minerals

MDPI

Article

Toward Robust Mineral Prospectivity Mapping: A Transformer-Based Global–Local Fusion Framework with Application to the Xiadian Gold Deposit

Xiaoming Huang [1,2], Pancheng Wang [1,2] and Qiliang Liu [1,2,*]

[1] Key Laboratory of Metallogenic Prediction of Nonferrous Metals and Geological Environment Monitoring (Ministry of Education), School of Geosciences and Info-Physics, Central South University, Changsha 410083, China
[2] Hunan Key Laboratory of Nonferrous Resources and Geological Hazards Exploration, Changsha 410083, China
* Correspondence: qiliang.liu@csu.edu.cn

Abstract

As mineral exploration increasingly targets deeper and more geologically complex terrains, the need for reliable predictive models becomes critical to mitigating exploration risk and improving cost efficiency. Correspondingly, the effectiveness of deep mineral exploration strategies depends substantially on the effectiveness and precision of three-dimensional mineral prospectivity mapping (3D MPM) models. However, the inherent spatial non-stationarity—where ore grade variability changes across geological domains—and the strongly skewed distribution of high-grade samples present a dual challenge. Conventional methods, which primarily rely on mean-based regression, often struggle to adequately address this dual challenge, limiting their predictive performance in complex geological settings. To address these issues, this paper proposes a pinball-loss-guided, global–local fusion Transformer model within a unified framework for 3D MPM. It leverages a multi-head self-attention mechanism with global–local fusion to capture long-range dependencies and global geological contexts, while incorporating local feature extraction modules to adaptively model spatially varying mineralization controls, jointly optimized through a pinball loss function to address mineralization distribution skewness. The proposed framework was first rigorously evaluated using the Xiadian gold deposit as a case study. Bootstrap analysis of the ablation experiments confirmed its predictive performance in terms of quantile-specific accuracy and prediction interval (PI) calibration. Ten rounds of random data splits provided further confirmation of the model's stability. Subsequently, the validated model was applied to prospectivity mapping in unexplored regions, leading to the delineation of several high-potential exploration targets. Finally, comparative analyses with state-of-the-art machine learning methods were conducted, which further validated the competitive fitting capability of the proposed framework.

Keywords: Transformer; global–local fusion; 3D MPM; hydrothermal deposit; quantile-based regression

Check for updates

Academic Editor: Behnam Sadeghi

Received: 20 January 2026
Revised: 9 March 2026
Accepted: 17 March 2026
Published: 20 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Three-dimensional mineral prospectivity mapping (3D MPM) has gained significant attention in mineral exploration, which is essential for targeting concealed deposits and mitigating the high risks of deep-seeking exploration. However, its success depends entirely

on predictive model accuracy and effectiveness. Advancing methodologies to enhance the model's performance is therefore crucial—transforming data into discoveries and ensuring 3D MPM delivers tangible economic and operational outcomes.

3D MPM has evolved substantially since its inception in the 1980s [1], achieving extensive application through the integration of multi-source geological data within a volumetric spatial framework [2–6]. While these methodologies effectively establish quantitative relationships between known mineralization occurrences and geological features, their dependence on stationary assumptions and linear mathematical frameworks often constrains their capacity to fully represent the complex, non-stationary dynamics inherent in ore formation [7,8]. The genesis of mineral deposits is fundamentally a multi-stage, nonlinear process, typically involving the prolonged and heterogeneous interaction of geological, geochemical, and hydrodynamic factors across varying spatial and temporal scales [9–11]. Driven by extreme chemical gradients and locally favorable conditions, this process creates a highly skewed distribution of mineral concentration. As a result, high-grade ore is a statistical outlier against a broad low-grade background [12]. This characteristic fundamentally challenges the predictive model [13].

More recently, the rapid advancement of machine learning (ML)—particularly deep learning (DL)—has further transformed the field by enabling the capture of complex nonlinear relationships between predictor variables and mineralization, resulting in improved predictive capability and significantly enhanced accuracy [14–16]. However, these methods are limited to mean-based estimations, which fundamentally struggle to capture the strongly skewed distributions with extreme values inherent in mineralization [17]. A key consequence of these limitations is the underestimation of high-grade mineralization, as predictions tend to be oversimplified and fail to reflect its true spatial distribution [18].

The quantile-based model, introduced by Koenker & Bassett in 1978 [19], has gained broad adoption across disciplines from economics to medicine due to its ability to model complete conditional distributions through a specialized pinball loss function [20–23]. By estimating multiple quantiles, the quantile-based model does not merely predict a central "average" prospectivity but captures the full spectrum of mineralization potential. This is critically robust to the skewed distribution and extreme values inherent in ore-forming processes [24–28].

The integration of deep learning architectures with quantile-based learning principles [29–33] may present a powerful paradigm for addressing the two fundamental challenges in 3D MPM: spatial non-stationarity and the strongly skewed, heavy-tailed distribution of mineralization. The Transformer's self-attention mechanism, particularly when enhanced with relative positional encoding, dynamically weights relationships across spatial locations, naturally capturing localized processes and non-stationary patterns without requiring predefined spatial kernels. Furthermore, when integrated with the pinball loss, it can effectively model heavy-tailed distributions and extreme values, overcoming the limitations of traditional methods that assume stationarity and Gaussian distributions [29–33].

Leveraging the pinball loss, this study presents a global–local (G-L) fusion mechanism for 3D MPM within a unified Transformer framework. Specifically, the relative position encoder explicitly models local geometric constraints, while the fused attention mechanism adaptively aggregates information across multiple scales. The pinball loss function orchestrates this process by directly optimizing for quantile estimates, ensuring robust predictions that are less sensitive to skewed distributions and extreme values. The framework's efficacy is rigorously demonstrated through three key steps: a bootstrap-based ablation study and a stability test were performed to verify both the contribution of individual components and the overall stability of the proposed model; the identification of potential exploration

targets based on these insights; and a comparative analysis that highlights its competitive fitting capability against state-of-the-art machine learning methods.

2. Study Area and Data

The Xiadian gold deposit, a typical hydrothermal deposit, is located in Zhaoyuan City, Shandong Province, within the core of the Jiaodong Peninsula's gold province, one of China's most important gold districts (Figure 1). Geologically, it is controlled by the Zhaoping Fault, a subsidiary structure of the Tan-Lu Fault Zone. The ore bodies are mainly hosted in the sericite–quartz–pyrite alteration zone along the contact between Mesozoic Linglong granite and Archean metamorphic rocks. It is a typical altered rock-type gold deposit formed by Mesozoic magmatic–hydrothermal activity around 120 million years ago, characterized by disseminated mineralization and clear alteration zoning, strictly controlled by fault structures [17,18,34,35].

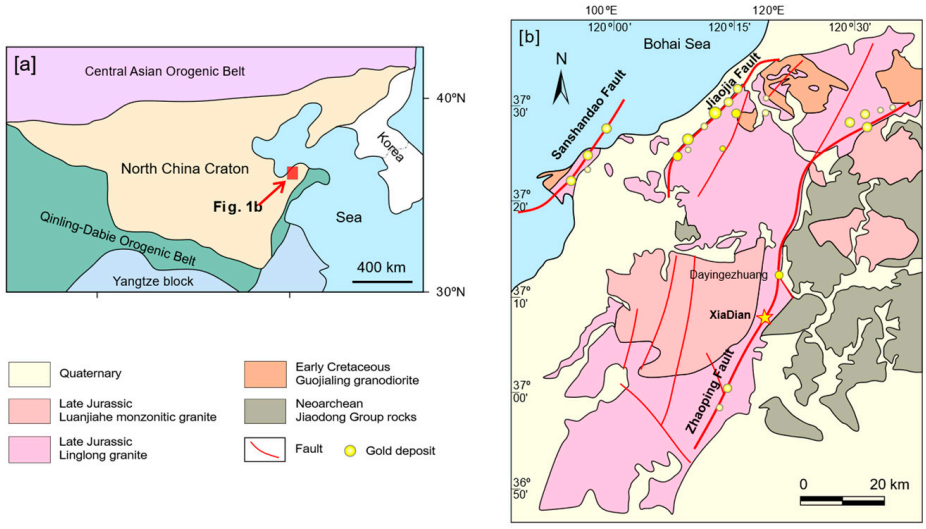
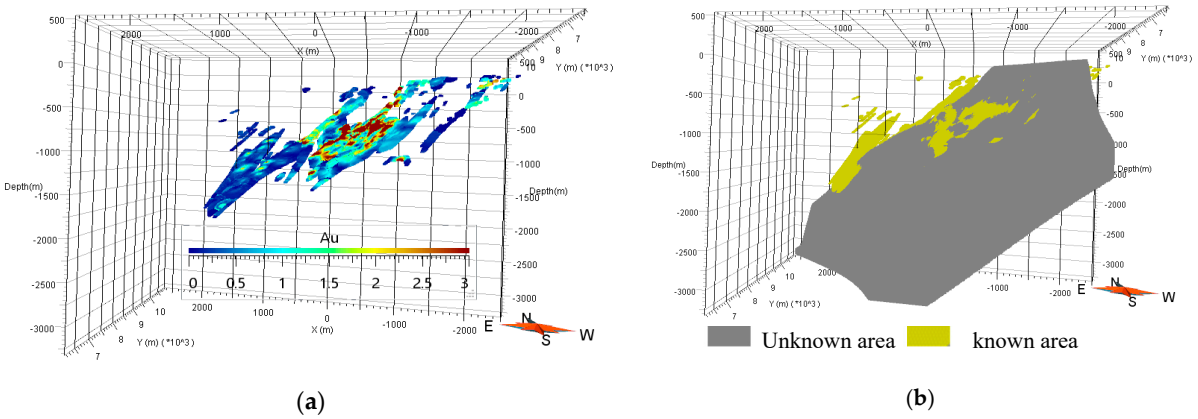


Figure 1. Geological map of the Xiadian gold deposit: (**a**) location; (**b**) strata and faults.

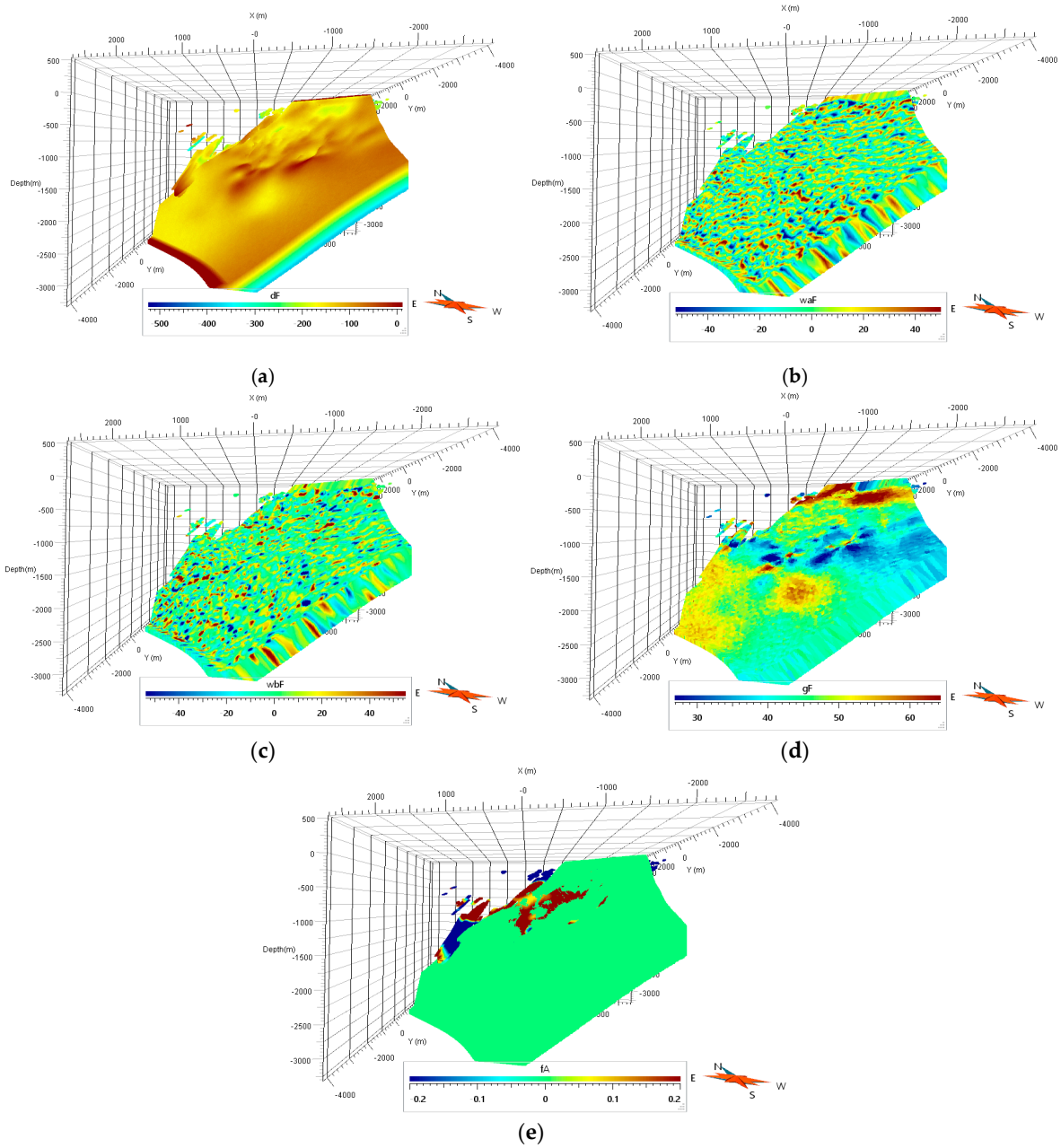
Data for this study were derived through analysis and extraction from the pre-constructed three-dimensional structural model and the mineralization model of the deposit. The study area is partitioned into known and unknown regions (Figure 2). The currently delineated ore bodies, designated as the known area, occur along secondary fault structures subsidiary to the main fault. Model training and comparison were performed using data exclusively from the known zone, which was partitioned into training and testing subsets in an 8:2 ratio. The area directly underlying and parallel to the known ore zones—interpreted as the downward projection of ore-hosting structures—is identified as a critical deep exploration target adjacent to existing mineralization (Figure 2b).

