Towards Large-Scale Heterogeneous Data Organization for Scientific Foundation Models: A Nuclear Fusion Case Study

Anonymous authors
Paper under double-blind review

Abstract

Training effective foundation models requires massive and organized datasets, yet scientific domains such as nuclear fusion present unique challenges due to largely heterogeneous and sparse data. Here we characterize the data used in developing such a model: with over 20 sensor types spanning 5 orders of magnitude in sampling rate, mixed tensor structures (point measurements, spectrograms, images), and nonstationary physics. We analyze our input complexity and discuss trade-offs between temporal context and frequency resolution. Our analysis provides a template for representing multi-modal fluctuation data at scale, with implications for both multi-modal control systems and nuclear fusion.

1 Introduction

Foundation models have transformed machine learning via their ability to learn general representations from large-scale data (Bommasani et al., 2021). While scaling laws are well-characterized for language and vision (Kaplan et al., 2020; Zhai et al., 2022), scientific domains present distinct data organization challenges that remain.

Tokamak data offers an interesting case study. Experiments can generate rich multivariate text, time series, and images, measuring fusion behavior across spatial, temporal, and spectral dimensions (Churchill, 2025). The data is inherently heterogeneous as sensors can sample from a range of low frequency amplitude-focused scalars to high frequency fluctuation measurements (Gohil et al., 1991). This data can also be sparse, in which enabling certain actuators during experiments can disable nearby sensor pickup even if data is still present. Heterogeneity, combined with turbulent non-stationary physics, makes standard data loading difficult (Surko & Slusher, 1983).

Prior work has addressed fusion prediction tasks individually: disruption forecasting (Vega et al., 2022), profile prediction (Char et al., 2024), and cross-diagnostic inference (Abbate et al., 2024; Jalalvand et al., 2025). More recent studies have investigated methods of reducing the latent space of spectral data (Bouchiat et al., 2025). However, each approach employs task-specific preprocessing, hindering the development of a scalable dataset for standardized models.

Here, we characterize fusion data heterogeneity and its implications for foundation model training. First, we categorize 23 candidate measurements from the DIII-D tokamak by sampling rate, tensor structure, and role in physics; second, we show approximate scaling relationships relating window size to model input complexity; third, we identify key design trade-offs for data loading pipelines.

2 The Fusion Data Landscape

2.1 Shared Measurements and Input Estimations

We analyze sensors and actuators from across several tokamaks, namely DIII-D and guided by similar structures in other devices including KSTAR which are representative of modern fusion facilities (Fenstermacher et al., 2022; Lee et al., 1999). A proposed method of enabling cross-device modeling with the same architecture can use relative encoding for sensor arrays, which are shared characteristics across devices. For example, interferometers on both DIII-D and KSTAR form a

common multichannel configuration (Carlstrom et al., 1988; Lee et al., 2016), while electron cyclotron emission (ECE) often follows a linear array structure (Austin & Lohr, 2003; Jeong et al., 2010). Magnetics also often follow a gridded pattern, though in toroidal coordinates instead of Cartesian (Strait, 2006). While device sizes and operating magnetic field strengths can greatly influence the range of observed phenomena, measurement rates across devices are often similar.

Table 1 summarizes the heterogeneous structure.

Table 1: Heterogeneous fusion data: 23 diagnostics across 6 physics categories which fall into 12 more general categories given a 100ms window.

Category	Representative	Rate (Hz)	Shape	Elements
Fluctuations	Magnetics, BES	100–2000	(C, F, T)	365,714
Kinetic	Thomson, CER	0.04–0.2	(C, T)	6,089
Equilibrium	MSE, I_p	1.5–2.5	(C, T)	3,133
Heat/Fuel	ECH, NBI, Gas	2–100	(C, T)	2,605
Radiation	ECE, D_α	50–1667	(C, F, T)	136,867
Imaging	Divertor Camera	1	$(H, W, 1)$	65,536
			Total (100ms window)	**∼398,500**

Some key observations include:

1. **Extreme sampling rate diversity.** Rates span roughly 5 orders of magnitude (0.04 Hz for Thomson scattering to 2000 Hz for high-resolution magnetics) (Ponce-Marquez et al., 2010; Strait, 2006). This discourages naive concatenation between measurements since resampling to a common rate either discards high-frequency physics or explodes low-frequency channel dimensions.

2. **Mixed tensor structures.** Data includes 2D time series (C, T), 3D spectral tensors (C, F, T), and images $(H, W, 1)$ (Chen et al., 2026). Unlike vision (fixed $224 \times 224 \times 3$) or language (variable-length 1D sequences), fusion models may require architectures to handle heterogeneous tensor ranks simultaneously which is more akin to time series foundation models (Talukder et al., 2025), but with a required accuracy of audio models (Kumar et al., 2023). This encourages the adoption targeting spectral or video data.

