

Reply to Anonymous Referee #2' comments and suggestions

The manuscript presents an attribution-guided DCN-Transformer framework for mineral prospectivity mapping, a relevant modelling problem within the scope of GMD. Combining deformable convolution, attention, and model explanation may be useful. However, substantial concerns currently limit confidence in the central claims of local-global modelling, predictive generalization, interpretability, and methodological novelty. The manuscript needs a more rigorous evaluation and a substantially more complete, precise description before its conclusions can be assessed reliably.

Re: Thank you for your interest. We have addressed all your comments and suggestions. I hope you are satisfied with this revision.

The abstract is substantially too long and reads more like an extended introduction and methods summary than a conventional abstract. It should be shortened to a concise, self-contained version accounting for the research question, the specific methodological contribution, the study setting/data, the evaluation design, the principal quantitative result with appropriate qualification, and the main supported conclusion. All the background details, claims, and unsupported statements should be removed.

Re: The abstract has been rewritten.

Mineral prospectivity mapping (MPM) delineates areas favorable for a given deposit type by integrating multi-source geoscientific data, but deep learning models for this task are widely regarded as black boxes. We present an interpretable hybrid deep network, DCN-Attribution-Transformer, that learns irregular spatial patterns while exposing the contribution of each evidence layer rather than merely post-processing it. The model couples a deformable convolutional network (DCN) for adaptive spatial feature extraction with a Transformer classification head. A DeepLIFT-style attribution branch measures how far each input deviates from a learned background baseline, and these contribution scores are fed back as bias terms into the Transformer attention, so that factor relevance both explains and guides the prediction. We apply the framework to granite-type Li-Be-Nb-Ta prospectivity mapping in the Nanling region, South China, using 39 geochemical elements and oxides together with fault, Yanshanian-granite, and ore-controlling-strata layers. Under five-fold cross-validation, the proposed model achieved the highest mean AUC of 0.958, accuracy of 0.887,

and F1-score of 0.886, exceeding the DCN-Transformer, standalone DCN, Transformer, and standard CNN-based variants. The attribution branch reproduces known metallogenic controls and provides pixel-level, per-deposit factor contributions. These results indicate that embedding DeepLIFT-style attribution into the attention mechanism improves both predictive performance and geological interpretability of data-driven MPM.

It is difficult to read the manuscript because the geological setting chapter (2) is not connected with the methodology. The authors should clearly explain how the geological setting and data characteristics motivate the modelling choices. Partly this is addressed in the introduction, but due to the length of the data chapter, the connection is lost.

Re: We agree. We have added a new paragraph at the end of Section 2 that directly links each architectural component to the specific geological and data characteristics of the Nanling region.

These geological characteristics and data characteristics directly motivate the modelling choices adopted in this study. The complex, multi-phase tectonic evolution and the irregular spatial distribution of granitic, faults, and geochemical anomalies necessitate a model capable of adapting to spatially variable and geometrically irregular patterns, hence the adoption of deformable convolution. Furthermore, ore-controlling factors (42 evidence layers) and their varying importance across different deposit types calls for an interpretable attribution mechanism that can quantify factor contributions, motivating the DeepLIFT inspired attribution branch. These contribution scores can be used to modulate the key values within the Transformer's attention mechanism, which enhances the Transformer to focus on the most diagnostically relevant feature interactions.

The manuscript presents a potentially useful integration of deformable convolution, Transformer-based spatial modeling, and attribution-guided attention for mineral prospectivity mapping. However, to my opinion its scientific novelty is currently moderate because the principal architectural and interpretability components are largely established. The authors should clarify how the attribution module is genuinely distinct from DeepLIFT, define the geological prior knowledge incorporated into the model, and moderate the novelty claims unless these issues are resolved.

Re: We have revised the manuscript to clarify the novelty of our attribution module, define the nature of the incorporated knowledge, and moderate our novelty claims.

