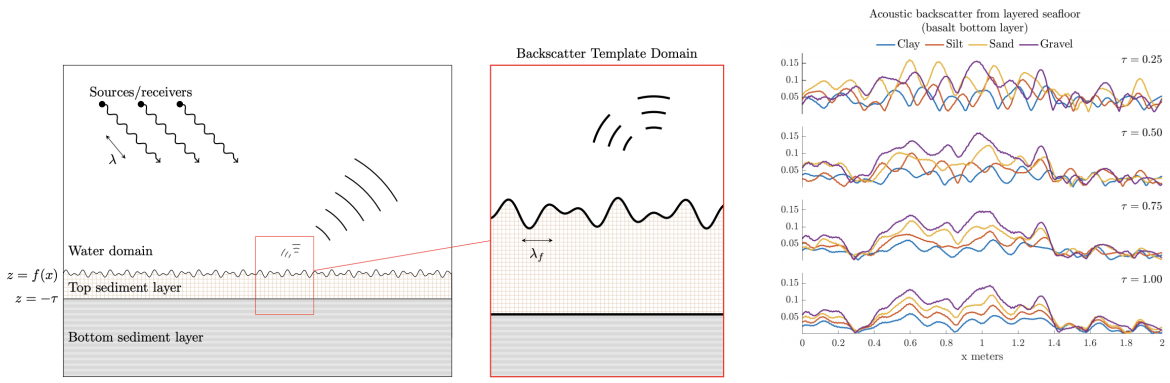


Figure 3: (left) High-frequency scattering from a two-layer seafloor with rough water-sediment interface given by $z = f(x)$. The acoustic wavelength λ of the incoming source ping is comparable to the spatial wavelength λf of the water-sediment interface. (right) Approximate backscatter signals corresponding to environments with varying sediments and top layer thicknesses (τ).

2019 report) and for which the SoS decomposition is explicitly known, corresponds to a valid certificate of optimality for the non decimated problem. The Gram matrix formulation of this ansatz reads as $\mathcal{P}_U^{\perp} + X$ where X is a correction that expands as $X = \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\varepsilon$ where the precise definition of $\mathcal{A}$ can be found in [12] and where ϵ encodes the deviation between the polynomial $p(\theta)$, which satisfies the interpolation condition but for which no sum of squares decomposition is known, and the polynomial p_U whose coefficients are encoded by the positive semidefinite projector $\mathcal{P}_U^{\perp}$ (and which thus has a sum of squares decomposition following from the eigenvalues decomposition of this projector) but for which the interpolation condition is not satisfied. The correction $\mathcal{P}_U^{\perp} + X$ satisfies the interpolation condition by definition of the operator $\mathcal{A}$ (see the 2018-2019 report). To prove that the resulting certificate also satisfies the remaning SoS condition (and hence to prove recovery of the measure), it suffices to show that the smallest eigenvalue of the Gram matrix $\mathcal{P}_U^{\perp} + X$ remains non negative. This can be done by controlling the smallest non zero eigenvalue of the operator $\mathcal{A}\mathcal{A}^*$ as well as the deviation ε. Those two steps are tackled in [12].

4 Results and Discussion

4.1 Can graph neural networks count substructures?

In our research article [7] we prove that unfortunately most mainstream graph neural networks cannot express functions that count substructures non-trivial substructures. Our results are summarized as follows:

1. We establish that neither Message Passing Neural Networks (MPNNs) [13] nor 2nd-order Invariant Graph Networks (2-IGNs) [14] can count any connected *induced subgraph* of 3 or more nodes. For any such pattern, we prove this by constructing a pair of graphs that provably cannot be distinguished by any MPNN or 2-IGN but with different induced-subgraph-counts of the given pattern. This result points at an important class of simple-looking tasks that are provably hard for classical GNN architectures.

2. We show that MPNNs and 2-IGNs can count *subgraphs* that are star-shaped, thus generalizing the results in [15] to incorporate node and edge features. We also show that k-WL and k-IGNs can count

subgraphs and *induced subgraphs* of size-k, which provides an intuitive understanding of the hierarchy of k-WL's in terms of increasing power in counting substructures.

3. While a negative result for general k-WL is difficult to obtain, we show that T iterations of k-WL is unable to count *induced subgraphs* that we call path patterns of $(k+1)2^T$ or more nodes. The result is relevant since real-life GNNs are often shallow, and also demonstrates an interplay between k and depth.

4. To exploit the fact that substructures present themselves in local neighborhoods, we present a novel GNN architecture that we call *Local Relation Pooling*, with inspirations from [16]. We empirically demonstrate that it performs on synthetic graphs as well as real both induced-subgraph-count and subgraph-count effectivel graphs, while also achieving competitive performance on real molecular datasets.