3. **Spectral dominance.** Magnetic and density fluctuation spectrograms comprise 87% of total elements, yet represent only 5 of 23 unique modalities. This imbalance may affect both compute allocation and representation learning since models may overfit to spectral features while underweighting scalar quantities critical for understanding physics.

4. **Noise and Non-stationarity.** Thermal noise can greatly obscure coherent structures. Turbulent transport causes statistical properties to evolve on ∼10ms timescales, while global equilibrium evolves over ∼1s (Luce, 2005). Standard normalization (z-score, min-max) computed over full discharges may fail to capture this multi-scale structure.

2.2 EXAMPLE DATA VIEWS

Measurement identification important to prediction are often sparse and difficult to observe. For illustrative clarity, synthetic data is shown to represent the kind of data often analyzed in the context of nuclear fusion (Figure 1).

3 DESIGN CONSIDERATIONS

3.1 WINDOW SIZE TRADE-OFFS

The choice of our time window w fundamentally shapes the learning problem. The frequency resolution trade-off may play an important role in the captured dynamics:
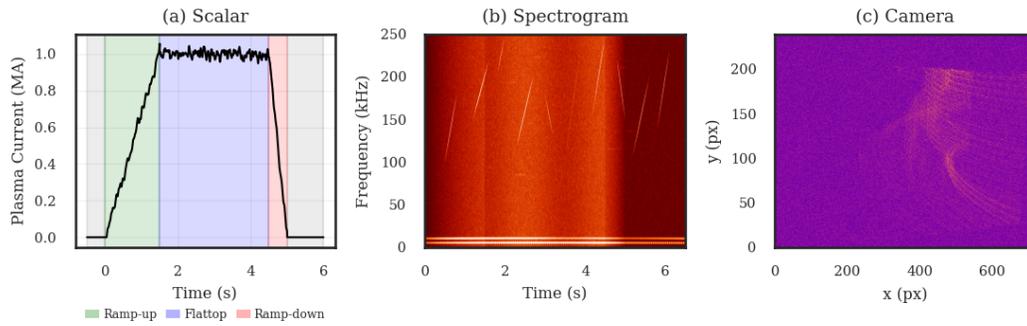
Figure 1: Synthetic examples of the three primary data modalities in fusion diagnostics: (a) representative scalar time series (plasma current), (b) spectrogram (frequency vs. time) where chirps and low frequency modes can both be present at a given time, and (c) camera images in which pixel-level changes can indicate different physics.
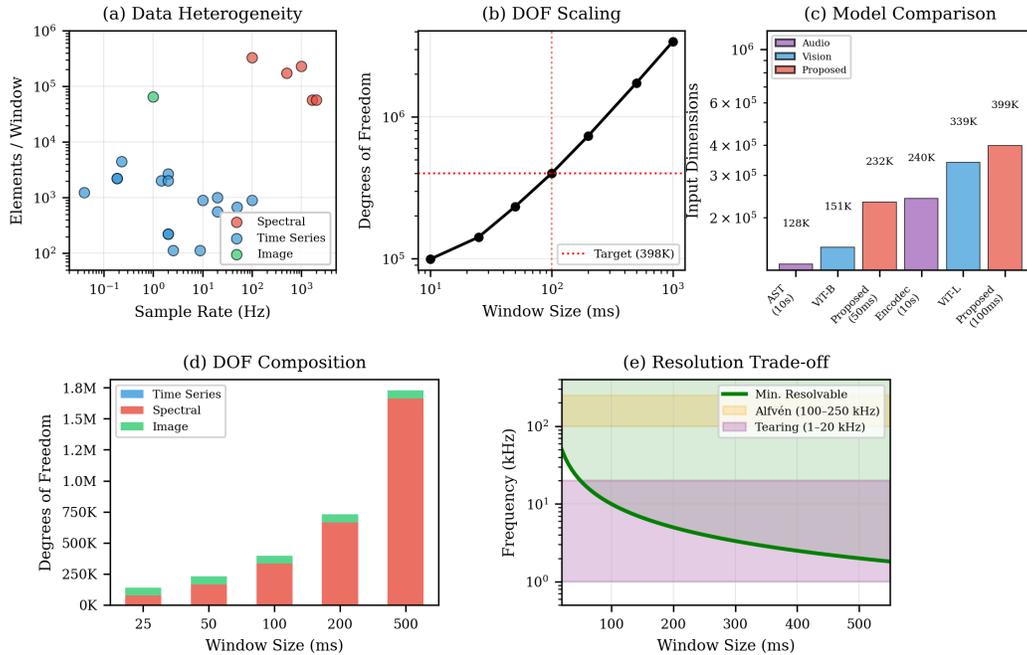

Figure 2: Window size analysis. (a) Data heterogeneity across sampling rates and modalities. (b) Degrees of freedom scale approximately linearly with window size. (c) Model input dimension comparison with existing audio and vision architectures (Gong et al., 2021; Baevski et al., 2020; Dosovitskiy et al., 2020). (d) DOF composition by data type. (e) Frequency resolution trade-off: larger windows capture lower frequencies but sacrifice temporal localization.

