SPECTRALTF-MAE: A PRE-TRAINED SPECTRAL FOUNDATION MODEL FOR TOKAMAK FUSION PLASMA DIAGNOSTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

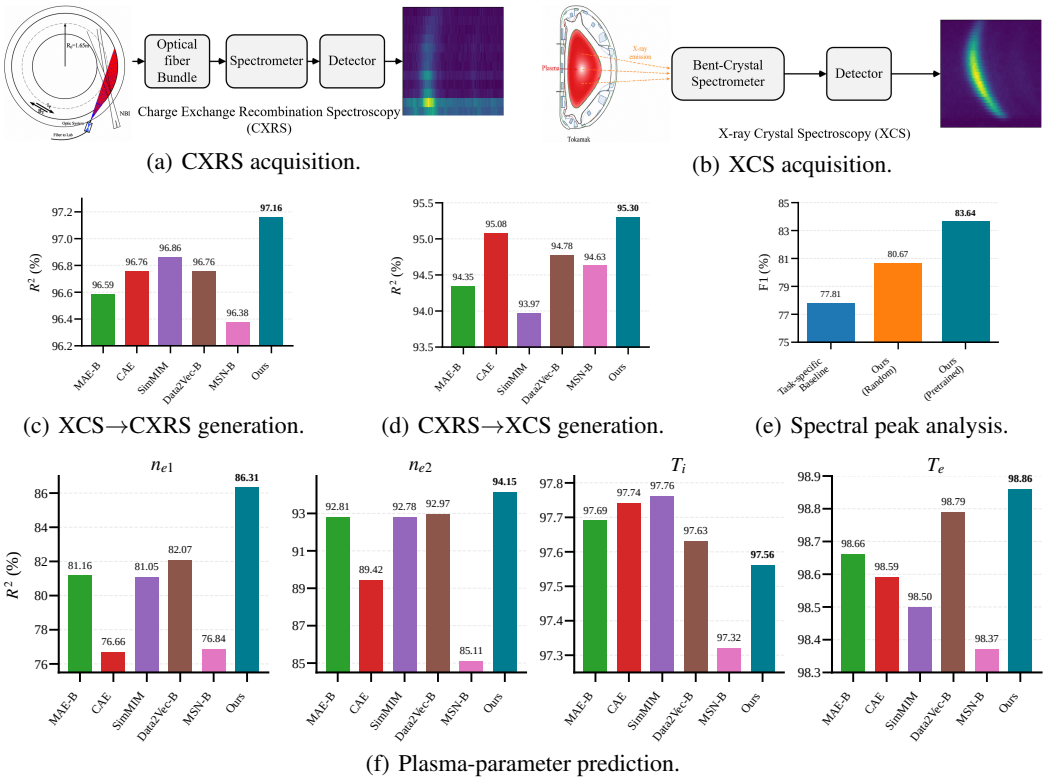
Spectroscopic diagnostics provide rich information for characterizing plasma states in tokamak fusion experiments, while existing learning-based approaches are often designed for individual tasks and rely heavily on labeled data. In this work, we propose SpectralTF-MAE, a self-supervised spectral foundation model for tokamak plasma diagnostics. The model is pretrained on paired charge-exchange recombination spectroscopy (CXRS) and X-ray crystal spectroscopy (XCS) measurements through masked spectral reconstruction. To better capture diagnostically important spectral structures, we introduce a peak-aware masking strategy together with a peak-enhanced reconstruction objective, encouraging the model to focus on characteristic spectral regions. A shared Transformer encoder is employed to learn transferable representations across different diagnostic modalities. The pretrained model is further adapted to several downstream tasks, including plasma-parameter prediction, cross-diagnostic spectral generation, and spectral peak analysis. Experiments on EAST diagnostic data demonstrate that the proposed framework achieves competitive performance across different tasks and consistently benefits from self-supervised spectral pretraining. These results indicate the potential of SpectralTF-MAE as a general and transferable representation learning framework for plasma spectroscopic diagnostics.

1 INTRODUCTION

Artificial Intelligence (AI) empowered controlled nuclear fusion has attracted growing attention from researchers (Humphreys et al., 2020; Degrave et al., 2022; Seo et al., 2024). Spectroscopic diagnosis constitutes one of the core approaches for plasma analysis. On the EAST large-science facility, commonly used spectroscopic diagnostic systems acquire two major categories of spectra: active spectra and passive spectra. Specifically, charge-exchange recombination spectroscopy (CXRS) is an active diagnostic technique that employs external neutral beams to induce charge-exchange reactions and obtain plasma parameters such as ion temperature and rotation velocity, as illustrated in Fig. 1 (a). In contrast, X-ray crystal spectroscopy (XCS) is a passive diagnostic technique that measures intrinsic X-ray emission from plasma and provides information related to electron temperature, impurity behavior, and plasma conditions, as shown in Fig. 1 (b). Existing studies generally adopt small-scale models for spectral interpretation (Chai et al., 2019; Lin et al., 2024; 2025; Saura et al., 2025; Li et al., 2025), which evidently fail to fully exploit the potential of large-scale spectral datasets. Spectral perception tailored for controlled-nuclear-fusion tasks remains an open and pressing challenge.

With the proposal of the Transformer (Vaswani et al., 2017) and the rapid advancement of large-model technologies (Bommasani et al., 2021), a variety of large models for text, vision and even multimodal data have emerged (Devlin et al., 2019; Dosovitskiy et al., 2020; Radford et al., 2021). Leveraging self-supervised learning, these models capture inherent patterns from massive training data and achieve substantial performance improvements across downstream tasks. Several existing studies (Su et al., 2026; Liu et al., 2025) have investigated large-model-based learning strategies for spectral data, such as SpectraLLM (Su et al., 2026) and MolSpectra (Liu et al., 2025). Nevertheless, these approaches have not performed dedicated pre-training on controlled-fusion spec-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) CXRS acquisition. (b) XCS acquisition.

(c) XCS→CXRS generation. (d) CXRS→XCS generation. (e) Spectral peak analysis.

(f) Plasma-parameter prediction.

Figure 1: (a-b). Demonstration of different principles for spectral-data acquisition; (c-f). Performance comparison of downstream tasks based on our newly proposed spectral foundation model.

tral data. Consequently, there remains a lack of domain-specific large models tailored for this type of data.

Through a systematic investigation of the spectral applications in controlled nuclear fusion research, we summarize three core requirements for plasma spectral analysis. **First**, accurate predictions of plasma temperature (T_i/T_e) profiles and electron density (n_e) from measured spectral data are essential. **Second**, active and passive spectra show inconsistent correspondence, and the prevalent missing data in active spectra severely limits downstream diagnostic performance. **Third**, quantitative analysis of spectral peak intensity and spectral broadening is required to provide reliable data support for subsequent plasma parameter inversion and follow-up research tasks. Existing studies (Chai et al., 2019; Samuell et al., 2021; Lin et al., 2024; 2025; Lee et al., 2026; Saura et al., 2025; Li et al., 2025) typically design independent network architectures for individual sub-tasks and train corresponding surrogate models using small-scale networks. Such paradigms fail to fully explore the potential of massive spectral data and cannot leverage the powerful data fitting and feature mining capabilities of large models.

In this paper, we propose a novel pre-trained spectral foundation model for tokamak fusion plasma diagnostics based on Masked Auto-Encoder (MAE) (He et al., 2022), termed SpectralTF-MAE. As shown in Fig. 2, given the XCS and CXRS spectral input, we first perform spectral-peak detection and adaptive masking on it. Different from the random masking strategy adopted in vanilla MAE, our adaptive masking implements automatic thresholding over high-amplitude spectral-peak regions. This physics-aware masking operation prioritizes the most informative peak components for corruption, which forces the model to learn underlying physical correlations of fusion spectra rather than trivial background patterns. Then, we project the remaining spectral patches into tokens using a linear layer and extract the deep representations using a shard spectral Transformer encoder network. After that, the XCS and CXRS decoder networks are utilized for masked spectral reconstruction. We conduct experiments on three tasks to validate the effectiveness of our pre-trained spectral foundation model SpectralTF-MAE, including cross-spectral generation (i.e., XCS

$\rightarrow$ CXRS, CXRS $\rightarrow$ XCS), T_i/T_e profile prediction, electron-density (n_e) prediction, and spectral peak analysis.

To sum up, the main contributions of this paper can be summarized as follows:

1). We propose a novel foundation model for controllable-fusion spectral data, termed SpectralTF-MAE, which builds on the Masked Auto-Encoder (MAE) self-supervised learning framework and adopts a physics-aware masking strategy to incentivize the model to learn intrinsic plasma spectral characteristics.

2). Based on the proposed spectral foundation model, we develop a suite of downstream tasks covering spectral generation and analysis, including cross-spectral generation, T_i/T_e profile prediction, electron-density (n_e) prediction, and spectral-peak analysis, which establishes a solid model benchmark for large model driven controlled nuclear fusion spectral analysis.

3). Extensive experiments with comprehensive quantitative and qualitative analysis on the newly collected spectral dataset fully validated the effectiveness and generalization of our proposed pretrained spectral foundation model.

2 Related Works

Foundation models learn transferable representations through large-scale pretraining and downstream adaptation. Transformer-based masked or predictive self-supervised learning has been successful in language and vision (Vaswani et al., 2017; Devlin et al., 2019; He et al., 2022; Bao et al., 2021; Xie et al., 2022; Chen et al., 2024; Baevski et al., 2022), and related ideas have been extended to hyperspectral and other scientific spectral data (Hong et al., 2021; Cong et al., 2022; Hong et al., 2024; Wang et al., 2025; Parker et al., 2024; Leung and Bovy, 2024). These studies motivate reusable spectral representations, but plasma spectroscopy remains comparatively underexplored.

Machine learning has also been widely investigated for magnetic-confinement fusion, including disruption prediction, diagnostic inference, and plasma control (Humphreys et al., 2020; Rea et al., 2018; Montes et al., 2019; Degrave et al., 2022; Seo et al., 2024). More recently, reusable pretrained models have begun to appear in fusion research (Churchill, 2025; Yang et al., 2026). Existing fusion foundation-model studies primarily emphasize heterogeneous diagnostic time series and global plasma-state representations, whereas our work focuses on high-dimensional XCS and CXRS measurements and transferable spectral representations.

Spectroscopic diagnostics provide measurements of plasma temperature, density, impurity states, and rotation (Weller, 2010; Mazon et al., 2025). Learning-based methods have been developed for temperature and rotation inference from fusion spectra (Chai et al., 2019; Samuell et al., 2021; Lin et al., 2024; 2025; Lee et al., 2026), but these approaches are typically task-specific and fully supervised. In contrast, SpectralTF-MAE learns a shared representation from paired XCS and CXRS measurements and transfers it to plasma-parameter prediction, cross-diagnostic generation, and spectral peak analysis. Additional related-work discussion is provided in Appendix A.

