# PROGRESSIVE-MASKING MIL DISCOVERS DISRUPTION PRECURSOR EVENT CHAINS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Tokamak disruptions are abrupt losses of plasma confinement that can damage the device. Reliable labels usually record only whether a discharge disrupts, while precursor timing is rarely annotated. We treat each discharge as a bag of 5-ms tokens and use multiple-instance learning (MIL) with a split head that scores each token from its own window. Local scores and representations thereby have a common meaning across discharges, supporting a shared vocabulary of candidate events. A bag classifier can rely on late signatures, so *progressive masking* removes its identified evidence and all subsequent tokens, and retraining drives successive models toward earlier candidates. A frozen event grammar and shared vocabulary organize these candidates into six families and event chains. On 1,816 J-TEXT discharges, twelve rounds place the earliest disruption-associated event at a median of 185 ms before disruption on the development pool and 165 ms on the evaluation set, against 65 and 60 ms for one MIL model. The share of disruption-associated events inside the final 100 ms falls from 70% to 33%, and multi-event chains rise from 3.9% to 23%. A fixed truncation schedule and a contextual scoring head also yield earlier, deeper chains, so these effects do not require evidence-guided cutoffs or local scoring. With the last 30 ms of each length-matched evaluation record hidden, PM-MIL reaches AUPRC 0.86 and AUROC 0.92, against 0.83 and 0.91 for attention MIL and 0.68 and 0.87 for one model. This is a truncation-sensitivity experiment rather than a warning-time test. Onset attribution shows that events rising from low risk start mainly on magnetic-fluctuation and equilibrium signals and that the frequent transition from family C to family D adds magnetic-fluctuation evidence. Event presence and family composition carry predictive information. Shuffling order weakens the Markov chain classifier, but geometry-adjusted controls do not establish a consistent additional gain from directed adjacency.

## 1 INTRODUCTION

Nuclear fusion releases energy when light nuclei combine. Doing so on Earth requires heating matter into a plasma and confining it long enough for reactions to occur. A tokamak confines a toroidal plasma with magnetic fields, and each experimental pulse is a *discharge*. A *disruption* is an abrupt loss of confinement that terminates the discharge, with rapid loss of thermal energy and decay of the plasma current. The resulting thermal and electromagnetic loads limit the operation of present devices (de Vries et al., 2011) and are a design-driving concern for reactor-scale tokamaks (Hender et al., 2007). Disruptions therefore have to be predicted early enough to be avoided or mitigated.

Disruption prediction is a multivariate time-series problem in which diagnostics record the evolving plasma and a model estimates whether the discharge will end disruptively. Deep networks predict disruptions with high accuracy (Kates-Harbeck et al., 2019), random-forest predictors run in real time (Rea et al., 2019), and Vega et al. (2022) review these machine-learning approaches. Physics-guided features make them interpretable (Shen et al., 2023), and transfer between devices is being pursued (Zheng et al., 2023). Three problems remain. First, the only reliable label is whether the discharge disrupted. The onset, duration and identity of precursor activity are seldom annotated, and manual annotation is unreliable, because a uniform "pre-disruptive window" writes an assumption into the labels, per-discharge expert labeling does not scale, and multi-stage, overlapping or reversing activity has no single correct time stamp. Second, a classifier trained on these discharge-level



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.

labels can satisfy its objective from the last discriminative interval, and it need not represent the earlier changes or the sequence connecting them. We call this *terminal-evidence concentration* and measure it below. In a standard MIL model, 70% of the events found in disruptive discharges start within the final 100 ms. Third, a good predictor is therefore not an account of how a disruption develops, which is what avoidance requires and what a real-time predictor needs as training signal.

This paper lets the model find the precursors itself and organizes what it finds into chains. (i) We treat a discharge as a bag of 5-ms tokens whose only label is whether it disrupted, and train a multiple-instance learning (MIL) model that must score tokens locally while classifying the bag, so the model localizes the precursor events it relies on without any time-level label. (ii) *Progressive masking* removes the evidence the model found together with the entire remainder of the discharge, and retrains from scratch on the shortened prefix, restricting successive fits to earlier observations. (iii) After freezing all rounds, an event grammar converts model outputs into candidate events, every event is encoded by every frozen local encoder to give a shared representation, and clustering yields candidate event families. Assembling the families per discharge gives an *event chain* ending in the observed DIS or ND terminal, summarized by an absorbing Markov chain. Because a single event is a chain of depth one, every method produces chains, and we compare methods on the chains' temporal *concentration*, *depth* and *accuracy*, not only on classification.

Our contributions are (1) a MIL procedure that discovers precursor candidates from discharge-level labels, with local scores and representations that support a shared candidate-event vocabulary across discharges, and progressive masking that exposes earlier evidence, (2) a common vocabulary and chain evaluation protocol for comparing discovery methods, and (3) evidence on 1,816 J-TEXT discharges that progressive shortening with retraining moves the median earliest disruptive event from 65 to 185 ms on the development pool and from 60 to 165 ms on the evaluation set, reduces terminal concentration from 70% to 33%, and raises multi-event chains from 3.9% to 23%. Onset-level attribution identifies the input signals that start events and those that change at family transitions. Controls distinguish this mechanism from the choice of cutoff or scoring head. Further chain classifiers separate information in event count, family composition and first-order adjacency, while a nominal suffix-removal experiment tests classification sensitivity to missing late observations. These retrospective candidates can support future training and validation of real-time predictors.

## 2 Related work

**Disruption prediction, interpretation and avoidance.** Deep predictors (Kates-Harbeck et al., 2019), random-forest predictors with feature-level explanations that run in real time (Rea et al., 2019) or on several devices (Montes et al., 2019), physics-guided interpretable features (Shen et al., 2023), survival formulations of the time to disruption (Tinguely et al., 2019) and cross-device transfer (Zheng et al., 2023) all train on discharge-level labels and evaluate warning time against a deadline, the setting that Vega et al. (2022) survey. DECAF (Sabbagh et al., 2023) builds event chains for disruptions from physics rules with hand-set thresholds and is the closest precedent for the chain view, whereas the chains here are learned from the discharge label alone and their families are candidates rather than named instabilities. Unsupervised maps of the operational space (Pau et al., 2019) and rule extraction from a self-organizing predictor (Setzu et al., 2026) describe static disruption-prone regions rather than the succession of events within a discharge. Closed-loop avoidance of a tearing instability by reinforcement learning (Seo et al., 2024) needs to know which transition to act on, which is what event chains are meant to expose.

**Multiple-instance learning for time series.** Attention-based MIL aggregates instance evidence into a bag prediction (Ilse et al., 2018). MILLET makes time-series classification locally interpretable through conjunctive pooling of instance scores (Early et al., 2024) and TimeMIL adds time-aware pooling (Chen et al., 2024). Our backbone pools attention times instance scores as MILLET does, with the one change that the instance classifier reads local rather than contextual features, and we ablate that change. Weakly supervised temporal localization in video also learns instance scores from clip-level labels (Paul et al., 2018), and the nominal suffix-removal evaluation below is related to early classification of time-series work (Xing et al., 2012). We add the search induced by repeatedly withholding evidence, plus a vocabulary that makes the resulting models' events comparable, and we run embedding-level attention MIL and a TimeMIL-style aggregator as controls.

**Discriminative evidence erasing.** Adversarial erasing mines additional object regions under image-level supervision (Wei et al., 2017). Iterative-Winners-Out removes the most discriminative video segments and retrains (Zeng et al., 2019) and is the closest precedent. Progressive masking differs in what is removed, which is not a local interval but the interval *and its entire future*, so later fits are restricted to earlier prefixes of the discharge. Section 4 tests this temporal restriction against local erasing and against a fixed shortening schedule that consults no model.

## 3 Discovering event chains from discharge-level labels

### 3.1 Problem and MIL backbone

Let $\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{N}$, where $X_i = (x_{i0}, \ldots, x_{i,K_i-1})$ is a sequence of multivariate windows (tokens) and $y_i \in \{0, 1\}$ marks a non-disruptive (ND) or disruptive (DIS) discharge. No event boundaries, event classes or transition labels are supplied. For the visible prefix $V_i^{(r)} = \{0, \ldots, b_i^{(r)} - 1\}$ at round $r$, a local encoder produces $h_{ik}^{(r)}$ from the token alone, and a bidirectional Transformer produces contextual features $z_{ik}^{(r)}$. The classifier reads *local* features and the attention head reads *contextual* features,

$$\ell_{ik}^{(r)} = f_r(h_{ik}^{(r)}) \in \mathbb{R}^2, \quad a_{ik}^{(r)} = \frac{\exp u_r(z_{ik}^{(r)})}{\sum_{j \in V_i^{(r)}} \exp u_r(z_{ij}^{(r)})}, \quad p_i^{(r)} = \mathrm{softmax}\Big( \sum_{k \in V_i^{(r)}} a_{ik}^{(r)} \ell_{ik}^{(r)} \Big)_{\mathrm{DIS}}. \tag{1}$$

The split forces each token's disruption score to be justified by the token itself, while the attention decides where the bag looks. Training minimizes $\mathcal{L} = \mathcal{L}_{\mathrm{focal}} + \lambda_{\mathrm{ND}} \mathcal{L}_{\mathrm{ND}} + \lambda_h \mathcal{L}_{\mathrm{fwd},h} + \lambda_z \mathcal{L}_{\mathrm{fwd},z}$, a bag focal loss, the mean disruptive token score on ND bags, and next-token representation-prediction losses on the two branches ($\lambda_{\mathrm{ND}} = 0.3$, $\lambda_h = 0.005$, $\lambda_z = 0.003$), with a stochastic attention regularizer that drops the highest-attention tokens during fitting (Appendix A.2). The contextual branch's prediction residual $\rho_{ik}$ is the change signal used by the event grammar.

### 3.2 Progressive masking

Initialize $b_i^{(0)} = K_i$. After fitting round $r$, the mask-candidate procedure selects, in each DIS bag that the model classifies positively, the latest eligible evidence boundary $\widehat{s}_i^{(r)}$ (a residual peak followed by a rise of local risk, Appendix A.2). The next cutoff is

$$b_i^{(r+1)} = \min\{b_i^{(r)}, \widehat{s}_i^{(r)}\}, \tag{2}$$

so the selected evidence *and its entire suffix* are removed. Round $r+1$ is trained from a fresh initialization on the remaining prefixes, and ND bags are length-matched to the DIS prefixes so that shrinking positive bags do not create a length shortcut. A discharge whose model output no longer supports a positive decision is retired by a fixed rule, and no boundary is invented to force progress. The sequence stops when class support is insufficient, when the validation criterion fails, or when masks cease to advance. The cutoff restricts the optimization input and does not assert that a true precursor starts there. Figure 1 summarizes the procedure.

### 3.3 From model outputs to candidate events

After freezing the round sequence, we replay the full record of every discharge through every model with its own normalizer and apply one *event grammar* to all of them. Two signals are combined, the local disruption risk $P(\mathrm{DIS} \mid h_k)$ and the contextual prediction residual $\rho_k$, a change signal that is large when the plasma leaves the trajectory the model expected. A residual peak (exceeding the shot-round mean by $1.75\sigma$) that is followed by a sustained high-risk episode (above the model's ND-reference 0.99 quantile for at least three tokens, released at the 0.90 quantile) defines an *event* whose start is the trigger token, and a new peak inside an already high-risk episode ends the current event and starts the next. High risk without a trigger, or a trigger never confirmed by high risk, is not an event. The only additional filter is a per-model bag gate set on ND discharges at 5% shot-level false-positive rate, and ND discharges are scanned with exactly the same rule. No time-level label enters at any point (Appendix A.3).

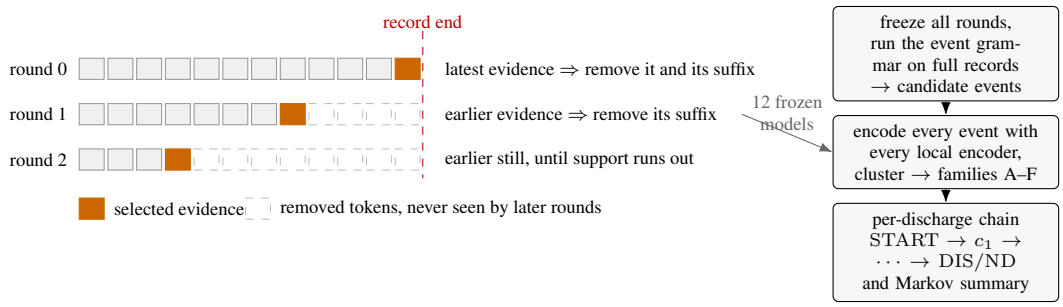



Figure 1: Left, progressive masking. Each round's model locates the latest evidence it relies on in a disruptive discharge, that evidence and its whole suffix are removed, and the next round is trained afresh on the remaining prefix, restricting it to earlier observations. Right, once the rounds are frozen an event grammar extracts candidate events, a shared vocabulary makes events from different models comparable, and each discharge becomes a chain ending in its observed terminal, DIS or ND.