- **Alfvén eigenmodes** (100-250 kHz): Requires $w \geq 1$ms for adequate spectral resolution, but benefit from longer context to track frequency chirping (Zeeland et al., 2005).

- **Tearing modes** (1–20 kHz): Evolve slowly, permitting $w \geq 100$ms windows that capture mode locking dynamics (La Haye et al., 2002).

Based on task requirements in prior work, a 50ms baseline is recommended for modeling a good amount of the kinetics. However, given a major goal of autoregressive prediction, this would require 100 accurate prediction steps for an average shot that is 5000ms. Increasing the window to $w = 100$ms as a baseline halves the auto-regressive roll-out depth, potentially stabilizing training, yielding an input sequence length in the range of a large ViT model at ~398K inputs.

3

3.2 Precomputation vs. On-the-Fly Processing

Given the current TB scale of combined fusion data on hand, we must decide what to precompute. An important consideration should be made for fluctuation measurements, especially depending on the architecture they are to be potentially used in (AST which requires patch spectrogram inputs vs. wav2vec which operates on raw waveforms) (Gong et al., 2021; Baevski et al., 2020).

Precompute: (1) Spectrograms from raw waveforms (reduces 2000 Hz $\rightarrow$ 500 Hz $\times$ 128 freq bins) with additional preprocessing such as clipping and filtering; (2) Spatial interpolation to common grids.

On-the-fly: (1) Window extraction and padding; (2) Normalization (requires batch statistics); (3) Augmentations (time shifts, channel dropout).

In addition, methods such as wavelet transforms and multiscale transforms have been shown to improve performance. However, these may introduce additional overhead for baseline tasks.

3.3 Handling Sparsity and Missing Data

Fusion diagnostics exhibit structured sparsity: sensors fail, diagnostics operate intermittently, and device configurations vary across facilities. We propose:

- **Channel masking**: Positional embeddings can indicate the input and measurement ID, with a sparse input and reconstructed output with masked variables indicating that the measurement is not available, while keeping computation tractable (Baade et al., 2022).

- **Cross-device alignment**: Mapping diagnostics to common categories (Table 1) rather than device-specific names, facilitating transfer between DIII-D, KSTAR, EAST, and future ITER data. This also includes mapping the condition of the device during experiments, as documented in log files, which can greatly affect the results of experiments even with the same actuation and measurements.

4 Conclusion

We have characterized the heterogeneous data landscape of nuclear fusion and identified key challenges for foundation model training: extreme sampling rate diversity, mixed tensor structures, spectral data dominance, and physics-driven non-stationarity. Our window size scaling analysis and design recommendations provide a template for organizing scientific data at scale.

Future work includes training a foundation model to identify critical considerations: (1) developing full scaling laws across model size and data volume; (2) benchmarks on downstream tasks such as Alfvén eigenmode classification, ELM prediction, and tearing mode forecasting; (3) cross-device transfer experiments.

References

J. Abbate, E. Fable, B. Grierson, A. Pankin, G. Tardini, and E. Kolemen. Large-database cross-verification and validation of tokamak transport models using baselines for comparison. *Physics of Plasmas*, 31(4):042506, April 2024. ISSN 1070-664X. doi: 10.1063/5.0190908. URL https://doi.org/10.1063/5.0190908.

M. E. Austin and J. Lohr. Electron cyclotron emission radiometer upgrade on the DIII-D tokamak. *Review of Scientific Instruments*, 74(3):1457–1459, March 2003. ISSN 0034-6748. doi: 10.1063/1.1530387. URL https://doi.org/10.1063/1.1530387.

Alan Baade, Puyuan Peng, and David Harwath. MAE-AST: Masked Autoencoding Audio Spectrogram Transformer, March 2022. URL https://arxiv.org/abs/2203.16691v1.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, October 2020. URL http://arxiv.org/abs/2006.11477. arXiv:2006.11477 [cs].

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah 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 Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models, August 2021. URL https://arxiv.org/abs/2108.07258v3.