The known and unknown zones were discretized into 103,758 and 5,686,520 cubic voxels with dimensions of 10 m. Units in the known zone were populated with quantified gold grades and ore-controlling factors. In contrast, units in the unknown zone were assigned derived features of ore-controlling factors, with their gold grades to be predicted. Gold grade (Au) was selected as the dependent variable, and five ore-controlling factors (dF, waF, wbF, gF, and fV) were extracted as independent variables according to their spatial relationship with the alteration zone and fault. The spatial distributions of these variables are demonstrated in Figure 3. Their calculations proceed as follows [35].



(**a**) (**b**)

Figure 2. Spatial distribution of (**a**) gold grade in known area; (**b**) known and unknown areas.



(**a**) (**b**)

(**c**) (**d**)

(**e**)

Figure 3. Spatial distribution of (**a**) dF, (**b**) waF, (**c**) wbF, (**d**) gF, and (**e**) fA of the whole study area.

Let the fault surface $\mathcal{S}$ be discretized into a collection of m triangulated facets, denoted as $\{s_1, s_2, \ldots, s_m\}$. For each voxel v intersected by n drill hole samples, its gold grade $\mathrm{G}(v)$ is computed as a length-weighted average of the sample grades within the voxel.

$$\mathrm{G}(v) = \frac{\sum_{i=1}^{n} l_i \cdot g_i}{\sum_{i=1}^{n} l_i} \tag{1}$$

where l_i and g_i represent the length and grade of sample i within the voxel. For voxels without sample intersections, values are assigned through interpolation.

dF for voxel v is defined as the minimum Euclidean distance to the fault surface $\mathcal{S}$:

$$dF(v) = \min_{s \in \boldsymbol{\mathcal{S}}} \|v - s\| \tag{2}$$

gF for voxel v represents the slope value of its nearest neighboring unit s in the fault surface $\mathcal{S}$:

$$gF(v) = \mathrm{slope}(\arg\min_{s \in \boldsymbol{\mathcal{S}}} \|v - s\|) \tag{3}$$

For a given search radius r, the morphological trend of the fault surface is represented as:

$$\mathcal{T}(r) = MorphoFilter(\mathcal{S}, r) \tag{4}$$

waF and wbF for voxel v quantify the minimum distance to the trend at the radius, 120 m and 240 m, respectively.

$$\mathrm{waF}(v) = \min\|v, \mathcal{T}(120)\|, \ \mathrm{wbF}(v) = \min\|v, \mathcal{T}(240)\| \tag{5}$$

fA represents the alteration intensity at voxel v. It is computed using distinct rules based on spatial coincidence with alteration zones: if a voxel is located within an alteration zone, its intensity combines its alteration intensity with a weighted sum from surrounding units; otherwise, it is calculated solely via inverse distance squared weighting of neighboring alteration intensities. The formula is as follows:

$$\mathrm{fV}(v) = \begin{cases} S(v) + \dfrac{\sum_{A_j \in \mathcal{N}(v,r), A_j \neq \mathrm{v}} w(d_j) \cdot S(A_j)}{\sum_{A_j \in \mathcal{N}(v,r), A_j \neq \mathrm{v}} w(d_j)}, & if\ v \in \mathcal{A} \\ \dfrac{\sum_{A_j \in \mathcal{N}(v,r)} w(d_j) \cdot S(A_j)}{\sum_{A_j \in \mathcal{N}(v,r)} w(d_j)}, & if\ v \notin \mathcal{A} \end{cases} \tag{6}$$

where $S(\cdot)$ denotes the alteration intensity of the alteration unit, $\mathcal{A}$ is the collection of alteration units, $\mathcal{N}(v, r)$ is the neighborhood of voxel v with radius r, d_j represents the distance to the j^{th} unit in the neighborhood of voxel v, and $w(\cdot)$ is the weight.

The important contributions of the variables described above and their importance in mineralization prediction have been validated in recent studies [18,36].

3. Methods

3.1. Overview of Transformer Regression

This study employs a standard Transformer regression model [31,37], which utilizes the encoder stack to predict continuous numerical values (Figure 4). The input features are encoded and projected to a latent dimension, expanded with positional information, and regularized with dropout before Transformer processing. The core of the model is a 3-layer Transformer encoder with 8 attention heads. The multi-head self-attention mechanism captures complex dependencies between all encoded positions, effectively modeling their geometric relationships. The Transformer outputs are processed through a 2-layer multi-layer perception (MLP) with ReLU and dropout, and then projected via

9 parallel linear heads to predict each quantile. The predictions are concatenated to form the final output. The model was trained for 50 epochs using the AdamW optimizer with a WarmupCosineSchedule learning rate scheduler and a batch size of 32.

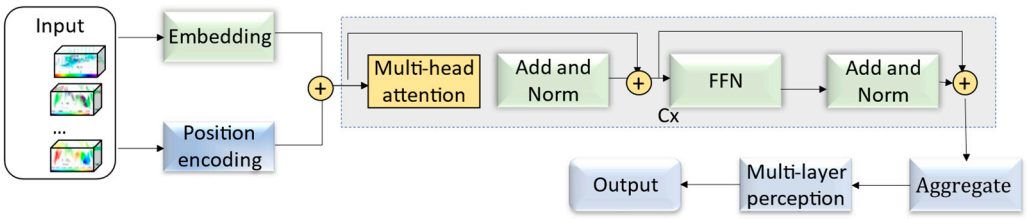


Figure 4. The standard Transformer regression architecture. The colors are used for visual distinction only and do not convey any scientific meaning.

The entire process can be summarized as:

$$\hat{\mathbf{y}} = \mathrm{MLP}\Big(\mathrm{Aggregate}\Big(\mathrm{Encoder}^{(C)}(\mathrm{Proj}(\mathbf{X}) + \mathbf{P})\Big)\Big) \tag{7}$$

where $\mathbf{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ represents the feature vectors of n sample points, $\hat{\mathbf{y}}$ is the corresponding predictor vector, $\mathbf{P}$ is the positional encoding vector, and $\mathrm{Encoder}^{L}$ denotes the composition of C encoder layers, which implements the multi-head self-attention and feedforward network operations with residual connections and layer normalization. The Encoder's computation is structured as follows.

A learnable linear projection layer maps the raw features to dense vector representations as follows:

$$Proj(\boldsymbol{x}_i) = \boldsymbol{W}_e \cdot normalize(\boldsymbol{x}_i) + \boldsymbol{b}_e \tag{8}$$

where $\mathbf{x}_i$ denotes the raw features of point i, $\mathbf{W}_e$ is a learnable projection matrix, and $\mathbf{b}_e$ is the bias vector.

The initial input to the Transformer encoder is then:

$$\boldsymbol{h}_i^{(0)} = \mathrm{Proj}(\boldsymbol{x}_i) + \boldsymbol{P}_i \tag{9}$$

This formulation allows self-attention to jointly capture content patterns and spatial relationships.

In the Transformer encoder, the multi-head attention mechanism processes the initial input $\boldsymbol{h}^{(0)}$ through C consecutive layers to produce the final encoded representation. For layer c with input $\boldsymbol{h}^{(c-1)}$, each attention head t computes:

$$Head_t^{(c)} = \mathrm{Attention}\Big(\boldsymbol{h}^{(c-1)} W_t^{Q,c}, \boldsymbol{h}^{(c-1)} W_t^{K,c}, \boldsymbol{h}^{(c-1)} W_t^{V,c}\Big) \tag{10}$$

where $\mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{Softmax}(\frac{\mathrm{QK}^T}{\sqrt{d_k}})V$, $W_t^{Q,c}, W_t^{K,c} \in \mathbb{R}^{d_{model} \times d_k}$ and $W_t^{V,c} \in \mathbb{R}^{d_{model} \times d_v}$ are learnable projection matrices for layer c.

Here, $\sqrt{d_k}$ scales the dot products to stabilize gradient flow. The outputs of all D heads are then concatenated and linearly projected as:

$$MultiHead^{(c)} = Concat\Big(head_1^{(c-1)}, head_2^{(c-1)}, \cdots, head_D^{(c-1)}\Big)\mathbf{W}^{O,c} \tag{11}$$

This output is then passed through a feed-forward network and residual connections to produce the layer output $\boldsymbol{h}^{(C)}$ as follows:

$$\boldsymbol{h}'^{(c)} = \mathrm{LayerNorm}\Big(\boldsymbol{h}^{(c-1)} + \mathrm{Dropout}\Big(MultiHead^{(c)}\Big)\Big) \tag{12}$$

$$FFN^{(c)}\left(\boldsymbol{h}'^{(c)}\right) = Relu\left(\boldsymbol{h}'^{(c)}W_1 + b_1\right)W_2 + b_2 \tag{13}$$

$$\boldsymbol{h}^{(c)} = \text{LayerNorm}\left(\boldsymbol{h}'^{(c)} + Dropout\left(FFN^{(c)}\left(\boldsymbol{h}'^{(c)}\right)\right)\right) \tag{14}$$

where W_1, W_2, b_1, and b_2 are learnable parameters.

The encoder stack repeats this process for C layers, transforming the initial $\boldsymbol{h}^{(0)}$ into the final contextualized representation $\boldsymbol{h}^{(C)}$, which captures rich, hierarchical dependencies across the input sequence.

3.2. Proposed Method

Our proposed architecture integrates both global and local attention mechanisms within the Transformer framework for mineralization prediction in 3D space. The model employs a spatial encoding scheme where 3D coordinates are explicitly incorporated as positional encodings, and processes input points through parallel pathways: a standard global self-attention branch captures long-range spatial dependencies across all points, while a novel local attention branch, based on the k-nearest neighbor search, focuses on fine-grained geometric structures within local neighborhoods. The outputs from both branches are adaptively fused through a gating mechanism, and the combined representations are processed by multi-layer regression heads to produce final predictions. This dual-branch design enables the model to effectively leverage both broad spatial context and detailed local patterns for accurate coordinate-based regression.

3.2.1. Relative Position Encoding

The relative position encoding scheme is employed in the local attention mechanism. It is defined by the Euclidean distance between points in 3D space. For each query–key pair within a local neighborhood, their pairwise distance is mapped to a learnable bias term that directly modulates the attention score. This distance-based bias functionally prioritizes spatially closer points, enforcing a strong locality inductive bias that ensures the attention mechanism focuses on geometrically proximate neighbors, which is fundamental for capturing fine-grained local patterns in 3D data.

In this study, the distance to bias is discretized by employing a bucket-encoding function. For two points with the coordinates $\boldsymbol{c}_i$ and $\boldsymbol{c}_j$, and a Euclidean distance $d_{ij} = \left\|\boldsymbol{c}_i - \boldsymbol{c}_j\right\|_2$, the mapping function is as follows:

$$\text{bucket}_{ij} = \left\lfloor \frac{d_{ij}}{d_{max}} \cdot N_{\text{buckets}} \right\rfloor \tag{15}$$

$$b_{ij} = \text{B}\left[\text{bucket}_{ij}\right] \cdot \mathbf{w}_b \tag{16}$$

where d_{max} is the maximum considered distance, N_{buckets} is the number of buckets, $\mathbf{B}$ is a learnable bucket embedding matrix, $\mathbf{w}_b$ is a learnable projection vector, and b_{ij} is the attention bias term.