Because the contextual branch is bidirectional and the latch waits for confirmation, this extraction is retrospective. The stored record of a discharge runs from 0.15 s after breakdown to the end time labeled in the J-TEXT database, which for a DIS discharge is the disruption time, and the *earliest-event lead* reported below is the distance from the event start to that time (Appendix A.1).

### 3.4 A SHARED VOCABULARY ACROSS ROUNDS

Latent coordinates of independently fitted encoders need not agree, so instead of describing each event with its source model only, we encode *every event with every frozen local encoder*. For event interval $e$ we average the local features of each model over the interval and standardize each model's block,

$$\bar{h}_e^{(r)} = \frac{1}{|e|}\sum_{k\in e} h_k^{(r)}, \qquad \tilde{h}_e^{(r)} = \frac{\bar{h}_e^{(r)} - \mu_r}{s_r}, \qquad v_e = \frac{1}{\sqrt{R}}\left[\tilde{h}_e^{(0)}, \ldots, \tilde{h}_e^{(R-1)}\right], \tag{3}$$

with $\mu_r$ the event-population mean of block $r$ and $s_r$ its root-mean-square radius, so that every encoder receives equal weight without assuming coordinate-wise correspondence. With $R = 12$ models of 190 dimensions, $v_e$ has 2,280 dimensions. K-means in this space with $K = 6$ gives the working vocabulary, chosen after exploratory comparison of representations and resolutions. Families are named A–F in order of decreasing median DIS lead, and Shapley attributions of the local classifier margin describe the signal pattern each family responds to (Appendix A.4).

Overlapping discoveries from different rounds are consolidated by one rule. Each token is owned by its *earliest* covering round (the least-masked, hence most authoritative, model), and contiguous tokens with the same family are merged into one *mainline segment*. Under this rule the chains of a single MIL model are exactly the round-0 subset of the construction, which is what makes the comparison in Section 5 possible.

### 3.5 EVENT CHAINS AND THEIR MARKOV SUMMARY

For each event-bearing discharge, the ordered mainline families define START $\to c_{i1} \to \cdots \to c_{iJ_i} \to y_i$. Counting consecutive transitions and normalizing outgoing counts gives a descriptive first-order matrix in which DIS and ND are absorbing, and the step is the *next retained event* rather than a fixed time interval. Writing $P = \left[\begin{smallmatrix} Q & B \\ 0 & I \end{smallmatrix}\right]$, $B$ gives direct terminal transitions and $(I-Q)^{-1}B$ eventual absorption (Grinstead & Snell, 1997).

### 3.6 MEASURING CHAINS BY CONCENTRATION, DEPTH AND ACCURACY

A single event is a chain of depth one, so every discovery method yields chains. We compare them on three quantities, all computed under the shared vocabulary and consolidation rule. **Concentration** measures how strongly the discovered events cluster at the terminal end of disruptive discharges,

through the share of DIS events starting within the final 100 ms of the record and through the median earliest-event lead. **Depth** is the number of retained events per event-bearing discharge, summarized by the share of chains with at least two events and by the number of family-to-family transitions. **Accuracy** asks whether the chain is informative about disruption. It is measured by an out-of-fold chain classifier that scores each chain by the class-conditional first-order Markov likelihood ratio fitted on the training folds (Appendix B.3). We report its AUROC on event-bearing chains, which asks whether chain *content* carries information about disruption beyond the mere existence of an event, and its AUPRC on the whole development pool, in which a discharge without events enters as an empty chain and existence itself counts.

## 4 EXPERIMENTAL SETTING

**Data.** The J-TEXT data set used here contains 1,816 discharges (412 DIS, 1,404 ND) with 27 physics-guided features on a 10-kHz grid grouped into five diagnostic groups (magnetic fluctuations, radiation, density, equilibrium, vertical profile shape). Fifty samples form one 5-ms token. A stratified random evaluation set of 182 discharges (39 DIS, 143 ND) is never used for discovery-model training, checkpoint selection, stopping, event thresholds or the vocabulary. For whole-discharge classification, only the ensemble weights and decision threshold of Eq. 4 are fitted on it, and cross-fitted scores are reported alongside. The development pool holds the remaining discharges, and after excluding nine DIS records shorter than 50 ms it contains 1,625 discharges (364 DIS, 1,261 ND). Each record runs from 0.15 s to the labeled end of the discharge, so DIS records (median 256 ms) are shorter than ND records (median 348 ms), and the visible length of a record is itself a label cue that the evaluation must neutralize (Section 5.1). Event thresholds and the vocabulary are fitted on the development pool and then frozen for the replay on the evaluation set. A corrected high-frequency feature rule is applied throughout (Appendix A.1).

**Rounds and final score.** Each round selects the checkpoint with the largest within-round validation AUPRC (AP below) and is admitted if that AP exceeds the validation DIS fraction. The run admits twelve models and stops when the eleven remaining DIS discharges cannot supply the minimum ten training plus five validation discharges. For whole-discharge classification the twelve full-record bag probabilities $q_{ir}$ are combined as

$$w_r = \frac{\max(\mathrm{AP}_r - \pi_0, 0)}{\sum_j \max(\mathrm{AP}_j - \pi_0, 0)}, \qquad S_i = \textstyle\sum_r w_r q_{ir}, \qquad \pi_0 = 39/182, \tag{4}$$

where $\mathrm{AP}_r$ is measured on the evaluation set. The threshold $\tau$ is set on the ND scores of the evaluation set at empirical FPR below 5%. Cross-fitted scores refit the weights on four folds of twenty stratified five-fold partitions and score the fifth. AUPRC is the primary metric. Bootstrap intervals condition on the fitted predictions. For truncation sensitivity, each ND record is cropped to a random window with a sampled DIS length, capped by its own available length. We average per-model probabilities over three ND crops, then calculate metrics. The last nominal $\Delta$ ms are removed subject to retaining the first 50 ms, so short records undergo less removal. We retain all 182 discharges and use the original full-record ensemble weights unless stated otherwise. Chains use full records, frozen thresholds and the shared vocabulary.

**Methods compared.** M1, *Single MIL*, is round 0 of our run, one MIL model on full records. M5 is *PM-MIL*, all twelve rounds. Four controls each remove one ingredient of PM-MIL. M0-repeat refits round 0 twelve times without masking and aggregates the fits with the same rule (re-fitting and ensembling alone). M2, *Masked MIL*, trains one model on fixed truncated-prefix views (truncation as augmentation, no re-fitting). M3, an attention-erasing/Winners-Out adaptation, erases one block of five highest-contribution tokens per round from every DIS discharge the model still supports but keeps later observations visible (no suffix removal). M4, fixed-schedule suffix truncation, removes 25 ms more per round from every supported DIS discharge (progressive shortening with retraining, no evidence and no learned schedule). Both keep the screening, retirement and stopping rules of PM-MIL (Appendix C). Two aggregators from the related work are trained on the same front-end and encoder with the bag focal loss only. R1 is embedding-level gated attention MIL (Ilse et al., 2018), localized by its attention, and R2 is a class-token Transformer aggregator with a learnable positional embedding after TimeMIL (Chen et al., 2024), localized by the class-token attention. Both signals enter the same event grammar. All are specified in Appendix C. PM-MIL, M2, M3 and

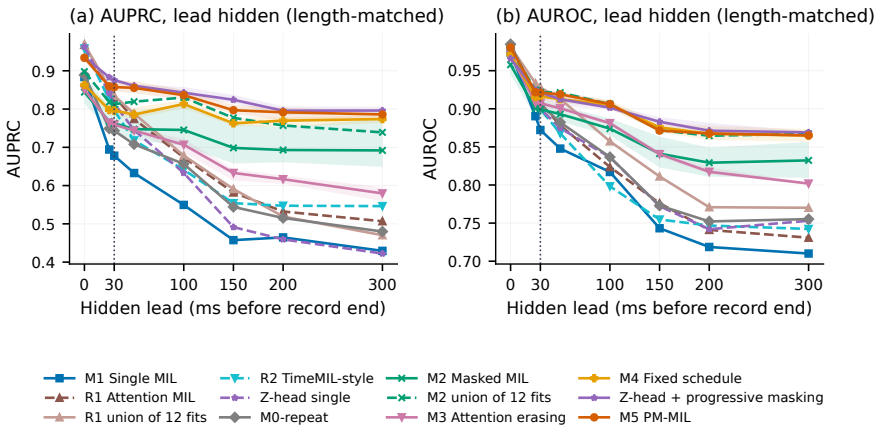



Figure 2: Classification on the evaluation set under nominal suffix removal $\Delta$, with a 50-ms minimum retained prefix. Actual removal is smaller for short records. ND windows are randomly length-matched, and three crop probabilities are averaged before scoring. (a) AUPRC and (b) AUROC, with seed means and ranges. Original full-record ensemble weights are retained. The dotted line marks the 30-ms setting of Table 1.

M4 are run with three seeds each and M0-repeat uses twelve further seeds. The reference PM-MIL run is seed 42, and all vocabulary and chain analyses use it. A split-head ablation (Z-head) keeps every loss and reads the instance classifier from $z_{ik}$ instead of $h_{ik}$, as a single model (one seed) and under the same progressive masking (three seeds), and twelve M2 fits from different seeds are pooled like M0-repeat.

## 5 Results

### 5.1 Whole-discharge classification

Record length is a label cue on J-TEXT. A logistic regression on length reaches AUPRC 0.858 on the unmatched evaluation records (Appendix B.6). We therefore also evaluate randomly length-matched ND windows, on which the length classifier reaches AUPRC 0.453 and AUROC 0.567 with the last 30 ms hidden (prevalence 0.214), so matching removes most but not all of the length cue. Table 1 reports classification with the last 30 ms of every record hidden, which leaves at least 50 ms of every evaluation record, and Figure 2 the range from 0 to 300 ms. PM-MIL reaches AUPRC 0.857 [0.763, 0.937] and AUROC 0.921, against 0.833 and 0.907 for R1, 0.797 and 0.900 for R2 and 0.678 and 0.872 for Single MIL. At 30 ms the paired intervals separate PM-MIL from Single MIL only, and from 100 ms onward also from M0-repeat, R1, the R1 union, R2, the single Z-head model, single-model M2 and M3 (Appendix B.3), conditional on fitted predictions. M4 reaches 0.796, the M2 union 0.813 and progressive Z-head 0.876, with paired intervals that span zero, so training on shortened records supports truncation robustness across several designs. The ranking from nominal 50 ms onward is similar without length matching. On the 1,625 development discharges, where the models were trained and the events are found, AUPRC at 30 ms is 0.951 for PM-MIL and 0.886 for Single MIL and lies between 0.869 and 0.964 across all methods (Table 1).

### 5.2 Progressive masking moves evidence earlier and de-concentrates it

Table 2 gives the chain quantities on the pool, where the twelve models were fitted, and on the evaluation discharges replayed with frozen thresholds and vocabulary. On the pool the median earliest DIS event moves from 65 ms [IQR 25, 130] for Single MIL to 185 ms [101, 255] for PM-MIL, the difference of medians over resampled discharges is 105 to 140 ms (Appendix B.3), and the share of DIS events starting in the final 100 ms falls from 70% to 33%. On the evaluation set, the same replay gives 60 against 165 ms and 66% against 34%, so the earlier and less concentrated

Table 1: Classification with the last 30 ms of every record hidden, on the length-matched evaluation records (182 discharges, 39 DIS) and on the 1,625 development discharges (pool columns). Brackets are conditional 95% bootstrap intervals from 2,000 stratified shot resamples, and three-seed rows give seed means. Cross-fitted weights are refitted at 30 ms over twenty five-fold partitions.

| Method | AUPRC, 30 ms [95% CI] | AUROC, 30 ms [95% CI] | Cross-fitted AUPRC | Pool AUPRC | Pool AUROC |
|---|---|---|---|---|---|
| Record length only | .453 | .567 | — | — | — |
| M1 Single MIL | .678 [.55, .84] | .872 [.79, .94] | .678 | .886 | .928 |
| R1 Attention MIL | .833 [.73, .92] | .907 [.84, .97] | .833 | .876 | .921 |
| R1 union of 12 fits | .835 [.73, .92] | .928 [.87, .97] | .834 | .932 | .966 |
| R2 TimeMIL-style | .797 [.69, .89] | .900 [.84, .95] | .797 | .869 | .925 |
| Z-head single | .814 [.71, .90] | .900 [.83, .96] | .814 | .896 | .945 |
| M0-repeat | .743 [.62, .89] | .906 [.83, .96] | .744 | .915 | .950 |
| M2 Masked MIL | .763 [.64, .88] | .899 [.82, .96] | .763 | .884 | .949 |
| M2 union of 12 fits | .813 [.69, .93] | .922 [.84, .98] | .812 | .943 | .975 |
| M3 Attention erasing | .759 [.63, .90] | .907 [.83, .97] | .758 | .941 | .977 |
| M4 Fixed schedule | .796 [.67, .93] | .916 [.84, .97] | .790 | .952 | .976 |
| Z-head + prog. masking | .876 [.80, .94] | .921 [.86, .97] | .870 | .964 | .978 |
| M5 PM-MIL (3 seeds) | .857 [.76, .94] | .921 [.84, .98] | .840 | .951 | .975 |