3 Our Proposed Approach

3.1 Overview

We propose a self-supervised spectral representation learning framework for plasma diagnostics, consisting of multi-diagnostic spectral pretraining and downstream adaptation. As illustrated in Fig. 2, paired charge-exchange recombination spectroscopy (CXRS) and X-ray crystal spectroscopy (XCS) measurements are jointly modeled through masked spectral reconstruction without requiring plasma-parameter labels. The two spectral modalities are embedded into a shared latent space and processed by a Transformer encoder (Vaswani et al., 2017) to capture both modality-specific structures and cross-diagnostic correlations. To incorporate plasma-specific spectral characteristics, we introduce a peak-aware masked modeling strategy that emphasizes informative emission regions and encourages the preservation of physically relevant spectral features. After pretraining, the learned representation is adapted to downstream tasks, including plasma-parameter inference from XCS spectra and cross-diagnostic spectral generation between CXRS and XCS. This frame-

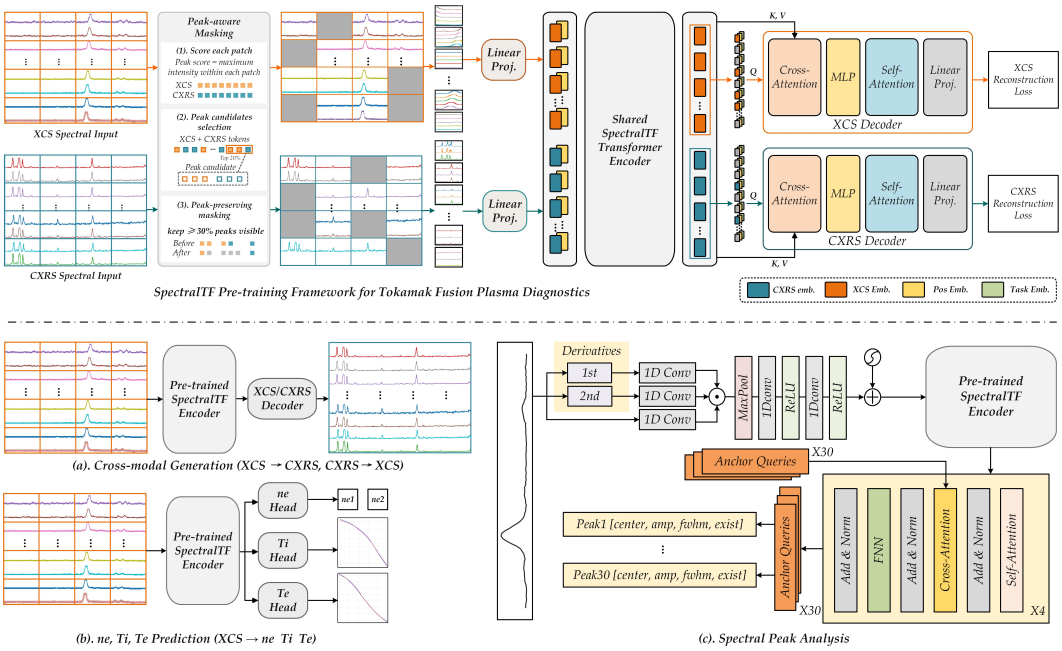


Figure 2: An overview of our proposed pre-trained spectral foundation model (upper panel) and downstream tasks (lower panel) for tokamak fusion plasma diagnostics.

work enables the utilization of paired unlabeled spectral data to learn transferable representations for diverse plasma diagnostic applications.

3.2 Pretraining Stage

The pretraining stage learns a shared spectral representation from paired CXRS and XCS measurements without plasma-parameter labels. As shown in Fig. 2, both spectra are normalized and converted into modality-specific tokens. Peak-aware masking defines the reconstruction targets, while the remaining visible tokens are jointly encoded by a shared Transformer. Modality-specific decoders then reconstruct the two diagnostics from the shared latent representation, and a peak-enhanced objective places additional emphasis on the characteristic spectral regions.

Let $\widetilde{\mathbf{X}}^m$ denote the normalized spectrum for modality $m \in \{c, x\}$. Modality-specific rectangular patches are projected into a common latent space with positional embeddings. To emphasize diagnostically informative responses, the maximum normalized intensity within each patch is used as a peak score, and a sample-specific quantile threshold identifies peak candidates. These candidates are preferentially selected as reconstruction targets while part of the peak-related context remains visible.

The visible CXRS and XCS tokens are concatenated and processed by a shared Transformer encoder,

$$\mathbf{Z} = \mathcal{F}_{\text{enc}}([\mathcal{V}^c; \mathcal{V}^x; \mathbf{g}]), \tag{1}$$

where $\mathbf{g}$ is a learnable global token. Separate modality-specific decoders reconstruct CXRS and XCS from the shared representation. For each diagnostic, reconstruction is optimized over the masked region $\mathcal{M}^m$, with additional weight assigned to the predefined characteristic region Ω_m. The overall objective is

$$\mathcal{L}_{\text{pre}} = \sum_{m \in \{c,x\}} \left(\mathcal{L}^m_{\text{rec}} + \lambda_{\text{peak}} \mathcal{L}^m_{\text{peak}} \right). \tag{2}$$

Peak-aware masking determines which spectral structures are preferentially reconstructed, whereas the peak-enhanced loss increases optimization emphasis on characteristic spectral regions. Full tokenization, masking, reconstruction, and loss definitions are given in Appendix B.

3.3 Downstream Tasks

After self-supervised pretraining, the learned representation is adapted to three task families: plasma-parameter prediction from XCS measurements, bidirectional cross-diagnostic spectral generation between XCS and CXRS, and spectral peak analysis. These tasks evaluate physical-parameter inference, cross-diagnostic correspondence, and spectral-component analysis, respectively.

• **Plasma-Parameter Prediction.** For temperature prediction, T_i and T_e are radial-profile regression targets represented at 32 radial positions. For electron density, two measurements are used: n_{e1} denotes the solid-source measurement and n_{e2} is obtained from POINT. After consistency-based screening, 5,330 valid samples are retained. The complete XCS spectrum is encoded and passed to independent task-specific regression heads,

$$\widehat{\mathbf{y}}_k = \mathcal{H}_k(\mathbf{f}_x), \qquad k \in \{n_{e1}, n_{e2}, T_i, T_e\}. \tag{3}$$

• **Cross-Diagnostic Spectral Generation.** Given a complete source spectrum, the pretrained encoder and target-diagnostic decoder reconstruct the complete target spectrum,

$$\widehat{\mathbf{X}}^t = \mathcal{D}_t(\mathcal{F}_{\text{enc}}(\mathbf{X}^s)), \qquad s \neq t. \tag{4}$$

XCS-to-CXRS and CXRS-to-XCS directions are fine-tuned independently, with supervision concentrated on the characteristic peak region of the target diagnostic.

• **Spectral Peak Analysis.** The downstream peak-analysis model decomposes an input spectrum into multiple components and predicts the center, amplitude, full width at half maximum (FWHM), and existence probability of each peak. To isolate the contribution of pretraining, the original task-specific pipeline is retained and only its Transformer encoder is replaced by the pretrained SpectralTF encoder. Additional task details are provided in Appendix C.

4 Experiments

4.1 Datasets and Evaluation Metrics

The spectral dataset contains paired CXRS and XCS measurements acquired during EAST experiments from 2021 to 2023. After synchronization and quality control, 15,303 valid paired samples are retained. The complete CXRS input is represented as a 1024×12 spectral matrix and XCS is spatially integrated to a 35×487 representation. Training-only augmentation jointly shifts paired spectra along the wavelength direction and adds Poisson noise; validation and test samples are not augmented.

Regression and spectral-generation tasks are evaluated using MAE, RMSE, and R^2. For T_i and T_e, metrics aggregate all test samples and all 32 radial positions. For cross-diagnostic generation, metrics are computed in the normalized characteristic region of the target diagnostic. Spectral peak analysis uses detection, localization, shape, and reconstruction metrics. Detailed preprocessing and metric definitions are provided in Appendix D.

4.2 Implementation Details

The framework was implemented in PyTorch (Paszke et al., 2019), and all experiments were conducted on an NVIDIA GeForce RTX 3090 GPU. SpectralTF-MAE-Base is used as the default configuration throughout the study. SpectralTF-MAE-Large is additionally evaluated on the two main quantitative tasks, namely bidirectional cross-diagnostic spectral generation and plasma-parameter prediction; unless otherwise stated, the remaining analyses and ablation studies use the Base configuration.

All models are optimized using AdamW (Loshchilov and Hutter, 2017). Self-supervised pretraining is conducted for 800 epochs with a batch size of 128 and an effective learning rate of 5×10^{-5}. Plasma-parameter prediction and each cross-diagnostic generation direction are fine-tuned for 200 epochs with a batch size of 32 and an effective learning rate of 6.25×10^{-6}. During pretraining, approximately 80% of the combined CXRS–XCS tokens are masked, whereas complete spectra

Table 1: Comparison of methods on bidirectional cross-diagnostic spectral generation. For XCS-to-CXRS generation, MAE and RMSE are calculated in the normalized CXRS characteristic peak region; for CXRS-to-XCS generation, they are calculated in the normalized XCS characteristic peak region. All R^2 values are multiplied by 100.

Method	XCS → CXRS			CXRS → XCS		
	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑
ResNet-50 (He et al., 2016)	0.00683	0.01398	96.90	0.02866	0.04759	90.04
ResNet-101 (He et al., 2016)	0.00677	0.01403	96.88	0.02880	0.04912	89.39
ViT-Small (Dosovitskiy et al., 2020)	0.01540	0.02625	89.08	0.06026	0.11271	44.13
ViT-Base (Dosovitskiy et al., 2020)	0.01552	0.02655	88.83	0.06012	0.11262	44.21
ViT-Large (Dosovitskiy et al., 2020)	0.01522	0.02603	89.26	0.05993	0.11254	44.29
MAE-Base (He et al., 2022)	0.00706	0.01466	96.59	0.02215	0.03583	94.35
MAE-Large (He et al., 2022)	0.00697	0.01431	96.76	0.02171	0.03531	94.52
CAE (Chen et al., 2024)	0.00730	0.01429	96.76	0.01974	0.03342	95.08
SimMIM (Xie et al., 2022)	0.00716	0.01407	96.86	0.02306	0.03701	93.97
Data2Vec-Base (Baevski et al., 2022)	0.00705	0.01429	96.76	0.02055	0.03444	94.78
Data2Vec-Large (Baevski et al., 2022)	0.00692	0.01429	96.77	0.01985	0.03381	94.90
MSN-Base (Assran et al., 2022)	0.00751	0.01510	96.38	0.02088	0.03491	94.63
MSN-Large (Assran et al., 2022)	0.00698	0.01379	96.99	0.01991	0.03351	94.99
SpectralTF-MAE-Base	0.00632	**0.01339**	97.16	0.01796	0.03269	95.30
SpectralTF-MAE-Large	**0.00623**	0.01364	**97.30**	**0.01760**	**0.03193**	**95.45**

are used without masking for downstream adaptation. The complete pretrained backbone is fine-tuned rather than frozen. For plasma-parameter prediction, targets are standardized using statistics estimated exclusively from the training set.

Model performance is monitored on the validation set, and the checkpoint achieving the best validation R^2-based criterion is retained for evaluation on the held-out test set. Hyperparameters are selected using the validation set only.

4.3 Comparison With Other Methods

The baselines include conventional convolutional and Transformer backbones as well as representative self-supervised pretraining methods. ResNet-50 and ResNet-101 follow the residual-learning architecture of He et al. (2016), and the ViT variants follow the Vision Transformer design (Dosovitskiy et al., 2020). The self-supervised baselines include MAE (He et al., 2022), CAE (Chen et al., 2024), SimMIM (Xie et al., 2022), data2vec (Baevski et al., 2022), and MSN (Assran et al., 2022). To make the model-scale comparison explicit, Base and Large variants are reported for ViT, MAE, data2vec, and MSN where applicable, alongside the Base and Large configurations of SpectralTF-MAE. Each baseline is initialized from a pretrained checkpoint released by its original authors. Downstream adaptation is controlled across methods by using the same task-specific datasets, downstream heads, fine-tuning strategy, and evaluation protocol. This setup preserves each method's published pretrained initialization while holding the downstream training conditions fixed.

• **Results on Bidirectional Spectral Generation.**

Table 1 compares the evaluated methods for both XCS-to-CXRS and CXRS-to-XCS spectral generation. The two generation directions are fine-tuned independently using the corresponding source–target pairs.

