Response to Reviewer 2's comments

Reviewer's comments are in black. Responses are in blue.

The manuscript proposes a transformer-based approach for estimating Doppler velocity from ISR spectra, trained entirely on synthetic data generated from standard ISR theory. The idea is interesting, but several scientific and methodological issues weaken the validity of the results. The current version does not show sufficient evidence that the method generalizes beyond the highly idealized training configuration, nor does it justify the architectural design or provide the necessary details for reproducibility.

The current version does not meet the level of rigor required for publication. Major revisions are needed. Until these issues are resolved, the reliability of the scientific conclusions remains uncertain.

Response: We have modified the manuscript extensively and hope we have addressed the issues raised by the reviewer. As the method requires only the theoretical model and the observed spectra, it is applicable to all cases where the least-squares fitting technique is used. Training for the model is general and does not have any restrictions beyond the incoherent scatter theory.

Below I summarize the major concerns:

1. Oversimplified synthetic data

The model is trained exclusively on synthetic data generated under very restrictive assumptions. Equation (1) produces only smooth and monotonic profiles with limited variability. This excludes many structures that commonly appear in real ISR observations, such as sharp gradients, inversion layers, and localized irregularities.

Since the network never sees more realistic patterns during training, it is not possible to conclude that it will behave reliably when applied to real conditions. In addition, the evaluation metric used in the real-data analysis (second differences in height and time) strongly favors smooth profiles. A model trained on artificially smoothed data will naturally perform well under such a metric, even if the estimated Doppler velocities are biased or physically incorrect.

The manuscript does not discuss these limitations, and no experiments are presented to test the robustness of the method to deviations from the synthetic assumptions. Therefore, the reported improvements cannot be interpreted as evidence that the transformer is learning meaningful Doppler information. Instead, they appear to reflect the smoothness that is already imposed by the synthetic data.

The manuscript also refers to other works for parameter bounds, but these bounds should be explicitly stated here. The reader does not know the limits used to

generate the synthetic data.

Response: The synthetic spectra cover all practical combinations. In ISR applications, the synthesized data with noise added are observations without instrumental problems. The smooth profile assumption is within the 1.5 km height resolution. Any variation with a scale larger than 1.5 km is completely resolved. 1.5 km is a well-accepted fine resolution for ISR applications.

The instrument limit at Arecibo is 300 m, which is practically a limit for all ISRs. If the instrument-limited resolution of 300 m in the output is required, one must use the height-unaware model. While we see in Figure 2 that AI height-unaware model can do better than LSF, it does not offer the same advantage as the height-aware model if a reduced resolution is acceptable.

One way to think about Equation (1) is to compare it with traditional ways to do averages. In traditional approaches, all the heights are averaged with a fixed set of weights (often equal weights). Equation (1) allows the weights to vary dynamically and thus makes optimization possible. We would like to point out that the most essential information needed for training the model is the theoretical spectra. While our tests show that Equation (1) works reasonably well for the purpose of context awareness, there can be other forms of the equation that may work even better.

In short, the AI model does not assume or require anything more than the LSF fitting technique.

2. Extremely large transformer without justification

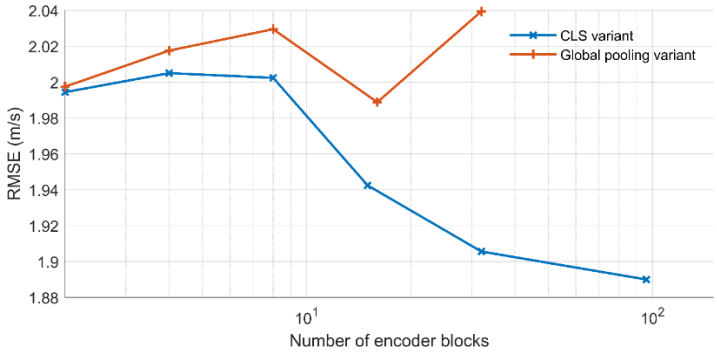
The proposed architecture is a 31-layer transformer with approximately 100 million parameters. This is extremely large considering that the input consists of only five altitude bins and 101 frequency channels. The manuscript does not justify why such a large model is necessary. The vertical variability is limited and the spectral dimension is fixed, under these circumstances simpler architectures (e.g., shallow transformers, small CNNs, or MLPs) could probably achieve similar performance.

The manuscript does not include comparisons with simpler models and does not analyze model efficiency. As a result, it is only demonstrated that “a very large transformer can work,” not that this architecture is appropriate nor optimal. Without justification, the use of such a large model raises concerns that it may overfit the limited synthetic data distribution rather than learn robust physical behavior.

Response: The task addressed here is not a generic regression from a small input, but the extraction of a weak Doppler signal embedded in spectrally structured noise.