Table 2: Chain quantities on the 1,625 development discharges and on the 182 evaluation discharges (thresholds and vocabulary frozen). Lead is the median earliest DIS event lead, the next columns give the shares of DIS events in the final 100 ms and of chains with two or more events, and AUROC is that of the out-of-fold Markov chain classifier with a bootstrap interval (reference run for PM-MIL, seed means otherwise, paired differences in Appendix B.3). R2 has too few events for chains.

| | Development pool | | | | Evaluation set | | | |
|---|---|---|---|---|---|---|---|---|
| Method | Lead ms | Final 100 ms | Depth $\geq 2$ | AUROC [95% CI] | Lead ms | Final 100 ms | Depth $\geq 2$ | AUROC [95% CI] |
| M1 Single MIL | 65 | 70% | 3.9% | .49 [.38, .60] | 60 | 66% | 4.5% | .58 [.32, .83] |
| R1 Attention MIL | 15 | 100% | 2.9% | .41 | — | — | — | — |
| R1 union of 12 fits | 50 | 76% | 12.2% | .55 | 50 | 72% | 12.1% | .43 |
| Z-head single | 100 | 53% | 9.4% | .47 | 70 | 72% | 10.3% | .27 |
| M0-repeat | 150 | 39% | 11.1% | .53 [.46, .60] | 138 | 36% | 9.3% | .62 [.45, .79] |
| M2 Masked MIL | 112 | 47% | 4.0% | .48 [.40, .57] | 148 | 36% | 9.3% | .41 [.20, .66] |
| M2 union of 12 fits | 160 | 36% | 16.2% | .55 | 165 | 32% | 7.8% | .39 |
| M3 Attention erasing | 166 | 34% | 10.7% | .53 [.49, .58] | 160 | 26% | 6.9% | .56 [.41, .70] |
| M4 Fixed schedule | 186 | 33% | 23.5% | .62 [.58, .66] | 186 | 29% | 19.8% | .64 [.51, .76] |
| Z-head + prog. masking | 203 | 27% | 34.0% | .73 | 198 | 29% | 29.3% | .60 |
| M5 PM-MIL (3 seeds) | 185 | 33% | 23.2% | .71 [.66, .77] | 165 | 34% | 17.9% | .63 [.48, .78] |

evidence is not a memory of the training discharges, although the lead there is 20 ms smaller. Twelve unmasked fits already move the earliest event to 150 ms on the pool and 138 ms on the evaluation set, since the union takes the earliest event over twelve models, so about two thirds of the pool gain is the union and the remaining 35 ms (paired difference 20 to 50 ms) is due to masking. The fixed schedule (M4) reaches 186 ms on the pool and on the evaluation set, with paired differences from PM-MIL that cover zero, local erasing (M3) 166 and 160 ms, and the single M2 model 112 and 148 ms on fewer DIS discharges. Progressive shortening with retraining therefore carries the lead gain, and the placement of the cutoff adds nothing at the median.

### 5.3 DEEPER CHAINS AND A RICHER VOCABULARY

On the pool the reference run's 2,697 source events consolidate into 655 mainline segments on 512 discharges (350 DIS, 162 ND), against 376 segments on 360 discharges for round 0 alone, all of which survive inside the PM-MIL chains. The share of chains with two or more events rises from 3.9% to 23.2% over three seeds, or from 3.6% to 16.4% when a chain must contain two distinct families, and family-to-family transitions from 16 to 132, of which 88 join different families. The unmasked ensemble reaches 11.1% with 54 transitions, the R1 union 12.2% and twelve pooled M2 models 16.2%, so repeated fits also produce multi-event chains, at a lower rate than progressive shortening. Local erasing (M3) stays at 10.7% [8.4, 12.5], the single M2 model at 4.0%, and the fixed schedule (M4) reaches 23.5% [22.8, 24.3] with 122 to 133 transitions. Progressive suffix removal and retraining produce deeper chains than these single-model controls, and Table 14 lists how many discharges each union covers. Neither evidence-chosen cutoffs nor the split head is required, since the Z-head runs under progressive masking reach 34.0% on the pool and 29.3% on the evaluation set, with events in 246 ND discharges against 162 and most of the extra depth on ND chains (Appendix C).

Table 6 (Appendix A.4) lists the six candidate families and their signal patterns. Progressive masking changes the composition. The share of segments in the three earliest families A–C rises from 19.4% to 38.8% and the largest-family share falls from 55.6% to 43.8%. Family C, whose events start a median 215 ms before the disruption, receives 156 of its 195 segments from rounds $\geq 1$. The same rounds also supply most of its ND events. The DIS fraction of family-C events falls from 0.95 for round 0 to 0.50 to 0.80 for rounds 3 to 7 and 0.25 for rounds 10 and 11 (Appendix B.4), so the late rounds, whose full-record AUPRC is lowest, contribute both the earliest DIS events and most of the ND events of this family. The A–C share of the controls follows the same order, 27% for M0-repeat, 26% for M3, 35% for M4, 39% for PM-MIL and 15% for single M2.

### 5.4 CHAIN CONTENT CARRIES INFORMATION ABOUT DISRUPTION

The Markov chain classifier of Table 2 combines event count, family identity and order, and methods cover different discharges. We therefore fit fixed-feature classifiers on matched cohorts of the development pool and apply them to the evaluation chains (Appendix B.5). The count, geometry, family and adjacency ladder was fixed before these classifiers were fitted, the family-only and fraction classifiers were added afterwards, and no multiplicity correction is applied, so single intervals are descriptive. For the reference run, event count alone reaches AUROC 0.865 on all 182 evaluation discharges, but 0.542 on the 61 event-bearing discharges. Much of its discrimination thus comes from event presence. The family-only classifier, which adds the six family counts to event count without geometry, reaches 0.754, an increment of 0.213 [0.094, 0.338] over count alone, with positive increments for the other two seeds (Table 13). Event composition therefore adds information beyond event number. Count plus record and event geometry reaches event-bearing AUROC 0.907, and adding family counts reaches 0.937, a paired increment of 0.031 [-0.002, 0.074]. Full directed adjacency reaches 0.880, with no consistent geometry-adjusted gain across the three seeds. The separate Markov shuffle control preserves each shot's count and family multiset. Its AUROC on the evaluation set falls from 0.713 to 0.692, a difference of 0.021 [0.002, 0.049], while the AP difference interval includes zero. This is limited evidence for order sensitivity, including the first and last families, rather than a robust independent adjacency gain. Only 5 evaluation chains contain distinct families. On identical PM-MIL/M4 cohorts, the Markov advantage persists on the pool but not on the evaluation set, and geometry-adjusted classifiers do not distinguish the methods.

The vocabulary itself is a representation-learning result. Table 5 (Appendix A.4) compares vocabularies built with the same clustering and chain classifier from raw features, attribution profiles, the contextual features of the same models and the cross-encoded local features used here. Only the local representation yields chains whose content predicts disruption at AUROC above 0.7, and with $K = 4$ or $K = 8$ the PM-MIL chain AUROC stays within 0.70 to 0.72 and Single MIL at chance.

C→D is the most frequent chain with exactly two events (47 of 123) and ends in DIS in 42 cases. In the reference development run, 110 of 132 multi-event chains end in DIS (83.3%), compared with 240 of 380 single-event chains (63.2%). Thus greater event count is associated with a higher disruption fraction in this event-bearing cohort. These are retrospective group fractions, not a prospective risk update after C.

**What starts an event and what changes at a transition.** Attributing the local-margin change at each event's first token to the 27 input channels (Appendix A.5) shows that the 743 rises from low to high risk (0.69 reference standard deviations) are carried mainly by magnetic-fluctuation (0.38) and equilibrium (0.19) channels. The change from C to D raises the magnetic contribution at the onset by 1.26 [1.02, 1.50] and the equilibrium contribution by 0.38 [0.32, 0.44] over 50 transitions, and D to C lowers the magnetic contribution by 0.40 [0.09, 0.72].

## 6 Discussion and limitations

Refitting twelve unmasked models accounts for much of the full-record classification gain and earlier lead but leaves shallower chains. A fixed schedule reproduces the lead, de-concentration and depth of PM-MIL on the pool and on the evaluation set. Progressive shortening with retraining is therefore a supported mechanism. Event-guided cutoffs remove the need to choose a step, but M4 was run with one step of 25 ms, so this is a convenience rather than a demonstrated advantage. The twelve-fit M2 union also produces multi-event chains and truncation robustness (Sections 5.1 and 5.2), so progressive retraining is favored for chain depth but is not necessary for either.

The related-work aggregators score highly on full records and concentrate on late tokens. They differ in both their heads and auxiliary losses, whereas Z-head isolates the scoring head. Progressive Z-head reproduces the principal classification and chain effects and also finds seed-stable locations. Its scores can depend on the surrounding record, which does not show that its events cannot recur elsewhere. The split head instead gives an architectural guarantee that a fixed local input has the same score and representation under a fixed model and normalizer. This common meaning supports a shared candidate-event vocabulary. Because the complete detector still uses contextual residuals, record-level thresholds and confirmation, neither head establishes context-invariant detection or physical identity.

**Candidate disruption-avoidance strategies.** The chains suggest three candidate intervention targets. The early displacement-associated families A and C (Table 6) point to position or equilibrium regulation before instability develops. The frequent C→D chain (Section 5.4), at which the magnetic-fluctuation contribution to the onset rises, points to stabilization during C to prevent entry into the enhanced magnetic-fluctuation pattern D, in the spirit of feedback avoidance (Seo et al., 2024). The timed event representation points to intervention timing adapted to event identity and elapsed duration rather than a common waiting time, without assuming that longer duration is more dangerous. They nominate control channels and transitions, not actuator commands, and do not show that C causes D, that D is a locked mode or that a backward transition means recovery, and retrospective leads are not actuator response windows, so these are observational hypotheses rather than validated control policies.

Only 39 disruptions in the evaluation set limit comparisons among progressive-shortening methods. Event lead is retrospective. The 50-ms prefix floor makes nominal suffix removal differ from actual removal, while contextual features and event confirmation also preclude an operational warning-time claim. Count, type and adjacency are distinct information sources, and comparisons of chain classifiers must use the same discharges and folds. Candidate families still require physical validation against expert-annotated tearing, locking and impurity signatures, particularly for events on ND discharges. Warm starts, a comparison with feature-based predictors and a second device were not tested.

## 7 Conclusion

A model given only discharge-level labels can locate candidate precursor evidence, and progressive shortening with retraining moves that evidence earlier on development and evaluation records. Local scoring and representation provide a common basis for a cross-discharge candidate vocabulary. Event presence is predictive, family composition adds information beyond count, and the evidence for order effects is limited. Classification remains informative after nominal suffix removal, and the chains nominate avoidance targets that still need physical and control validation.

### AI USE STATEMENT

This work used generative AI assistants (Claude and Codex, and one further assistant for a simulated review round), as declared at abstract submission. Among the uses that require disclosure, they helped design and gave feedback on research methodology and experiments, namely proposing and specifying the controls, the ablations and the evaluation protocols and running a simulated review round whose comments the authors adjudicated, they implemented methods, namely the training, evaluation and analysis procedures, and they helped interpret results, namely drafting readings of the measured numbers that the authors then checked against the measurements. They were not used to generate synthetic data, to develop theoretical models or conceptual frameworks, to formulate or prove mathematical claims, to clean or reformat the dataset, to assist with translation, or to support qualitative or thematic analysis. Among the uses for which disclosure is recommended, they also drafted parts of the manuscript, edited it for readability, searched for and summarized related work, formatted references and prepared figures. They were not the source of any experimental measurement. We have reviewed all AI-assisted work. Every claim, number and citation was verified by the human authors against the measurements and the cited sources, the authors take responsibility for the final content of this work, and the assistants are not authors.

### ETHICS STATEMENT

The discharges analyzed here come from the J-TEXT tokamak database and record machine diagnostics only, so the work involves no human subjects, no personal data and no crowdsourcing. The data were obtained under the access rules of the operating institution and cannot be redistributed publicly, which is why the reproducibility statement offers them under a data-use agreement rather than as a download. The method locates candidate precursor events and does not control the device. The chains are retrospective candidates that still need physical validation, so the proposed avoidance candidates must not be treated as validated control policies, used to trigger mitigation or used to certify an operating limit without that validation, and the limitations section states where the evidence stops. The authors declare no competing interests, and the AI use statement lists every use of generative AI in this work.

