Natural Resources Canada
Ressources naturelles Canada

GEOLOGICAL SURVEY OF CANADA
OPEN FILE 9318e

Vector representations of the Canadian Geoscience Foundation model

M. Parsa

2025

GEOLOGICAL SURVEY OF CANADA
OPEN FILE 9318e

Vector representations of the Canadian Geoscience Foundation model

M. Parsa

2025

Recommended citation

Parsa, M., 2025. Vector representations of the Canadian Geoscience Foundation model; Geological Survey of Canada, Open File 9318e, 1 .zip file. https://doi.org/10.4095/pxvyky08xx

Publications in this series have not been edited; they are released as submitted by the author.

ISSN 2816-7155
ISBN 978-0-660-79985-8
Catalogue No. M183-2/9318E-PDF
https://doi.org/10.4095/pxvyky08xx

1. Introduction

The term “foundation model” refers to large-scale machine learning models that are trained on massive amounts of unlabeled data leveraging self-supervised learning techniques (Bommasani et al., 2021). Foundation models are developed on the basis of a specific type of deep neural network architecture, namely Transformers (Vaswani et al., 2017), and form the basis for a wide range of downstream applications across diverse fields, such as natural language processing (Brown et al., 2020), decision making (Yang et al., 2023), computer vision (Yuan et al., 2021), and multi-modal learning (Li et al., 2024). These models offer adoptability to multiple downstream tasks with little or no additional training and require much less labeled data for fine-tuning on specific tasks. These models can potentially outperform traditional task-specific models due to their scale and pretraining diversity (Bommasani et al., 2021).

The Geological Survey of Canada (GSC) has recently developed several data-driven, geospatial predictive models (e.g., Parsa et al., 2024a; 2025a), with more expected to follow. Nevertheless, several challenges hinder the advancement of this initiative. First, the geoscientific phenomena, such as mineralization, targeted by geospatial predictive modelling tasks are typically rare events, which limits the effectiveness of conventional data-driven algorithms (Parsa, 2021). Second, training models from scratch for each task is both resource-intensive and time-consuming (Parsa and Cumani, 2025). The Canadian Geoscience Foundation model (CGF) offers a promising solution to the aforementioned challenges by providing a general-purpose, adaptable AI model trained on extensive, diverse, pan-Canadian geoscientific datasets. Parsa et al. (2025b) introduced the first version of the CGF—a transformer-based model trained on pan-Canadian geochronological, geological, and geophysical data. This report presents the CGF model, outlines the development process, and aims to make the model publicly accessible.

This report is structured into four sections. Following this introduction, the Methods section describes the input data and the model development process. The third section presents the released model and provides guidelines as to how use the model.

2. Methods

2.1. Input Data

The dataset used to train the CGF is summarized in Table 1. In short, the dataset includes geophysical data, namely aeromagnetic and gravity data, derivatives of geophysical data, geochronological data indicating the minimum and maximum ages of bedrock geology units, and the embedding vectors derived from processing bedrock geology text data using a domain-adapted large language model developed and described by Parsa et al. (2025b). The geophysical and geochronological data were made publicly available in earlier publications (Parsa et al., 2024b; 2025c).

Table 1. Data used for developing the CGF. Data were converted to a matrix of 1,820,346 rows, representing H3 cells covering Canada, as and 790 columns.

1st vertical derivative of gravity data
Bouguer gravity
Horizontal gradient magnitude of gravity data
Vectors corresponding to gravity worms
1st vertical derivative of magnetic data
Horizontal gradient magnitude of aeromagnetic data
Reduced to pole aeromagnetic data
Tilt derivative of magnetic data
minimum age (geochronology)
maximum age (geochronology)
Vectors corresponding to magnetic worms
Bedrock geology map of Canada converted into numeric embedding vectors leveraging language modeling

The original data were collected in the form of raster maps and subsequently transformed into a multi-dimensional vector format (Parsa et al., 2025c). This transformation involved applying zonal statistics to assign raster values to individual H3 level 7 cells. H3 is a global, hierarchical hexagonal gridding system, with each level 7 cell covering approximately five square kilometers. For further details on the dataset used in the development of the CGF, readers are referred to Parsa et al. (2025b) and Parsa et al. (2025c).

The dataset was fed into the CGF as a matrix of 1,820,346 H3 cells as rows and 790 columns, where rows correspond to the individual H3 cells covering Canada and columns represent individual datasets (Table 1). Vectors corresponding to proximity to gravity and magnetic worms constitute 12 different columns in the data matrix, whereas numeric embedding vectors derived from language modeling correspond to 768 columns. The rest of entries in Table 1 appear as single columns in the data matrix.