Each point is responded as follows:

(1) How the attribution module is distinct from DeepLIFT:

We have substantially rewritten Section 3.1 to clarify the relationship between our attribution network and the original DeepLIFT method. The key distinction is as follows:

Original DeepLIFT (Shrikumar et al., 2017) is a post-hoc attribution method that propagates contribution scores backward through a trained network via multiplicative rules, comparing each neuron's activation against a reference activation. It is applied after training and does not affect the model's learning process. The proposed attribution network embeds the DeepLIFT principle of reference-baseline differencing into a feed-forward, trainable branch that is optimized end-to-end alongside the prediction task. The differential representation directly implements the DeepLIFT concept of quantifying feature contribution by deviation from a reference baseline. However, rather than propagating scores backward through a fixed network, our approach learns the attribution mapping jointly with the feature extraction and classification components, enabling the attribution scores to serve as learnable priors that actively shape the Transformer's attention mechanism during training. This integration creates a dual-role attribution mechanism: (a) post-hoc interpretability, the trained attribution scores quantify factor contributions for individual predictions; and (b) ante-hoc interpretability, attribution scores modulate the Key values in the Transformer attention, injecting data-driven attribution priors into the global feature learning process. This dual role is the primary novelty of our attribution design, going beyond the standard post-hoc application of DeepLIFT.

(2) Definition of the geological prior knowledge:

We have clarified in the revised Section 3.1 that the attribution priors are learned from the training data rather than being directly specified by geological experts. The reference baseline is computed as the mean of the encoded training-set features, representing the typical non-mineralized background condition. High attribution scores therefore correspond to geochemical and geological anomalies relative to this background. While these scores are data-driven, they naturally align with established ore-controlling patterns (e.g., proximity to major faults, Yanshanian granites, and W-Sn-Li-Be element associations), as validated by the metallogenic model of the Nanling region. We have revised the manuscript to avoid describing the learned scores as "geological prior knowledge" without qualification; instead, we characterize them as "data-driven attribution priors" that bridge data-driven learning and geological understanding.

(3) Moderation of novelty claims:

We have moderated our novelty claims in the manuscript. In the Abstract and Introduction, we now frame the contribution as a novel integration and extension of established components, the end-to-end integration of DeepLIFT-based attribution into a DCN-Transformer hybrid architecture, where the attribution scores serve both interpretability and attention modulation roles, rather than claiming novelty in the individual components themselves. We acknowledge that DCN, Transformer, and DeepLIFT are established methods, and our contribution lies in their synergistic integration and the dual-role attribution mechanism. Besides, we have removed or moderated all excessive claims throughout the manuscript. Specifically, "state-of-the-art" and "new benchmark" have been removed from the Results and Conclusions; and the Introduction now positions the work as an integration of established components (DCN, Transformer, DeepLIFT) with attribution-guided attention mechanism.

References:

Shrikumar, A., Greenside, P., and Kundaje, A.: Learning important features through propagating activation differences. In International conference on machine learning, pp. 3145-3153, 2017.

The explanation method needs a clearer and more accurate description. Equations (1)-(4) appear to define a learned branch that identifies features differing from the average training data. This does not look like the established DeepLIFT method, although the manuscript calls it DeepLIFT-based. The manuscript later also uses Integrated Gradients, but it is unclear what each approach contributes. The authors should clearly explain what each method does, what result it produces, and how it influences the Transformer. They should also state whether the claimed geological prior knowledge comes from independent geological expertise or is learned from the data. Learned scores should not automatically be described as geological prior knowledge. Finally, the authors should show that the explanations are reliable, for example by checking whether they remain similar across repeated training runs and after reasonable changes to the input data.

Re: We thank the reviewer for this detailed comment. We have made revisions to address each point.

The responses are organized below:

(1) Clarification that Equations (1)–(4) implement the DeepLIFT principle:

We have substantially rewritten Section 3.1 to clarify the relationship between our attribution network and the original DeepLIFT method.