### REPRODUCIBILITY STATEMENT

The appendix specifies the training loop, the event grammar and consolidation, the representation and vocabulary, the controls and every evaluation protocol, and it distinguishes fit-time masking, event extraction, local explanations and the final ensemble. The J-TEXT diagnostic data cannot be released publicly, but they are available from the authors under a data-use agreement. The code will be made public on GitHub.

## REFERENCES

Xiwen Chen, Peijie Qiu, Wenhui Zhu, Huayu Li, Hao Wang, Aristeidis Sotiras, Yalin Wang, and Abolfazl Razi. TimeMIL: Advancing multivariate time series classification via a time-aware multiple instance learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7190–7206, 2024. URL https://proceedings.mlr.press/v235/chen24af.html.

P. C. de Vries, M. F. Johnson, B. Alper, P. Buratti, T. C. Hender, H. R. Koslowski, V. Riccardo, and JET-EFDA Contributors. Survey of disruption causes at JET. *Nuclear Fusion*, 51(5): 053018, 2011. doi: 10.1088/0029-5515/51/5/053018. URL https://iopscience.iop.org/article/10.1088/0029-5515/51/5/053018.

Joseph Early, Gavin KC Cheung, Kurt Cutajar, Hanting Xie, Jas Kandola, and Niall Twomey. Inherently interpretable time series classification via multiple instance learning. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2311.10049.

Charles M. Grinstead and J. Laurie Snell. *Introduction to Probability*. American Mathematical Society, second revised edition, 1997. URL https://chance.dartmouth.edu/teaching_aids/books_articles/probability_book/pdf.html.

T. C. Hender, J. C. Wesley, J. Bialek, A. Bondeson, A. H. Boozer, R. J. Buttery, et al. Chapter 3: MHD stability, operational limits and disruptions. *Nuclear Fusion*, 47(6):S128–S202, 2007. doi: 10.1088/0029-5515/47/6/S03. URL https://iopscience.iop.org/article/10.1088/0029-5515/47/6/S03.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Julian Kates-Harbeck, Alexey Svyatkovskiy, and William Tang. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature*, 568:526–531, 2019. doi: 10.1038/s41586-019-1116-4. URL https://www.nature.com/articles/s41586-019-1116-4.

K. J. Montes, C. Rea, R. S. Granetz, R. A. Tinguely, N. Eidietis, O. M. Meneghini, D. L. Chen, B. Shen, B. J. Xiao, K. Erickson, and M. D. Boyer. Machine learning for disruption warnings on Alcator C-Mod, DIII-D, and EAST. *Nuclear Fusion*, 59(9):096015, 2019. doi: 10.1088/1741-4326/ab1df4.

A. Pau, A. Fanni, S. Carcangiu, B. Cannas, G. Sias, A. Murari, F. Rimini, and JET Contributors. A machine learning approach based on generative topographic mapping for disruption prevention and avoidance at JET. *Nuclear Fusion*, 59(10):106017, 2019. doi: 10.1088/1741-4326/ab2ea9. URL https://iopscience.iop.org/article/10.1088/1741-4326/ab2ea9.

Sujoy Paul, Sourya Roy, and Amit K. Roy-Chowdhury. W-TALC: Weakly-supervised temporal activity localization and classification. In *European Conference on Computer Vision (ECCV)*, 2018. doi: 10.1007/978-3-030-01225-0_35.

C. Rea, K. J. Montes, K. G. Erickson, R. S. Granetz, and R. A. Tinguely. A real-time machine learning-based disruption predictor in DIII-D. *Nuclear Fusion*, 59(9):096016, 2019. doi: 10.1088/1741-4326/ab28bf. URL https://iopscience.iop.org/article/10.1088/1741-4326/ab28bf.

S. A. Sabbagh, J. W. Berkery, Y. S. Park, J. Butt, J. D. Riquezes, J. G. Bak, R. E. Bell, L. Delgado-Aparicio, S. P. Gerhardt, C. J. Ham, J. Hollocombe, J. W. Lee, J. Kim, A. Kirk, J. Ko, W. H. Ko, L. Kogan, B. P. LeBlanc, J. H. Lee, A. Thornton, and S. W. Yoon. Disruption event characterization and forecasting in tokamaks. *Physics of Plasmas*, 30(3):032506, 2023. doi: 10.1063/5.0133825.

Jaemin Seo et al. Avoiding fusion plasma tearing instability with deep reinforcement learning. *Nature*, 626:746–751, 2024. doi: 10.1038/s41586-024-07024-9. URL https://www.nature.com/articles/s41586-024-07024-9.

Samuele Setzu, Enrico Aymerich, Alessandra Fanni, Fabio Pisano, Giuliana Sias, Barbara Cannas, JET Contributors, and WPTE Team. A procedure for rule extraction from a self-organising plasma disruption predictor for JET. *Scientific Reports*, 16:16931, 2026. doi: 10.1038/s41598-026-38318-9. URL https://www.nature.com/articles/s41598-026-38318-9.

Chengshuo Shen, Wei Zheng, Yonghua Ding, Xinkun Ai, Fengming Xue, Yu Zhong, Nengchao Wang, Li Gao, Zhipeng Chen, Zhoujun Yang, Zhongyong Chen, Yuan Pan, and J-TEXT team. IDP-PGFE: An interpretable disruption predictor based on physics-guided feature extraction. *Nuclear Fusion*, 63:046024, 2023. doi: 10.1088/1741-4326/acbe0f. URL https://arxiv.org/abs/2208.13197.

R. A. Tinguely, K. J. Montes, C. Rea, R. Sweeney, and R. S. Granetz. An application of survival analysis to disruption prediction via random forests. *Plasma Physics and Controlled Fusion*, 61 (9):095009, 2019. doi: 10.1088/1361-6587/ab32fc.

Jesús Vega, Andrea Murari, Sebastián Dormido-Canto, Giuseppe A. Rattá, Michela Gelfusa, and JET Contributors. Disruption prediction with artificial intelligence techniques in tokamak plasmas. *Nature Physics*, 18:741–750, 2022. doi: 10.1038/s41567-022-01602-2. URL https://www.nature.com/articles/s41567-022-01602-2.

Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1568–1576, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Wei_Object_Region_Mining_CVPR_2017_paper.html.

Zhengzheng Xing, Jian Pei, and Philip S. Yu. Early classification on time series. *Knowledge and Information Systems*, 31:105–127, 2012. doi: 10.1007/s10115-011-0400-x.

Runhao Zeng, Chuang Gan, Peihao Chen, Wenbing Huang, Qingyao Wu, and Mingkui Tan. Breaking winner-takes-all: Iterative-winners-out networks for weakly supervised temporal action localization. *IEEE Transactions on Image Processing*, 28(12):5797–5808, 2019. doi: 10.1109/TIP.2019.2922108. URL https://peihaochen.github.io/files/publications/Breaking.pdf.

Wei Zheng, Fengming Xue, Zhongyong Chen, et al. Disruption prediction for future tokamaks using parameter-based transfer learning. *Communications Physics*, 6:181, 2023. doi: 10.1038/s42005-023-01296-9. URL https://www.nature.com/articles/s42005-023-01296-9.

# A Method details and data

## A.1 Data and reference run

The run analyzed here is a single-stage sequence with checkpoints selected by validation AP, followed by the local-representation vocabulary and the excess-AP soft ensemble. All reported events, families and ensemble weights come from this run and its seed replicates.

The data set has 412 DIS and 1,404 ND discharges. The reference split contains 373 DIS/1,261 ND in the development pool and 39 DIS/143 ND in the evaluation set. The event replay excludes nine short DIS records, leaving 364 DIS/1,261 ND in development and all 182 evaluation discharges. Inputs are stored engineering-feature records that start 0.15 s after breakdown and end at the labeled end time of the discharge, which for DIS discharges is the disruption time recorded in the database (the record end is the last sample before it) and for ND discharges the end of the stored discharge. Preprocessing takes the common available channel length and retains at most the first 5,150 samples, a cap reached by two DIS discharges (3 ms and 1.4 ms lost). In the development pool the DIS records have median length 256 ms (10th to 90th percentiles 109 to 379 ms) and the ND records 348 ms (343 to 450 ms). Lead times are measured to the record end, which for DIS discharges is the disruption time.

The feature correction sets the derived dominant Mirnov frequency and amplitude to their physical-zero encodings where the stored frequency exceeds 20 kHz, retaining the pre-existing storage normalization. This is a deterministic correction to derived features, not a remeasurement of raw waveforms. Per-round normalization uses training-visible inputs only.

## A.2 Training details

Algorithm 3 states the procedure. Each round solves the MIL objective on the visible prefixes only, every round starts from a fresh initialization, the visible set of a disruptive discharge can only shrink, and the loop ends when no admissible model can be fitted or no mask advances, so it terminates after at most $\max_i K_i$ rounds in principle and after twelve in the present run.

The backbone has five grouped local convolutional branches (one per diagnostic group, kernel and stride of 50 samples) totaling 190 dimensions. The local classifier has a 64-dimensional GELU hidden layer, and the contextual encoder has three Transformer layers with five heads. Training uses learning rate $5 \times 10^{-4}$, batch size two with gradient accumulation over sixteen batches, at most 200 epochs, early-stopping patience thirty and ten warm-up epochs. The loss is a bag focal loss (focusing parameter 2, positive-class weight 0.25) plus a suppression term on disruptive token scores of ND bags plus next-token representation-prediction losses on the local and contextual branches, each comparing a predicted next representation with a stop-gradient target. Every round starts from

Table 3: Algorithm 1. Progressive-masking MIL (training loop).

**Input** bags $X_i$ with labels $y_i$, cutoffs $b_i^{(0)} = K_i$, round limit $R_{\max}$
**for** $r = 0, 1, \ldots, R_{\max} - 1$
split the active discharges into training and validation, fit the normalizer on visible tokens
initialize $f_r, u_r$ at random and minimize the MIL loss on prefixes $V_i^{(r)}$ (ND bags length-matched)
select the checkpoint with maximal validation AP, admit it if AP exceeds the validation DIS fraction, else stop
set the ND bag gate at 5% empirical FPR on the pool
**for** each DIS discharge $i$ classified positive above the gate
$\widehat{s}_i^{(r)} \leftarrow$ latest eligible residual peak followed by a local-risk rise, $b_i^{(r+1)} \leftarrow \min\{b_i^{(r)}, \widehat{s}_i^{(r)}\}$
retire discharges by the false-negative and no-candidate rules
**if** no mask advanced for the set patience **then** stop
**Output** frozen models $\{f_r, u_r\}$ with their normalizers, used only through full-record replay afterwards

random initialization, and each checkpoint is paired with its own normalizer. A stochastic attention regularizer drops the highest-attention 10% of visible tokens with probability 0.4 during fitting. It is used in the shared-backbone controls but omitted by R1 and R2, and is distinct from persistent suffix masking. Checkpoint selection maximizes validation AP, and validation prevalence is recomputed within each round. The completed twelve-round fit took approximately 82 minutes. Table 4 lists the active discharges, the AP of every round on the evaluation set and the ensemble weights. From round 8 on fewer than 65 DIS discharges remain, so the late rounds are fitted on small cohorts and reach low full-record accuracy, which the weights reflect.

Table 4: The twelve rounds of the PM-MIL reference run (seed 42). Active DIS discharges are those still masked and refitted in the round, split into training and validation. Evaluation AP is the full-record AUPRC of that round's single model on the 182 evaluation discharges, for PM-MIL and for the fixed schedule (M4) with the same seed, whose round 0 is the same model. Pool AP is the same quantity on the 1,625 development discharges. The last column is the excess-AP ensemble weight of Eq. 4.

| Round | Active DIS | Train DIS | Val DIS | Eval. AP PM | Eval. AP M4 | Pool AP PM | Pool AP M4 | Weight $w_r$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 364 | 309 | 55 | .857 | .857 | .989 | .989 | .168 |
| 1 | 344 | 292 | 52 | .836 | .806 | .929 | .968 | .163 |
| 2 | 308 | 261 | 47 | .753 | .825 | .769 | .918 | .141 |
| 3 | 278 | 236 | 42 | .676 | .835 | .814 | .911 | .121 |
| 4 | 229 | 194 | 35 | .474 | .725 | .710 | .770 | .068 |
| 5 | 182 | 154 | 28 | .491 | .698 | .540 | .844 | .072 |
| 6 | 151 | 128 | 23 | .515 | .674 | .578 | .798 | .079 |
| 7 | 122 | 103 | 19 | .447 | .652 | .470 | .728 | .061 |
| 8 | 65 | 55 | 10 | .412 | .658 | .445 | .738 | .052 |
| 9 | 43 | 36 | 7 | .388 | .617 | .390 | .701 | .046 |
| 10 | 32 | 27 | 5 | .225 | .458 | .303 | .538 | .003 |
| 11 | 17 | 12 | 5 | .315 | .553 | .296 | .639 | .026 |

**Training-time mask candidates.** Residual peaks use the shot-round $\mu + 1.75\sigma$ rule. Let $p_-$ be the preceding two-token risk mean and $p_+$ the largest mean over two-token windows starting one, two or three tokens after the peak, truncated at the visible endpoint. Eligibility requires $p_+ > p_- + 0.05$, or both means above 0.5, and an escape rule admits a residual z-score above 3.5 when $p_+ > p_-$ and $p_+ > 0.20$. The latest eligible peak is selected. If none exists, a fallback uses the latest start of at least two consecutive tokens above $\mathrm{mean}(p) + \mathrm{sd}(p)$ with at least 40% above-threshold occupancy to

the visible endpoint. A mask update requires a positive bag decision at the checkpoint's validation-F1 threshold and passage of the empirical 5% ND gate. For the $n$th consecutive false-negative decision a DIS discharge is retained only when $n \leq 5$ and its bag score exceeds $\max(0.03, \gamma_n \theta_{\mathrm{val}})$ with $\gamma_n = 0.3[1 - 1/\log(n+e)]$. Additional retirement occurs when the positive-without-candidate counter reaches four or fewer than ten visible tokens remain. These operational rules do not establish that evidence has been exhausted.