The CGF operates exclusively on numerical data; however, bedrock geology data is typically presented in the textual format. As a result, the text must be converted into numerical format. Two methods have been used for this purpose: one-hot encoding (See Parsa et al., 2024a) and language modeling (Parsa et al., 2025a, b). While one-hot encoding is straightforward, it risks discarding important contextual information. In contrast, language modeling can capture and preserve most of the information embedded in the text. Parsa et al. (2025b) developed a domain-specific large language model by fine-tuning the original BERT model (Bidirectional Encoder Representations from Transformers: Devlin et al., 2019) using extensive geoscience text corpora. The language model transforms any input text, regardless of length, into a corresponding 768-dimensional embedding vector. Text data assigned to H3 cells from bedrock geology polygons was input into the language model, leading to a matrix of 1,820,346 H3 cells as rows and 768 numeric embedding vectors as columns.

2.2. Modeling Approach

The CGF was trained using a self-supervised approach known as masked regression modeling, in which 15% of the input data was masked and the model was tasked with imputing the masked values. The effectiveness of this approach was evaluated using a loss function, with the model achieving a validation loss of 0.65.

The CGF comprises several deep neural networks, including a regressor, Transformer layers, and a multilayer perceptron (MLP) head. The Transformer layers project each input data instance into a 790-dimensional numeric vector, which is then compressed by the MLP head into a 10-dimensional representation. This compact representation facilitates the efficient storage and processing of CGF embeddings. Readers are referred to Parsa et al. (2025b) for details on the modeling approach.

3. Vector Representations of the Model

"The 50-dimensional representation vectors are provided in a comma-separated values file titled 'Data.csv'. This file contains 1,820,346 rows, each corresponding to a unique H3 cell in Canada, and 11 columns. The first column contains the H3 addresses, which users can utilize for mapping purposes. The remaining 50 columns, labeled $R1$ to $R10$, represent the vector embeddings generated by the CGF for each H3 cell.

Acknowledgements

This report is part of Geological Survey of Canada's Critical Minerals and Geoscience Data (CMGD) Program.

REFERENCES

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., 2021. On the opportunities and risks of foundation models. ArXiv Prepr. ArXiv210807258.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. Presented at the Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pp. 4171–4186.

Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., Gao, J., 2024. Multimodal foundation models: From specialists to general-purpose assistants. Found. Trends® Comput. Graph. Vis. 16, 1–214.

Parsa, M., 2021. A data augmentation approach to XGboost-based mineral potential mapping: an example of carbonate-hosted ZnPb mineral systems of Western Iran. J. Geochem. Explor. 228, 106811.

Parsa, M., Cumani, R., 2025. Class Label Representativeness in Machine Learning-Based Mineral Prospectivity Mapping. Nat. Resour. Res. 1–25.

Parsa, M., Lawley, C.J., Cawood, T., Martins, T., Cumani, R., Zhang, S.E., Thompson, A., Schetselaar, E., Beyer, S., Lentz, D.R., 2025a. Pan-Canadian Predictive Modeling of Lithium–Cesium–Tantalum Pegmatites with Deep Learning and Natural Language Processing. Nat. Resour. Res. 1–30.

Parsa, M., Cumani, R., Fam, H.J.A., Tawbe, B., 2025b. Large Language Models and Geoscience Transformers for Predictive Mapping of Canadian Critical Minerals. Nat. Resour. Res. 1–30. https:// /doi.org/10.1007/s11053-025-10564-0

Parsa, M., Lawley, C.J., Cumani, R., Schetselaar, E., Harris, J., Lentz, D.R., Zhang, S.E., Bourdeau, J.E., 2024. Predictive Modeling of Canadian Carbonatite-Hosted REE+/− Nb Deposits. Nat. Resour. Res. 1–25.

Parsa, M., Lawley, C.J.M., Cumani, R., Schetselaar, E., Harris, J., Lentz, D.R., Zhang, S.E. and Bourdeau, J.E., 2024b. Data for Prospectivity Mapping of Canadian Carbonatite-hosted REE+/-Nb Deposits. Geological Survey of Canada.

Parsa, M., Lawley, C.J.M., Schetselaar, E., Martins, T., and Thompson, A., 2025c. Multicomponent geospatial data for predictive mapping of Canadian critical minerals; Geological Survey of Canada, Open File 9270, 1 .zip file. https://doi.org/10.4095/p80q4bhma6

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Yang, S., Nachum, O., Du, Y., Wei, J., Abbeel, P., Schuurmans, D., 2023. Foundation models for decision making: Problems, methods, and opportunities. ArXiv Prepr. ArXiv230304129.

Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., 2021. Florence: A new foundation model for computer vision. ArXiv Prepr. ArXiv211111432.