Kouroche Bouchiat, Nathaniel Chen, Peter Steiner, Azarakhsh Jalalvand, and Egemen Kolemen. Towards a Foundation Model for Fusion: Multimodal Representation Learning of Plasma State and Control. 2025.

T. N. Carlstrom, D. R. Ahlgren, and J. Crosbie. Real-time, vibration-compensated CO2 interferometer operation on the DIII-D tokamak. *Review of Scientific Instruments*, 59(7):1063–1066, July 1988. ISSN 0034-6748. doi: 10.1063/1.1139726. URL https://doi.org/10.1063/1.1139726.

Ian Char, Youngseog Chung, Joseph Abbate, Egemen Kolemen, and Jeff Schneider. Full Shot Predictions for the DIII-D Tokamak via Deep Recurrent Networks, April 2024. URL http://arxiv.org/abs/2404.12416. arXiv:2404.12416 [physics].

Nathaniel Chen, Cheolsik Byun, Azarakhsh Jalalvand, Sangkyeun Kim, Andrew Rothstein, Filippo Scotti, Steve Allen, David Eldon, Keith Erickson, and Egemen Kolemen. Regulation compliant AI for fusion: explainable image-based feedback control of divertor detachment in DIII-D tokamak. *Nuclear Fusion*, 66(3):036014, February 2026. ISSN 0029-5515. doi: 10.1088/1741-4326/ae3972. URL https://doi.org/10.1088/1741-4326/ae3972.

R. Michael Churchill. AI foundation models for experimental fusion tasks. *Frontiers in Physics*, 12, February 2025. ISSN 2296-424X. doi: 10.3389/fphy.2024.1531334. URL https://www.frontiersin.org/journals/physics/articles/10.3389/fphy.2024.1531334/full.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, October 2020. URL https://arxiv.org/abs/2010.11929v2.

M. E. Fenstermacher, for the DIII-D. Team, J. Abbate, S. Abe, T. Abrams, M. Adams, B. Adamson, N. Aiba, T. Akiyama, P. Aleynikov, E. Allen, S. Allen, H. Anand, J. Anderson, Y. Andrew, T. Andrews, D. Appelt, R. Arbon, N. Ashikawa, A. Ashourvan, M. Aslin, Y. Asnis, M. Austin, D. Ayala, J. Bak, I. Bandyopadhyay, S. Banerjee, K. Barada, L. Bardoczi, J. Barr, E. Bass, D. Battaglia, A. Battey, W. Baumgartner, L. Baylor, J. Beckers, M. Beidler, E. Belli, J. Berkery, T. Bernard, N. Bertelli, M. Beurskens, R. Bielajew, S. Bilgili, B. Biswas, S. Blondel, J. Boedo, I. Bogatu, R. Boivin, T. Bolzonella, M. Bongard, X. Bonnin, P. Bonoli, M. Bonotto, A. Bortolon, S. Bose, N. Bosviel, S. Bouwmans, M. Boyer, W. Boyes, L. Bradley, R. Brambila, D. Brennan, S. Bringuier, L. Brodsky, M. Brookman, J. Brooks, D. Brower, G. Brown, W. Brown, M. Burke,