Across the two model scales, SpectralTF-MAE attains the best result for every metric in the bidirectional spectral-generation comparison. For XCS-to-CXRS generation, SpectralTF-MAE-Large achieves the lowest MAE (0.00623) and the highest R^2 (97.30), whereas SpectralTF-MAE-Base gives the lowest RMSE (0.01339). This complementary behavior indicates that increasing model

Table 2: Comparison of methods on plasma-parameter prediction from XCS measurements. The T_i and T_e metrics are calculated by aggregating all test samples and all 32 radial positions. All R^2 values are multiplied by 100.

2.6pt

Method	n_{e1}			n_{e2}			T_i			T_e		
	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑
ResNet-50 (He et al., 2016)	0.29697	0.41943	77.18	0.16697	0.26488	89.15	44.38147	66.45265	97.18	98.92274	165.28667	97.86
ResNet-101 (He et al., 2016)	0.29142	0.42682	76.37	0.18265	0.30272	85.83	48.54834	73.08930	96.59	110.23382	185.75727	97.30
ViT-Small (Dosovitskiy et al., 2020)	0.26455	0.37971	81.30	0.15322	0.25030	90.31	42.46420	65.57747	97.26	86.27301	142.69225	98.41
ViT-Base (Dosovitskiy et al., 2020)	0.24610	0.34939	84.17	0.12798	0.20237	93.67	39.81028	61.63069	97.58	76.46207	125.69500	98.77
ViT-Large (Dosovitskiy et al., 2020)	0.22526	0.34130	84.89	0.12143	0.20529	93.48	**38.71456**	59.66490	97.73	73.02069	130.84304	98.66
MAE-Base (He et al., 2022)	0.25405	0.38112	81.16	0.13886	0.21567	92.81	40.36044	60.22436	97.69	78.77065	130.91341	98.66
MAE-Large (He et al., 2022)	0.23208	0.34132	86.09	0.12361	0.20091	93.75	47.79674	172.89531	88.65	81.62728	153.79097	98.28
CAE (Chen et al., 2024)	0.31417	0.42417	76.66	0.17557	0.26151	89.42	40.08760	59.56808	97.74	87.15660	134.33110	98.59
SimMIM (Xie et al., 2022)	0.28651	0.38220	81.05	0.15139	0.21604	92.78	39.90261	**59.24226**	**97.76**	83.63551	138.52274	98.50
Data2Vec-Base (Baevski et al., 2022)	0.25115	0.37178	82.07	0.13043	0.21320	92.97	40.22185	60.93765	97.63	73.58009	124.37146	98.79
Data2Vec-Large (Baevski et al., 2022)	0.22278	0.33778	85.20	0.11900	0.19804	94.01	38.84260	59.28787	97.76	70.51602	121.79372	98.84
MSN-Base (Assran et al., 2022)	0.29429	0.42256	76.84	0.19804	0.31028	85.11	44.40505	64.83239	97.32	88.98394	144.31597	98.37
MSN-Large (Assran et al., 2022)	0.30508	0.42095	77.02	0.20727	0.30502	85.61	50.12944	70.63700	96.82	89.60778	142.97821	98.40
SpectralTF-MAE-Base	0.21281	0.32483	86.31	0.11008	0.19448	94.15	39.10889	61.79632	97.56	**69.18898**	**121.00328**	**98.86**
SpectralTF-MAE-Large	**0.19365**	**0.31588**	**87.06**	**0.10532**	**0.19356**	**94.21**	42.79847	65.12049	97.30	77.02085	135.12647	98.57

Table 3: Comparison on spectral peak analysis. Exact-K, F1, and weak-peak recall are reported in percentage. Reconstruction R^2 values are multiplied by 100.

Method	Exact-K ↑	F1 ↑	Weak Recall ↑	Center MAE ↓	Recon. R^2 ↑	NRMSE ↓
Task-specific Baseline	98.88	77.81	22.22	0.9631	98.54	0.02579
SpectralTF-MAE-Base (Random)	98.92	80.67	19.05	0.9646	**99.37**	**0.01584**
SpectralTF-MAE-Base (Pretrained)	**98.92**	**83.64**	**23.81**	**0.8174**	99.22	0.01767

capacity improves some aspects of the forward generation task but does not uniformly reduce all error measures. For the reverse CXRS-to-XCS direction, SpectralTF-MAE-Large achieves the best results on all three metrics, with an MAE of 0.01760, an RMSE of 0.03193, and an R^2 of 95.45. Overall, the Base and Large variants consistently recover the characteristic target-diagnostic response, while the Large model provides the clearest advantage in the CXRS-to-XCS direction.

• Results on Plasma-Parameter Prediction.

Table 2 reports performance for two electron-density quantities (n_{e1} and n_{e2}), ion temperature (T_i), and electron temperature (T_e).

Taken together, the Base and Large variants of SpectralTF-MAE achieve the best result on 9 of the 12 reported parameter-prediction metrics. SpectralTF-MAE-Large performs best on both electron-density quantities, obtaining the lowest MAE and RMSE and the highest R^2 for both n_{e1} and n_{e2}. In contrast, SpectralTF-MAE-Base remains the strongest model for T_e, achieving the best MAE, RMSE, and R^2 in that target. For T_i, the best individual metrics are obtained by the competing baselines: ViT-Large gives the lowest MAE, while SimMIM gives the lowest RMSE and the highest R^2. These results show that the effect of scaling is target dependent: the Large configuration is particularly beneficial for electron-density prediction, whereas the Base configuration is better for T_e, and neither SpectralTF-MAE scale dominates the T_i metrics.

For the compact ablation tables below, the mean parameter-prediction score is defined as

$$R^2_{\text{Param}} = \frac{R^2_{n_{e1}} + R^2_{n_{e2}} + R^2_{T_i} + R^2_{T_e}}{4}. \tag{5}$$

• Results on Spectral Peak Analysis.

Table 3 compares our task-specific peak-analysis baseline with SpectralTF-MAE using random and pretrained initialization.

Pretrained SpectralTF-MAE improves F1, weak-peak recall, and peak-center localization relative to both the task-specific baseline and the randomly initialized variant. The randomly initialized model,

Table 4: Effect of pretrained initialization on representative downstream tasks.

Model	Init.	XCS→CXRS Peak R^2	CXRS→XCS Peak R^2	Param. Mean R^2
SpectralTF-MAE-Base	Random	95.70	94.91	86.03
SpectralTF-MAE-Base	**Pretrained**	**97.16**	**95.30**	**94.22**

Table 5: Component analysis of the peak-centered pretraining strategy.

Setting	Peak-aware Masking	Peak-enhanced Loss	XCS→CXRS R^2	Param. Mean R^2
Vanilla masked pretraining	✗	✗	96.38	90.25
+ Peak-enhanced loss	✗	✓	96.71	91.69
+ Peak-aware masking	✓	✗	96.98	93.24
Full model	✓	✓	**97.16**	**94.22**

however, obtains slightly better waveform-level reconstruction metrics, including reconstruction R^2 and NRMSE. The effect of pretraining is therefore more evident in peak detection and localization than in uniform improvement across all reconstruction-oriented metrics.

4.4 Ablation Study

• Effect of Pretrained Initialization.

This experiment isolates the effect of self-supervised pretraining by comparing the same SpectralTF-MAE-Base architecture with random initialization and with the pretrained checkpoint. Spectral peak analysis is evaluated separately in Table 3.

Pretrained initialization improves all three representative downstream scores relative to random initialization. The gain is largest for plasma-parameter prediction, with additional improvements in both cross-diagnostic generation directions. These results support the transferability of the learned spectral representation across distinct downstream tasks.

• Component Analysis.

We further evaluate the two peak-centered components used during self-supervised pretraining: peak-aware masking and the peak-enhanced reconstruction loss. To limit experimental cost, XCS-to-CXRS generation and plasma-parameter prediction are used as representative downstream tasks.

Each peak-centered component improves the two representative downstream scores when added individually to vanilla masked pretraining. Combining peak-aware masking with the peak-enhanced loss yields the highest scores on both tasks, making the full configuration the best-performing combination among those evaluated.

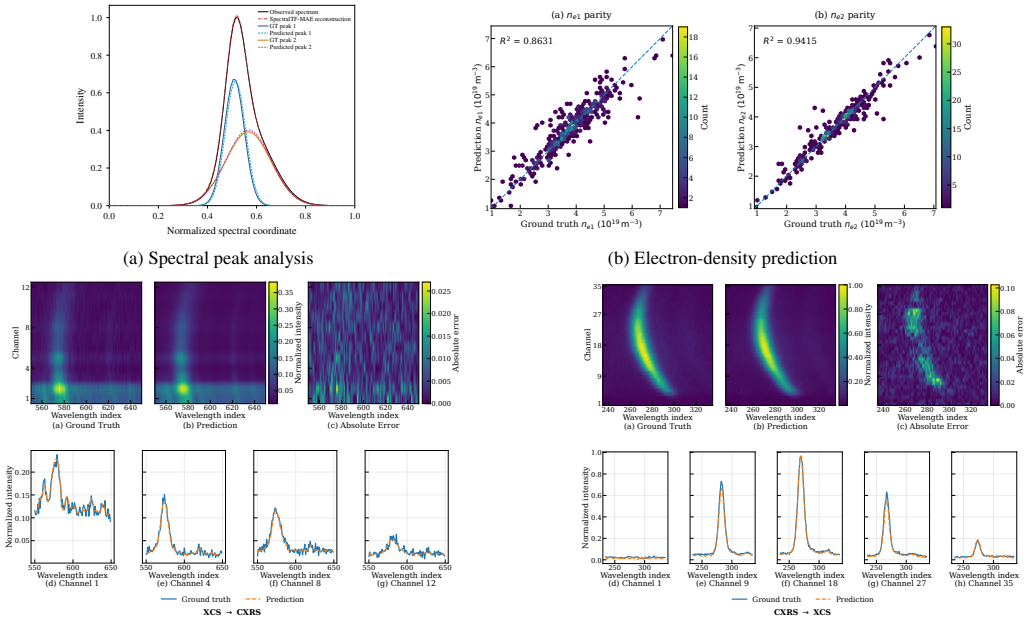
• Data Efficiency.

To limit experimental cost, data efficiency is evaluated on two representative downstream tasks: XCS-to-CXRS spectral generation and plasma-parameter prediction. Identical labeled-data subsets are used for the randomly initialized and pretrained variants, with $D_{10} \subset D_{25} \subset D_{50} \subset D_{100}$.

Table 6: Data efficiency of spectral pretraining on representative downstream tasks.

Labeled Data	XCS→CXRS R^2		Parameter Mean R^2	
	Scratch	Pretrained	Scratch	Pretrained
10%	84.88	95.94	58.72	69.60
25%	94.33	96.90	70.95	81.38
50%	95.56	97.03	79.59	89.23
100%	95.70	97.16	86.03	94.22

Across all labeled-data fractions, the pretrained model outperforms the randomly initialized counterpart on both representative tasks. The gain is especially pronounced in the low-data regime for XCS-to-CXRS generation and remains clear for plasma-parameter prediction across all fractions, indicating improved label efficiency after self-supervised pretraining.



(a) Spectral peak analysis

(b) Electron-density prediction

(c) XCS-to-CXRS generation

(d) CXRS-to-XCS generation

Figure 3: Representative qualitative results for the three downstream task families: (a) spectral peak decomposition, (b) electron-density prediction for n_{e1} and n_{e2}, (c) XCS-to-CXRS spectral generation, and (d) CXRS-to-XCS spectral generation.