### A.3 Event grammar and consolidation

The frozen-model replay uses full records without training masks or ND length matching. A model-specific ND-calibrated bag threshold screens event extraction at an empirical 5% FPR (63 of 1,261 ND discharges per round). The union of twelve catalogs admits 162 ND discharges (12.8%), and this gate is neither the final ensemble threshold nor a bound on the union. Token risk uses the local classifier, with upper and lower levels at the per-model ND token 0.99 and 0.90 quantiles and three-token persistence. The contextual residual at source token $k$ predicts target $k + 1$ and is assigned to the target, its squared components are summed, and the threshold is the shot-round $\mu + 1.75\sigma$ within the searchable region, which excludes the first three tokens. A trigger is latched until a subsequent high-risk episode is confirmed, with no finite maximum delay. Events can also start from residual peaks inside an already high-risk episode (1,954 of the 2,697 source events), so an earlier timestamp can denote the boundary of already high-risk activity rather than a newly confirmed rise.

All 2,697 original intervals remain available, and exact duplicates give 2,356 clustering objects whose memberships map back to every source event. Each token takes the family of its earliest covering source round, and contiguous equal-family ownership is merged without bridging gaps. For round-0 events alone this rule reduces to the events themselves with contiguous same-family merging, which yields the 376 Single-MIL segments of Section 5. The original trigger of an event is kept even when its mainline boundary is cropped.

### A.4 Representation and local explanation

All twelve models encode all intervals with their own normalizers (normalized inputs clipped to $[-10, 10]$). Block centering and total-RMS scaling are fitted to event descriptors. K-means uses $K = 6$, twenty initializations, a 300-iteration limit and seed 42, and candidate comparisons included raw and shot-centered features, local and contextual representations, alternative resolutions and clustering algorithms. Descriptor-space silhouette is 0.264 (minimum family silhouette 0.175, and 2.7% of objects lie on a non-positive boundary). These are internal-geometry measures, not physical validation. The frozen vocabulary is built from the encoders of the PM-MIL reference run. Rebuilding it for every twelve-model method from that method's own accepted-round encoders, with the same descriptors, centering, K-means with $K = 6$, fifty initializations and a 1,000-iteration limit, and the same chain classifier, gives event-bearing chain AUROC .72 for PM-MIL (three runs .72, .73, .72), .58 for the fixed schedule, .56 for attention erasing, .55 for the unmasked ensemble and .70 for the Z-head progressive runs (shared vocabulary .73), the same order as under the shared vocabulary, so the shared vocabulary does not favor the run it was built from. Clustering the DIS objects alone with the same protocol and assigning the ND objects to the nearest DIS cluster places 95.7% of them inside that cluster's 95th-percentile radius over its DIS members, and the ND events of the reference run are reproduced by independent seeds at the same rate above chance as its DIS events (Appendix C), supporting descriptor overlap and above-chance reproducibility without establishing physical event identity.

Local Shapley values explain the local classifier margin $g_r(x) = \ell_{r,\mathrm{DIS}}(x) - \ell_{r,\mathrm{ND}}(x)$ of a 5-ms window. To compare models on common channels the target is $G(x) = \frac{1}{R} \sum_r g_r(x)/\sigma_{r,\mathrm{ND}}$, and by Shapley linearity its attribution is the same scaled average of per-model attributions. Estimation is Monte-Carlo interventional with paired forward and reversed channel permutations, and replacing a semantic channel replaces its full window with the corresponding channel of a sampled ND reference window (256 references, 256-window sensitivity batch, reference universe excluding detected ND event positions). Numerical checks give maximum discrepancies of $1.5 \times 10^{-5}$ for the fused-affine forward equivalence and $4.6 \times 10^{-5}$ for additivity. Family profiles use equal-shot weighting (Figure 9), and the descriptions in Table 6 combine the top-3 raw-signal deltas (event-window mean minus same-shot record median) with the top-3 attributions. Figure 8 shows one complete discharge.

Table 5: Vocabulary source and chain accuracy on the 512 PM-MIL chains. Each candidate is clustered from the same 2,356 objects with the same K-means protocol and its own stability-selected $K$, and scored by out-of-fold Markov and ordered-logistic chain classifiers with scaling and clustering refitted inside each fold. The last column is the minimum ARI with the full solution when one encoder is left out.

| Event descriptor | $K$ | Markov AUROC | Ordered AUROC | Leave-one-encoder ARI |
|---|---|---|---|---|
| Raw features, shot-centered | 5 | .622 | .615 | — |
| Raw features, pre-event delta | 4 | .612 | .627 | — |
| Attribution profile (SHAP only) | 6 | .632 | .639 | — |
| Contextual $z$, twelve encoders | 5 | .666 | .674 | .40 |
| Local $h$, twelve encoders (this work) | 6 | **.707** | **.724** | .83 |

Table 6: Candidate event families of the shared vocabulary (development pool). Descriptions summarize the top raw-signal deltas and local-margin attributions (Appendix A.4) and name signal patterns. Segments are mainline counts for Single MIL → PM-MIL, lead is the PM-MIL median DIS segment lead, and next→DIS uses all PM-MIL visits of the family as denominator. Here $dx$ and $dy$ denote horizontal and vertical displacement, Mirnov the magnetic-fluctuation amplitude, $m$ and $n$ the poloidal and toroidal mode numbers, $q_a$ the edge safety factor, and SXR and AXUV the soft-X-ray and extreme-ultraviolet arrays.

| | Family | Candidate signal pattern | Objects | Segments | Lead (ms) | Next→DIS |
|---|---|---|---|---|---|---|
| early | A | $dy/dx$ displacement excursion with $I_p$ and $B_t$ weighted in the local margin | 131 | 16 → 35 | 265 | 12/35 |
| | B | small $dx$ shift with $m$-number and frequency change and a $q_a$-weighted margin | 114 | 18 → 24 | 235 | 16/24 |
| | C | $dx$ rise with Mirnov-amplitude and $m$-number decrease | 492 | 39 → 195 | 215 | 45/195 |
| | D | Mirnov-amplitude rise with mode-number change | 1,248 | 209 → 287 | 160 | 197/287 |
| late | E | vertical-profile variance rise with $n$-number change and C III increase | 264 | 64 → 77 | 85 | 57/77 |
| | F | SXR core/profile variance rise with AXUV-array variance increase | 107 | 30 → 37 | 80 | 23/37 |

### A.5 Onset attribution

Each event of the reference run starts at a token $k$ where the forward residual peaks, either with a rise from low to high risk (743 events) or inside a high-risk episode (1,954 events). For the model of the event's source round, the exact local margin of the family profiles is evaluated on the windows $k-3$ to $k$, and the onset target $m_k - \frac{1}{3}(m_{k-3} + m_{k-2} + m_{k-1})$ is attributed to the 27 channels by the antithetic permutation Shapley estimator with two disjoint batches of 256 ND reference windows. Because Shapley values are linear in the function, the attribution of the target is the same combination of the window attributions, and the channel sum reproduces the target to within $10^{-5}$. Attributions are divided by the standard deviation of the reference margin of the round. The two reference batches give group-level attributions with correlation 0.999. Table 7 gives shot-balanced group means for the rises from low risk, per family, and the change of the onset attribution from one mainline segment to the next for family transitions with at least five cases, with 95% intervals from 2,000 stratified discharge resamples in Section 5.4. Family D onsets are carried by

mode frequency and poloidal mode number, with 0.57 from the magnetic group, whereas family B and E onsets are carried by the horizontal displacement, with equilibrium contributions of 0.49 and 0.32. The 1,954 events that begin inside a high-risk episode show a mean margin change of -0.09, so their start is set by the contextual forward residual rather than by a local change. These attributions describe model dependence, not physical causation.

Table 7: Onset attribution of the reference run by physics group, in reference standard deviations of the local margin. Top, events that begin with a rise from low to high risk, per family (jump is the margin change). Bottom, change of the onset attribution from one mainline segment to the next along chains, per family transition.

| | $n$ | Jump | MHD | Radiation | Density | Equilibrium | $Z$ profile |
|---|---|---|---|---|---|---|---|
| A | 68 | 0.24 | +0.12 | +0.00 | +0.00 | +0.10 | +0.00 |
| B | 31 | 0.67 | +0.08 | +0.02 | +0.06 | +0.49 | +0.02 |
| C | 117 | 0.56 | +0.20 | +0.03 | +0.01 | +0.20 | +0.12 |
| D | 364 | 0.85 | +0.57 | +0.08 | +0.02 | +0.15 | +0.04 |
| E | 111 | 0.67 | +0.19 | −0.02 | +0.03 | +0.32 | +0.14 |
| F | 52 | 0.49 | +0.32 | +0.03 | +0.03 | +0.08 | +0.02 |
| C→D | 50 | – | +1.26 | +0.19 | +0.03 | +0.38 | +0.08 |
| D→C | 15 | – | −0.40 | +0.11 | +0.10 | −0.05 | +0.01 |
| C→E | 9 | – | +0.67 | −0.13 | +0.17 | −0.28 | −0.35 |

# B Additional numerical checks

## B.1 Ensemble weights and fixed operating points

The twelve ensemble weights are listed in Table 4. The full-record precision–recall curves are in Figure 6. The exact ensemble threshold is $\tau = 0.3155069$ (strict, ties ND), and the round-0 threshold is $0.3614504$. Both give seven false positives among the 143 ND discharges of the evaluation set. At 4/143 and 1/143 ND false positives, TPR is 74.4% and 28.2% for Single MIL, 87.2% and 59.0% for M0-repeat, 61.5% and 39.3% for M2, 68.4% and 45.3% for M3, 76.9% and 53.8% for M4 and 76.9% and 60.7% for PM-MIL (seed means). Exact AP values are 0.9033811 (ensemble) and 0.8569957 (round 0), and the confusion matrices $(TN, FP, FN, TP)$ on the evaluation set are $(136, 7, 8, 31)$ and $(136, 7, 3, 36)$. These are unmatched full-record classification results. A separate frozen event replay supplies the chain results on the evaluation set.

## B.2 Ensemble false negatives

Table 8 lists the eight DIS discharges of the evaluation set below the ensemble threshold $\tau = 0.316$ with their round-0 probability, the mean over rounds 1 to 11 and the ensemble score. Six exceed the Single-MIL threshold 0.361 at round 0. Taking the maximum over rounds instead of the weighted average gives AP 0.819 and the same TPR, and the plain mean gives AP 0.820, so among fixed rules the excess-AP average is the best and the remedy is a learned combination. Table 4 gives the per-round AP on the evaluation set of the reference run and of the fixed schedule with the same seed. Under the fixed schedule rounds 2 to 11 are stronger full-record classifiers, the weights are spread more evenly, and the unmatched ensemble reaches .844 against .903.