Original DeepLIFT (Shrikumar et al., 2017) is a post-hoc attribution method that propagates contribution scores backward through a trained network via multiplicative rules, comparing each neuron's activation against a reference activation. It is applied after training and does not affect the model's learning process. The proposed attribution network embeds the DeepLIFT principle of reference-baseline differencing into a feed-forward, trainable branch that is optimized end-to-end alongside the prediction task. The differential representation directly implements the DeepLIFT concept of quantifying feature contribution by deviation from a reference baseline. However, rather than propagating scores backward through a fixed network, our approach learns the attribution mapping jointly with the feature extraction and classification components, enabling the attribution scores to serve as learnable priors that actively shape the Transformer's attention mechanism during training. This integration creates a dual-role attribution mechanism: (a) post-hoc interpretability, the trained attribution scores quantify factor contributions for individual predictions; and (b) ante-hoc interpretability, attribution scores modulate the Key values in the Transformer attention, injecting data-driven attribution priors into the global feature learning process. This dual role is the primary novelty of our attribution design, going beyond the standard post-hoc application of DeepLIFT.

(2) Removal of Integrated Gradients and clarification of DeepLIFT's dual contribution:

We agree with the reviewer that the simultaneous use of DeepLIFT and Integrated Gradients created confusion about each method's role. We have removed Integrated Gradients entirely from the manuscript and unified all attribution analysis under the DeepLIFT-based attribution branch. DeepLIFT is inherently a local interpretability method that computes per-sample attribution scores by propagating activation differences; this is precisely what produces the per-deposit factor contribution results. This simplifies the interpretability framework and eliminates the ambiguity. The revised manuscript now clearly describes the two contributions of the DeepLIFT-based attribution branch:

Post-hoc interpretability: The attribution scores generated by the branch quantify the contribution of each ore-controlling factor to individual predictions. At the global scale, aggregated scores yield an overall factor importance ranking. At the local scale, per-pixel scores reveal the specific factors driving predictions at individual mineral deposit locations, enabling spatially explicit interpretation.

Ante-hoc attention modulation: The attribution-weighted feature representation is projected to produce the Key values for the Transformer attention mechanism (Section 3.3). This means the Transformer's attention is not based solely on raw feature similarity but is modulated by data-driven

attribution scores, enhancing the prominence of geologically anomalous regions in the attention distribution.

Accordingly, Section 3.5 now describes only two visualization techniques (Grad-CAM and Transformer attention mapping), and the Integrated Gradients reference (Sundararajan et al., 2017) has been removed from the bibliography.

(3) Definition of the geological prior knowledge:

We have clarified in the revised Section 3.1 that the attribution priors are learned from the training data rather than being directly specified by geological experts. The DeepLIFT reference was defined as the mean of the standardized, labeled non-mineralized samples in the training set. Attribution values therefore quantify each feature's contribution to mineralization prediction relative to this empirical non-mineralized reference state. High attribution scores therefore correspond to geochemical and geological anomalies relative to this background. While these scores are data-driven, they naturally align with established ore-controlling patterns (e.g., proximity to major faults, Yanshanian granites, and W-Sn-Li-Be element associations), as validated by the metallogenic model of the Nanling region. We have revised the manuscript to avoid describing the learned scores as "geological prior knowledge" without qualification; instead, we characterize them as "data-driven attribution priors" that bridge data-driven learning and geological understanding.

(4) Reliability of the explanations:

The key factors identified by the attribution analysis-proximity to major crustal faults, Yanshanian granites, and Li-Be-Nb geochemical associations-are consistent with the well-established metallogenic model of the Nanling region, where Li-Be-Nb-Ta mineralization is genetically linked to highly evolved Yanshanian granites and structurally controlled by NE-NNE trending faults. This independent geological validation supports the reliability of the learned attributions.

References:

Shrikumar, A., Greenside, P., and Kundaje, A.: Learning important features through propagating activation differences. In International conference on machine learning, pp. 3145-3153, 2017.

The manuscript should include a systematic ablation study separating the contributions of the deformable convolution, Transformer module, attribution-guided attention, and attribution regularization. The authors should add component-removal experiments and compare the complete model with appropriate reduced variants to demonstrate the effectiveness and necessity of each

proposed component. Without such analysis, the claimed methodological improvements and scientific novelty remain insufficiently supported.