4.5 Qualitative Results

To complement the quantitative evaluation, Fig. 3 shows representative results from all three downstream task families. For spectral peak analysis, the reconstructed spectrum follows the observed spectrum and the predicted components remain consistent with the corresponding ground-truth peaks in position and overall line shape. For plasma-parameter prediction, the n_{e1} and n_{e2} parity plots show that most predictions lie close to the identity line over the observed density range. For cross-diagnostic generation, both XCS-to-CXRS and CXRS-to-XCS examples recover the dominant target-diagnostic response structures with localized residual errors. These complementary examples are consistent with the quantitative comparisons and illustrate transfer across spectral-component analysis, parameter inference, and bidirectional spectral generation. Additional qualitative results, including peak-center localization, temperature-profile errors, and the complete density analysis, are provided in Appendix F.

5 Conclusion

In this work, we presented SpectralTF-MAE, a self-supervised spectral foundation model for tokamak plasma diagnostics that learns transferable representations from paired CXRS and XCS measurements. By combining multi-diagnostic masked pretraining with peak-aware masking and a peak-enhanced reconstruction objective, the framework emphasizes diagnostically informative spectral structures while capturing shared correlations across the two modalities. The pretrained representation was evaluated on three complementary downstream tasks: bidirectional cross-diagnostic spectral generation, plasma-parameter prediction for T_i, T_e, and n_e, and spectral peak analysis. Experimental results on EAST data demonstrate strong transfer performance across these tasks, with particularly clear gains in cross-diagnostic generation and electron-density prediction, while pretraining also improves peak detection and localization. The ablation and data-efficiency studies further show that the proposed pretraining strategy provides consistent benefits over random initialization and remains effective when fewer labeled samples are available. Overall, these results support self-supervised spectral pretraining as a promising route toward reusable models for plasma spectroscopy and provide a foundation for extending spectral representation learning to broader diagnostic systems and fusion devices.

LLM USAGE

We used LLMs to improve the grammar, phrasing, and readability of this manuscript. We also used LLMs as a search tool to help us find related works, but we independently reviewed, interpreted, and summarized all identified works.

REFERENCES

Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European conference on computer vision*, pages 456–473. Springer, 2022.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629. IEEE, 2023.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International conference on machine learning*, pages 1298–1312. PMLR, 2022.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

M Bitter, KW Hill, B Stratton, AL Roquemore, D Mastrovito, SG Lee, JG Bak, MK Moon, UW Nam, G Smith, et al. Spatially resolved spectra from a new x-ray imaging crystal spectrometer for measurements of ion and electron temperature profiles. *Review of Scientific Instruments*, 75(10):3660–3665, 2004.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021. URL `https://arxiv.org/abs/2108.07258`.

Baoyue Chai, Yingying Li, Ze Chen, Wei Tao, Yixuan Zhou, Shifeng Mao, Zhengping Luo, Yi Yu, Bo Lyu, and Minyou Ye. Fast estimation of ion temperature from east charge exchange recombination spectroscopy using neural network. *Plasma Science and Technology*, 21(10):105103, 2019.

Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132(1):208–223, 2024.

R Michael Churchill. Ai foundation models for experimental fusion tasks. *Frontiers in Physics*, 12: 1531334, 2025.

RM Churchill, B Tobias, Y Zhu, DIII-D team, et al. Deep convolutional neural networks for multi-scale time-series classification and application to tokamak disruption prediction using raw, high temporal resolution diagnostic data. *Physics of Plasmas*, 27(6), 2020.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022.

Pengming Feng, Kaihan Wang, Jian Guan, Guangjun He, and Shichao Jin. Spectral masked autoencoder for few-shot hyperspectral image classification. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, pages 5041–5044. IEEE, 2023.

Diogo R Ferreira, Pedro J Carvalho, and Horácio Fernandes. Deep learning for plasma tomography and disruption prediction from bolometer data. *IEEE Transactions on Plasma Science*, 48(1): 36–45, 2019.

RJ Fonck, DS Darrow, and KP Jaehnig. Determination of plasma-ion velocity distribution via charge-exchange recombination spectroscopy. *Physical Review A*, 29(6):3288, 1984.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, June 2022.

KW Hill, ML Bitter, SD Scott, A Ince-Cushman, M Reinke, JE Rice, P Beiersdorfer, M-F Gu, SG Lee, Ch Broennimann, et al. A spatially resolving x-ray crystal spectrometer for measurement of ion-temperature and rotation-velocity profiles on the alcator c-mod tokamak. *Review of Scientific Instruments*, 79(10), 2008.

Danfeng Hong, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot. Spectralformer: Rethinking hyperspectral image classification with transformers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–15, 2021.

Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, Antonio Plaza, Paolo Gamba, Jon Atli Benediktsson, and Jocelyn Chanussot. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5227–5244, 2024. doi: 10.1109/TPAMI.2024.3362475.

David Humphreys, Ana Kupresanin, Mark D Boyer, John Canik, CS Chang, Eric C Cyr, Robert Granetz, Jeffrey Hittinger, Egemen Kolemen, Earl Lawrence, et al. Advancing fusion with machine learning research needs workshop report. *Journal of Fusion Energy*, 39(4):123–155, 2020.

A Ince-Cushman, JE Rice, M Bitter, ML Reinke, KW Hill, MF Gu, E Eikenberry, Ch Broennimann, S Scott, Y Podpaly, et al. Spatially resolved high resolution x-ray spectroscopy for magnetically confined fusion plasmas. *Review of Scientific Instruments*, 79(10), 2008.

RC Isler. An overview of charge-exchange spectroscopy as a plasma diagnostic. *Plasma Physics and Controlled Fusion*, 36(2):171–208, 1994.

Julian Kates-Harbeck, Alexey Svyatkovskiy, and William Tang. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature*, 568(7753):526–531, 2019.

SangKyeun Kim, Ricardo Shousha, SeongMoo Yang, Qiming Hu, SangHee Hahn, Azarakhsh Jalalvand, J-K Park, Nikolas Christopher Logan, Andrew Oakleigh Nelson, Y-S Na, et al. Highest fusion performance without harmful edge energy bursts in tokamak. *Nature communications*, 15 (1):3990, 2024.

Adam Kit, AE Järvinen, Sven Wiesen, Yoeri Poels, Lorenzo Frassinetti, and JET Contributors. Developing deep learning algorithms for inferring upstream separatrix density at jet. *Nuclear Materials and Energy*, 34:101347, 2023.

JK Lee, WH Ko, HH Lee, B Kim, YH Lee, GW Shin, J Kim, and JM Lee. Neural network-based analysis of charge exchange spectra in kstar. *Fusion Engineering and Design*, 222:115518, 2026.

SG Lee, JG Bak, M Bitter, K Hill, UW Nam, YJ Kim, and MK Moon. Research and development of x-ray imaging crystal spectrometers for kstar. *Review of scientific instruments*, 75(10):3693–3695, 2004.

Henry W Leung and Jo Bovy. Towards an astronomical foundation model for stars with a transformer-based model. *Monthly Notices of the Royal Astronomical Society*, 527(1):1494–1520, 2024.

Boting Li, Elijah Martin, Michael Churchill, and Ahmed Diallo. Prediction of electric and magnetic fields from spectral data using machine learning algorithms for doppler-free saturation spectroscopy diagnostics. *Journal of Quantitative Spectroscopy and Radiative Transfer*, page 109710, 2025.

Da-Wei Li, Alexandar L Hansen, Chunhua Yuan, Lei Bruschweiler-Li, and Rafael Brüschweiler. Deep picker is a deep neural network for accurate deconvolution of complex two-dimensional nmr spectra. *Nature communications*, 12(1):5229, 2021.

Yingying Li, Yixuan Zhou, Di Jiang, Xiaojun Zhou, Yi Yu, Jia Fu, Bo Lyu, Yuejiang Shi, Guosheng Xu, Minyou Ye, et al. Concept design of ultrafast charge exchange recombination spectroscopy on east tokamak. *Fusion Engineering and Design*, 146:522–525, 2019.

YY Li, J Fu, B Lyu, XW Du, CY Li, Y Zhang, XH Yin, Y Yu, QP Wang, M Von Hellermann, et al. Development of the charge exchange recombination spectroscopy and the beam emission spectroscopy on the east tokamak. *Review of Scientific Instruments*, 85(11), 2014.

Junyan Lin, Feng Gao, Xiaochen Shi, Junyu Dong, and Qian Du. Ss-mae: Spatial–spectral masked autoencoder for multisource remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.

Zichao Lin, Hongming Zhang, Fudi Wang, Cheonho Bae, Jia Fu, Yongcai Shen, Shuyu Dai, Yifei Jin, Dian Lu, Shengyu Fu, et al. Prediction of plasma rotation velocity and ion temperature profiles in east tokamak using artificial neural network models. *Nuclear Fusion*, 64(10):106061, 2024.

Zichao Lin, Hongming Zhang, Fudi Wang, Jia Fu, Yongcai Shen, Shuyu Dai, Hailin Zhao, Kaixuan Ye, Guangle Lin, Yifei Jin, et al. Development of a neural network-based model for electron temperature inference via modelled and experimental argon spectra measured by x-ray crystal spectrometer with an extension to tungsten spectra on east. *Nuclear Fusion*, 65(11):116035, 2025.

Shaozhen Liu, Yu Rong, Deli Zhao, Qiang Liu, Liang Wang, et al. Molspectra: Pre-training 3d molecular representation with multi-modal energy spectra. In *International Conference on Learning Representations*, volume 2025, pages 86733–86751, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

B Lyu, FD Wang, XY Pan, J Chen, J Fu, YY Li, M Bitter, KW Hill, LF Delgado-Aparicio, N Pablant, et al. Upgrades of imaging x-ray crystal spectrometers for high-resolution and high-temperature plasma diagnostics on east. *Review of Scientific Instruments*, 85(11), 2014.

Didier Mazon, George Vayakis, Michael Walsh, Gunsu Yun, S-H Hong, B Peterson, MH Aumeunier, Arnaud Bultel, C Klepper, D Rasmussen, et al. Diagnostics: Chapter 8 of the special issue: on the path to tokamak burning plasma operation. *Nuclear Fusion*, 65(11):113001, 2025.

Arsenty D Melnikov, Yuri P Tsentalovich, and Vadim V Yanshole. Deep learning for the precise peak detection in high-resolution lc–ms data. *Analytical chemistry*, 92(1):588–592, 2020.

Shaheer Mohamed, Maryam Haghighat, Tharindu Fernando, Sridha Sridharan, Clinton Fookes, and Peyman Moghadam. Factoformer: Factorized hyperspectral transformers with self-supervised pretraining. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2023.

Kevin Joseph Montes, Cristina Rea, RS Granetz, Roy Alexander Tinguely, N Eidietis, OM Meneghini, DL Chen, Biao Shen, BJ Xiao, Keith Erickson, et al. Machine learning for disruption warnings on alcator c-mod, diii-d, and east. *Nuclear Fusion*, 59(9):096015, 2019.

Liam Parker, Francois Lanusse, Siavash Golkar, Leopoldo Sarra, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Geraud Krawezik, Michael McCabe, Rudy Morel, et al. Astroclip: a cross-modal foundation model for galaxies. *Monthly Notices of the Royal Astronomical Society*, 531 (4):4990–5011, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

Christina Rea, KJ Montes, KG Erickson, RS Granetz, and RA Tinguely. A real-time machine learning-based disruption predictor in diii-d. *Nuclear Fusion*, 59(9):096016, 2019.

Cristina Rea, RS Granetz, K Montes, Roy Alexander Tinguely, N Eidietis, Jeremy M Hanson, and B Sammuli. Disruption prediction investigations using machine learning tools on diii-d and alcator c-mod. *Plasma Physics and Controlled Fusion*, 60(8):084004, 2018.