K. Burrell, K. Butler, R. Buttery, I. Bykov, P. Byrne, A. Cacheris, K. Callahan, J. Callen, G. Campbell, J. Candy, J. Canik, P. Cano-Megias, N. Cao, L. Carayannopoulos, T. Carlstrom, W. Carrig, T. Carter, W. Cary, L. Casali, M. Cengher, G. Cespedes Paz, R. Chaban, V. Chan, B. Chapman, I. Char, A. Chattopadhyay, R. Chen, J. Chen, X. Chen, X. Chen, J. Chen, M. Chen, J. Chen, Z. Chen, M. Choi, W. Choi, G. Choi, L. Chousal, C. Chrobak, C. Chrystal, Y. Chung, R. Churchill, M. Cianciosa, J. Clark, M. Clement, S. Coda, A. Cole, C. Collins, W. Conlin, A. Cooper, J. Cordell, B. Coriton, T. Cote, J. Cothran, A. Creely, N. Crocker, C. Crowe, B. Crowley, T. Crowley, D. Cruz-Zabala, D. Cummings, M. Curie, D. Curreli, A. Dal Molin, B. Dannels, A. Dautt-Silva, K. Davda, G. De Tommasi, P. De Vries, G. Degrandchamp, J. Degrassie, D. Demers, S. Denk, S. Depasquale, E. Deshazer, A. Diallo, S. Diem, A. Dimits, R. Ding, S. Ding, W. Ding, T. Do, J. Doane, G. Dong, D. Donovan, J. Drake, W. Drews, J. Drobny, X. Du, H. Du, V. Duarte, D. Dudt, C. Dunn, J. Duran, A. Dvorak, F. Effenberg, N. Eidietis, D. Elder, D. Eldon, R. Ellis, W. Elwasif, D. Ennis, K. Erickson, D. Ernst, M. Fasciana, D. Fedorov, E. Feibush, N. Ferraro, J. Ferreira, J. Ferron, P. Fimognari, D. Finkenthal, R. Fitzpatrick, P. Fox, W. Fox, L. Frassinetti, H. Frerichs, H. Frye, Y. Fu, K. Gage, J. Galdon Quiroga, A. Gallo, Q. Gao, A. Garcia, M. Garcia Munoz, D. Garnier, A. Garofalo, A. Gattuso, D. Geng, K. Gentle, D. Ghosh, L. Giacomelli, S. Gibson, E. Gilson, C. Giroud, F. Glass, A. Glasser, D. Glibert, P. Gohil, R. Gomez, S. Gomez, X. Gong, E. Gonzales, A. Goodman, Y. Gorelov, V. Graber, R. Granetz, T. Gray, D. Green, C. Greenfield, M. Greenwald, B. Grierson, R. Groebner, W. Grosnickle, M. Groth, H. Grunloh, S. Gu, W. Guo, H. Guo, P. Gupta, J. Guterl, W. Guttenfelder, T. Guzman, S. Haar, R. Hager, S. Hahn, M. Halfmoon, T. Hall, K. Hallatschek, F. Halpern, G. Hammett, H. Han, E. Hansen, C. Hansen, M. Hansink, J. Hanson, M. Hanson, G. Hao, A. Harris, R. Harvey, S. Haskey, E. Hassan, A. Hassanein, D. Hatch, R. Hawryluk, W. Hayashi, W. Heidbrink, J. Herfindal, J. Hicok, D. Hill, E. Hinson, C. Holcomb, L. Holland, C. Holland, E. Hollmann, J. Hollocombe, A. Holm, I. Holmes, K. Holtrop, M. Honda, R. Hong, R. Hood, A. Horton, L. Horvath, M. Hosokawa, S. Houshmandyar, N. Howard, E. Howell, D. Hoyt, W. Hu, Y. Hu, Q. Hu, J. Huang, Y. Huang, J. Hughes, T. Human, D. Humphreys, P. Huynh, A. Hyatt, C. Ibanez, L. Ibarra, R. Icasas, K. Ida, V. Igochine, Y. In, S. Inoue, A. Isayama, O. Izacard, V. Izzo, A. Jackson, G. Jacobsen, A. Jaervinen, A. Jalalvand, J. Janhunen, S. Jardin, H. Jarleblad, Y. Jeon, H. Ji, X. Jian, E. Joffrin, A. Johansen, C. Johnson, T. Johnson, C. Jones, I. Joseph, D. Jubas, B. Junge, W. Kalb, R. Kalling, C. Kamath, J. Kang, D. Kaplan, A. Kaptanoglu, S. Kasdorf, J. Kates-Harbeck, P. Kazantzidis, A. Kellman, D. Kellman, C. Kessel, K. Khumthong, E. Kim, H. Kim, J. Kim, S. Kim, J. Kim, H. Kim, K. Kim, C. Kim, W. Kimura, M. King, J. King, J. Kinsey, A. Kirk, B. Kiyan, A. Kleiner, V. Klevarova, R. Knapp, M. Knolker, W. Ko, T. Kobayashi, E. Koch, M. Kochan, B. Koel, M. Koepke, A. Kohn, R. Kolasinski, E. Kolemen, E. Kostadinova, M. Kostuk, G. Kramer, D. Kriete, L. Kripner, S. Kubota, J. Kulchar, K. Kwon, R. La Haye, F. Laggner, H. Lan, R. Lantsov, L. Lao, A. Lasa Esquisabel, C. Lasnier, C. Lau, B. Leard, J. Lee, R. Lee, M. Lee, M. Lee, Y. Lee, C. Lee, J. Lee, S. Lee, M. Lehnen, A. Leonard, E. Leppink, M. Lesher, J. Lestz, J. Leuer, N. Leuthold, X. Li, K. Li, E. Li, G. Li, L. Li, Z. Li, J. Li, Y. Li, Z. Lin, D. Lin, X. Liu, J. Liu, Y. Liu, T. Liu, Y. Liu, C. Liu, Z. Liu, C. Liu, D. Liu, A. Liu, D. Liu, A. Loarte-Prieto, L. Lodestro, N. Logan, J. Lohr, B. Lombardo, J. Lore, Q. Luan, T. Luce, T. Luda Di Cortemiglia, N. Luhmann, R. Lunsford, Z. Luo, A. Lvovskiy, B. Lyons, X. Ma, M. Madruga, B. Madsen, C. Maggi, K. Maheshwari, A. Mail, J. Mailloux, R. Maingi, M. Major, M. Makowski, R. Manchanda, C. Marini, A. Marinoni, A. Maris, T. Markovic, L. Marrelli, E. Martin, J. Mateja, G. Matsunaga, R. Maurizio, P. Mauzey, D. Mauzey, G. Mcardle, J. Mcclenaghan, K. Mccollam, C. Mcdevitt, K. Mckay, G. Mckee, A. Mclean, V. Mehta, E. Meier, J. Menard, O. Meneghini, G. Merlo, S. Messer, W. Meyer, C. Michael, C. Michoski, P. Milne, G. Minet, A. Misleh, Y. Mitrishkin, C. Moeller, K. Montes, M. Morales, S. Mordijck, D. Moreau, S. Morosohk, P. Morris, L. Morton, A. Moser, R. Moyer, C. Moynihan, T. Mrazkova, D. Mueller, S. Munaretto, J. Munoz Burgos, C. Murphy, K. Murphy, C. Muscatello, C. Myers, A. Nagy, G. Nandipati, M. Navarro, F. Nave, G. Navratil, R. Nazikian, A. Neff, G. Neilson, T. Neiser, W. Neiswanger, D. Nelson, A. Nelson, F. Nespoli, R. Nguyen, L. Nguyen, X. Nguyen, J. Nichols, M. Nocente, S. Nogami, S. Noraky, N. Norausky, M. Nornberg, R. Nygren, T. Odstrcil, D. Ogas, T. Ogorman, S. Ohdachi, Y. Ohtani, M. Okabayashi, M. Okamoto, L. Olavson, E. Olofsson, M. Omullane, R. Oneill, D. Orlov, W. Orvis, T. Osborne, D. Pace, G. Paganini Canal, A. Pajares Martinez, L. Palacios, C. Pan, Q. Pan, R. Pandit, M. Pandya, A. Pankin, Y. Park, J. Park, J. Park, S. Parker, P. Parks, M. Parsons, B. Patel, C. Pawley, C. Paz-Soldan, W. Peebles, S. Pelton, R. Perillo, C. Petty, Y. Peysson, D. Pierce, A. Pigarov, L. Pigatto, D. Piglowski, S. Pinches, R. Pinsker, P. Piovesan, N. Piper, A. Pironti, R. Pitts, J. Pizzo, U. Plank, M. Podesta, E. Poli, F. Poli, D. Ponce, Z. Popovic, M. Porko-