## B.3 Bootstrap intervals

Shot-level bootstrap with 2,000 stratified resamples of the 1,625 development discharges gives, for Single MIL and PM-MIL respectively, median earliest lead 55 to 75 ms and 170 to 200 ms, final-100-ms share 65 to 75% and 28 to 36%, multi-event share 2 to 6% and 22 to 29%, and event-bearing chain AUROC .38 to .59 and .66 to .77 (the last from resampling fixed out-of-fold predictions). Paired differences from joint resampling are 105 to 140 ms in median lead, 34 to 43 percentage points in the final-100-ms share and 18 to 26 percentage points in the multi-event share. Against the

Table 8: DIS discharges of the evaluation set missed by the PM-MIL ensemble.

| Discharge | $q_0$ | mean $q_{1..11}$ | max $q_{4..11}$ | $S$ |
|---|---|---|---|---|
| 1053042 | .958 | .109 | .141 | .299 |
| 1058040 | .812 | .149 | .243 | .315 |
| 1054187 | .764 | .075 | .020 | .249 |
| 1053761 | .657 | .076 | .018 | .230 |
| 1057556 | .629 | .112 | .224 | .243 |
| 1057474 | .463 | .151 | .335 | .268 |
| 1057172 | .091 | .158 | .276 | .146 |
| 1055374 | .001 | .182 | .353 | .202 |

PM-MIL reference run, M0-repeat differs by $-50$ to $-20$ ms in median lead, $+3.6$ to $+10.2$ points in the final-100-ms share and $-18.7$ to $-10.5$ points in the multi-event share. For the three M4 seeds the paired lead difference is $-10$ to $+15$, $-10$ to $+17.5$ and $-10$ to $+15$ ms and the multi-event difference $-6.4$ to $+0.5$, $-6.3$ to $+1.5$ and $-5.4$ to $+2.4$ points, while PM-MIL seeds 43 and 44 differ from the reference run by $-7.5$ to $+0.9$ and $-8.4$ to $-0.6$ points. For M3 the paired lead difference is $-35$ to $-5$, $-37.5$ to $-5$ and $-35$ to $-5$ ms and the multi-event difference $-17.3$ to $-9.2$, $-21.2$ to $-13.6$ and $-18.5$ to $-10.2$ points, and for M2 the lead difference is $-80$ to $-30$, $-105$ to $-70$ and $-100$ to $-68$ ms and the multi-event difference $-27.5$ to $-18.9$, $-26.3$ to $-18.3$ and $-24.2$ to $-15.5$ points. Table 9 gives the paired differences of classification on the evaluation set and Table 10 those of the chain classifier. On the evaluation set, the chain AUROC of every method lies inside the interval of the reference run, and the pool margin of PM-MIL over the fixed schedule is not reproduced by 39 disruptions.

Table 9: Paired differences in AUPRC on the length-matched evaluation records (2,000 common shot-level resamples, 95% intervals), at nominal suffix-removal settings $\Delta$ with a 50-ms minimum retained prefix.

| PM-MIL minus | $\Delta = 0$ | 30 ms | 100 ms | 200 ms | 300 ms |
|---|---|---|---|---|---|
| M4 fixed schedule | [−.01, +.17] | [−.03, +.16] | [−.03, +.09] | [−.02, +.06] | [−.03, +.05] |
| M3 attention erasing | [−.01, +.19] | [−.01, +.19] | [+.02, +.23] | [+.07, +.27] | [+.08, +.30] |
| M2 Masked MIL | [+.01, +.18] | [−.00, +.19] | [+.02, +.15] | [+.04, +.16] | [+.03, +.16] |
| M2 union of 12 fits | [−.04, +.14] | [−.05, +.14] | [−.04, +.06] | [−.02, +.10] | [−.04, +.13] |
| progressive Z-head | [−.05, +.04] | [−.08, +.04] | [−.07, +.06] | [−.06, +.05] | [−.07, +.04] |
| M0-repeat | [−.07, +.16] | [−.01, +.22] | [+.06, +.29] | [+.16, +.38] | [+.17, +.41] |
| M1 Single MIL | [−.05, +.17] | [+.04, +.29] | [+.13, +.40] | [+.20, +.44] | [+.23, +.47] |
| R1 attention MIL | [−.09, +.01] | [−.06, +.11] | [+.05, +.28] | [+.13, +.38] | [+.14, +.40] |
| R1 union of 12 fits | [−.09, +.01] | [−.06, +.11] | [+.04, +.27] | [+.13, +.39] | [+.15, +.43] |
| R2 TimeMIL-style | [−.07, +.02] | [−.00, +.13] | [+.10, +.30] | [+.13, +.36] | [+.13, +.36] |

### B.4 Transition accounting and chain classifiers

For PM-MIL, START exits by family are $(27, 22, 166, 214, 51, 32)$, summing to 512, and outgoing family totals are $(35, 24, 195, 287, 77, 37)$, summing to 655. DIS exits are $(12, 16, 45, 197, 57, 23)$ and ND exits $(12, 5, 76, 49, 10, 10)$, and the remaining 143 transitions connect families, 91 of them between different families. For Single MIL, START exits are $(15, 18, 32, 206, 60, 29)$ summing to 360, outgoing totals $(16, 18, 39, 209, 64, 30)$ summing to 376, DIS exits $(12, 16, 33, 187, 55, 22)$ and ND exits $(4, 2, 1, 16, 6, 6)$, and 16 transitions connect families. Eventual DIS absorption for PM-MIL, $(I-Q)^{-1}R_T$, is $(.547, .765, .512, .785, .839, .705)$ for A–F, with START-weighted absorption $350/512$.

The Markov chain classifier of Table 2 fits a class-conditional first-order transition table over START, the six families and a generic END within each training fold, with additive smoothing

Table 10: Out-of-fold chain AUROC on event-bearing chains and the paired difference of the PM-MIL reference run minus the method (shot-level bootstrap on the union of the two cohorts, 2,000 resamples, 95% intervals), on the development pool and on the evaluation discharges. The reference run scores .71 [.66, .77] on the pool and .63 [.48, .78] on the evaluation set.

| | Development pool | | Evaluation set | |
|---|---|---|---|---|
| Method | AUROC | Reference − method | AUROC | Reference − method |
| PM-MIL seeds 43 and 44 | .71 | [−.05, +.05] | .69 | [−.19, +.06] |
| M4 fixed schedule | .62 | [+.05, +.14] | .64 | [−.11, +.09] |
| M3 attention erasing | .53 | [+.13, +.24] | .56 | [−.04, +.20] |
| M2 Masked MIL | .48 | [+.14, +.33] | .41 | [−.03, +.46] |
| M0-repeat | .53 | [+.11, +.26] | .62 | [−.12, +.15] |
| M1 Single MIL | .49 | [+.11, +.35] | .58 | [−.13, +.25] |

0.5. Test scores are log-likelihood ratios plus the fitted class prior. Actual DIS/ND endpoints never enter the input. On event-bearing development cohorts, Single MIL ($n = 360$, 325 DIS) reaches AUROC 0.487 and AP 0.903 at prevalence 0.903, while PM-MIL ($n = 512$, 350 DIS) reaches AUROC 0.714 and AP 0.803 at prevalence 0.684. High AP alone therefore does not isolate information in order. On the whole pool ($n = 1{,}625$, 364 DIS), the respective AUROCs are 0.924 and 0.936 and APs 0.846 and 0.780. Refitting representation scaling and K-means within each fold gives PM-MIL AUROC 0.707 and AP 0.792. These older results use their original folds. The additional matched ablations below distinguish count, geometry, family identity and adjacency explicitly and apply classifiers fitted on the development pool to the evaluation shots, unlike Table 2, whose evaluation-set classifier is fitted out of fold within the evaluation cohort.

### B.5 Separating event count, family identity and order

We fixed the analysis protocol before computing these ablations and use only frozen chains, source-event intervals and shot audits. No discovery model, event threshold or vocabulary is refitted. The primary populations are the reference run's event-bearing development and evaluation cohorts. Empty chains are retained in a secondary all-shot analysis. Seeds 43 and 44 are reported separately. Comparisons against M4 and the M2 union each use the pairwise intersection of event-bearing shots, with identical labels and folds for both methods.

Count features are indicators for depths 0, 1, 2, 3 and at least 4, plus log one plus depth. Geometry adds log record duration, covered fraction, log one plus covered duration, normalized first-event position and last covered endpoint, and event presence. Complete-record geometry is a retrospective control, not a prospective warning feature. Family counts add six coordinates. Direction adds counts of increases, decreases and self transitions in the A–F index, together with signed and absolute displacement. This index was ordered by development DIS median lead and does not encode physical progression or recovery. First and last families are added separately before the 36 directed internal transition counts. The final model therefore tests first-order adjacency, not arbitrary long-range order. Scaling is fitted on training rows only, and all stages use L2 logistic regression with $C = 1$, no class reweighting and no tuning. More features still mean more fitted parameters, so a negative increment does not prove that the information is absent.

Development classifiers use five repetitions of stratified five-fold splits, shared across features and methods. Metrics are averaged over repeat-specific predictions. For the evaluation set, coefficients and scaling are fitted once on the development cohort and frozen. Discovery and vocabulary fitting already used development data, so its out-of-fold designation applies only to the chain classifier. The evaluation set was also inspected while the analyses were developed, so it is not an untouched test set.

The Markov control uses the original smoothing rule, generic START and END, and no observed DIS/ND endpoint. Twenty within-shot random permutations preserve every shot's event count and family multiset. Both training and test chains are permuted, and metrics are averaged across permutations rather than across prediction probabilities. This contrast includes first- and last-family

information. Intervals use 2,000 paired stratified shot bootstraps and condition on the already fitted discovery models, vocabulary, classifiers, folds and permutations. They omit retraining uncertainty and are not adjusted for the multiple reported contrasts.

Table 11: Reference-run information ablation on identical event-bearing cohorts. The first six rows cumulatively add logistic features. The two Markov rows are separate classifiers. Development has 512 shots and 0.684 DIS prevalence, the evaluation set has 61 shots and 0.574 prevalence. Coefficients for the evaluation set are fitted only on the development pool.

| | Development OOF | | Evaluation, frozen | |
|---|---|---|---|---|
| Features | AUROC | AP | AUROC | AP |
| Count | 0.567 | 0.732 | 0.542 | 0.609 |
| + record and event geometry | 0.913 | 0.961 | 0.907 | 0.939 |
| + family counts | 0.911 | 0.956 | 0.937 | 0.959 |
| + vocabulary direction | 0.912 | 0.957 | 0.931 | 0.948 |
| + first and last families | 0.908 | 0.954 | 0.918 | 0.941 |
| + directed adjacency | 0.908 | 0.951 | 0.880 | 0.919 |
| Markov ordered | 0.713 | 0.799 | 0.713 | 0.736 |
| Markov shuffled | 0.704 | 0.793 | 0.692 | 0.718 |

Table 12: Paired increments on the evaluation set for the reference event-bearing cohort, with conditional 95% intervals. The feature and permutation definitions are those of Table 11.

| Contrast | AUROC difference | AP difference |
|---|---|---|
| Geometry over count | 0.365 [0.249, 0.478] | 0.330 [0.238, 0.394] |
| Family over geometry | 0.031 [-0.002, 0.074] | 0.020 [-0.00042, 0.046] |
| Direction over family | -0.007 [-0.032, 0.013] | -0.010 [-0.034, 0.005] |
| Adjacency and endpoints over family | -0.057 [-0.128, -0.003] | -0.040 [-0.080, -0.007] |
| Adjacency over endpoints | -0.037 [-0.078, -0.005] | -0.023 [-0.047, -0.004] |
| Markov ordered over shuffled | 0.021 [0.002, 0.049] | 0.017 [-0.011, 0.046] |

Event presence matters. On all 182 evaluation discharges, with prevalence 0.214, count alone reaches AUROC 0.865 and AP 0.569, whereas on event-bearing shots it reaches 0.542 and 0.609. The multi-event group of the evaluation set contains 8 DIS shots among 12, compared with 27 among 49 single-event shots. Only 5 chains contain distinct families, and the 16 internal transitions comprise 4 increases, 1 decrease and 11 self transitions. This sharply limits conclusions about direction.

Seed 43 has 58 event-bearing evaluation shots, prevalence 0.569 and 5 distinct-family chains. Family counts improve over geometry by AUROC 0.045 [0.008, 0.096], direction over family by -0.018 [-0.056, 0.006], and full adjacency with endpoints over family by -0.004 [-0.032, 0.021]. Ordered versus shuffled Markov gives 0.807 versus 0.782, difference 0.024 [0.006, 0.051], with AP difference 0.026 [0.003, 0.056].

Seed 44 has 60 event-bearing evaluation shots, prevalence 0.550 and 4 distinct-family chains. Family counts improve over geometry by AUROC 0.034 [-0.025, 0.098], direction over family by 0.030 [0.007, 0.064], and full adjacency with endpoints over family by 0.011 [-0.019, 0.053]. Ordered versus shuffled Markov gives 0.843 versus 0.824, difference 0.019 [-0.00012, 0.051], with AP difference 0.030 [0.003, 0.069].

The ordered Markov point estimate exceeds its shuffled counterpart for all three seeds, but the geometry-adjusted direction and adjacency increments vary across seeds. The evidence supports limited order sensitivity and family information, without establishing a stable conditional advantage of full directed adjacency.

The comparison with M4 uses 440 common development shots and 54 common evaluation shots, with DIS prevalences 0.786 and 0.648. PM-MIL minus comparator Markov AUROC is 0.060 [0.014,

0.107] on the pool and -0.011 [-0.078, 0.061] on the evaluation set. The corresponding full directed logistic differences are 0.010 [-0.024, 0.042] and -0.046 [-0.126, 0.020].