Re: To isolate the contribution of the attribution branch, the prospectivity maps produced by three comparative architectures, a standalone DCN, a standalone Transformer, and a DCN-Transformer hybrid without the attribution component, were generated under identical training and inference settings. Besides, to separate the effect of deformable convolution from that of the attribution-guided attention, a standard CNN, a CNN-Transformer without attribution and a CNN-Attribution-Transformer, are further compared and evaluated with the above DCN-based architectures. The details can be found in Section 4.1 and 4.2.

The input is a 9 x 9 x 42 neighbourhood (chapter 4) centered on each target pixel. With the reported 1 km grid, direct attention is consequently confined to approximately a 9 km local window; it cannot directly learn relationships among locations hundreds of kilometres apart, as repeatedly claimed in the manuscript. Producing a full-map prediction or visualizing map-wide scores does not itself establish that the model attended jointly to distant locations. The authors should document the tokenization and full-map inference procedure, including how window-level attention is mapped to regional outputs. I think one should either define and consistently use a window-scale meaning of "global", or demonstrate that the model incorporates the regional-scale context.

Re: We agree with the reviewer on this critical point. The input is indeed a 9x9 neighborhood window, and the Transformer self-attention operates within this window (approximately 9 km at 1 km grid resolution), not across hundreds of kilometers. Our original use of "global" was inaccurate.

We have therefore made the following substantial revisions:

(1) We have changed the manuscript title from "Considering mineralization local-global geological features: An interpretable DCN-Transformer hybrid model with attribution for mineral prospectivity mapping" to "An interpretable DCN-Transformer hybrid model with attribution for mineral prospectivity mapping," removing the local-global framing.

(2) We have removed all claims about the model learning "long-range," "global-scale," or "hundreds of kilometers" relationships throughout the abstract, introduction, methods, results, discussion, and conclusion. The Transformer is now consistently described as modeling "spatial dependencies" or "contextual relationships" within the input neighborhood window.

(3) We have reframed the manuscript core contribution from "capturing local and global features" to "providing multi-faceted interpretability" for deep learning-based MPM. The focus is now on how the DeepLIFT-based attribution network, Grad-CAM, and attention maps collectively explain the model decision-making process at both regional and deposit-specific scales.

In the results authors describes 23 positive and 23 negative locations, augments these to 196 samples per class, and only then partitions the data 8:2 into training and validation sets. This sequence risks placing augmented versions, or strongly overlapping neighborhoods, of the same original location in both subsets and can materially inflate reported performance. I would recommend to split original deposits/locations before any augmentation and apply augmentation only inside each training fold. And a completely untouched outer test set, with the split identifiers disclosed, is needed for final performance claims.

Re: We agree.

Prior to model training, data augmentation, which was adopted in Yang and Zuo (2024), was performed. However, to avoid augmented versions, or strongly overlapping neighborhoods of the same original location in both subsets, we split original deposits before any augmentation. The original dataset was partitioned into training and validation sets at an 8:2 ratio. This augmentation strategy results in 314 (157 positive and negative samples, respectively) and 78 (39 positive and negative samples, respectively) samples for training and validation, respectively.

The reported AUC improvement over the DCN-Transformer baseline is 0.960 versus 0.951, which is a relatively small difference.

At the same time, no repeated-run variability, confidence interval or statistical test analysis is reported. The manuscript should provide per-run results over multiple seeds, uncertainty intervals, raw numerical metric tables, and ROC and precision-recall curves. Ideally any paired comparison must be performed on a genuinely independent test set. Until such evidence is supplied, claims of "statistically significant" improvement, "state-of-the-art" performance, and a "new benchmark" should be substantially moderated.

Re: We agree with that statistical validation is essential. We have addressed this comment by implementing five-fold cross-validation with confidence intervals, and raw numerical metric tables. We have also moderated our claims throughout the manuscript. The specific changes are as follows:

We employed a 5-fold cross-validation strategy to randomly partitioned original deposits into five equally sized folds before data augmentation. In each iteration, four folds were used for training and the remaining fold served as the validation set. This process was repeated five times such that every sample participated in both training and validation exactly once. All models were trained under identical hyperparameter configurations across all folds to ensure fair comparison. The final performance metrics are reported as the mean $\pm$ standard deviation (SD) across the five folds, with 95% confidence intervals (95% CI) derived from the t-distribution. For the proposed model, the mean AUC is 0.958 (SD = 0.018, 95% CI: 0.94 – 0.976), and for the DCN-Transformer, the mean AUC is 0.947 (SD = 0.027, 95% CI: 0.92 – 0.974).