lab, G. Porter, C. Powers, S. Powers, R. Prater, Q. Pratt, I. Pusztai, J. Qian, X. Qin, O. Ra, T. Rafiq, T. Raines, R. Raman, J. Rauch, A. Raymond, C. Rea, M. Reich, A. Reiman, S. Reinhold, M. Reinke, R. Reksoatmodjo, Q. Ren, Y. Ren, J. Ren, M. Rensink, J. Renteria, T. Rhodes, J. Rice, R. Roberts, J. Robinson, P. Rodriguez Fernandez, T. Rognlien, A. Rosenthal, S. Rosiello, J. Rost, J. Roveto, W. Rowan, R. Rozenblat, J. Ruane, D. Rudakov, J. Ruiz Ruiz, R. Rupani, S. Saarelma, S. Sabbagh, J. Sachdev, J. Saenz, S. Saib, M. Salewski, A. Salmi, B. Sammuli, C. Samuell, A. Sandorfi, C. Sang, J. Sarff, O. Sauter, K. Schaubel, L. Schmitz, O. Schmitz, J. Schneider, P. Schroeder, K. Schultz, E. Schuster, J. Schwartz, F. Sciortino, F. Scotti, J. Scoville, A. Seltzman, S. Seol, I. Sfiligoi, M. Shafer, S. Sharapov, H. Shen, Z. Sheng, T. Shepard, S. Shi, Y. Shibata, G. Shin, D. Shiraki, R. Shousha, H. Si, P. Simmerling, G. Sinclair, J. Sinha, P. Sinha, G. Sips, T. Sizyuk, C. Skinner, A. Sladkomedova, T. Slendebroek, J. Slief, R. Smirnov, J. Smith, S. Smith, D. Smith, J. Snipes, G. Snoep, A. Snyder, P. Snyder, E. Solano, W. Solomon, J. Song, A. Sontag, V. Soukhanovskii, J. Spendlove, D. Spong, J. Squire, C. Srinivasan, W. Stacey, G. Staebler, L. Stagner, T. Stange, P. Stangeby, R. Stefan, R. Stemprok, D. Stephan, J. Stillerman, T. Stoltzfus-Dueck, W. Stonecipher, S. Storment, E. Strait, D. Su, L. Sugiyama, Y. Sun, P. Sun, Z. Sun, A. Sun, D. Sundstrom, C. Sung, J. Sungcoco, W. Suttrop, Y. Suzuki, T. Suzuki, A. Svyatkovskiy, C. Swee, R. Sweeney, C. Sweetnam, G. Szepesi, M. Takechi, T. Tala, K. Tanaka, X. Tang, S. Tang, Y. Tao, R. Tao, D. Taussig, T. Taylor, K. Teixeira, K. Teo, A. Theodorsen, D. Thomas, K. Thome, A. Thorman, A. Thornton, A. Ti, M. Tillack, N. Timchenko, R. Tinguely, R. Tompkins, J. Tooker, A. Torrezan De Sousa, G. Trevisan, S. Tripathi, A. Trujillo Ochoa, D. Truong, C. Tsui, F. Turco, A. Turnbull, M. Umansky, E. Unterberg, P. Vaezi, P. Vail, J. Valdez, W. Valkis, B. Van Compernolle, J. Van Galen, R. Van Kampen, M. Van Zeeland, G. Verdoolaege, N. Vianello, B. Victor, E. Viezzer, S. Vincena, M. Wade, F. Waelbroeck, J. Wai, T. Wakatsuki, M. Walker, G. Wallace, R. Waltz, W. Wampler, L. Wang, H. Wang, Y. Wang, H. Wang, Z. Wang, H. Wang, Z. Wang, Y. Wang, G. Wang, S. Ward, M. Watkins, J. Watkins, W. Wehner, Y. Wei, M. Weiland, D. Weisberg, A. Welander, A. White, R. White, S. Wiesen, R. Wilcox, T. Wilks, M. Willensdorfer, H. Wilson, A. Wingen, M. Wolde, M. Wolff, K. Woller, A. Wolz, H. Wong, S. Woodruff, M. Wu, Y. Wu, S. Wukitch, G. Wurden, W. Xiao, R. Xie, Z. Xing, X. Xu, C. Xu, G. Xu, Z. Yan, X. Yang, S. Yang, T. Yokoyama, R. Yoneda, M. Yoshida, K. You, T. Younkin, J. Yu, M. Yu, G. Yu, Q. Yuan, L. Zaidenberg, L. Zakharov, A. Zamengo, S. Zamperini, M. Zarnstorff, E. Zeger, K. Zeller, L. Zeng, M. Zerbini, L. Zhang, X. Zhang, R. Zhang, B. Zhang, J. Zhang, J. Zhang, L. Zhao, B. Zhao, Y. Zheng, L. Zheng, B. Zhu, J. Zhu, Y. Zhu, Y. Zhu, M. Zsutty, and M. Zuin. DIII-D research advancing the physics basis for optimizing the tokamak approach to fusion energy. *Nuclear Fusion*, 62(4):042024, April 2022. ISSN 0029-5515. doi: 10.1088/1741-4326/ac2ff2. URL https://doi.org/10.1088/1741-4326/ac2ff2.