The comparison with the M2 union uses 402 common development shots and 45 common evaluation shots, with DIS prevalences 0.823 and 0.711. PM-MIL minus comparator Markov AUROC is 0.025 [-0.041, 0.098] on the pool and -0.026 [-0.126, 0.069] on the evaluation set. The corresponding full directed logistic differences are 0.006 [-0.040, 0.051] and -0.060 [-0.171, 0.034].

**Classifiers added after the protocol.** Two classifiers were added after the prespecified analysis. The family-only classifier compares event count alone with event count plus the six family counts, both without record geometry (Table 13). It separates the marginal information in family composition from the geometry-adjusted contrast above. The fraction classifier adds per-shot fractions of vocabulary-forward, vocabulary-backward and self transitions to the direction stage, with zero fractions when a chain has no transition. The cohorts, folds, fixed penalty and paired resamples are unchanged, and no feature was selected by performance on the evaluation set.

Table 13: Family-only classifier on the frozen evaluation chains, comparing event count alone with event count plus the six family counts. Both omit record geometry. Paired increments have conditional 95% intervals. Seed-specific event-bearing cohorts differ.

| Seed | Count AUROC | + family AUROC | AUROC increment | AP increment |
|---|---|---|---|---|
| 42 | 0.542 | 0.754 | 0.213 [0.094, 0.338] | 0.159 [0.089, 0.238] |
| 43 | 0.489 | 0.808 | 0.319 [0.168, 0.459] | 0.261 [0.148, 0.358] |
| 44 | 0.486 | 0.804 | 0.318 [0.193, 0.441] | 0.206 [0.100, 0.330] |

For the reference run, the family-only classifier gives development AUROC 0.712 and AP 0.798, and AP 0.769 on the evaluation set. The fraction classifier, which adds the directional fractions to geometry, count, families and direction counts, gives AUROC 0.930 on the evaluation set, a difference from the family stage of -0.008 [-0.033, 0.012]. The family comparison shows information beyond count in all three seeds, whereas the conditional directional comparison remains inconsistent. These associations do not establish that family or direction information is independent of unmeasured operating conditions.

### B.6 Record length and operating-point baselines

Records start 0.15 s after breakdown and end at the disruption for DIS discharges and at the end of the stored discharge for ND discharges, so the length of a record is a label cue. On the development pool the DIS records have median length 256 ms (10th to 90th percentiles 109 to 379 ms) and the ND records 348 ms (343 to 450 ms). A logistic regression on the record length alone, fitted on the pool, reaches AUPRC .858 and AUROC .857 on the evaluation set. Under nominal suffix removal with a retained-prefix floor, the fitted length classifier empirically keeps AUPRC .858 at nominal removal settings up to 200 ms and falls to .244 only at 300 ms, where most DIS records have reached the ten-token floor. A classifier on the operating point, the mean plasma current, toroidal field, line-averaged density and edge safety factor over the first 50 ms of each record, reaches .436 (AUROC .666), and the two together .841, so the operating point adds nothing to the length and these operating-point features do not exclude other scenario cues.

The matched evaluation draws three independent DIS lengths per ND discharge, caps each by the available ND record, samples a contiguous window and averages its three per-model probabilities before applying the ensemble weights and calculating metrics. This reduces length imbalance but does not enforce equal finite-sample distributions. On the same three crops the length-only classifier reaches AUPRC 0.442, 0.463 and 0.424 and AUROC 0.544, 0.594 and 0.617 at nominal 0, 100 and 200 ms, above the prevalence of 0.214, so matching removes most but not all of the length cue. With the last 30 ms hidden it reaches AUPRC 0.453 and AUROC 0.567. DIS records are unchanged. For a cropped record of $T$ samples at 10 kHz, the procedure retains $K = \max(T - 10\Delta, \min(500, T))$ samples. Thus 50 ms is the minimum retained input, not a warning horizon. Among 39 DIS records, actual removal is below nominal for 0, 0, 2, 7, 13, 16 and 32 records at $\Delta = 0, 25, 50, 100, 150, 200, 300$ ms. Token counts give exact affected counts but bound individual

durations to 5-ms intervals. The shortest DIS record is 80.0–84.9 ms long, so only 30.0–34.9 ms can be removed. All 182 discharges remain in each evaluation. Figure 3 shows the unmatched counterpart, where PM-MIL reaches AUPRC .896 on full records and .761 at nominal 100 ms, compared with .716 for M4, .597 for R1 and .498 for Single MIL. Differences may reflect crop length, crop location and averaging, and do not by themselves exclude nuisance dependence.

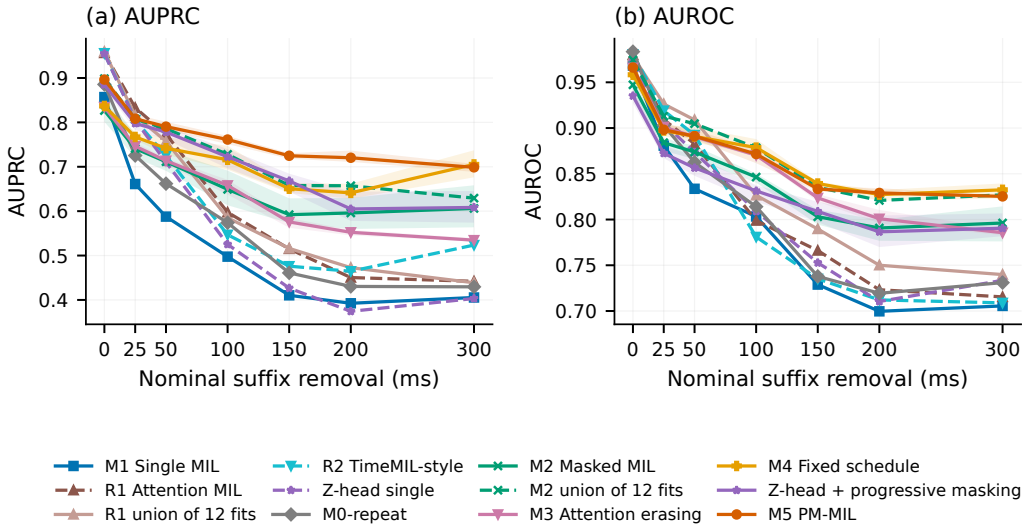



Figure 3: Classification under nominal suffix removal on unmatched records, retaining at least 50 ms as in Figure 2. (a) AUPRC and (b) AUROC with seed means and ranges. The unmatched record-length classifier has AUPRC .858 at nominal settings up to 200 ms.

### B.7 Event budgets

Table 14 lists the mainline segments and event-bearing discharges of each method, since lead and depth are properties of a union of detectors whose coverage differs between methods.

Table 14: Event budget of each method under the shared vocabulary and consolidation rule, seed means where three seeds exist. Segments are mainline segments, and DIS and ND are the event-bearing discharges out of 364 DIS and 1,261 ND in the development pool and 39 DIS and 143 ND in the evaluation set.

| | Development pool | | | Evaluation set | | |
|---|---|---|---|---|---|---|
| Method | Segments | DIS | ND | Segments | DIS | ND |
| M1 Single MIL | 376 | 325 | 35 | 47 | 33 | 11 |
| R1 union of 12 fits | 361 | 223 | 88 | 37 | 26 | 7 |
| Z-head single | 421 | 342 | 43 | 43 | 32 | 7 |
| M0-repeat | 515 | 349 | 112 | 59 | 36 | 18 |
| M2 Masked MIL | 272 | 228 | 32 | 28 | 21 | 5 |
| M2 union of 12 fits | 530 | 337 | 115 | 55 | 35 | 16 |
| M3 Attention erasing | 522 | 349 | 114 | 56 | 35 | 18 |
| M4 Fixed schedule | 619 | 346 | 145 | 71 | 34 | 23 |
| Z-head + prog. masking | 862 | 360 | 246 | 87 | 36 | 29 |
| M5 PM-MIL (3 seeds) | 651 | 345 | 174 | 73 | 34 | 26 |

## C SPECIFIED CONTROLS AND REMAINING EMPIRICAL SCOPE

The comparison is restricted to the methods and quantities of Tables 1 and 2. M1 is the exact round-0 checkpoint. Every control uses the same corrected inputs, evaluation membership, event grammar, vocabulary and consolidation rule, and no cell is filled from an older experiment with different inputs, membership or taxonomy.

**M0-repeat.** Twelve independent fits of round 0 (seeds 100 to 111) with no masking, the same validation rule and admission gate, aggregated by the excess-AP soft vote and evaluated by the same event grammar, vocabulary assignment and consolidation as PM-MIL (member index takes the place of the round index in the ownership rule). It measures what repeated re-initialization contributes on its own, both to the ensemble score and to the chains, and is the correct single-model-versus-ensemble control for the classification comparison.

**Masked MIL (M2).** One shared-backbone model trained on fixed truncated-prefix views with depth grid $\{0, 10, 30, 50, 100, 150, 200, 250, 300, 400\}$ ms, retaining at least ten tokens. DIS views keep their DIS label with each discharge's total weight shared across its valid depths, and ND views are length-matched. No preceding model chooses the next depth.

**Union of twelve M2 fits.** Twelve M2 models from seeds 100 to 111, pooled like M0-repeat, reach AUPRC .898 on the full length-matched evaluation records, .830 at 100 ms and .757 at 200 ms, median earliest lead 160 ms on the pool [150, 180] and 165 ms on the evaluation set, multi-event shares 16.2% and 7.8%, and chain AUROC .55 [.49, .62] and .39, so pooled truncation augmentation supports classification after nominal suffix removal and produces some multi-event chains, fewer than progressive masking.

**Split-head ablation (Z-head).** The backbone with the instance classifier reading $z_{ik}$, which is conjunctive pooling over contextual instance scores as in MILLET (Early et al., 2024), every loss unchanged and the grammar reading the $z$-based token risk, run as a single model with seed 42 and under evidence-guided progressive masking with seeds 42, 43 and 44 (seed 43 stopped after eleven rounds for lack of validation support). The single model reaches length-matched AUPRC .963 on full records and .633 at 100 ms, median earliest lead 100 ms on the pool and 70 ms on the evaluation set, final-100-ms shares 53% and 72%, multi-event shares 9.4% and 10.3% and chain AUROC .47 and .27. Under progressive masking the three seeds reach AUPRC .936, .843 at 100 ms and .796 at 200 ms, median earliest lead 203 ms on the pool (seeds 195 to 215) and 198 ms on the evaluation set (185 to 210), final-100-ms shares 27% and 29%, multi-event shares 34.0% (29.8 to 40.6) and 29.3% (22.4 to 36.4), 256 and 22 family-to-family transitions, and chain AUROC .73 (.72 to .75) and .60 (.50 to .67). Their unions admit events in 246 of 1,261 ND discharges (211 to 277, 19.5%) against 162 (12.8%) for PM-MIL and 29 of the 143 in the evaluation set against 26, so their event-bearing cohorts hold more ND chains, from which the chain classifier learns. Split by label for seed 42, its DIS chains have a multi-event share of 35% against 31% for the PM-MIL reference run and its ND chains 27% against 14%, so most of the additional depth lies on ND discharges. Whether an event on an ND discharge is spurious or genuine activity that did not end in disruption cannot be decided from the discharge label, and the chain classifier uses ND chains as its negatives, so the larger union is a difference in event budget rather than a measured error rate. The classification threshold on the evaluation set fixes the shot-level false-positive rate at 5% for every run (7 of 143 ND discharges), so the larger union does not change the classifier. The paired differences of seed 42 from the PM-MIL reference run are 0 to $+25$ ms in median lead and $+1.6$ to $+10.1$ points in multi-event share. Every reported chain and classification effect of progressive masking is therefore reproduced without the split head. What differs is where the events are. Table 15 scores the events of one run by the token risk of another, as the share of the 25-ms window beginning at the event trigger at or above the scoring run's ND-calibrated upper level (the level the grammar uses), against random windows of the same shots. Cross-head support measures agreement between detectors on the same discharges, not transfer to different discharges. The PM-MIL and M4 seed-42 runs also share round 0, so this is not a fully independent comparison. On ND events, Z-head support under PM-MIL is .30 against .25 at random, while M4 support is .53 against .44. Table 16 separately shows above-chance cross-seed reproduction for both heads, averaging +0.19 for PM-MIL and +0.14 for Z-head on ND discharges. These results motivate evaluating recurrence but do not establish non-transferability of contextual

events. Local scoring is retained for its common input-to-score and input-to-representation mapping, while event selection remains contextual. A raw-feature check compares a 25-ms onset window with up to 100 ms of preceding data, excluding events with less than 20 ms of prehistory or incomplete onset windows, using the maximum standardized change across the 27 stored engineering features. Across methods, 15 to 19% of eligible events exceed the same-shot random 95th percentile across methods. This limited check does not establish physical identity or rank the methods.