While the proposed model contains approximately 100 million parameters, this scale is comparable to widely used base configurations of modern transformer architectures (e.g., ViT-Base) and is not unusually large by current standards.

During model development, we examined the effect of transformer depth on Doppler velocity estimation accuracy. As shown in the blue curve below, the blue curve (CLS variant) observes performance gain up to 100 transformer blocks and perhaps beyond. The 31-block architecture in the present work was selected to balance accuracy and computational cost.



3. Incorrect characterization as physics-informed ML

The manuscript states that the approach "aligns with broader definitions of physics-informed machine learning." This is not accurate. A physics-informed model usually incorporates physical constraints directly into the architecture or the loss function (for example, PDEs or conservation laws). In this work, the model is a conventional transformer without any embedded physical constraint. The physics appears only through the synthetic spectra used for training. This distinction is important to avoid giving the impression that the network structurally incorporates ISR physics.

Response: The statement referring to "broader definitions of physics-informed machine learning" has been removed in the revised manuscript. The proposed model does not incorporate physical constraints directly into the architecture or loss function; physics enters only through the synthetic training data.

4. Unclear definition of network input and output

The manuscript does not clearly describe the exact input and output of the proposed neural network. It is not explicitly stated how the 5 altitude bins and 101 frequency channels are arranged, normalized, or preprocessed before being passed into the Conv1D layer. Likewise, it remains unclear whether the network outputs a single

Doppler velocity for the entire 5-bin block or one velocity per altitude bin. This information is essential to understand the model, to evaluate the fairness of comparisons with LSF, and to reproduce the results. The authors should provide a precise description of the input tensor shape, preprocessing steps, tokenization, and the definition of the output quantity.

Response: The network input consists of ion-line spectra sampled at five adjacent altitude bins, each with 101 frequency channels (5×101). The network outputs a single Doppler velocity for the full 5-bin window. Using five adjacent heights provides vertical context, as opposed to the least-squares fitting baseline, which typically averages several heights into a single profile before estimating Doppler velocity.

5. Insufficient description of the architecture

Several essential architectural details are missing. The number of attention heads, dropout usage, kernel sizes, feedforward dimensions, activation functions, normalization layers, positional-encoding formulation, and parameter initialization are not specified. For a model of this size, these details have a significant impact on performance and reproducibility.

In addition, the processing of the ISR spectra before they are given to the network, how the spectra are grouped, normalized, interpolated, and reshaped into tokens, is described only partially. These omissions prevent independent replication and validation.

Response: The model description in the revised manuscript has been substantially expanded.

The model uses a standard transformer encoder architecture. Each block consists of multi-head self-attention with 4 heads (128 features per head) followed by a feed-forward network with width 4×d_model, consistent with the original transformer design. Layer normalization and residual connections follow the standard formulation. These details have now been explicitly stated in the revised manuscript.

6. Incomplete simulation and comparison methodology

Important elements of the training and evaluation pipeline are not well described: the noise model, sampling of SNR and bandwidth, instrumental effects, real-data preprocessing, and the exact LSF configuration. The real-data comparison relies only on smoothness-based metrics with no independent reference, making it difficult to assess true accuracy.

Response: The noise model, SNR and bandwidth sampling, instrumental effects, real-data preprocessing, and the exact LSF configuration are described in Section 2.

One may not necessarily agree on the methodology of error computation. What matters most in the current study is the comparison between the LSF and AI results. We use exactly the same metrics for AI and LSF.

7. Limited applicability to other ISR systems

The manuscript suggests that the method could be applied to any ISR. However, the current model is strongly tied to the specific configuration of the Arecibo 430 MHz system. The synthetic spectra used for training assume a fixed radar frequency, fixed altitude resolution, fixed frequency resolution, fixed bandwidth, and a specific SNR and spectral shape that correspond only to this station.

Other ISR systems (e.g., EISCAT, Millstone Hill, Jicamarca) use very different pulse codes, ambiguity functions, altitude sampling, integration schemes, and spectral characteristics. Because the model learns the distribution of spectra produced by the Arecibo configuration, it cannot be directly transferred to another system. Adapting the model would require generating a completely new synthetic dataset (including all the possible REAL profiles) consistent with the new radar and then retraining the model from the beginning.

For this reason, the claim of broad applicability is overstated, and the manuscript should clearly acknowledge that the method is not easily portable to other ISR facilities.

Response: The synthetic training data consist of generic ISR ion-line spectra and are not limited to a specific facility. Radar frequency, bandwidth, altitude resolution, and noise characteristics are explicitly configurable in the simulations, allowing the same framework to be adapted to other ISR systems by regenerating instrument-specific synthetic data and retraining the model. Such retraining is expected when adapting the model to other facilities and is straightforward. The main difference between Arecibo and other facilities is in the signal-to-noise ratio. Figure 2 gives a large range of normalized noise standard deviation that can be applied to other ISR facilities.