P. Gohil, K.H. Burrell, R.J. Groebner, J. Kim, W.C. Martin, E.L. McKee, and R.P. Seraydarian. The charge exchange recombination diagnostic system on the DIII-D tokamak. In *[Proceedings] The 14th IEEE/NPSS Symposium Fusion Engineering*, pp. 1199–1204 vol.2, September 1991. doi: 10.1109/FUSION.1991.218656. URL https://ieeexplore.ieee.org/document/218656.

Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer, July 2021. URL http://arxiv.org/abs/2104.01778. arXiv:2104.01778 [cs].

Azarakhsh Jalalvand, SangKyeun Kim, Jaemin Seo, Qiming Hu, Max Curie, Peter Steiner, Andrew Oakleigh Nelson, Yong-Su Na, and Egemen Kolemen. Multimodal super-resolution: discovering hidden physics and its application to fusion plasmas. *Nature Communications*, 16(1):8506, September 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-63492-1. URL https://www.nature.com/articles/s41467-025-63492-1.

S. H. Jeong, K. D. Lee, Y. Kogi, K. Kawahata, Y. Nagayama, A. Mase, and M. Kwon. Electron cyclotron emission diagnostics on KSTAR tokamak. *The Review of Scientific Instruments*, 81(10):10D922, October 2010. ISSN 1089-7623. doi: 10.1063/1.3491224.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, January 2020. URL https://arxiv.org/abs/2001.08361v1.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-Fidelity Audio Compression with Improved RVQGAN, October 2023. URL http://arxiv.org/abs/2306.06546. arXiv:2306.06546 [cs].