3.2.2. Distance-Decay Dropout

We integrate distance-decay dropout with k-NN (k = 25 in this study) to preferentially retain spatially proximal samples during training while down-weighting or discarding distant ones. Distance-decay dropout is a regularization technique that stochastically drops elements with a probability proportional to their distance from a focal point, thereby enforcing spatial locality in the learning process. It incorporates the inductive bias that the influence of an element often decays with increasing distance. This dynamic filtering forces the regression head to integrate the global and local information more adaptively, enhancing

its ability to capture stable distance-aware representations and improving generalization in predicting continuous values.

The core of distance-decay dropout is a function where the keep probability decays as a function of distance. A common formulation uses exponential decay:

$$p_{ij}^{(\text{dropout})} = p_{base} \cdot \left(1 + \sigma\left(\alpha \cdot \overline{d}\right)\right) \tag{17}$$

$$\sigma(\cdot) = \frac{1}{1 + e^{-x}} \tag{18}$$

where p_{base} is the base dropout probability, and $\overline{d}$ is the mean distance between element i and its surroundings.

The mask matrix is:

$$M_{ij} \sim \text{Bernoulli}\left(1 - p_{ij}^{(\text{dropout})}\right) \tag{19}$$

The dropout probability $\boldsymbol{p}^{(\text{dropout})}$ increases monotonically with the mean distance $\overline{d}$. Elements that are farther away have a higher probability of being dropped. The use of the sigmoid function ensures a smooth, differentiable transition of dropout probabilities as a function of distance, which is controlled by the scale parameter α. The dropout operation applied to the value vectors is as follows:

$$\mathbf{V} = \mathbf{V} \odot \mathbf{M} \tag{20}$$

3.2.3. Multi-Head Attention with Global–Local Fusion

Our global–local fusion attention for 3D MPM consists of two parallel streams: a global self-attention module capturing extensive spatial contexts, and a local attention module with relative position encoding focusing on proximate geometric structures. The two streams are merged via concatenation and linear projection, forming a unified representation that encapsulates both macro-level spatial arrangements and micro-level geometric details (Figure 5).

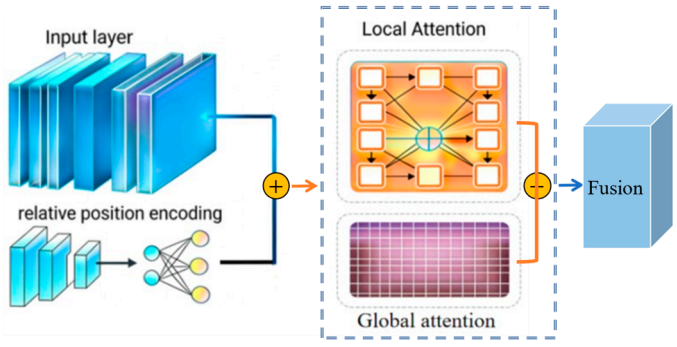


Figure 5. Multi-head attention mechanism with global–local fusion.

Our architecture processes 3D coordinate-based data through dual, complementary attention pathways. The global attention mechanism computes standard multi-head self-attention across all N points to capture long-range dependencies. For layer c with input $\boldsymbol{h}^{(c-1)}$, the global attention head t computes as:

$$\text{Global } head_i^{t,c} = \sum_{j=1}^{N} \text{Softmax}\left(\frac{\mathbf{q}_i \mathbf{k}_j^{\mathrm{T}}}{\sqrt{d_k}} + b_{ij}\right)\mathbf{v}_j \tag{21}$$

where $\boldsymbol{b}_{ij}$ is the relative bias, and $\mathbf{q}_i = \boldsymbol{h}_i^{(c-1)}\mathbf{W}_Q^t, \mathbf{k}_j = \boldsymbol{h}_j^{(c-1)}\mathbf{W}_K^t, \mathbf{v}_j = \boldsymbol{h}_j^{(c-1)}\mathbf{W}_V^t$ are the query, key, and value projections, respectively.

By projecting the hidden features $\boldsymbol{h}_j^{(c-1)}$ into query, key, and value spaces via learnable matrices $\mathbf{W}_Q^t$, $\mathbf{W}_K^t$, and $\mathbf{W}_V^t$, the model computes pairwise similarities across all positions i and j. The relative bias b_{ij} encodes positional information without relying on explicit position embeddings. After Softmax normalization, the resulting weights aggregate information from all input elements, enabling the model to focus on globally relevant contexts rather than only local neighborhoods.

Concurrently, the local attention mechanism restricts the receptive field to a local neighborhood $\mathcal{N}(i)$ around each point i, defined by its k-nearest neighbors in 3D space based on Euclidean distance. Crucially, it incorporates a distance-based relative position bias b_{ij} to steer the attention based on spatial proximity:

$$\text{Local } head_i^{t,c} = \sum_{j \in \mathcal{N}(i)} \text{Softmax}\left(\frac{\mathbf{q}_i \mathbf{k}_j^{\mathrm{T}}}{\sqrt{d_k}} + b_{ij}\right)\mathbf{v}_j \tag{22}$$

The outputs of these pathways are fused using a feature-wise gating mechanism. This gate dynamically calibrates the contribution from each pathway:

$$\alpha_i^{t,c} = \sigma\left(\mathbf{W}_g\left[\text{Global } head_i^{t,c}; \text{Local } head_i^{t,c}\right] + \boldsymbol{b}_g\right) \tag{23}$$

It takes the concatenated global and local features of each attention head, applies a linear transformation, and uses the sigmoid function to produce a weight $\alpha_i^{t,c}$ between 0 and 1. Specifically, $\alpha_i^{t,c} \approx 1$ indicates a preference for local fine-grained features, while $\alpha_i^{t,c} \approx 0$ prioritizes global contextual information. This mechanism dynamically balances the contribution of global and local features for each attention head across every layer.

The fused output for attention head t is:

$$\mathbf{F}_i^{t,c} = \alpha_i^{t,c} \odot \text{Global } head_i^{t,c} + \left(1 - \alpha_i^{t,c}\right) \odot \text{Local } head_i^{t,c} \tag{24}$$

where $[;]$ denotes concatenation, $\mathbf{W}_g$ and $\boldsymbol{b}_g$ are learnable parameters, σ is the sigmoid function, and $\odot$ is the Hadamard product. The resulting fused representation $\boldsymbol{F}_i^{t,c}$ synergistically combines holistic structural context with fine-grained geometric details, providing a comprehensive foundation for downstream regression tasks. The outputs of all D heads are then concatenated as:

$$MultiHead^{(c)} = Concat\left(\mathbf{F}_i^{1,c}, \mathbf{F}_i^{2,c}, \cdots, \mathbf{F}_i^{D,c}\right)\mathbf{W}^{O,c} \tag{25}$$

This output then serves as the input to Equation (12) in the standard Transformer framework.

3.2.4. Loss Function

This approach employs two key mechanisms to achieve a more robust estimation of the entire conditional distribution: the absolute value loss down-weights outliers by being less sensitive to large residuals than squared loss, while asymmetric weighting directly targets specific quantiles without requiring parametric error distributional assumptions, thus remaining valid under heterogeneity and heavy-tailed data. Instead of minimizing the sum of squared residuals, the loss function in this framework is as follows:

$$L_\tau(\boldsymbol{y}, \hat{\boldsymbol{y}}^\tau) = \sum_{i=1}^{N} (\tau - I(y_i < \hat{y}_i^\tau))(y_i - \hat{y}_i^\tau) \tag{26}$$

where N is the sample size, and y_i and $\hat{y}_i^{\tau}$ are the true value and estimated value at quantile τ for sample i. $I(\cdot)$ equals 1 if the condition holds, and 0 otherwise.

The quantile loss $L_{\tau}(\boldsymbol{y}, \hat{\boldsymbol{y}}^{\tau})$ imposes a linear penalty proportional to the error magnitude, preventing extreme values from disproportionately influencing learning. By enabling simultaneous estimation of multiple quantiles, it characterizes the full conditional distribution of the target variable beyond a single central tendency. A larger τ penalizes underpredictions more heavily, while a smaller τ emphasizes overpredictions—this asymmetry enables effective targeted quantile prediction. Furthermore, the resulting prediction intervals (PIs) explicitly communicate uncertainty ranges, providing inherently robust and interpretable outputs for risk-sensitive decision-making in mineral exploration targeting.

3.3. Model Implementation Details

Building on the methodological foundation presented above, the following details the hyperparameter specifications and computational cost involved in model implementation.

3.3.1. Hyperparameters

The models were implemented using Python 3.13 and PyTorch 2.6.0, and trained on an Intel Core i7-8550U CPU (1.80 GHz) with 16GB RAM. The detailed hyperparameter settings are provided in Table 1.

Table 1. Hyperparameter configuration.

Component	Parameter (Abbr.)	Value
Transformer	Normalization	LayerNorm
	Layers/Hidden dim(d_{model})/Heads	3/128/8
	Feed-forward dimension	256
MLP Head	Layer dimensions	$128 \rightarrow 256 \rightarrow 128$
	Activation/Dropout rate	ReLU/0.1
Optimization	Optimizer/LR schedule	AdamW/WarmupCosine
	Peak LR/Weight decay	1×10^{-5}/0.01
	Epochs/Batch size (B)	50/32
Distance-Decay	K-NN neighborhood size (k)	25
Data	Train/test/total count	80%/20%/103758

3.3.2. Computational Cost Analysis

Based on the hyperparameters described above, we conducted a comprehensive analysis of the computational complexity of the proposed model from both time and memory perspectives (Table 2). The detailed analysis is provided below.

Table 2. Computational cost.

n (Points)	Memory	Time/Epoch	Time/50 Epochs	Feasibility
1000	15 MB	14 s	12 min	feasible
10,000	152 MB	140 s	1.94 h	feasible
50,000	7.6 GB	11.7 min	9.7 h	feasible
100,000	15.2 GB	23.3 min	19.4 h	feasible
200,000	30.4 GB	46.3 min	38.9 h	infeasible

Time Complexity Analysis: The forward pass comprises three main components: (1) The feature encoder, implemented as a two-layer MLP with dimensions progressing from input dimension $d = 5$ to [128, 256, 128], contributes $O\left(Bd_{\text{model}}^2\right)$ operations, specifically

$B(5 \times 128 + 128 \times 256 + 256 \times 128) = 66,176B$ floating-point operations (FLOPs). (2) The core computational burden resides in the $L = 3$ layers of global–local fusion attention modules, where each layer comprises: (a) query–key-value projections requiring $3Bd^2_{\text{model}}$ FLOPs, (b) windowed local attention computation costing $BHd_{\text{head}}W$ with local window size $W = 10$, (c) dense relative position encoding adding BHd_{head} FLOPs, and (d) feed-forward networks contributing $8Bd^2_{\text{model}}$ FLOPs. (3) The output processor and $Q = 9$ quantile prediction heads collectively add $4Bd^2_{\text{model}} + 9d^2_{\text{model}}$ FLOPs.

Aggregating all components yields the total time complexity of $O(LBd^2_{\text{model}} + LBHd_{\text{head}} + B)$. For a representative configuration with $N = 100,000$ points, $L = 3$ layers, $B = 32$, $d_{model} = 64$, and $H = 8$ heads, each epoch takes approximately 23.3 min. Consequently, training for 50 epochs requires about 19.4 h (Table 2), demonstrating that the model can be trained on standard CPU hardware, albeit with a more substantial time investment.

Spatial Complexity Analysis: The parameter count reveals a highly skewed distribution across model components. The relative position encoding table constitutes the largest single module with $H \times (2W + 1)^3 \times d_{\text{head}} = 8 \times 21^3 \times 16 \approx 1.18 \times 10^6$ parameters, accounting for approximately 85% of the total 4.39 million trainable parameters. The remaining parameters are distributed among the three attention layer feed-forward networks (0.39 M), the feature encoder (0.07M), the quantile heads (0.09 M), and various projection layers (0.08 M). During training, the memory footprint scales as $O(BLd_{\text{model}} + BH)$, where the attention score matrices of size $B \times H$ serve as the primary memory bottleneck, including all components—model parameters, activations, gradients, and optimizer states. The total memory requirement for the typical configuration with batch size $B = 32$ is approximately 152 MB. For the training size of 100,000 samples, the memory requirement reaches approximately 15.2 GB, which approaches the practical upper limit of our hardware (Intel Core i7-8550U with 16GB RAM) (Table 2).