ML Reinke, YA Podpaly, M Bitter, IH Hutchinson, JE Rice, L Delgado-Aparicio, C Gao, M Greenwald, K Hill, NT Howard, et al. X-ray imaging crystal spectroscopy for use in plasma transport research. *Review of Scientific Instruments*, 83(11), 2012.

CM Samuell, AG Mclean, CA Johnson, F Glass, and AE Jaervinen. Measuring the electron temperature and identifying plasma detachment using machine learning and spectroscopy. *Review of scientific instruments*, 92(4), 2021.

N Saura, R Guirlet, M Koubiti, O Peyrusse, C Desgranges, S Mazzi, Sadruddin Benkadda, WEST Team, et al. Machine learning-based tungsten spectroscopy analysis in the west tokamak. *Physics of Plasmas*, 32(8), 2025.

Linus Scheibenreif, Michael Mommert, and Damian Borth. Masked vision transformers for hyperspectral image classification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2166–2176. IEEE, 2023.

Jaemin Seo, SangKyeun Kim, Azarakhsh Jalalvand, Rory Conlin, Andrew Rothstein, Joseph Abbate, Keith Erickson, Josiah Wai, Ricardo Shousha, and Egemen Kolemen. Avoiding fusion plasma tearing instability with deep reinforcement learning. *Nature*, 626(8000):746–751, 2024.

Yuejiang Shi, Fudi Wang, Baonian Wan, Manfred Bitter, Sanggon Lee, Jungyo Bak, Kennith Hill, Jia Fu, Yingying Li, Wei Zhang, et al. Imaging x-ray crystal spectrometer on east. *Plasma Physics and Controlled Fusion*, 52(8):085014, 2010.

Yunyue Su, Jiahui Chen, Zao Jiang, Zhenyi Zhong, Liang Wang, Qiang Liu, and Zhaoxiang Zhang. Spectrallm: Uncovering the ability of llms for molecule structure elucidation from multi-spectra. In *International Conference on Learning Representations*, volume 2026, pages 25079–25117, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Di Wang, Meiqi Hu, Yao Jin, Yuchun Miao, Jiaqi Yang, Yichu Xu, Xiaolei Qin, Jiaqi Ma, Lingyu Sun, Chenxing Li, et al. Hypersigma: Hyperspectral intelligence comprehension foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6427–6444, 2025.

Vladimir Weinzettl, Gaurav Shukla, Joydeep Ghosh, Radek Melich, Radomir Panek, Matej Tomes, Martin Imrisek, Diana Naydenkova, Josef Varju, Tiago Pereira, et al. High-resolution spectroscopy diagnostics for measuring impurity ion temperature and velocity on the compass tokamak. *Fusion Engineering and Design*, 96:1006–1011, 2015.

Arthur Weller. Diagnostics for magnetic confinement fusion research. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 623(2):801–805, 2010.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 9643–9653. IEEE, 2022.

Zongyu Yang, Zhenghao Yang, Wenjing Tian, Jiyuan Li, Xiang Sun, Guohui Zheng, Songfen Liu, Niannian Wu, Rongpeng Li, Zhaohe Xu, et al. Fusionmae, a self-supervised pretrained model to optimize and simplify diagnostic and control of fusion plasma. *Communications Physics*, 2026.

Minyou Ye, Yingying Li, Yi Yu, Yuejiang Shi, Bo Lyu, Jia Fu, Xuewei Du, Xianghui Yin, Yi Zhang, Qiuping Wang, et al. Toroidal charge exchange recombination spectroscopy on east. *Fusion Engineering and Design*, 96:1017–1020, 2015.

XL Zhang, ZYH He, ZF Cheng, Wei Yan, YB Dong, Yi Liu, Wei Deng, BZ Fu, ZB Shi, YP Zhang, et al. X-ray imaging crystal spectrometer (xics) diagnostic on the hl-3 tokamak. *Fusion Engineering and Design*, 208:114674, 2024.

Yi Zhang, Yingying Li, Jia Fu, Xianghui Yin, Bo Lyu, Yuejiang Shi, Xuewei Du, Qiuping Wang, Yi Yu, Minyou Ye, et al. Calibration of the toroidal charge exchange recombination spectroscopy system on east. *Fusion Engineering and Design*, 96:840–843, 2015.

A Detailed Related Works

This section reviews prior work in three areas most closely related to our study: foundation models, artificial intelligence for nuclear fusion, and spectroscopy-based plasma diagnostics. Broader perspectives on machine learning for fusion and on diagnostic techniques for magnetic-confinement plasmas are available in existing reviews and community reports (Humphreys et al., 2020; Weller, 2010; Mazon et al., 2025).

A.1 Large Foundation Models

Foundation models are typically pretrained on broad datasets to learn transferable representations that can be adapted to diverse downstream tasks (Bommasani et al., 2021). The Transformer provides the self-attention architecture underlying many modern pretrained models (Vaswani et al., 2017), while BERT demonstrated the effectiveness of masked self-supervised pretraining for learning transferable language representations (Devlin et al., 2019). This pretraining-and-adaptation paradigm was subsequently extended to vision through Vision Transformers (Dosovitskiy et al., 2020) and a range of masked or predictive self-supervised approaches, including MAE (He et al., 2022), BEiT (Bao et al., 2021), SimMIM (Xie et al., 2022), Context Autoencoder (CAE) (Chen et al., 2024), Masked Siamese Networks (MSN) (Assran et al., 2022), data2vec (Baevski et al., 2022), masked spatiotemporal autoencoding (Feichtenhofer et al., 2022), and joint-embedding predictive learning (Assran et al., 2023).

These ideas have also been extended to structured spectral and remote-sensing data. SpectralFormer introduced Transformer-based modeling for hyperspectral imagery (Hong et al., 2021), while SatMAE extended masked pretraining to temporal and multispectral satellite observations (Cong et al., 2022). Subsequent studies investigated masked vision Transformers for hyperspectral classification (Scheibenreif et al., 2023), spatial–spectral masked autoencoding (Lin et al., 2023), factorized hyperspectral Transformers with self-supervised pretraining (Mohamed et al., 2023), and spectral masked autoencoding for few-shot hyperspectral learning (Feng et al., 2023). More recently, SpectralGPT (Hong et al., 2024) and HyperSIGMA (Wang et al., 2025) developed foundation-model approaches for spectral remote sensing. Beyond remote sensing, AstroCLIP explored cross-modal foundation modeling for astronomical observations (Parker et al., 2024), while Leung and Bovy developed a self-supervised Transformer for stellar spectra and physical-parameter inference (Leung and Bovy, 2024). Collectively, these studies demonstrate the potential of reusable representations for high-dimensional spectral measurements; however, foundation modeling for plasma spectroscopy remains comparatively underexplored.

A.2 AI for Nuclear Fusion

Machine learning has been applied extensively to magnetic-confinement fusion for disruption prediction, diagnostic inference, and plasma control (Humphreys et al., 2020). Representative disruption studies include machine-learning approaches on DIII-D and Alcator C-Mod (Rea et al., 2018), multi-machine warning systems spanning Alcator C-Mod, DIII-D, and EAST (Montes et al., 2019), real-time prediction on DIII-D (Rea et al., 2019), and deep convolutional models using high-temporal-resolution diagnostic signals (Churchill et al., 2020). Deep learning has also been applied to plasma tomography and disruption prediction from bolometer measurements (Ferreira et al., 2019), while cross-device approaches have demonstrated the potential for generalization in disruptive-instability prediction (Kates-Harbeck et al., 2019).

Beyond prediction, learning-based methods have been extended to plasma control and operation optimization. Deep reinforcement learning has been demonstrated for magnetic control of tokamak plasmas (Degrave et al., 2022) and for avoidance of tearing instabilities (Seo et al., 2024); recent work has also integrated machine learning into high-performance tokamak operation strategies (Kim et al., 2024). For diagnostic inference, deep-learning models have been developed to estimate quantities such as upstream separatrix density at JET (Kit et al., 2023). These studies illustrate the breadth of AI applications in fusion experiments, although most models remain tailored to individual tasks or diagnostic channels.

More recently, reusable pretrained models have attracted growing attention in fusion research. Churchill (2025) discussed the potential of foundation models for exploiting unlabeled fusion data,

while FusionMAE (Yang et al., 2026) introduced masked self-supervised pretraining over multiple HL-3 diagnostic and operational signals and transferred the learned representation to several downstream tasks. Recent fusion foundation-model studies have primarily emphasized heterogeneous diagnostic time series and global plasma-state representations. In contrast, our work focuses on high-dimensional XCS and CXRS measurements and investigates whether a shared spectral representation can transfer across plasma-parameter inference and cross-diagnostic spectral generation.

A.3 Spectroscopy-Based Plasma Diagnostics

Spectroscopic diagnostics provide key measurements of electron and ion temperatures, density, impurity states, and plasma rotation in magnetic-confinement experiments (Weller, 2010; Mazon et al., 2025). Charge-exchange recombination spectroscopy (CXRS) infers ion temperature and rotation from Doppler broadening and spectral-line shifts (Fonck et al., 1984; Isler, 1994). CXRS systems, together with calibration and high-speed acquisition schemes, have been developed on EAST (Li et al., 2014; Ye et al., 2015; Zhang et al., 2015; Li et al., 2019), and related high-resolution spectroscopic diagnostics have been implemented on devices such as COMPASS (Weinzettl et al., 2015).

X-ray crystal spectroscopy (XCS/XICS) provides complementary high-resolution measurements of impurity X-ray emission. Spatially resolved crystal spectrometers have been developed to measure ion and electron temperatures and plasma rotation (Bitter et al., 2004; Ince-Cushman et al., 2008; Hill et al., 2008). Related systems have been implemented on KSTAR (Lee et al., 2004), EAST (Shi et al., 2010; Lyu et al., 2014), and HL-3 (Zhang et al., 2024), and X-ray imaging crystal spectroscopy has also been applied to plasma-transport studies (Reinke et al., 2012).

Conventional spectroscopic analysis typically relies on spectral fitting, line-ratio analysis, atomic models, or inversion procedures. Machine learning has therefore been investigated as a data-driven alternative for rapid spectral inference. Chai et al. (2019) used neural networks to estimate ion temperature from EAST CXRS measurements, while Samuell et al. (2021) inferred electron temperature and plasma detachment from EUV/VUV spectra. Lin et al. developed neural-network models for predicting ion temperature and plasma rotation from EAST XCS measurements (Lin et al., 2024) and subsequently investigated electron-temperature inference from XCS spectra (Lin et al., 2025). Neural-network-based analysis of charge-exchange spectra has also been demonstrated on KSTAR (Lee et al., 2026). Outside fusion, deep neural networks have been applied to automatic peak picking and spectral deconvolution in NMR and high-resolution LC–MS data (Li et al., 2021; Melnikov et al., 2020), further illustrating the potential of learning-based spectral-structure analysis.

Despite these advances, existing AI-based plasma-spectroscopy methods are predominantly task-specific and fully supervised, with separate models trained for individual diagnostic quantities. Consequently, spectral structures shared across diagnostics and downstream tasks are not explicitly exploited. In contrast, this work learns a reusable representation from paired XCS and CXRS measurements through masked self-supervised pretraining and transfers the pretrained model to multiple downstream tasks, including electron-density prediction, ion- and electron-temperature prediction, and cross-diagnostic spectral generation.

B Pretraining Details

The pretraining stage learns a shared spectral representation from paired CXRS and XCS measurements without plasma-parameter labels. As shown in Fig. 2, both spectra are normalized and converted into modality-specific tokens. Peak-aware masking defines the reconstruction targets, while the remaining visible tokens are jointly encoded by a shared Transformer. Modality-specific decoders then reconstruct the two diagnostics from the shared latent representation, and a peak-enhanced objective places additional emphasis on the characteristic spectral regions.