Overall, the authors should position the work explicitly against close prior work on CNN-Transformer mineral prospectivity mapping, deformable convolution for geochemical anomaly recognition, and interpretable or attention-branch MPM models. A comparison table should identify the exact new component and its intended benefit. Controlled ablations should isolate the effects of DCNv2, Transformer fusion, and attribution guidance, rather than only comparing whole architectures. This is necessary to demonstrate novelty, give proper credit to related work, and establish that the contribution is more than a new combination of existing modules.

Re: We agree. We have strengthened the positioning of our work against closely related prior studies, a comparison with CNN-based models.

We expanded comparison with CNN-based models. In addition to the previously compared DCN, Transformer, and DCN-Transformer models, we have now included three additional baselines: standalone CNN, CNN-Transformer (without attribution), and CNN-attribution-Transformer. These additions enable controlled comparisons that directly isolate the effect of each architectural choice:

- DCN vs. CNN (AUC: 0.934 vs. 0.832): Isolates the contribution of deformable convolution over standard convolution for local feature extraction.
- DCN vs. DCN-Transformer (AUC: 0.934 vs. 0.947): Isolates the contribution of deformable convolution over standard convolution for local feature extraction.
- DCN-Transformer vs. CNN-Transformer (AUC: 0.947 vs. 0.826): Demonstrates that the Transformer mechanism provides a consistent improvement.
- DCN-Attribution-Transformer vs. CNN-Attribution-Transformer (AUC: 0.958 vs. 0.855): Confirms that the deformable convolution contributes more even when the attribution mechanism is present.

- CNN-Attribution-Transformer vs. CNN-Transformer (AUC: 0.855 vs. 0.826): Demonstrates that the attribution mechanism provides a consistent improvement regardless of whether DCN or CNN is used as the local backbone, confirming the generalizability of the attribution guidance.

The specific performance indicators can be found in Table S1 in Supporting Information file.

Figure 4 should use a common color scale bar, and the symbols repeat for all 4 panels, so it is better to place one legend for all panels

Re: Revised.

In Eq. (11), the term "FTN" is undefined and appears to be a typographical error for "TN" (True Negative). In addition, the expression for the expected agreement, P_e , uses $n/2$ for both classes. This is valid only if the evaluation set contains exactly equal numbers of positive and negative samples; otherwise, the standard expression based on the observed class and prediction totals should be used.

Re: We agree. As you can see in Section 4, the positive and negative samples maintain a balance in our study.

P19L408 - The statement that the threshold-dependent metrics were calculated at an "optimal probability cutoff" is also insufficient. Accuracy, precision, recall, F1, Kappa, and MCC all depend on this cutoff, whereas AUC does not. The authors should state the numerical cutoff, the criterion used to select it, and the data used for selection. The authors should also report the resulting confusion matrix for the test set.

Re: We agree. Mineral prospectivity mapping is formulated as a binary classification task (prospective vs. non-prospective). Because the positive (mineral deposit) and negative (non-deposit) samples in both the training and test sets are fully balanced through random-drop data augmentation, no class-prior adjustment is required, and the conventional probability threshold of 0.5 was adopted throughout the experiments, consistent with common practice in mineral prospectivity mapping studies.

Would be good to supplement the radar chart in Fig. 5 with a numerical table of all metrics and their uncertainty. The sample count underlying every reported metric should be stated.

Re: The numerical table corresponding to radar chart has been supplemented in Table S1 in the supplementation file.

The manuscript should be edited throughout for grammar as well as for imprecise or promotional language. In particular, "global", "geological prior knowledge", "DeepLIFT-based", "interpretability", "statistically significant", and "state-of-the-art" should be used only where the method and evidence support those terms.

Re: These statements have been fully revised throughout the manuscript.