4.2 Can machine learning be effective for seabed classification?

We employed a model-based approach for designing training and test datasets of acoustic templates for capturing the relevant physics of representative patches of the seafloor. At low frequencies, this is accomplished with normal mode propagation, and at higher frequencies, local modeling on smaller computational domains enables fast, parallelizable simulations. An underlying assumption is the spatial stationarity of the seafloor, which is reasonable in situations where roughness statistics are similar over an area larger than the ensonified area. In this study we performed sediment classification in a two-layer seafloor, varying both geoacoustic parameters (sound speed, density) and geometric parameters (roughness, thickness) in the training and test data.

In low-frequency models, standard machine learning classifiers, such as logistic regression models and support vector machines, outperformed traditional matched-field processing techniques, especially when the test data had a low signal-to-noise ratio. Our results indicate that signals from certain classes have a higher likelihood of misclassification, namely silt and clay. On the other hand, predictions of signals from gravel and sand classes are more likely to be correct. For backscatter data, the standard machine learning classifiers demonstrated poor accuracy and did not generalize well to test environments with added noise and perturbed parameters. Some of the deep learning classifiers, namely AlexNet and GoogleNet, adapted to 1D signals, were more costly to train but demonstrated higher accuracy and improved generalization. We illustrate the performance of one of the classifiers in Figure 4. Our results also indicate that the layer thickness can have a significant influence on misclassification rates. Further investigation must place an emphasis on resolving the layer thickness with high accuracy.

Our results indicate the promise of machine learning and deep learning for the difficult problem of geoacoustic classification. The results from our simulations highlight the need to test models for a broad spectrum of environments to ensure generalization. Producing a well-performing deep learning model requires thorough experimental design. Several new directions can be explored with this framework, for example, finding elastic properties of sediments and incorporating the influence of the material type on the statistical properties of the seafloor roughness.

4.3 Solving relaxations over measures with compressed SDP

In [12], we provide a novel proof for the resolution of super-resolution through semidefinite programming that does not rely on the Fejér-Riesz theorem and gives an explicit form for the sum of squares decomposition of the optimality certificate. We hope to be able to combine this proof with an additional sampling step, in order to certify recovery of the measure by a more efficient compressed semidefinite program as indicated

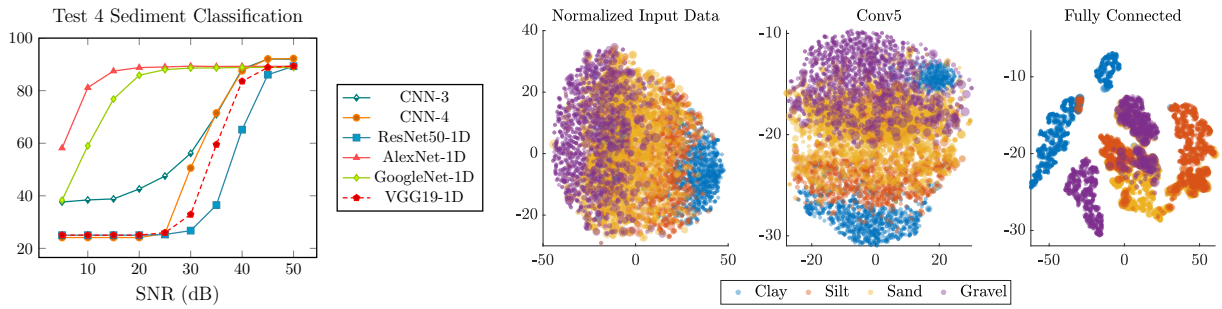


Figure 4: (left) We show the accuracy as a function of the SNR for different deep learning classifiers evaluated on a test set generated with slightly different parameters than the training set. We observe that all classifiers have similar performance at high SNR, but some of them generalize well to lower SNR settings, whereas other don't. This phenomenon is not theoretically understood. (right) t-SNE visualization of the activations in three different layers of the trained AlexNet-1D classifier applied to test data from the Test 4 environment.

in the 2018-2019 report. Our proof relies on the numerical computation of a lower bound on the singular values of a truncated matrix. The theoretical estimate on the truncation currently prevents this bound to be computed and we are working on a refined version of the proof that will make it possible to further reduce the memory needed to complete this last step.

4.4 Research articles that acknowledge the grant in this period

The research articles listed here were not listed in last year's report.

- **Published articles and accepted for publication:** [17], [7], [11], [18].
- **Preprints:** [19], [20], [12].