B.1 Multi-diagnostic Spectral Tokenization

Let $\mathbf{X}^c \in \mathbb{R}^{H_c \times W_c}$ and $\mathbf{X}^x \in \mathbb{R}^{H_x \times W_x}$ denote paired CXRS and XCS measurements, respectively, with the two dimensions corresponding to the sampling-position and wavelength axes. Each

spectrum is independently normalized using sample-wise min–max normalization:

$$\widetilde{\mathbf{X}}^m = \frac{\mathbf{X}^m - \min_{h,w} X^m_{h,w}}{\max_{h,w} X^m_{h,w} - \min_{h,w} X^m_{h,w} + \epsilon}, \qquad m \in \{c, x\}. \tag{6}$$

Unlike natural images, the spectral tensors are anisotropic: their two axes represent distinct physical dimensions and have substantially different sampling resolutions. We therefore use modality-specific rectangular patches elongated along the wavelength axis, following the general patch-based Transformer formulation used in ViT and spectral Transformer models (Dosovitskiy et al., 2020; Hong et al., 2021) while adapting the patch geometry to the sampling characteristics of CXRS and XCS.

For each modality, a modality-specific input adapter projects spectral patches into a common latent space, and a two-dimensional positional embedding encodes their sampling-position and wavelength locations. The i-th token is written as

$$\mathbf{t}^m_i = \mathcal{E}_m\left(\widetilde{\mathbf{X}}^m_i\right) + \mathbf{p}^m_i, \tag{7}$$

where $\mathcal{E}_m$ denotes the modality-specific patch projection and $\mathbf{p}^m_i$ is the corresponding positional embedding. Although CXRS and XCS use different patch geometries, both are mapped into the same latent space for joint modeling.

B.2 Peak-aware Spectral Masking

Uniform random masking assigns the same selection probability to all spectral patches, whereas the dominant diagnostic responses in the considered plasma spectra are concentrated around characteristic emission peaks. We therefore introduce peak-aware masking, which preferentially includes high-response patches among the reconstruction targets while retaining part of the peak-related information as visible context.

For each patch, its maximum normalized intensity is used as the peak score:

$$s^m_i = \max_{(h,w)\in\mathcal{P}^m_i} \widetilde{X}^m_{h,w}, \tag{8}$$

where $\mathcal{P}^m_i$ denotes the i-th patch of modality m. For each paired sample, the CXRS and XCS scores are concatenated and used to compute a sample-specific quantile threshold:

$$\tau = Q_q\left([\mathbf{s}^c; \mathbf{s}^x]\right). \tag{9}$$

Patches with scores above τ are regarded as peak candidates.

During masking, peak candidates are preferentially selected as reconstruction targets, while the fraction of masked peak candidates is constrained so that part of the characteristic spectral information remains visible to the encoder. The remaining masking positions are sampled from non-peak tokens. This design increases the reconstruction emphasis on informative spectral structures without removing all peak-related context.

B.3 Joint Spectral Representation and Reconstruction

After masking, the visible CXRS and XCS tokens are concatenated and processed by a shared Transformer encoder. A learnable global token is included to aggregate information across the paired diagnostics:

$$\mathbf{Z} = \mathcal{F}_{\mathrm{enc}}\left([\mathcal{V}^c; \mathcal{V}^x; \mathbf{g}]\right), \tag{10}$$

where $\mathcal{V}^c$ and $\mathcal{V}^x$ denote the visible token sets and $\mathcal{F}_{\mathrm{enc}}$ is the shared encoder. Because tokens from both diagnostics interact within the same self-attention layers, the encoder can jointly model intra-diagnostic spectral dependencies and cross-diagnostic correlations.

Separate output adapters are used to reconstruct CXRS and XCS:

$$\widehat{\mathbf{X}}^c = \mathcal{D}_c(\mathbf{Z}), \qquad \widehat{\mathbf{X}}^x = \mathcal{D}_x(\mathbf{Z}). \tag{11}$$

For each diagnostic, mask tokens and positional information are restored at their original patch locations. Modality-specific reconstruction queries then attend to the shared encoder representation through cross-attention:

$$\mathbf{Q}_{\mathrm{att}}^{m} = \mathrm{CrossAttn}\left(\mathbf{Q}^{m}, \mathbf{Z}\right), \qquad m \in \{c, x\}. \tag{12}$$

The resulting query features are further refined by the decoder and projected back to spectral patches. The two diagnostics therefore share a common latent representation while retaining modality-specific reconstruction pathways.

B.4 Peak-enhanced Reconstruction Objective

For each diagnostic, the basic reconstruction objective is the mean squared error computed over the masked spectral region. Let $\mathcal{M}^m$ denote the two-dimensional region corresponding to the masked patches:

$$\mathcal{L}_{\mathrm{rec}}^{m} = \frac{1}{|\mathcal{M}^m|} \sum_{(h,w)\in\mathcal{M}^m} \left(\widehat{X}_{h,w}^{m} - \widetilde{X}_{h,w}^{m}\right)^2. \tag{13}$$

Because the characteristic emission response occupies only a subset of the wavelength axis, we additionally define a characteristic wavelength-index interval I_m for each diagnostic. The corresponding two-dimensional peak region is

$$\Omega_m = \{(h, w) \mid w \in I_m\}. \tag{14}$$

The peak-enhanced loss is computed over the intersection of the characteristic region and the masked reconstruction region:

$$\mathcal{L}_{\mathrm{peak}}^{m} = \frac{1}{|\Omega_m \cap \mathcal{M}^m|} \sum_{(h,w)\in\Omega_m\cap\mathcal{M}^m} \left(\widehat{X}_{h,w}^{m} - \widetilde{X}_{h,w}^{m}\right)^2. \tag{15}$$

If no masked element falls within the predefined characteristic region for a given sample and modality, the corresponding peak-loss term is set to zero.

The overall pretraining objective combines the masked reconstruction loss with the peak-enhanced term:

$$\mathcal{L}_{\mathrm{pre}} = \sum_{m\in\{c,x\}} \left(\mathcal{L}_{\mathrm{rec}}^{m} + \lambda_{\mathrm{peak}}\mathcal{L}_{\mathrm{peak}}^{m}\right). \tag{16}$$

Peak-aware masking determines which spectral structures are preferentially selected for reconstruction, whereas the peak-enhanced loss assigns additional optimization weight to predefined characteristic regions. Together, these components constitute the peak-centered self-supervised objective used for multi-diagnostic spectral representation learning.

C Downstream Task Details

After self-supervised pretraining, the learned representation is adapted to downstream diagnostic tasks through task-specific fine-tuning. We consider three task families: plasma-parameter prediction from XCS measurements, including temperature-profile and electron-density inference; cross-diagnostic spectral generation between XCS and CXRS; and spectral peak analysis. Together, these tasks evaluate the pretrained representation in terms of physical-parameter inference, cross-diagnostic correspondence, and spectral-component analysis.

• Plasma-Parameter Prediction.

Spectroscopic diagnostics provide an important route for obtaining plasma kinetic parameters. In particular, ion temperature (T_i) and electron temperature (T_e) are fundamental for characterizing plasma confinement, turbulent transport, and stability, and established spectroscopic analysis can recover temperature information from spectral-line shapes and intensity responses (Bitter et al., 2004; Ince-Cushman et al., 2008; Hill et al., 2008). However, conventional spectral fitting and inversion procedures can be computationally expensive for large-scale or rapid analysis. Recent neural-network approaches have therefore learned direct mappings from EAST spectra to T_i and

T_e profiles (Lin et al., 2024; 2025), demonstrating the feasibility of fast data-driven temperature inference. These approaches are typically developed and trained separately for individual targets, however, and do not provide a reusable pretrained spectral representation.

Electron-density prediction constitutes a complementary and more exploratory task. Unlike temperature quantities with established spectroscopic signatures, electron density is not determined from raw XCS detector counts through a comparably direct mapping; its spectral dependence is entangled with line intensities and other plasma and atomic conditions. This makes conventional direct inversion from XCS measurements to n_e difficult. We therefore investigate whether a data-driven model can learn these latent correlations directly from experimental measurements. Accordingly, our plasma-parameter benchmark contains two complementary settings: established temperature-profile prediction for T_i and T_e, and exploratory electron-density prediction for n_e.

For temperature prediction, both T_i and T_e are formulated as radial-profile regression targets, with each profile represented at 32 radial positions. For electron density, two measurements are available: n_{e1} corresponds to the solid-source measurement, whereas n_{e2} is obtained from POINT. Their consistency is used for data screening, and samples showing substantial disagreement between the two measurements are excluded as invalid. After screening, a total of 5,330 valid samples are retained for the downstream plasma-parameter prediction task.

For each valid sample, the complete XCS spectrum is processed by the pretrained spectral encoder, while the reconstruction decoders used during pretraining are removed. The global representation and local XCS tokens are pooled and concatenated into a fused feature vector $\mathbf{f}_x$, which is passed to independent task-specific regression heads:

$$\widehat{\mathbf{y}}_k = \mathcal{H}_k(\mathbf{f}_x), \qquad k \in \{n_{e1}, n_{e2}, T_i, T_e\}, \tag{17}$$

where $\mathcal{H}_k$ denotes the task-specific regression head for either electron-density measurement, ion-temperature profiles, or electron-temperature profiles.

This task tests whether the self-supervised spectral representation encodes plasma-state information that is useful for quantitative parameter inference. Under the same downstream training setting, pretrained initialization improves the mean R^2 for plasma-parameter prediction from 86.03 to 94.22 (values multiplied by 100), corresponding to an absolute improvement of 8.19 percentage points over random initialization.

• Cross-Diagnostic Spectral Generation.

We next evaluate whether the pretrained representation captures transferable correlations between spectroscopic diagnostics through cross-diagnostic spectral generation. Given a complete source spectrum, all source tokens are passed to the pretrained encoder without masking, and the target-diagnostic decoder reconstructs the complete target spectrum:

$$\widehat{\mathbf{X}}^t = \mathcal{D}_t\left(\mathcal{F}_{\mathrm{enc}}(\mathbf{X}^s)\right), \qquad s \neq t, \tag{18}$$

where $s, t \in \{x, c\}$ denote the source and target diagnostics, respectively.

Although the decoder produces the complete target spectral tensor, optimization is restricted to the predefined characteristic peak region of the target diagnostic. Let Ω_t denote the characteristic region defined during pretraining. The downstream spectral-generation loss is

$$\mathcal{L}_{\mathrm{gen}}^{s\to t} = \frac{1}{|\Omega_t|} \sum_{(h,w)\in\Omega_t} \left(\widehat{X}_{h,w}^t - \widetilde{X}_{h,w}^t\right)^2. \tag{19}$$

This formulation is consistent with the peak-centered pretraining strategy and concentrates downstream supervision on the spectral structures of primary diagnostic interest.

Both XCS-to-CXRS and CXRS-to-XCS directions are fine-tuned independently, providing complementary tests of whether the pretrained representation captures cross-diagnostic correspondence beyond modality-specific self-reconstruction.

• Spectral Peak Analysis.

We further evaluate representation transfer on a spectral peak-analysis task, where the downstream model decomposes an input spectrum into multiple components and predicts the center, amplitude,

full width at half maximum (FWHM), and existence probability of each peak. To isolate the contribution of pretraining, we retain the original task-specific peak-analysis pipeline and optimization objectives, replacing only its Transformer encoder with the pretrained SpectralTF encoder and fine-tuning it jointly with the downstream model. This controlled transfer setting tests whether representations learned from CXRS–XCS pretraining generalize to a structurally different spectral-analysis task.