R. J. La Haye, S. Günter, D. A. Humphreys, J. Lohr, T. C. Luce, M. E. Maraschek, C. C. Petty, R. Prater, J. T. Scoville, and E. J. Strait. Control of neoclassical tearing modes in DIII–D. *Physics of Plasmas*, 9(5):2051–2060, May 2002. ISSN 1070-664X, 1089-7674. doi: 10.1063/1.1456066. URL https://pubs.aip.org/pop/article/9/5/2051/1070453/Control-of-neoclassical-tearing-modes-in-DIII-D.

G. S. Lee, J. Kim, S. M. Hwang, C. S. Chang, H. Y. Chang, M. H. Cho, B. H. Choi, K. Kim, S. Jardin, G. H. Neilson, H. K. Park, W. Reiersen, J. Schmidt, K. Young, J. Schultz, L. Sevier, S. Y. Cho, J. H. Han, N. I. Hur, K. H. Im, S. Kim, J. Y. Kim, M. C. Kyum, B. J. Lee, D. K Lee, S. G. Lee, H. L. Yang, B. G. Hong, Y. S. Hwang, S. H. Kim, Y. J. Kim, J. Y. Lim, W. Namkung, K. H. Chung, and D. I. Choi. The design of the KSTAR tokamak. *Fusion Engineering and Design*, 46(2):405–411, November 1999. ISSN 0920-3796. doi: 10.1016/S0920-3796(99)00032-0. URL https://www.sciencedirect.com/science/article/pii/S0920379699000320.

K. C. Lee, J. W. Juhn, Y. U. Nam, Y. S. Kim, H. M. Wi, S. W. Kim, and Y. c. Ghim. The design of two color interferometer system for the 3-dimensional analysis of plasma density evolution on KSTAR. *Fusion Engineering and Design*, 113:87–91, December 2016. ISSN 0920-3796. doi: 10.1016/j.fusengdes.2016.10.008. URL https://www.sciencedirect.com/science/article/pii/S0920379616306664.

T. C. Luce. Development of Steady-State Advanced Tokamak Research in the DIII-D Tokamak. *Fusion Science and Technology*, 48(2):1212–1225, October 2005. ISSN 1536-1055, 1943-7641. doi: 10.13182/FST05-A1072. URL https://www.tandfonline.com/doi/full/10.13182/FST05-A1072.

D. M. Ponce-Marquez, B. D. Bray, T. M. Deterly, C. Liu, and D. Eldon. Thomson scattering diagnostic upgrade on DIII-D. *The Review of Scientific Instruments*, 81(10):10D525, October 2010. ISSN 1089-7623. doi: 10.1063/1.3495759.

E. J. Strait. Magnetic diagnostic system of the DIII-D tokamak. *Review of Scientific Instruments*, 77 (2):023502, February 2006. ISSN 0034-6748. doi: 10.1063/1.2166493. URL https://doi.org/10.1063/1.2166493.

C. M. Surko and R. E. Slusher. Waves and Turbulence in a Tokamak Fusion Plasma. *Science*, 221(4613):817–822, August 1983. doi: 10.1126/science.221.4613.817. URL https://www.science.org/doi/10.1126/science.221.4613.817.

Sabera Talukder, Yisong Yue, and Georgia Gkioxari. TOTEM: TOkenized Time Series EMbeddings for General Time Series Analysis, January 2025. URL http://arxiv.org/abs/2402.16412. arXiv:2402.16412 [cs].

J. Vega, A. Murari, S. Dormido-Canto, G. A. Rattá, and M. Gelfusa. Disruption prediction with artificial intelligence techniques in tokamak plasmas. *Nature Physics*, 18(7):741–750, July 2022. ISSN 1745-2481. doi: 10.1038/s41567-022-01602-2. URL https://www.nature.com/articles/s41567-022-01602-2.

M A Van Zeeland, G J Kramer, R Nazikian, H L Berk, T N Carlstrom, and W M Solomon. Alfvén eigenmode observations on DIII-D via two-colour CO_2 interferometry. *Plasma Physics and Controlled Fusion*, 47(9):L31–L40, September 2005. ISSN 0741-3335, 1361-6587. doi: 10.1088/0741-3335/47/9/L01. URL https://iopscience.iop.org/article/10.1088/0741-3335/47/9/L01.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling Vision Transformers. pp. 12104–12113, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Zhai_Scaling_Vision_Transformers_CVPR_2022_paper.