5 Conclusions

Work supported by this grant makes a theoretical contribution towards understanding the capabilities of deep learning on graphs. It further expands our premise that graph neural networks have important shortcomings and limitations to express basic functions on graphs. Our work was selected for publication in a top machine learning conference.

Regarding the work on sonar models. Our research shows that, for model-generated data, machine learning techniques exhibit better performance than classical geoacoustic inversion techniques. However, our results employ large amounts of model-generated data and has yet to be tested on real data. Moreover, our research shows that the quality of results vary substantially as a function of the deep learning architecture, and the differences may only be explicit when models are evaluated in a wide range of different testing environments. The next challenge we face is to train models on synthetic data, and apply them to real data. This has been successfully addressed in a similar sonar context [21] but many open questions remain regarding domain adaptation, transfer learning and generalization.

References

[1] Zhengdao Chen, Soledad Villar, Lei Chen, and Joan Bruna. On the equivalence between graph isomorphism testing and function approximation with gnns. In *Advances in Neural Information Processing Systems*, pages 15894–15902, 2019.

[2] Emmanuel J Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on pure and applied Mathematics*, 67(6):906–956, 2014.

[3] David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pages 2224–2232, 2015.

[4] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852, 2016.

[5] Haggai Maron, Heli Ben-Hamu, Hadar Serviansky, and Yaron Lipman. Provably powerful graph networks. In *Advances in Neural Information Processing Systems*, pages 2153–2164, 6 2019.

[6] Zhengdao Chen, Lisha Li, and Joan Bruna. Supervised community detection with line graph neural networks. *Internation Conference on Learning Representations*, 2019.

[7] Zhengdao Chen, Lei Chen, Soledad Villar, and Joan Bruna. Can graph neural networks count substructures? *Accepted for publication in NeurIPS*, 2020.

[8] Weichi Yao, Afonso S Bandeira, and Soledad Villar. Experimental performance of graph neural networks on random instances of max-cut. In *Wavelets and Sparsity XVIII*, volume 11138, page 111380S. International Society for Optics and Photonics, 2019.

[9] Noga Alon, Phuong Dao, Iman Hajirasouliha, Fereydoun Hormozdiari, and S Cenk Sahinalp. Biomolecular network motif counting and discovery by color coding. *Bioinformatics*, 24(13):i241–i249, 2008.

[10] Michael B Porter. The kraken normal mode program. Technical report, Naval Research Lab Washington DC, 1992.

[11] Christina Frederick, Soledad Villar, and Zoi-Heleni Michalopoulou. Seabed classification using physics-based modeling and machine learning. *Accepted for publication in The Journal of the Acoustical Society of America*, 2020.

[12] Augustin Cosse. Compressed super-resolution i: Maximal rank sum-of-squares. *arXiv preprint arXiv:2001.01644*, 2020.

[13] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1263–1272. JMLR. org, 2017.

[14] Haggai Maron, Ethan Fetaya, Nimrod Segol, and Yaron Lipman. On the universality of invariant networks. *arXiv preprint arXiv:1901.09342*, 5 2019.

[15] V Arvind, Frank Fuhlbrück, Johannes Köbler, and Oleg Verbitsky. On weisfeiler-leman invariance: Subgraph counts and related graph properties. *arXiv preprint arXiv:1811.04801*, 2018.

[16] Ryan L Murphy, Balasubramaniam Srinivasan, Vinayak Rao, and Bruno Ribeiro. Relational pooling for graph representations. *arXiv preprint arXiv:1903.02541*, 2019.

[17] Culver McWhirter, Dustin G Mixon, and Soledad Villar. Squeezefit: Label-aware dimensionality reduction by semidefinite programming. *IEEE Transactions on Information Theory*, 66(6):3878–3892, 2019.

[18] Efe Onaran and Soledad Villar. Efficient belief propagation for graph matching. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9060–9064. IEEE, 2020.

[19] Bianca Dumitrascu, Soledad Villar, Dustin G Mixon, and Barbara E Engelhardt. Optimal marker gene selection for cell type discrimination in single cell analyses. *BioRxiv*, page 599654, 2019.

[20] Andrew J Blumberg, Mathieu Carriere, Michael A Mandell, Raul Rabadan, and Soledad Villar. Mrec: a fast and versatile framework for aligning and matching point clouds with applications to single cell molecular data. *arXiv preprint arXiv:2001.01666*, 2020.

[21] David F Van Komen, Tracianne B Neilsen, Kira Howarth, David P Knobles, and Peter H Dahl. Seabed and range estimation of impulsive time series using a convolutional neural network. *The Journal of the Acoustical Society of America*, 147(5):EL403–EL408, 2020.