D Dataset and Evaluation Details

D.1 Datasets and Preprocessing

The spectral dataset used in this study was constructed from CXRS and XCS measurements acquired during EAST experiments from 2021 to 2023. Only measurements for which valid CXRS and XCS data were simultaneously available were selected and paired according to the corresponding discharge and acquisition time. After data synchronization and quality control, a total of 15303 valid paired samples were retained for subsequent model development.

These paired measurements cover a broad range of EAST plasma conditions and provide the experimental basis for spectral pretraining and downstream evaluation. The complete CXRS input is represented as a 1024×12 spectral matrix. Because only 8 CXRS channels were available for part of the dataset owing to experimental-configuration changes, these incomplete measurements are reconstructed using a separately trained channel-completion model and then combined with the complete-channel samples. XCS measurements are spatially integrated to a compact 35×487 representation.

After quality control, data augmentation is applied only to complete-channel samples in the training set. Paired CXRS and XCS spectra are shifted jointly along the wavelength direction to preserve cross-diagnostic correspondence, and Poisson noise is added to emulate measurement fluctuations. Validation and test samples, as well as channel-completed samples, are not augmented, preventing augmentation-induced information leakage across data splits. The resulting paired CXRS–XCS dataset is used for self-supervised spectral pretraining and downstream evaluation.

D.2 Evaluation Metrics

We evaluate the regression and spectral-generation tasks using three standard metrics: mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination (R^2). These metrics characterize average prediction error, sensitivity to larger deviations, and agreement between predicted and ground-truth values, respectively.

Given N target values, with y_i and $\widehat{y}_i$ denoting the ground truth and prediction, respectively, the MAE is defined as

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| y_i - \widehat{y}_i \right|. \tag{20}$$

The RMSE is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(y_i - \widehat{y}_i \right)^2}. \tag{21}$$

Because the error is squared before averaging, RMSE places greater weight on large deviations than MAE.

The coefficient of determination is calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^{N} \left(y_i - \widehat{y}_i \right)^2}{\sum_{i=1}^{N} \left(y_i - \overline{y} \right)^2}, \tag{22}$$

where $\overline{y}$ is the mean of the ground-truth values. Higher R^2 and lower MAE/RMSE indicate better predictive agreement. For readability, all R^2 values reported in the following experiments are multiplied by 100.

Table 7: Effect of XCS spectral patch size on representative downstream tasks. All R^2 values are multiplied by 100.

XCS Patch	XCS→CXRS $R^2 \uparrow$	CXRS→XCS $R^2 \uparrow$	Parameter Mean $R^2 \uparrow$
5×10	**97.20**	94.92	92.88
7×8	97.11	95.05	92.53
7×10	96.55	94.82	93.24
5×14	97.16	**95.30**	**94.22**
7×14	96.89	95.15	92.99

Table 8: Effect of CXRS patch size on downstream transfer. The XCS patch size is fixed to the configuration selected from the XCS patch study.

CXRS Patch	XCS→CXRS $R^2 \uparrow$	CXRS→XCS $R^2 \uparrow$	Parameter Mean $R^2 \uparrow$
6×8	96.80	95.21	93.27
3×16	**97.16**	**95.30**	**94.22**

For plasma-parameter prediction, model outputs are transformed from the standardized target space back to their original physical scales before evaluation. The T_i and T_e profiles each contain 32 radial positions; for each profile quantity, predictions from all test samples and radial positions are aggregated before the metrics are computed.

For cross-diagnostic spectral generation, metrics are computed in the normalized spectral space and restricted to the predefined characteristic region of the target diagnostic. Specifically, the CXRS characteristic region is $\Omega_c = [550, 650)$ and the XCS characteristic region is $\Omega_x = [235, 335)$. Predicted and ground-truth values from all test samples, sampling positions, and wavelengths within the corresponding region are aggregated to compute MAE, RMSE, and R^2. This protocol focuses the quantitative assessment on the spectral structures of primary diagnostic interest rather than the non-characteristic background.

For spectral peak analysis, we report Precision, Recall, F1 score, Exact-K accuracy, center MAE, amplitude normalized MAE (NMAE), FWHM relative error, and the reconstruction R^2 between the reconstructed spectrum and the ground-truth spectrum. Among these metrics, F1 and Exact-K quantify peak-detection performance, whereas the center, amplitude, and FWHM errors measure the accuracy of the estimated spectral-peak parameters.

E Additional Ablation Studies

• Effect of Spectral Patch Size.

We compare representative rectangular patch configurations while holding the remaining pretraining settings fixed. For each configuration, the visible-token budget is adjusted to maintain an approximately 80% global masking ratio. XCS patch sizes are first compared with the CXRS patch fixed; the CXRS patch size is then evaluated using the selected XCS configuration.

For XCS, the 5×10 patch gives a marginally higher XCS-to-CXRS R^2 than 5×14 (97.20 versus 97.16), whereas 5×14 achieves the best CXRS-to-XCS and parameter-prediction scores. We therefore select 5×14 as the default XCS configuration because it provides the best overall transfer performance across the evaluated tasks rather than the best score on every individual task. With the XCS patch fixed at 5×14, the 3×16 CXRS patch outperforms 6×8 on all three reported downstream metrics.

• Sensitivity to Global Masking Ratio.

Table 9: Sensitivity to the global masking ratio during spectral pretraining.

Mask Ratio	XCS→CXRS $R^2 \uparrow$	Parameter Mean $R^2 \uparrow$
60%	97.04	91.72
70%	97.18	92.59
80%	97.16	94.22
90%	97.12	92.28

Table 10: Effect of data augmentation during self-supervised pretraining on cross-diagnostic spectral generation. MAE and RMSE are calculated in the normalized target characteristic region, and R^2 values are multiplied by 100.

Pretraining Aug.	XCS → CXRS			CXRS → XCS		
	MAE ↓	RMSE ↓	R^2 ↑	MAE ↓	RMSE ↓	R^2 ↑
No	0.00809	0.01615	95.86	0.02080	0.03417	94.86
Yes	**0.00632**	**0.01339**	**97.16**	**0.01796**	**0.03269**	**95.30**

A 70% masking ratio gives a marginally higher XCS-to-CXRS R^2 than 80% (97.18 versus 97.16), whereas 80% yields the highest parameter-prediction score. We therefore adopt 80% as the default because it provides the best overall trade-off across the evaluated downstream tasks rather than being uniformly optimal for every task.

• **Effect of Pretraining Data Augmentation.**

Data augmentation is applied only during self-supervised pretraining; the fine-tuning, validation, and test data are unchanged. Both cross-diagnostic generation directions are used to assess its effect on downstream transfer.

Pretraining augmentation improves MAE, RMSE, and R^2 in both cross-diagnostic generation directions. The augmentation strategy is therefore retained in the default pretraining configuration.

F Additional Qualitative Results

To complement the quantitative evaluation, we visualize representative results for cross-diagnostic spectral generation, spectral peak analysis, and plasma-parameter prediction.

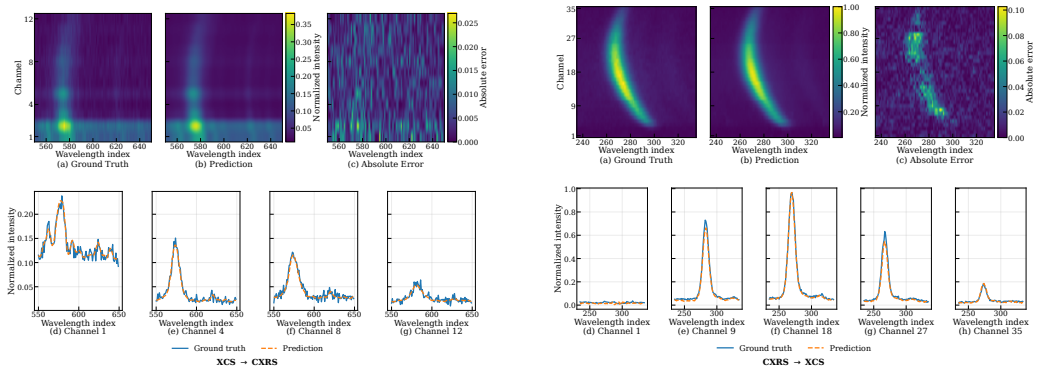
• **Cross-Diagnostic Spectral Generation.**

Fig. 4 presents representative generation results for both cross-diagnostic directions. Fig. 4 (a) shows XCS-to-CXRS generation, while Fig. 4 (b) shows the reverse CXRS-to-XCS generation. In both subfigures, the visualization compares the ground-truth target spectrum, the generated spectrum, and the corresponding absolute-error map, together with one-dimensional profiles at representative sampling positions.

As shown in Fig. 4 (a), the generated CXRS spectrum reproduces the dominant structures of the ground-truth spectrum, including the locations and overall shapes of the characteristic emission responses across sampling positions. The representative profiles also show close agreement in peak intensity and local line shape, while the remaining discrepancies are mainly localized around several strong responses.

Fig. 4 (b) presents the reverse CXRS-to-XCS direction. The generated XCS spectrum recovers the characteristic curved emission structure across the sampling dimension, and the representative profiles broadly agree in peak position, intensity, and local shape. Taken together, the two subfigures are consistent with the quantitative results in Table 1 and support transferable correspondence between the two diagnostics.

• **Spectral Peak Analysis.**



(a) XCS-to-CXRS generation.

(b) CXRS-to-XCS generation.

Figure 4: Visualization of bidirectional spectral generation: (a) XCS-to-CXRS generation and (b) CXRS-to-XCS generation. Each subfigure compares the ground-truth target spectrum, the generated spectrum, the absolute-error map, and representative spectral profiles.

To qualitatively examine the spectral decomposition capability, Fig. 5 presents a representative example in which the reconstructed spectrum is decomposed into individual peak components. The SpectralTF-MAE reconstruction closely follows the observed spectrum, while the predicted components are consistent with the corresponding ground-truth peaks in peak position and overall line shape.

To further assess spectral fidelity quantitatively, Fig. 6 compares the peak-center localization MAE of the baseline method and SpectralTF-MAE on paired samples with the correct peak count in both methods. Most points lie below the dashed diagonal, indicating that SpectralTF-MAE achieves lower peak-center localization error for a substantial fraction of the evaluated samples. As summarized in the figure, the median MAE decreases from 0.652 to 0.575 and the mean MAE decreases from 0.955 to 0.803, demonstrating improved spectral peak localization accuracy.

• **Plasma-Parameter Prediction.**

We further visualize plasma-parameter prediction from two complementary perspectives. For the temperature profiles, Fig. 7 (a) and Fig. 7 (b) show the relative-error distributions of T_e and T_i, respectively, as functions of the normalized radius ρ. Both subfigures also report the radial coverage within $\pm 10\%$ error and the corresponding marginal error distribution. To avoid unstable ratios when the ground-truth value approaches zero, relative error is displayed only where the magnitude of the ground-truth value is at least 3% of the maximum magnitude of the corresponding profile.

As shown in Fig. 7 (a) and (b), the displayed relative errors are concentrated near zero over most radial positions. The within-$\pm 10\%$ coverage is 86.7% for T_e and 76.1% for T_i, while a broader spread appears toward the outer radial region, especially for T_i. These distributions are consistent with the overall profile-level prediction accuracy reported quantitatively.

For the electron-density quantities, Fig. 8 (a) and Fig. 8 (b) show parity plots for n_{e1} and n_{e2}, respectively. Most predictions lie close to the identity line over the observed density range. The n_{e2} predictions in Fig. 8 (b) are more tightly concentrated around the identity line, whereas n_{e1} in Fig. 8 (a) exhibits greater dispersion for a subset of samples. Together with the temperature-profile results, these visualizations provide qualitative evidence that the pretrained XCS representation transfers to both scalar and profile-based plasma-parameter prediction.