4. Results

Using the Xiadian gold deposit as a case study, we first performed ablation studies with bootstrap resampling [38], which confirmed the contribution of each component and demonstrated the overall superiority of the proposed method. All experiments were conducted on a system equipped with an Intel Core i7-8550U CPU (1.80 GHz, 16GB RAM) using Python 3.13 and PyTorch 2.6.0. Subsequent stability tests and spatial distribution pattern analysis further corroborated the proposed model's stability and effectiveness. The model was then employed to delineate exploration targets in adjacent unexplored regions.

4.1. Ablation Study

To investigate the effectiveness of each component, we conducted ablation experiments by adding components one at a time. The model abbreviations after incremental and decremental modifications are presented in Table 3.

Table 3. Ablation experiments.

Model Abbreviation	Component		
	Global–Local Fusion	**Distance-Decay Dropout**	**Loss Function**
T-M	×	×	MSE
T	×	×	Pinball
T-GL	√	×	Pinball
T-D	×	√	Pinball
T-GL-D	√	√	Pinball

Note: A checkmark (√) indicates that the corresponding component is included in the model configuration, whereas a cross (×) indicates that it is not.

To conduct a comprehensive assessment of the models, we employed two complementary categories of metrics. The first category comprises the commonly used coefficient of determination (R^2) and four quantile-specific metrics: the pinball score (P_τ), pseudo R^2_τ, the quantile reliability score (R_τ), and hit rate. The second category focuses on the sharpness and calibration of the PIs using the mean PI width and the Winkler score. To complement these point estimates with an assessment of their reliability, we conducted bootstrap analysis on the test set to quantify performance uncertainty arising from sampling variability. Given the computational constraints that made multi-run training infeasible, this resampling approach offered an efficient alternative for estimating the sampling distribution of each metric. For each model, we generated 1000 bootstrap replicates, recalculating all metrics across all quantiles on every replicate. The standard deviations derived from this procedure—displayed as error bars in the chart—reflect the stability of each model's performance, allowing us to identify not only which models achieve the highest scores but also which do so consistently.

4.1.1. Overall Performance Comparison

To assess performance differences among the ablated models, we employed a hierarchical statistical approach. First, the Friedman test [39] was used to evaluate whether significant global differences existed. Upon detecting significance, the Nemenyi post hoc test [40] was applied for pairwise comparisons. Under the null hypothesis of equivalent median MAE across models, we further validated specific model-level distinctions using paired Wilcoxon signed-rank tests [39], supplemented by bootstrap-derived confidence intervals [38].

Table 4 presents the results of the Friedman test. The models exhibited significant differences ($\chi^2 = 6052.38$, $p < 0.001$) in median absolute error (Median AE), indicating that component selection had a substantial impact on predictive performance. The Friedman test was further employed to evaluate performance differences across quantiles. Significant differences were found across all nine quantiles (Q10–Q90, all $p < 0.001$). Notably, the test statistics increased progressively from the central quantiles (e.g., Q30: $\chi^2 = 4707.14$) to the extremes (e.g., Q90: $\chi^2 = 21212.56$), suggesting that model performance exhibits greater variability at the tails of the conditional distribution.

Table 4. Friedman test result.

Type	Type	Statistic	*p*-Value	Significant
Model comparison	Median MAE	6052.38	0	True
Quantile comparison	$\tau = 0.1$	9296.71	0	True
	$\tau = 0.2$	5364.71	0	True
	$\tau = 0.3$	4707.14	0	True
	$\tau = 0.4$	4811.03	0	True
	$\tau = 0.5$	6052.38	0	True
	$\tau = 0.6$	7616.15	0	True
	$\tau = 0.7$	11,129.21	0	True
	$\tau = 0.8$	15,631.17	0	True
	$\tau = 0.9$	21,212.56	0	True

Table 5 presents the results of paired Wilcoxon signed-rank tests with 95% bootstrapped confidence intervals (CIs) for all pairwise comparisons. All comparisons were found to be statistically significant, reflecting consistent and discernible differences in predictive accuracy across the evaluated models.

Table 5. Nemenyi post hoc test.

Model1	Model2	Wilcoxon *p*-Value	Mean Difference	CI_95 Lower	CI_95 Upper	Significant
T	T-GL	0.000	0.100	0.095	0.104	True
T	T-D	0.000	0.083	0.079	0.088	True
T	T-GL-D	0.000	0.115	0.109	0.120	True
T-GL	T-D	0.000	−0.017	−0.019	−0.014	True
T-GL	T-GL-D	0.000	0.015	0.012	0.018	True
T-D	T-GL-D	0.000	0.032	0.029	0.035	True

4.1.2. Quantile-Specific Accuracy Assessment

In this section, the evaluation of quantile-specific performance encompasses four metrics: the pinball score (P_τ), pseudo R^2_τ, the quantile reliability score (R_τ), and hit rate for mineralization enrichment, complemented by the coefficient of determination (R^2). The specific configurations and corresponding results are detailed below.

- R^2 and pseudo R^2

We use the commonly adopted R^2 and the pseudo R^2 to assess the goodness-of-fit of the models. The pseudo R^2 metric serves as a vital tool for evaluating the relative performance of quantile-based models, which, by their nature, do not rely on the Gaussian likelihood assumptions of ordinary least squares [18]. It is computed as one minus the ratio of the sum of weighted absolute deviations from the fitted quantile model to that of a naive model, typically one containing only an intercept. It is critical to note that, unlike R^2, this pseudo R^2 is a comparative gauge of quantile-specific adequacy, not a proportion of variance explained. A value near 1 signifies a strong model, but its absolute value is inherently lower and varies across quantiles. An increasing pseudo R^2 across quantiles indicates that the model's explanatory power strengthens at higher quantiles of the conditional distribution. The formula is as follows:

$$\text{Pseudo } R^1_\tau = 1 - \frac{\sum_{i=1}^{N} \rho_\tau\left(y_i - \hat{y}_i^\tau\right)}{\sum_{i=1}^{N} \rho_\tau\left(y_i - y_i^\tau\right)} \tag{27}$$

Figure 6 presents the R^2 and pseudo R^2 values for each model across different quantiles, with error bars indicating $\pm$ 1 standard deviation derived from bootstrap resampling. The two bar charts display the mean R^2 and pseudo R^2 values of four models with small error bars ($\pm$0.005–0.02 in R^2 and $\pm$0.003–0.004 in pseudo R^2) across quantiles, indicating high measurement precision. The non-overlapping error bars confirm statistically significant differences between all pairs of models. All models achieve their highest R^2 values around the 60th to 70th percentile, with declining fit towards the extremes (Figure 6a). This heterogeneity in model performance across quantiles reflects the varying influence of predictors on different segments of the conditional distribution. Examination of R^2 and pseudo R^2 values across quantiles reveals that models T-D, T-GL, and T-GL-D consistently outperformed model T, with peak performance achieved by model T-GL-D. This confirms that adding either the global–local fusion mechanism or distance-decay dropout enhances the model performance. The pseudo R^2 bar charts for four models in Figure 6b show an increasing trend with the quantiles, revealing that the explanatory power is more effective for predicting high-graded mineralization than low-graded or background ones. This reflects fundamental heteroscedasticity in the mineralization process, or that the predictors are more effective at explaining outcome variability among high-graded mineralization.

- Pinball score P_τ

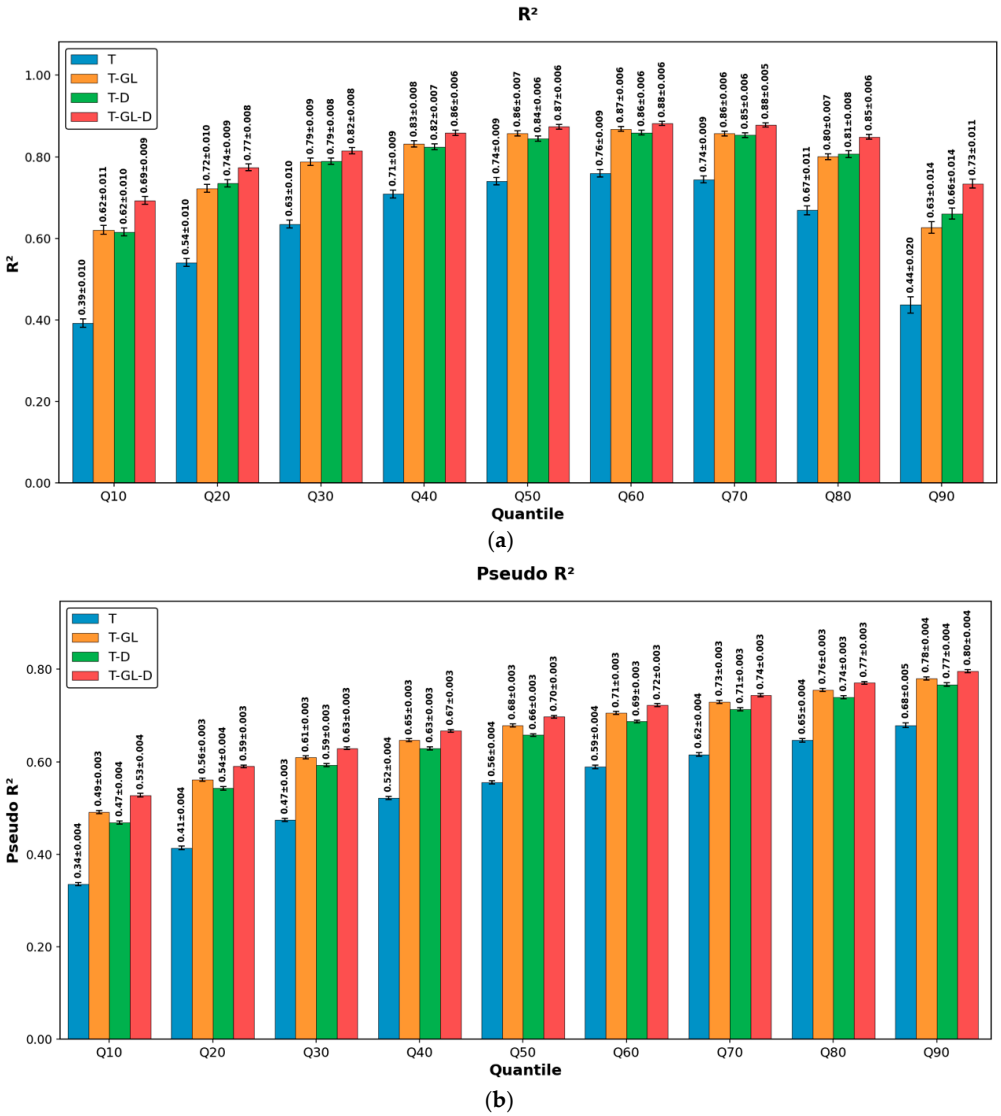


Figure 6. Comparison of (**a**) R^2 and (**b**) pseudo R^2 values with error bars. The proposed T-GL-D model yields higher R^2 and pseudo R^2 values than the other three models across all quantiles.

The pinball score is the primary evaluation metric for quantile-based models [17,41]. It asymmetrically penalizes prediction errors based on the target quantile τ—for values above the predicted quantile, the penalty is weighted by τ, while values below are weighted by $(1 - \tau)$. According to Equation (17), the pinball score is expressed as follows:

$$P_\tau = \overline{L}_\tau(\boldsymbol{y}, \hat{\boldsymbol{y}}^\tau) = \frac{1}{N} L(\tau) \tag{28}$$

Model comparison via pinball loss must be performed separately at each quantile level—where a lower loss indicates better accuracy—as the loss values across different τ are not directly comparable due to the τ-specific asymmetric error weighting inherent to the function.

Figure 7a shows the pinball score bar chart with error bars across quantiles for the four quantile-based models, respectively. All the error bars are small (±0.001–0.002), indicating high precision and low data dispersion. Pinball scores decrease with the inclusion of the global–local fusion mechanism or distance-decay dropout, and the highest performance across quantiles is achieved by combining all three components. This verifies that including

either the global–local fusion mechanism or distance-decay dropout boosts performance, further reinforcing the pattern observed in the R^2 and pseudo R^2 curves.

- Reliability score R_τ

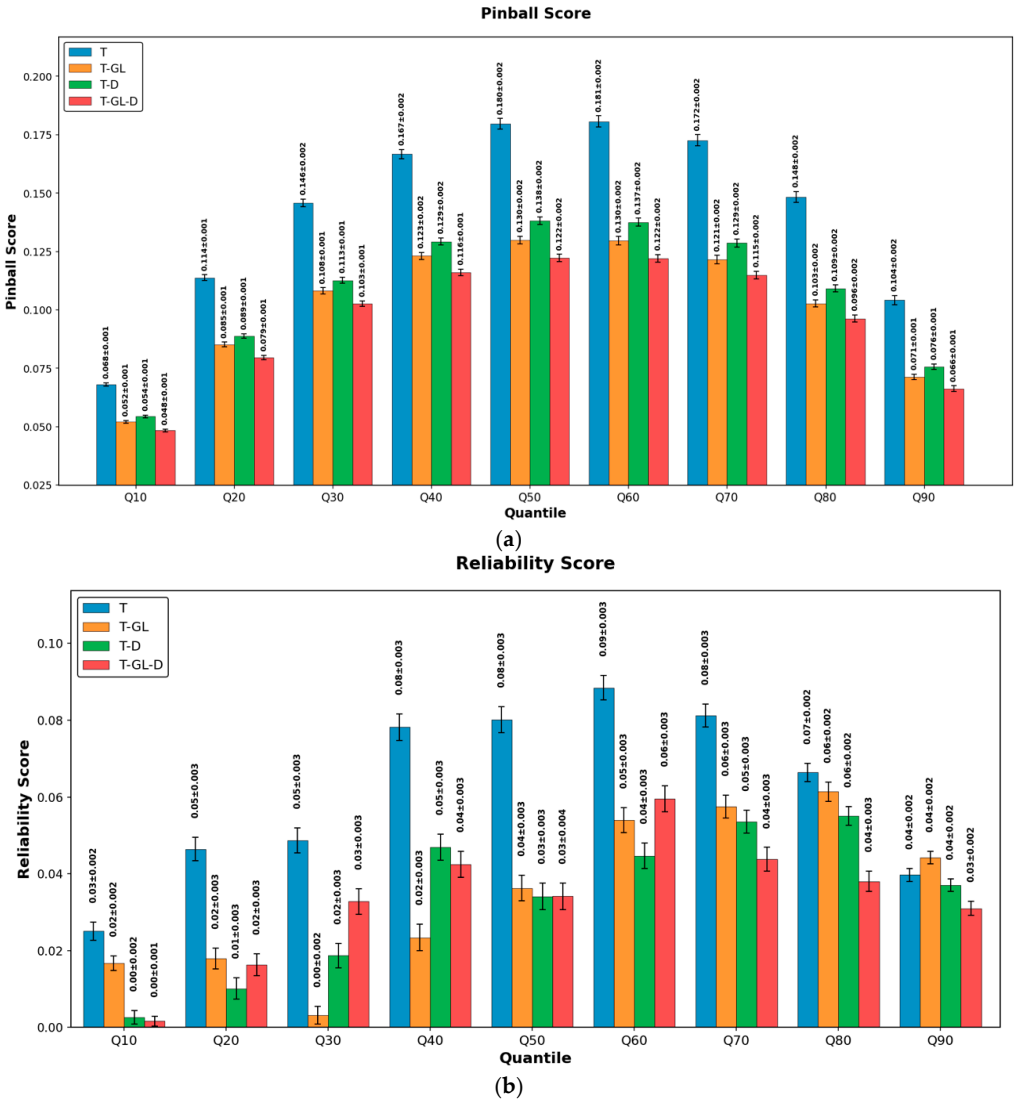


Figure 7. Comparison of (**a**) pinball score and (**b**) reliability score values with error bars.

Coverage probability and quantile reliability diagrams serve as critical diagnostic tools for assessing the calibration of quantile-based models. Coverage probability quantifies the empirical proportion of observations falling below a given predicted quantile, which should ideally match the nominal quantile level [17,41]. For a specified quantile level τ, it is defined as:

$$\text{Coverage}(\tau) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(y_i \leq \hat{y}_i^{\tau}) \tag{29}$$

where $\mathbf{1}(\cdot)$ is the indicator function.

The quantile reliability score quantifies the discrepancy between the target coverage rate τ and the actual proportion of observed values falling below the quantile τ, that is:

$$R_\tau = |\text{Coverage}(\tau) - \tau| \tag{30}$$

Better statistical consistency corresponds to a lower reliability score, where perfect calibration is defined as $R_\tau = 0$ across all quantiles τ.

Figure 7b illustrates the reliability score bar chart with small error bars (±0.001–0.005) for four models. All the error bars are small, indicating high precision and low data dispersion. The T-GL-D model exhibits lower reliability scores compared to the T model across quantiles, further confirming that the inclusion of the global–local fusion mechanism and distance-decay dropout enhances model effectiveness.

- Hit rate, Hr_τ

The hit rate is utilized to evaluate the success of mineralization enrichment predictions, a metric that is critical for subsequent target delineation. Given our focus on mineralization enrichment, defined as having a grade exceeding 1g/t, the following analysis targets these high-grade zones. The hit rate of the mineralization enrichment is defined as the conditional probability that the true grade is enriched, given that the model's prediction also meets the enrichment criterion [17]. The hit rate at the quantile τ is calculated as follows:

$$\mathrm{Hr}_\tau = \frac{\#\{i : \hat{y}(\tau|\boldsymbol{x}_i) \geq 1 \; and \; y_i \geq 1\}}{\#\{j : \hat{y}(\tau|\boldsymbol{x}_j) \geq 1\}} \tag{31}$$

Figure 8 displays the hit rate bar chart with error bars of models across quantiles. Error bars are small (±0.001–0.005), indicating high precision and low variability. The hit rate of the four models peaks at 0.1 quantile and subsequently declines as the quantile level increases. This reflects an inverse relationship between the quantile and the corresponding hit rate—while the model predicts higher-grade mineralization at elevated quantiles, its hit rate diminishes. The peak hit rate at the lowest quantile underscores the models' strength in identifying high-probability, high-grade targets, which is critical for optimizing exploration accuracy. The hit rate of the mean-based model shows the nearest value to that of the quantile-based models at the 60th to 70th percentiles, indicating a right-skewed distribution consistent with the R^2 pattern observed across quantiles. Meanwhile, the hit rates of the T-GL-D model exceed the values of other quantile-based models across quantiles, underscoring its improved predictive capability for mineralization enrichment.

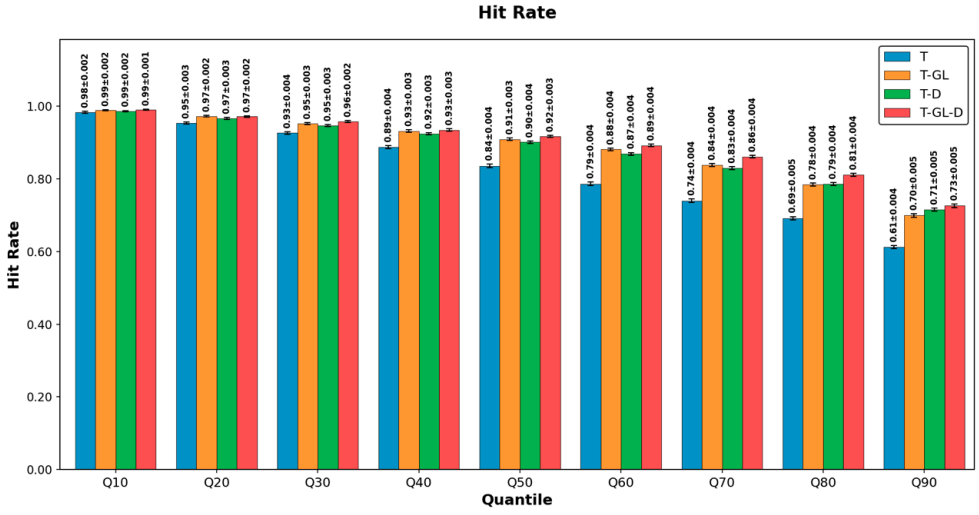


Figure 8. Comparison of the hit rate of the mineralization enrichment with error bars.

4.1.3. PI Calibration Evaluation

PIs are essential in uncertainty quantification as they provide a probabilistic range—rather than a single-point estimate—within which future observations are expected to fall [17,41]. In contrast to confidence intervals, which pertain to the uncertainty of model parameters, PIs quantify the expected range of future observable values. In this study, the

models were evaluated through sharpness (via mean PI width $\overline{W}_\alpha$) and overall interval quality (via mean PI score $\overline{S}_\alpha$) of the PIs, respectively.

For a given observation i, PI is defined as:

$$PI = \left[\hat{y}_i^{\frac{\alpha}{2}}, \hat{y}_i^{1-\frac{\alpha}{2}}\right] \tag{32}$$

where $(1-\alpha)$ is the PI level and $\hat{y}_i^{\frac{\alpha}{2}}$ and $\hat{y}_i^{1-\frac{\alpha}{2}}$ are the predicted values for observation i at quantiles $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$, respectively.

The mean PI width is denoted as:

$$\overline{W}_\alpha = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{u}_i - \hat{l}_i\right) \tag{33}$$

where $\hat{l}_i$ and $\hat{u}_i$ are the lower and upper bounds of the PI for the observation, and i directly measures the model's sharpness—narrower intervals indicate more precise estimates.

Meanwhile, the Winkler score [42], which evaluates the sharpness and calibration of PIs simultaneously, offers a comprehensive assessment of overall quality and effectiveness, as defined by:

$$\overline{S}_\alpha = \frac{1}{N}\sum_{i=1}^{N}\left(\left(\hat{u}_i - \hat{l}_i\right) + \frac{2}{\alpha}\left(\hat{l}_i - y_i\right)I\left\{y_i < \hat{l}_i\right\} + \frac{2}{\alpha}(y_i - \hat{u}_i)I\{y_i > \hat{u}_i\}\right) \tag{34}$$

In analysis, a low score combined with a narrow width represents optimal performance, indicating well-calibrated estimates. Conversely, a good score achieved with excessively wide intervals suggests conservative but less useful predictions, while a narrow width with poor scores reflects reliable probabilistic forecasts.

Figure 9 compares the mean PI width and Winkler score of the four models across PI levels of 20%, 40%, 60%, and 80%. The two bar charts display distinct mean values with non-overlapping error bars, confirming statistically significant differences between all model pairs. The narrow error bars (ranging from ± 0.001 to ± 0.006 in mean PI width and ± 0.008 to ± 0.025 in Winkler score) further indicate high measurement precision and low data dispersion. The proposed T-GL-D model outperforms the other three models in both metrics through all the PI intervals, demonstrating the lowest uncertainty. This reflects the proposed model's enhanced ability to capture underlying patterns and its stronger generalization capability.

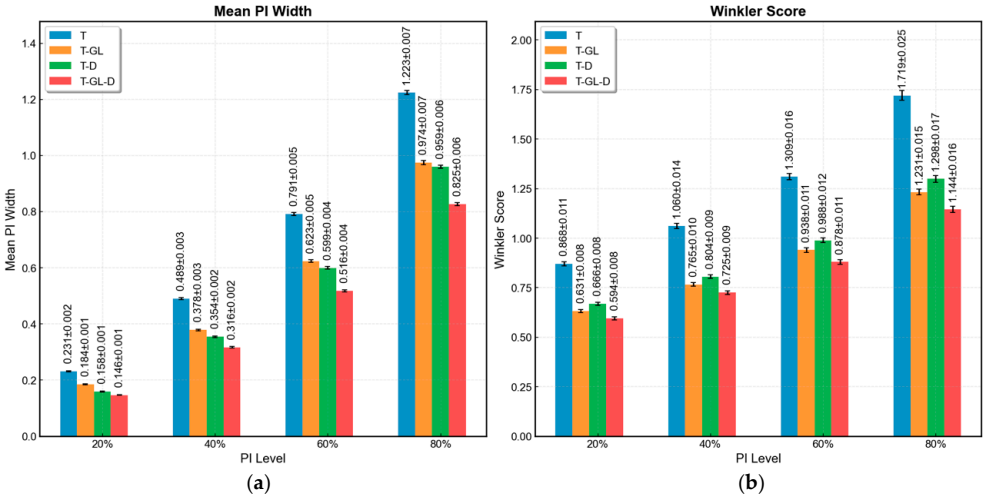


Figure 9. PI calibration by (**a**) mean PI width and (**b**) the Winkler score of PIs. The proposed T-GL-D model achieves the narrowest mean PI width ($\overline{W}_\alpha$) and the lowest Winkler score ($\overline{S}_\alpha$) across PI levels.