G Limitation Analysis

Although SpectralTF-MAE demonstrates effective transferability across multiple plasma diagnostic tasks, several limitations remain. First, the current pretraining framework relies on paired CXRS and XCS measurements acquired under corresponding plasma states. While such paired observations facilitate the learning of cross-diagnostic correlations, their availability may constrain the

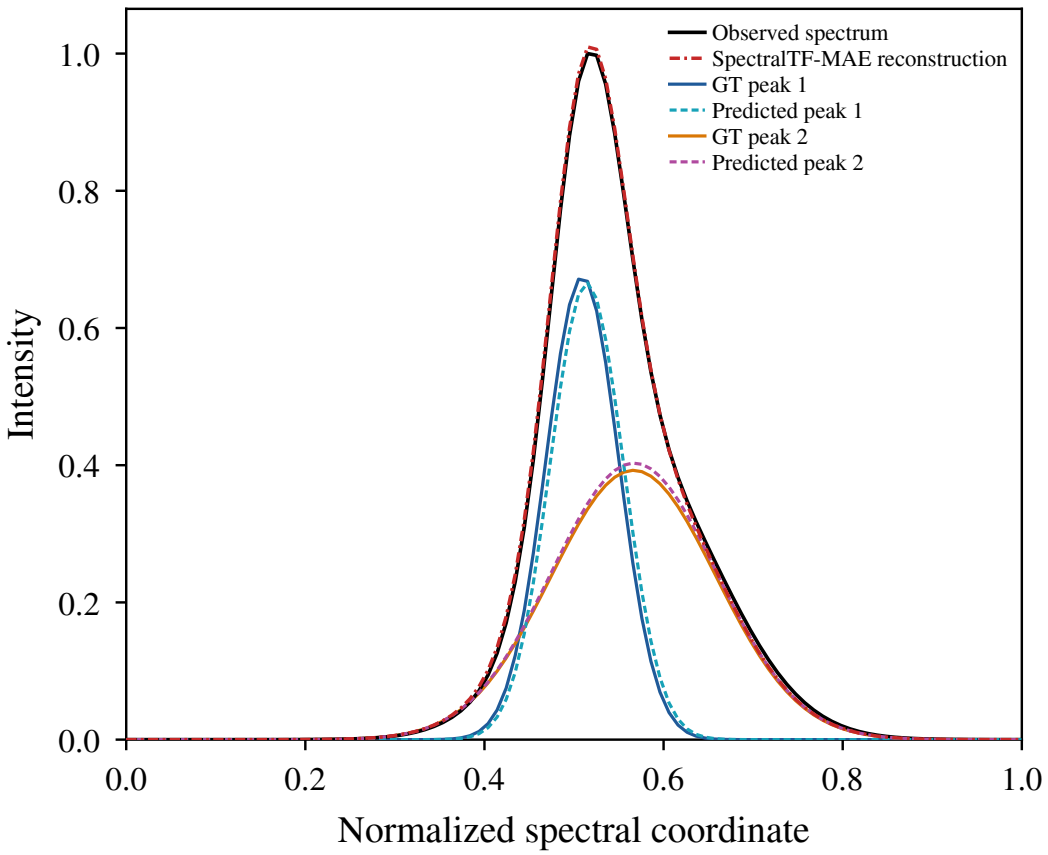


Figure 5: Visualization of spectral peak decomposition for a representative sample. The observed spectrum and SpectralTF-MAE reconstruction are shown together with the ground-truth peak components and their corresponding predicted components.

amount of usable pretraining data. Extending the framework to partially paired or unpaired diagnostic measurements would further improve its applicability. Second, the present experiments are conducted using data from EAST and the corresponding diagnostic configurations. The generalization capability of the learned spectral representations across different tokamaks, diagnostic systems, and operating conditions therefore requires further validation. Finally, the peak-enhanced objective and downstream spectral-generation supervision employ predefined characteristic spectral regions. Although these regions effectively capture the dominant plasma-induced spectral responses considered in this study, they may require adjustment when the spectral calibration, impurity species, or diagnostic configuration changes. Developing adaptive mechanisms for identifying diagnostically informative spectral regions represents an important direction for future work. In addition, the Base/Large comparisons show that increasing model capacity does not uniformly improve every target or metric, suggesting that a more systematic study of architecture scaling and task-specific capacity requirements is warranted.

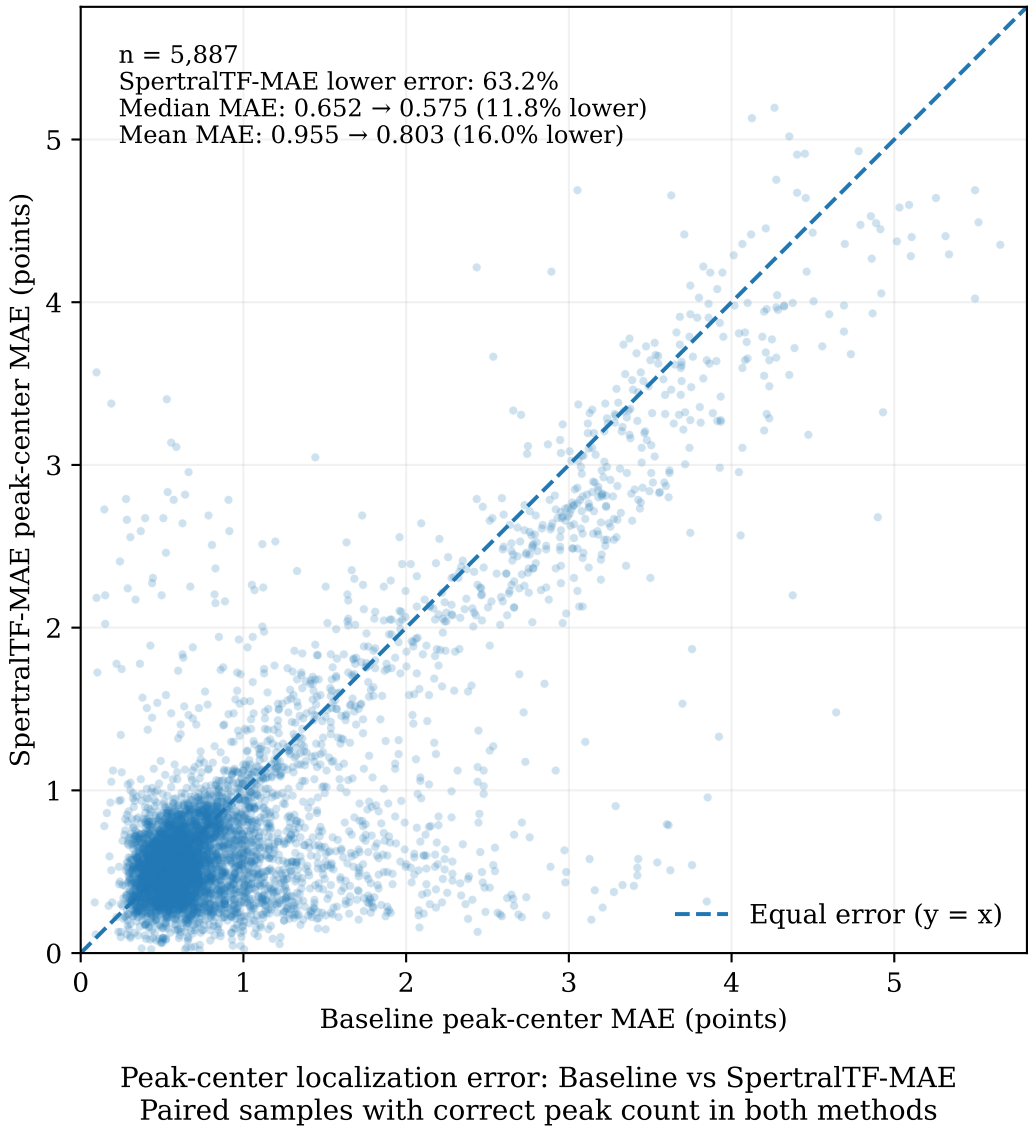
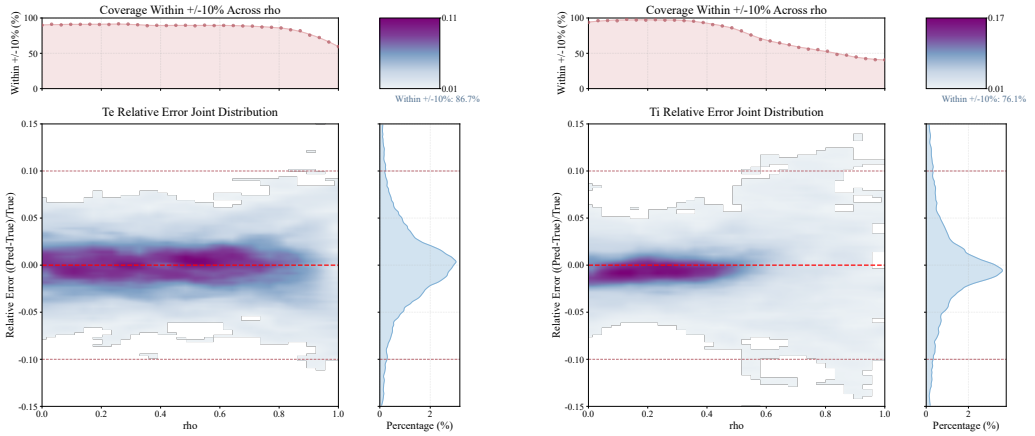


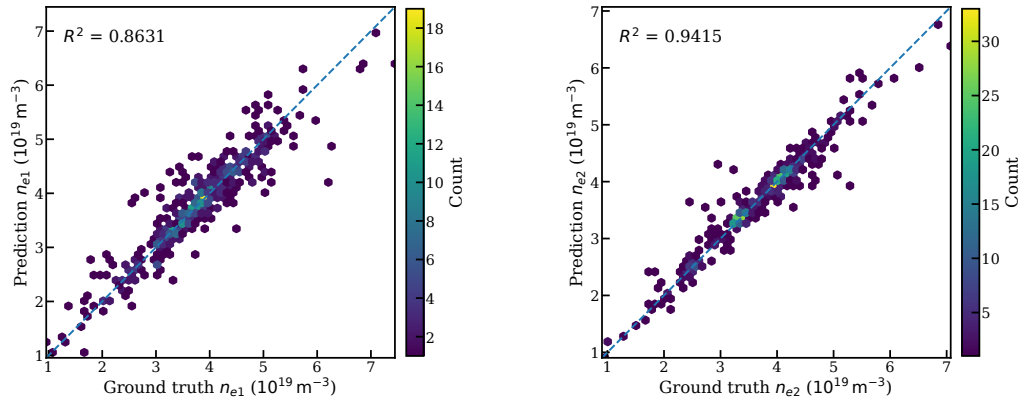
Figure 6: Peak-center localization error comparison between the baseline method and SpectralTF-MAE. Each point denotes a paired sample with the correct peak count in both methods, and the dashed diagonal indicates equal MAE.



(a) Electron-temperature (T_e) relative error.

(b) Ion-temperature (T_i) relative error.

Figure 7: Relative-error distributions of predicted plasma temperatures: (a) electron temperature T_e and (b) ion temperature T_i. Each subfigure shows the error distribution as a function of the normalized radius ρ, together with the within-$\pm 10\%$ coverage curve and the marginal error distribution.



(a) n_{e1} parity.

(b) n_{e2} parity.

Figure 8: Parity plots for electron-density prediction: (a) n_{e1} and (b) n_{e2}. The solid markers represent model predictions against the corresponding ground-truth values, and the dashed diagonal denotes ideal agreement.