4.2. Stability Test

The bootstrap-based ablation experiments described above demonstrate the comprehensive superiority of the proposed T-GL-D model across all evaluation metrics. While the bootstrap approach captures uncertainty arising from test set variability, it does not account for variability due to different random initializations or training dynamics. Nevertheless, the performance gap between the T-GL-D model and the other baselines is substantial and consistent across all bootstrap iterations, rendering the superiority of the T-GL-D model statistically evident even under this conservative estimation.

To assess the stability of the proposed T-GL-D model across different data splits, a comprehensive stability testing procedure was conducted. Specifically, the original dataset was randomly partitioned into training and testing subsets ten times using distinct random seeds, ensuring variability in the composition of each split. For each of the ten partitions, the model was retrained and evaluated, and the predictive performance metrics (e.g., pinball score and PI coverage) were recorded. To statistically verify the equivalence of model performance across these splits, the two one-sided test (TOST) [43] procedure was performed. The TOST framework was used to test whether the mean difference in performance between any two splits fell within a pre-specified equivalence margin, thereby confirming that the model's predictive capability remained consistent regardless of the random partitioning.

Model stability was assessed using the coefficient of variation (CV), calculated as the ratio of the standard deviation to the mean of each performance metric across 10 random data splits. Lower CV values indicate greater stability, with CV < 10% considered stable and CV < 5% indicating excellent stability.

The model was evaluated across ten distinct data splits, defined by the random seeds 42, 257,630, 314,837, 381,557, 462,922, 724,630, 861,405, 875,564, 923,476, and 605,210. The statistical analyses of the predictive results from these datasets are summarized below.

Tables 6–10 summarize the model performance metrics obtained from 10 different random data partitions. All evaluation metrics exhibited coefficients of variation (CV) below 5%, and the two one-sided test (TOST) for equivalence were significant ($p < 0.05$), confirming that different random splits did not significantly influence model performance.

Table 6. R^2 statistics.

Quantile	Mean	Std	Min	Max	CV
Q10	0.664	0.0167	0.6403	0.6926	2.52%
Q20	0.7556	0.0141	0.7333	0.774	1.87%
Q30	0.8076	0.0086	0.7934	0.8223	1.06%
Q40	0.8432	0.0087	0.8309	0.8588	1.03%
Q50	0.8621	0.0097	0.8449	0.8741	1.13%
Q60	0.8735	0.0076	0.8619	0.8835	0.87%
Q70	0.8704	0.0076	0.856	0.8812	0.87%
Q80	0.8387	0.0104	0.8196	0.8553	1.24%
Q90	0.7218	0.0175	0.68	0.7377	2.42%

Table 7. Pseudo R^2 statistics.

Quantile	Mean	Std	Min	Max	CV
Q10	0.5163	0.0058	0.5082	0.5277	1.12%
Q20	0.5826	0.0064	0.5731	0.5906	1.10%
Q30	0.6281	0.0041	0.6211	0.6363	0.65%
Q40	0.6632	0.0049	0.6563	0.6726	0.74%

Table 7. *Cont.*

Quantile	Mean	Std	Min	Max	CV
Q50	0.6916	0.0061	0.6826	0.6987	0.88%
Q60	0.7176	0.0059	0.7089	0.7289	0.82%
Q70	0.7432	0.0049	0.7366	0.7551	0.66%
Q80	0.7684	0.0034	0.7638	0.7759	0.44%
Q90	0.7962	0.0044	0.7907	0.8073	0.55%

Table 8. Pinball score statistics.

Quantile	Mean	Std	Min	Max	CV
Q10	0.0493	0.0008	0.0482	0.0507	1.65%
Q20	0.0807	0.0016	0.0788	0.0832	1.97%
Q30	0.1025	0.0014	0.1003	0.1043	1.32%
Q40	0.1166	0.0018	0.1142	0.1198	1.53%
Q50	0.1240	0.0030	0.1207	0.1290	2.43%
Q60	0.1235	0.0028	0.1209	0.1290	2.23%
Q70	0.1147	0.0019	0.1115	0.1178	1.68%
Q80	0.0969	0.0015	0.0951	0.0994	1.55%
Q90	0.0660	0.0012	0.0639	0.0677	1.89%

Table 9. Hit rate statistics.

Quantile	Mean	Std	Min	Max	CV
Q10	0.9879	0.0018	0.9851	0.9907	0.18%
Q20	0.9741	0.0027	0.9707	0.9788	0.28%
Q30	0.9572	0.0040	0.9512	0.9635	0.42%
Q40	0.9377	0.0068	0.9300	0.9489	0.73%
Q50	0.9184	0.0072	0.9104	0.9307	0.78%
Q60	0.8948	0.0086	0.8829	0.9087	0.96%
Q70	0.8578	0.0099	0.8430	0.8739	1.15%
Q80	0.8083	0.0100	0.7891	0.8207	1.24%
Q90	0.7361	0.0101	0.7217	0.7577	1.37%

Table 10. TOST equivalence testing results for prediction consistency.

Quantile	Equivalent Pairs	Ineq. Pairs	Mean d	Max d	Mean Range	Overall Mean	Rel. Diff.	Mean CV	ANOVA *p*	TOST
Q10	45/45	0/45	0.0118 *	0.0314 *	0.0282	0.6698	4.21%	1.3376	0.0334	✓
Q20	45/45	0/45	0.0077 **	0.0224 *	0.0221	0.7878	2.81%	1.2637	0.542	✓
Q30	45/45	0/45	0.0085 **	0.0234 *	0.0252	0.8767	2.88%	1.2197	0.3468	✓
Q40	45/45	0/45	0.0104 *	0.0253 *	0.0292	0.9581	3.05%	1.1885	0.1034	✓
Q50	45/45	0/45	0.0092 **	0.0244 *	0.0297	1.0257	2.90%	1.1628	0.2112	✓
Q60	45/45	0/45	0.01 *	0.0271 *	0.0339	1.1063	3.07%	1.1363	0.1591	✓
Q70	45/45	0/45	0.0161 *	0.0446 *	0.0596	1.1979	4.98%	1.1163	0.0001	✓
Q80	45/45	0/45	0.0122 *	0.041 *	0.0593	1.3180	4.50%	1.0983	0.0096	✓
Q90	45/45	0/45	0.0074 **	0.016 *	0.0264	1.5130	1.74%	1.0746	0.6274	✓

Note: d = Cohen's d effect size; Rel. Diff. = relative difference; CV = coefficient of variation; ANOVA *p*-values are reference only due to large sample sizes; all Cohen's d < 0.2 indicate negligible differences; and ✓ indicates passing the TOST equivalence test ($p < 0.05$). All file pairs across all quantiles passed the TOST equivalence test ($p < 0.05$), confirming that predictions from different data splits are statistically equivalent. Cohen's d values < 0.2 indicate negligible differences. CV = coefficient of variation. * $p < 0.05$, ** $p < 0.01$ (ANOVA results are provided for reference only, as they are sensitive to large sample sizes).

These results provided rigorous evidence that the model's performance was not contingent upon a particular data split, thereby demonstrating its stability across various data partitions.

4.3. Spatial Distribution

Building on the demonstrated effectiveness of the proposed model, a comparative analysis of the spatial distribution of predictions across different quantiles is conducted here to further validate its performance.

Figure 10 shows the spatial distributions of the true values and predicted mineralization enrichment at quantiles 0.1, 0.3, 0.5, 0.7, and 0.9. The 0.5 quantile prediction aligns most closely with the actual distribution, with deviations gradually growing as the quantile level moves away from this median value. This pattern reflects the model's ability to capture central tendency at the median quantile, where it best approximates the actual spatial footprint of mineralization. The spatial extent of the predicted enrichment zones expands progressively with increasing quantile levels, which is consistent with the inherent properties of quantile-based models. The predicted distributions at the 0.1 and 0.3 quantiles are largely encompassed within that of the 0.5 quantile. This suggests that predictions below the 0.5 quantile yield more concentrated and conservative spatial extents, suggesting higher stability in identifying high-probability mineralization zones. In contrast, the distributions at the 0.7 and 0.9 quantiles expand outward from the 0.5 quantile prediction, reflecting greater predictive uncertainty and a higher tendency toward overprediction as the quantile level increases.

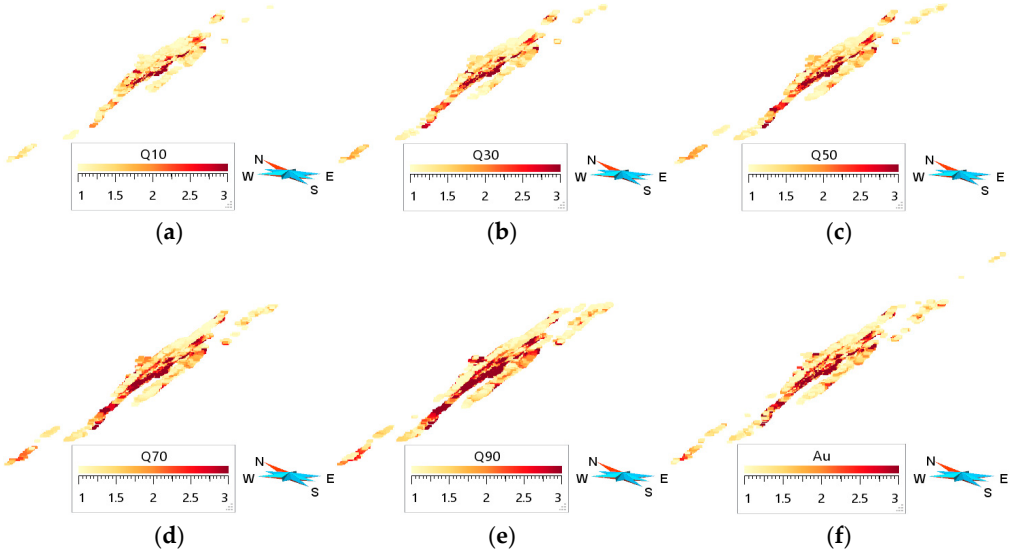


Figure 10. Spatial distribution of mineralization enrichment (grade $\geq$ 1g/t) at quantiles (**a**) 0.1, (**b**) 0.3, (**c**) 0.5, (**d**) 0.7, (**e**) 0.9, and (**f**) the true value.

The predicted spatial patterns across different quantiles are spatially coherent and align well with known mineralization enrichment zones, indicating that the model captures geologically meaningful signals. Moreover, while the predicted extent of high-grade mineralization expands with increasing quantiles, the high-probability core regions remain stable across quantiles, with only marginal variations at the boundaries. This consistency demonstrates that the model can provide reliable and decision-ready data for exploration targeting.

4.4. Target Delineation

The above analysis confirmed that the proposed model was both accurate and reliable. Beyond overall performance, the examination of hit rates and spatial patterns across quantiles specifically showed that its lower-quantile predictions offered a conservative basis for planning future exploration.

In this section, leveraging the established model, the predictions for the deep-seated unknown area encompassed the conservative 0.1 and 0.3 quantiles alongside the median (Figure 11). The corresponding targets are presented in Figure 12. These figures illustrate that the target locations predicted under different quantiles remain largely consistent. The target area is smallest at the 0.1 quantile and progressively enlarges with increasing quantile values, demonstrating the quantile-consistent property of the prediction model. Furthermore, the predicted target regions align well with the positions of identified enrichment zones in the known area, in accordance with established geological expectations.

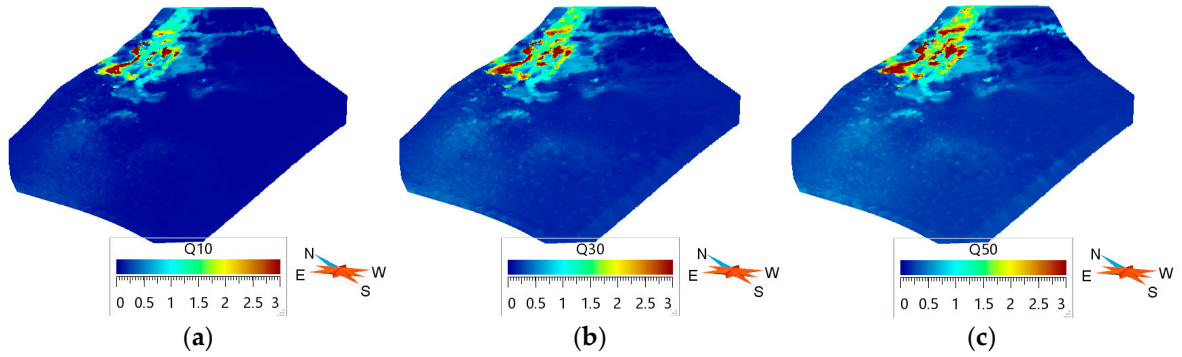


Figure 11. Predictions at quantiles (**a**) 0.1, (**b**) 0.3, and (**c**) 0.5.

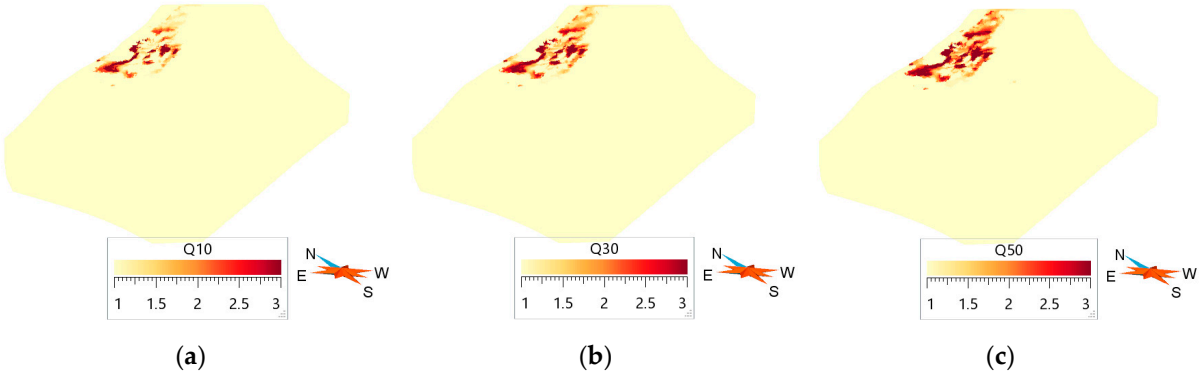


Figure 12. Exploration targets at quantiles (**a**) 0.1, (**b**) 0.3, and (**c**) 0.5.

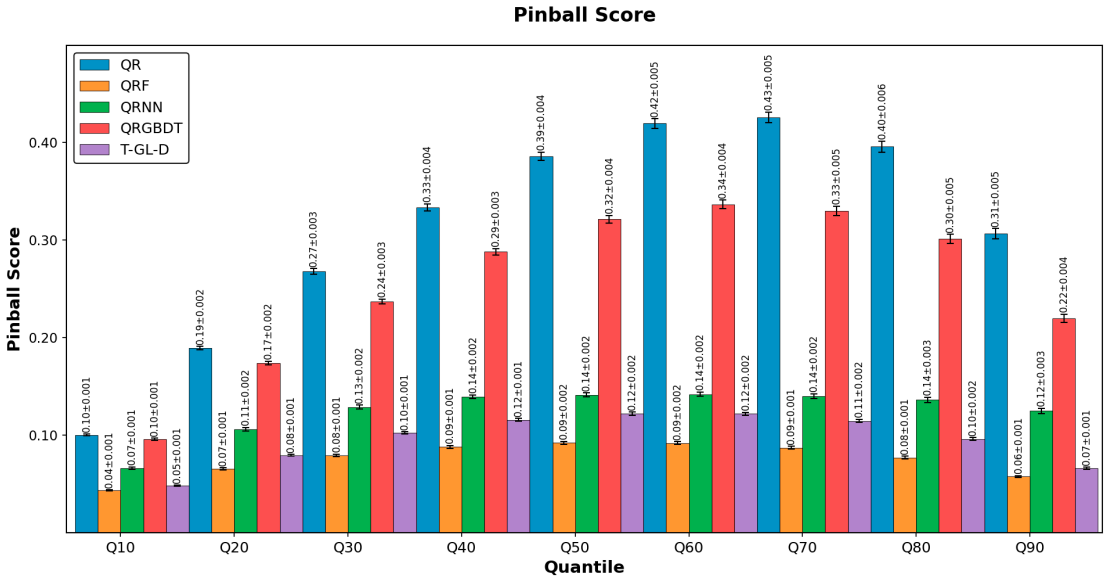
5. Discussion

To contextualize the predictive performance of the proposed T-GL-D model, we compare it with established machine learning and deep learning approaches for probabilistic forecasting, including standard Quantile Regression (QR), Quantile Regression Forest (QRF), Quantile Regression Neural Networks (QRNNs), and gradient boosting Decision Tree probabilistic methods (QRGBDT). The comparative evaluation examines the pinball score, mean R_τ, and pseudo R^2 across quantiles, as well as the mean PI width and Winkler score across PI levels, all under the same data split. Similarly, bootstrap analysis was conducted for model comparison.

Robust evaluation of quantile-based regression models demands multiple indicators, as single metrics capture only narrow performance dimensions, and a model optimal under one metric may be unreliable overall. Figures 13 and 14 depict the bar chart with error

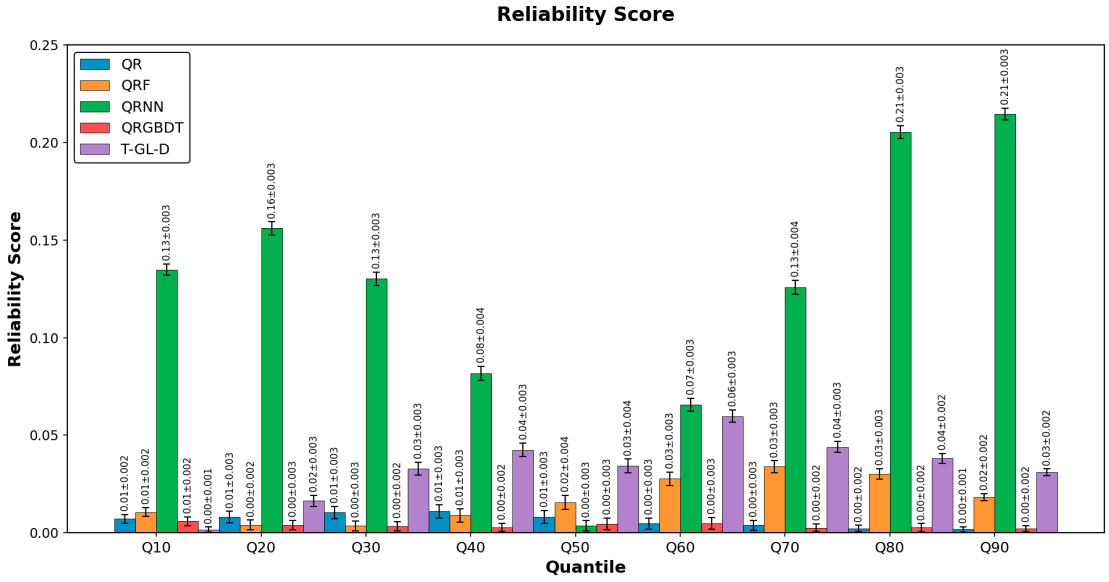
bars of the pinball score (P_τ), reliability score ($R\tau$), and pseudo R^2 across quantiles. The error bars are all small, indicating high measurement precision and low variability in these metrics. The superior accuracy of the QRF and T-GL-D models is evident across all quantiles, with both models achieving lower pinball scores, moderate reliability scores (R_τ), and higher pseudo R^2 values relative to the remaining three models. Figure 15 compares the Winkler score and mean PI width at the 20%, 40%, 60%, and 80% PI levels for the five models. It confirms that the QRF and T-GL-D models possess lower uncertainty, characterized by narrower mean PI widths and lower Winkler scores. The QRF and T-GL-D models exhibit marginal trade-offs in accuracy, sharpness, and uncertainty, reflecting their superior comprehensive performance. The proposed model in this study incorporates a global–local attention fusion mechanism and dropout strategy, which is specifically designed to enable predictions in unknown areas while balancing local fitting capability and generalization capacity. However, due to the random splitting of training and test sets employed in this study—a single-site validation approach—the inherent spatial challenge is effectively eliminated. When training and test points are randomly intermingled, each test sample is surrounded by geographically proximate training samples in both feature space and geographic space. Under this configuration, spatial prediction essentially reduces to spatial interpolation, a task for which QRF is particularly well-suited due to its localized learning mechanism. Consequently, the comparative evaluation conducted in this study exclusively under random data splitting may not fully capture the intended advantages of the proposed deep learning model. The key limitation is that validation remains confined to the training set domain, offering no assessment of model performance in genuinely unknown areas. Yet this extrapolation scenario is precisely what our model was designed to handle—and a known weakness of QRF approaches [44].

Nevertheless, the proposed model demonstrated performance comparable to QRF in interpolation tasks, confirming its effectiveness and accuracy in spatial interpolation. This suggests that the model successfully retains local fitting capability while incorporating mechanisms for enhanced generalization.



(**a**)

Figure 13. *Cont.*



(**b**)

Figure 13. Comparison of (**a**) pinball score and (**b**) mean reliability score values with error bars.

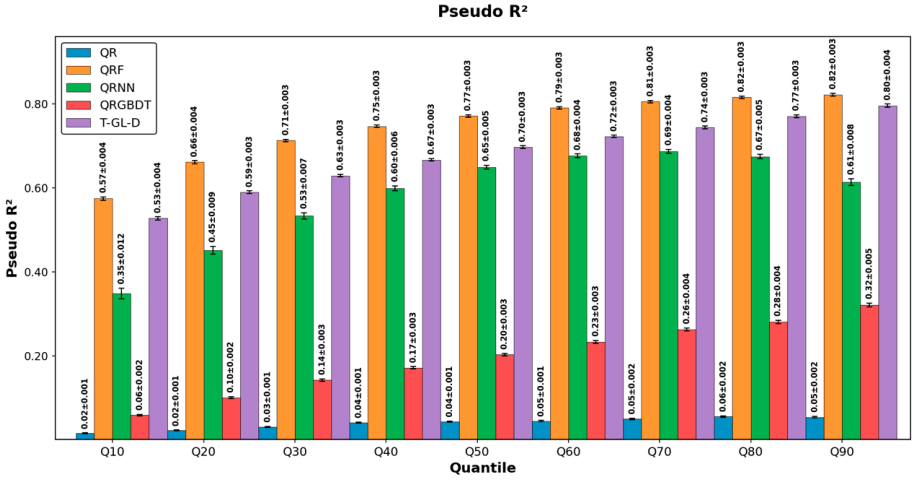
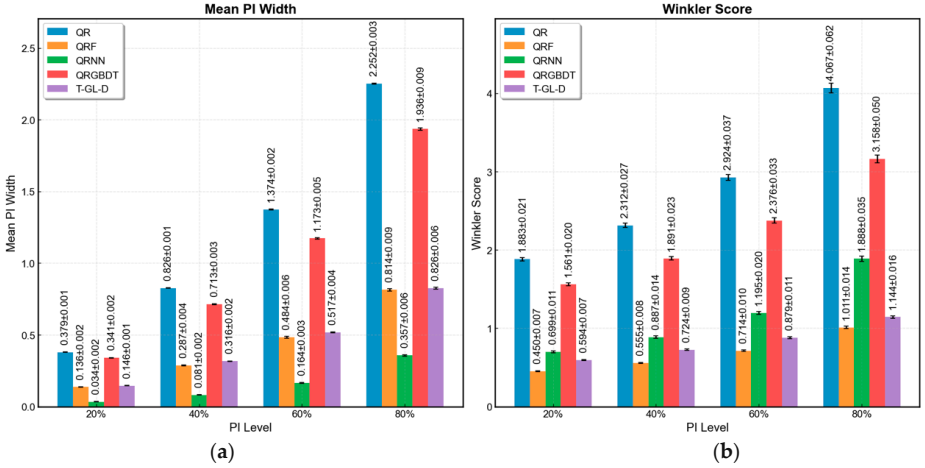


Figure 14. Comparison of pseudo R^2 values with error bars.



(**a**) (**b**)

Figure 15. Comparison of (**a**) mean PI width values and (**b**) Winkler score with error bars.

The target area delineation presented in this study was not subjected to validity testing, representing a limitation of the current work. Future work will extend the evaluation to multi-site validation using spatially disjoint training and test regions with GPU, thereby providing a more rigorous assessment of the model's capacity to predict in entirely unknown spatial contexts and confirming its generalization capability across heterogeneous geographic regions. Such an experimental design would better reveal the advantages of the global–local attention mechanism and dropout strategy in handling genuine spatial prediction challenges.

6. Conclusions

This study presents a Transformer-based framework for 3D mineral prospectivity mapping (MPM) in Xiadian gold deposits, designed to capture both structural controls and localized mineralization patterns inherent in complex geological datasets. The proposed architecture incorporates coordinate-aware relative position encoding and a global–local fused attention mechanism to enhance representational capacity and spatial sensitivity. By employing pinball loss within a probabilistic framework, the model transcends traditional MSE or MAE losses by directly capturing the uncertainty inherent in predictions. The ablation experiments, conducted on the Xiadian gold deposit, were rigorously evaluated using Bootstrap analysis. The results demonstrate the enhancement of each model component, as evaluated through quantile-specific accuracy metrics and PI calibration across multiple evaluation criteria. The spatial distribution of the model outputs exhibited a quantile-dependent expansion of predicted high-grade mineralization, which remained spatially coherent with known enrichment zones. The stability of the framework was further corroborated through ten iterations of random data partitioning. Applied to the unexplored area, the established framework generated a prospectivity model and delineated exploration targets across multiple probability quantiles. The resulting spatial distributions exhibit strong cross-quantile consistency and align closely with known geological principles, providing preliminary evidence for the model's reliability. Furthermore, comparative analysis against state-of-the-art methods confirms its competitive fitting ability.

This study demonstrates the potential of Transformer-based 3DMPM as a promising approach in complex geological settings. Nevertheless, as a preliminary attempt, multiple limitations need to be addressed in future research. To advance this direction, subsequent research should focus on four interconnected fronts: (1) architectural innovation, by deepening the integration of Transformers with multi-scale geological systems through a geology-informed attention mechanism that explicitly encodes domain knowledge and spatial relationships; (2) rigorous evaluation, by implementing multi-site validation and incorporating comprehensive uncertainty quantification to assess robustness and spatial generalizability; (3) enhanced physical consistency, by developing hybrid architectures and domain adaptation techniques that embed physical constraints to ensure geologically consistent predictions across diverse mineral systems; and (4) computational efficiency, by reducing computational cost through algorithmic improvements and expanded computational resources.

Author Contributions: Conceptualization, Q.L.; methodology, X.H.; software, X.H. and P.W.; validation, X.H.; formal analysis, X.H.; investigation, X.H.; data curation, X.H.; writing—original draft, X.H.; writing—review and editing, X.H. and Q.L.; visualization, X.H.; supervision, P.W. and Q.L.; project administration, Q.L.; funding acquisition, P.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been generously supported by the National Major Science and Technology Projects of China (No. 2024ZD1001904), the Natural Science Foundation of Hunan Province (No.

2024JJ8323, 2026JJ30010) and the 2025 College Students' Innovation and Entrepreneurship Training Program (University-level Cultivation Project) of Central South University (No. CXPY2025329).

Data Availability Statement: The datasets generated during the current study are not publicly available due to a confidentiality agreement.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mao, X.; Chen, G. The Xianghualing Sn-deposit: Its mathematical model and three-dimensional quantitative prognostication. *Geol. Prospect.* **1988**, *24*, 25–31.
2. Zuo, R.; Carranza, E.J.M. Support vector machine: A tool for mapping mineral prospectivity. *Comput. Geosci.* **2011**, *37*, 1967–1975. [CrossRef]
3. Dürr, S.; Levy, A.; Rothlisberger, U. Accurate prediction of transition metal ion location via deep learning. *bioRxiv* **2022**. [CrossRef]
4. Dürr, S.; Levy, A.; Rothlisberger, U. Metal3D: A general deep learning framework for accurate metal ion location prediction in proteins. *Nat. Commun.* **2023**, *14*, 2713. [CrossRef] [PubMed]
5. Li, X.H.; Yuan, F.; Zhang, M.M.; Jia, C.; Jowitt, S.M.; Ord, A.; Zheng, T.; Hu, X.; Li, Y. Three-Dimensional Mineral Prospectivity Modeling for Targeting of Concealed Mineralization within the Zhonggu Iron Orefield, Ningwu Basin, China. *Ore Geol. Rev.* **2015**, *71*, 633–654. [CrossRef]
6. Hoseinzade, Z.; Shojaei, M.; Khademi, F.; Mokhtari, A.R.; Saremi, M. Integration of deep learning models for mineral prospectivity mapping: A novel Bayesian index approach to reducing uncertainty in exploration. *Model. Earth Syst. Environ.* **2025**, *11*, 161. [CrossRef]
7. McCuaig, T.C.; Kreuzer, O.P.; Brown, W.M. Fooling ourselves—Dealing with model uncertainty in a mineral systems approach to exploration. In *Mineral Exploration and Research-Digging Deeper: Proceedings of the 9th Biennial SGA Meeting*; Society for Geology Applied to Mineral Deposits: Geneva, Switzerland, 2007; pp. 1435–1438.
8. Huang, J.; Deng, H.; Mao, X.; Wan, S.; Liu, Z. A Global-Local collaborative approach to quantifying spatial non-stationarity in three-dimensional mineral prospectivity modeling. *Ore Geol. Rev.* **2024**, *168*, 106069. [CrossRef]
9. Cheng, Q. Modeling local scaling properties for multiscale mapping. *Vadose Zone J.* **2008**, *7*, 525–532. [CrossRef]
10. Pohl, W. Metallogenic models as the key to successful exploration—A review and trends. *Miner. Econ.* **2022**, *35*, 373–408. [CrossRef]
11. Vigneresse, J.L. Addressing ore formation and exploration. *Geosci. Front.* **2019**, *10*, 1613–1622. [CrossRef]
12. Mooney, C.R. Spatial Modelling of Heavy-Tailed Mineral Grades Using a Spatial Point Process. Master's Thesis, University of Alberta, Edmonton, AB, Canada, 2015. [CrossRef]
13. Dutaut, R.V.; Marcotte, D. A new grade-capping approach based on coarse duplicate data correlation. *J. South. Afr. Inst. Min. Metall.* **2021**, *121*, 193–200. [CrossRef]
14. Zuo, R.; Carranza, E.J.M. Deep learning for mineral prospectivity mapping: A review. *Ore Geol. Rev.* **2021**, *128*, 103887.
15. Deng, H.; Zheng, Y.; Chen, J.; Yu, S.; Xiao, K.; Mao, X. Learning 3D mineral prospectivity from 3D geological models using convolutional neural networks: Application to a structure-controlled hydrothermal gold deposit. *Comput. Geosci.* **2022**, *161*, 105074. [CrossRef]
16. Ghyselincks, S.; Okhmak, V.; Zampini, S.; Turkiyyah, G.; Keyes, D.; Haber, E. Synthetic geology: Structural geology meets deep learning (Version 3). *J. Geophys. Res. Mach. Learn. Comput.* **2026**, *3*, e2025JH000986. [CrossRef]
17. Huang, J.; Wan, S.; Mao, W.; Deng, H.; Chen, J.; Tang, W. Risk-aware quantitative mineral prospectivity mapping with quantile-based regression models. *Nat. Resour. Res.* **2024**, *33*, 2433–2455. [CrossRef]
18. Huang, J.; Wan, S.; Deng, H.; Zhang, B.; Huang, X.; Mao, X. Quantifying Spatial and Statistical Heterogeneities in the Relationships Between Mineralization and its Determinants for Quantile–Specific 3D Mineral Prospectivity Mapping. *Nat. Resour. Res.* **2026**. *online first*. [CrossRef]
19. Koenker, R.; Bassett, G., Jr. Regression quantiles. *Econometrica* **1978**, *46*, 33–50. [CrossRef]
20. Azimli, A. The impact of COVID-19 on the degree of dependence and structure of risk-return relationship: A quantile regression approach. *Financ. Res. Lett.* **2020**, *36*, 101648. [CrossRef]
21. Hung, N.T. Green investment, financial development, digitalization and economic sustainability in Vietnam: Evidence from a quantile-on-quantile regression and wavelet coherence. *Technol. Forecast. Soc. Change* **2023**, *186*, 122185. [CrossRef]
22. Liu, F.; Umair, M.; Gao, J. Assessing oil price volatility co-movement with stock market volatility through quantile regression approach. *Resour. Policy* **2023**, *81*, 103375. [CrossRef]
23. Li, Z.; Patel, N.; Liu, J.; Kautish, P. Natural resources-environmental sustainability-socio-economic drivers nexus: Insights from panel quantile regression analysis. *Resour. Policy* **2023**, *86*, 104176. [CrossRef]

24. Castro, M.; Azevedo, C.; Nobre, J. A robust quantile regression for bounded variables based on the Kumaraswamy Rectangular distribution. *Stat. Comput.* **2024**, *34*, 74. [CrossRef]
25. Wang, K.; Zhang, D.; Sun, X. Robust Composite Quantile Regression with Large-scale Streaming Data Sets. *Scand. J. Stat.* **2025**, *52*, 736–755. [CrossRef]
26. Sottile, G.; Frumento, P. Robust estimation and regression with parametric quantile functions. *Comput. Stat. Data Anal.* **2022**, *171*, 107471. [CrossRef]
27. Beyaztaş, U.; Tez, M.; Shang, H. Robust scalar-on-function partial quantile regression. *J. Appl. Stat.* **2023**, *51*, 1359–1377. [CrossRef]
28. Zhang, J.; Yang, H. Bounded quantile loss for robust support vector machines-based classification and regression. *Expert Syst. Appl.* **2023**, *242*, 122759. [CrossRef]
29. Jia, Y.; Jeong, J.H. Deep learning for quantile regression under right censoring: DeepQuantreg. *Comput. Stat. Data Anal.* **2022**, *165*, 107323. [CrossRef]
30. Bailie, T.; Koh, Y.S.; Rampal, N.; Gobson, P.B. Quantile-regression-ensemble: A deep learning algorithm for downscaling extreme precipitation. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*; AAAI Press: Washington, DC, USA, 2024; Volume 38, pp. 21914–21922.
31. Wu, R.; Tian, J.; Yao, J.; Han, T.; Hu, C. Confidence-aware quantile Transformer for reliable degradation prediction of battery energy storage systems. *Reliab. Eng. Syst. Saf.* **2025**, *260*, 111019. [CrossRef]
32. Cannon, A.J. Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Comput. Geosci.* **2011**, *37*, 1277–1284. [CrossRef]
33. Taylor, J.W. A quantile regression neural network approach to estimating the conditional density of multiperiod returns. *J. Forecast.* **2000**, *19*, 299–311. [CrossRef]
34. Liu, Z.; Mao, X.; Jedemann, A.; Bayless, R.C.; Deng, H.; Chen, J.; Xiao, K. Evolution of pyrite compositions at the Sizhuang gold deposit, Jiaodong Peninsula, Eastern China: Implications for the genesis of Jiaodong-type orogenic gold mineralization. *Minerals* **2021**, *11*, 344. [CrossRef]
35. Mao, X.; Ren, J.; Liu, Z.; Chen, J.; Tang, L.; Deng, H.; Liu, C. Three-dimensional prospectivity modeling of the Jiaojia-type gold deposit, Jiaodong Peninsula, Eastern China: A case study of the Dayingezhuang deposit. *J. Geochem. Explor.* **2019**, *203*, 27–44.
36. Huang, J.; Liu, Z.; Deng, H.; Li, L.; Mao, X.; Liu, J. Exploring Multiscale Non-stationary Influence of Ore-Controlling Factors on Mineralization in 3D Geological Space. *Nat. Resour. Res.* **2022**, *31*, 3079–3100. [CrossRef]
37. Mao, W.; Liu, P.; Huang, J. SF-Transformer: A Mutual Information-Enhanced Transformer Model with Spot-Forward Parity for Forecasting Long-Term Chinese Stock Index Futures Prices. *Entropy* **2024**, *26*, 478. [CrossRef]
38. Zrimšek, U.; Štrumbelj, E. Quantifying uncertainty: All we need is the bootstrap? *J. Stat. Comput. Simul.* **2025**, *96*, 1009–1027. [CrossRef]
39. Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
40. Hollander, M.; Wolfe, D.A.; Chicken, E. *Nonparametric Statistical Methods*, 3rd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2013.
41. Yu, Y.; Wang, M.; Yan, F.; Yang, M.; Yang, J. Improved convolutional neural network-based quantile regression for regional photovoltaic generation probabilistic forecast. *IET Renew. Power Gener.* **2020**, *14*, 2712–2719. [CrossRef]
42. Winkler, R.L. A Decision-Theoretic Approach to Interval Estimation. *J. Am. Stat. Assoc.* **1972**, *67*, 187–191. [CrossRef]
43. Rubin, M. That's Not a Two-Sided Test! It's Two One-Sided Tests! *Significance* **2022**, *19*, 50–53.
44. Booker, D.J.; Whitehead, A.L. Inside or Outside: Quantifying Extrapolation Across River Networks. *Water Resour. Res.* **2018**, *54*, 6983–7003. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.