Table 15: Cross-head support of events on the development pool (Z-head rows are the progressive-masking runs, means over three seeds). Share of the 25-ms window beginning at an event trigger at or above the scoring run's per-round ND-calibrated upper risk level (maximum over its twelve rounds), and the same share for random windows of the same shots.

| Events of | Risk of | DIS events | DIS random | ND events | ND random |
|---|---|---|---|---|---|
| PM-MIL (reference) | PM-MIL (reference) | .83 | .79 | .57 | .44 |
| PM-MIL (reference) | Z-head (3 seeds) | .94 | .93 | .41 | .37 |
| Z-head (3 seeds) | PM-MIL (reference) | .73 | .73 | .30 | .25 |
| Z-head (3 seeds) | Z-head (3 seeds) | .95 | .95 | .57 | .46 |
| M4 fixed schedule | PM-MIL (reference) | .85 | .78 | .53 | .44 |
| M4 fixed schedule | Z-head (3 seeds) | .95 | .93 | .50 | .47 |
| Z-head seed $a$ | Z-head seed $b \neq a$ | .93 | .93 | .38 | .36 |

Table 16: Cross-seed reproduction of source events on the development pool. An event of the first run is reproduced when the second run has an event on the same discharge starting within 25 ms (start) or overlapping it (overlap). Chance shifts the second run's events on each discharge by a random circular offset (20 draws).

| | DIS events | | | ND events | | |
|---|---|---|---|---|---|---|
| Seeds | Start | Chance | Overlap | Start | Chance | Overlap |
| PM-MIL 42 vs 43 | .84 | .62 | .94 | .57 | .35 | .66 |
| PM-MIL 42 vs 44 | .83 | .62 | .94 | .52 | .34 | .61 |
| PM-MIL 43 vs 44 | .86 | .64 | .94 | .51 | .35 | .65 |
| M4 42 vs 43 | .91 | .69 | .96 | .72 | .50 | .78 |
| M4 42 vs 44 | .90 | .66 | .97 | .66 | .43 | .77 |
| Z-head 42 vs 43 | .86 | .67 | .96 | .47 | .33 | .58 |
| Z-head 42 vs 44 | .86 | .69 | .96 | .41 | .29 | .54 |
| Z-head 43 vs 44 | .87 | .69 | .95 | .50 | .33 | .60 |

**Attention erasing / Winners-Out adaptation (M3).** After each accepted round, every DIS discharge that the model still supports (predicted positive with bag probability at or above the ND-calibrated onset threshold) loses one contiguous block of five visible tokens with the largest summed contribution $a_{ik} \max(\ell_{ik,\mathrm{DIS}} - \ell_{ik,\mathrm{ND}}, 0)$, the same amount that M4 removes per round, while later observations outside the erased blocks remain visible. Original positions are preserved, forward-loss pairs crossing gaps are omitted, and ND records receive random holes matching the erased fraction of the active DIS discharges to avoid a label-specific missingness cue. A discharge that the model no longer classifies as disruptive is retired by the second-chance rule of PM-MIL, and a discharge with fewer than ten visible tokens left is retired at the floor. Over three seeds 32, 51 and 49 discharges were retired by the second-chance rule, all twelve rounds passed the admission gate, and the run never met the no-progress stop. This adapts the discriminative-erasing principle (Zeng et al., 2019) and does not reproduce the original video network.

**Fixed-schedule suffix truncation (M4).** After each accepted round, every supported DIS discharge loses 25 ms (five tokens) more from its end, ND bags are length-matched as in PM-MIL, and the screening, retirement and stopping rules are those of PM-MIL with the evidence boundary

replaced by the fixed step. The step was fixed before the run so that twelve rounds span 0 to 275 ms, about the median DIS record length of 256 ms, and it was not tuned to any result. For seed 42 the active DIS discharges fall from 364 at round 0 to 204 at round 8 and 140 at round 11, and 41, 45 and 60 discharges were retired by the second-chance rule over the three seeds. All twelve rounds passed the admission gate in every seed.

**Retirement rules (M3a, M4a).** Earlier versions of these two controls did not screen with the model. M3a copied the erasure budget and the retirement round of every discharge from the PM-MIL reference run, and M4a retired only at the ten-token floor. They reach AUPRC .896 [.887, .906] and .844 [.829, .858], median earliest lead 162 and 182 ms, multi-event shares 11.3% and 22.3%, and chain AUROC .64 and .65, so earlier and deeper chains are observed under both retirement rules, and the versions with the PM-MIL rules are the ones reported.

**Related-work aggregators (R1, R2).** R1 keeps the front-end, positional encoding and encoder of the backbone and replaces the split head by embedding-level gated attention pooling (Ilse et al., 2018), a bag embedding $\sum_k a_k z_k$ followed by a linear classifier. R2 prepends a learnable class token with a learnable positional embedding, applies one aggregator layer in which the class token attends to the instance embeddings, and classifies the class token, after TimeMIL (Chen et al., 2024) without its wavelet positional encoding. Both are trained with the bag focal loss only, without StKIM, ND suppression or forward-prediction losses, and their localization signal is the attention (head-averaged for R2) times the number of visible tokens, so that 1 means uniform attention. The change signal used as trigger is the squared jump of this signal over one token, and the ND-referenced thresholds and persistence are those of the grammar. R1 is run as a single fit, as twelve fits from different seeds whose events are pooled, and under the fixed schedule of M4 for twelve rounds. R2 is run as a single fit and under the fixed schedule. Under the fixed schedule R1 reaches AUPRC .893, median earliest lead 60 ms [IQR 20, 220], 61.7% of DIS events in the final 100 ms, events in 133 of 364 DIS discharges, a multi-event share of 17.7% and chain AUROC .59, and R2 yields four events over twelve rounds (AUPRC .863). Progressive shortening does move the attention of both aggregators earlier (the median position of the attention maximum on DIS discharges moves from the last token to 100 ms for R1 and 112 ms for R2 by round 11, and the attention mass in the final 100 ms from 96% and 75% to 51% and 52%), but the grammar cannot turn attention into events on most discharges, because ND bags carry attention peaks of the same size.

All controls share the evaluation split, event grammar, frozen vocabulary and consolidation rule. Progressive Z-head seed 43 admitted eleven models, while the other reported twelve-round runs admitted twelve. One twelve-round run takes 85 to 108 minutes on one GPU, one M0-repeat member 10 to 16 minutes and one M2 model 5 to 8 minutes. An operational warning-time column would additionally require causal-prefix replay with future-free features and confirmation delays, and the manuscript reports retrospective event lead only.

**Projection diagnostics.** Every method's events are encoded by the twelve frozen PM-MIL local encoders and assigned to the nearest of the six frozen centroids, so a method whose events fall far from all centroids would receive family labels that mean little. Table 17 reports, per method (mean over seeds), the source events and deduplicated objects, the median and 95th percentile of the distance to the assigned centroid in the 2,280-d shared space, the share of objects beyond the 95th percentile of the PM-MIL reference run (5% by construction for that run), and the family shares of the objects. The reference run's own objects reproduce their frozen labels exactly.

# D ADDITIONAL FIGURES

Table 17: Shared-vocabulary projection diagnostics for all evaluated runs. Family columns are the share of objects (%).

| Method | Events | Objects | Median dist. | P95 dist. | Beyond ref. P95 | A | B | C | D | E | F |
|---|---|---|---|---|---|---|---|---|---|---|---|
| R1 single fit | 36 | 36 | 0.59 | 1.26 | 11.1% | 19 | 0 | 17 | 50 | 14 | 0 |
| R1, 12-fit union | 1,398 | 1,030 | 0.83 | 1.44 | 12.4% | 9 | 2 | 12 | 40 | 29 | 8 |
| M0-repeat (pool) | 5,723 | 2,875 | 0.64 | 1.43 | 8.5% | 4 | 6 | 9 | 58 | 18 | 5 |
| M2 Masked MIL | 380 | 380 | 0.55 | 1.36 | 6.9% | 5 | 4 | 6 | 63 | 18 | 5 |
| M2, 12-fit union | 4,754 | 2,816 | 0.59 | 1.31 | 6.9% | 5 | 5 | 9 | 57 | 19 | 5 |
| M3 Attention erasing | 4,668 | 3,097 | 0.63 | 1.28 | 6.7% | 5 | 5 | 9 | 59 | 17 | 5 |
| M4 Fixed schedule | 4,246 | 3,092 | 0.58 | 1.22 | 5.7% | 6 | 5 | 14 | 57 | 14 | 4 |
| Z-head single | 683 | 683 | 0.64 | 1.30 | 6.9% | 3 | 6 | 18 | 57 | 13 | 3 |
| Z-head + prog. masking | 3,496 | 3,157 | 0.60 | 1.14 | 4.3% | 5 | 4 | 24 | 45 | 17 | 6 |
| PM-MIL, reference run | 2,697 | 2,356 | 0.57 | 1.18 | 5.0% | 6 | 5 | 21 | 53 | 11 | 5 |
| PM-MIL, seeds 43, 44 | 2,896 | 2,405 | 0.58 | 1.16 | 4.6% | 7 | 6 | 21 | 47 | 15 | 4 |

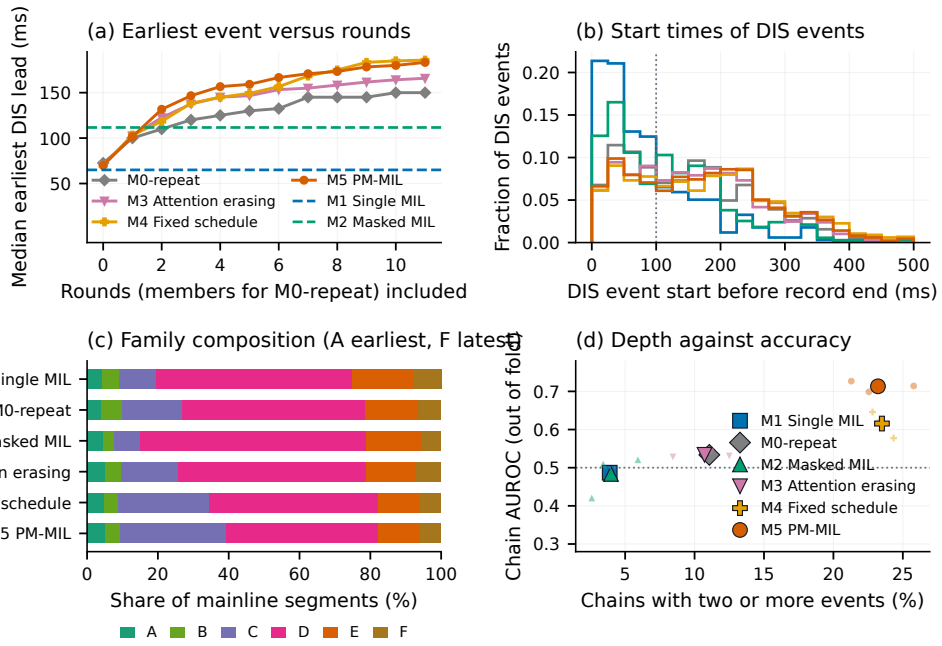



Figure 4: Chain-level comparison of the methods under the shared vocabulary and consolidation rule (development pool, seeds averaged). (a) Median earliest DIS event lead as rounds accumulate (members for M0-repeat), with Single MIL as a dashed line. (b) Start times of DIS mainline events. (c) Family composition of mainline segments, ordered from the earliest family A to the latest family F. (d) Share of chains with two or more events against the out-of-fold chain AUROC, with one small marker per seed and one large marker per method. Development pool. The Single-MIL versus PM-MIL detail is in Figure 7.

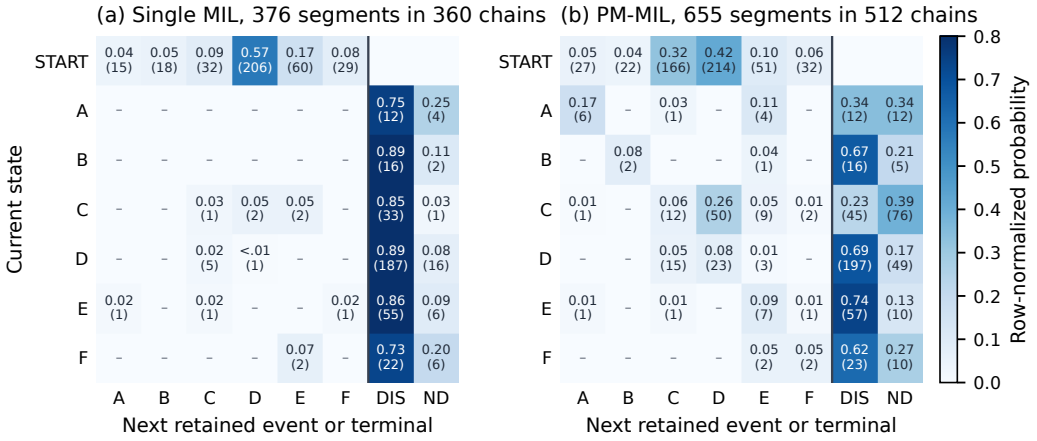
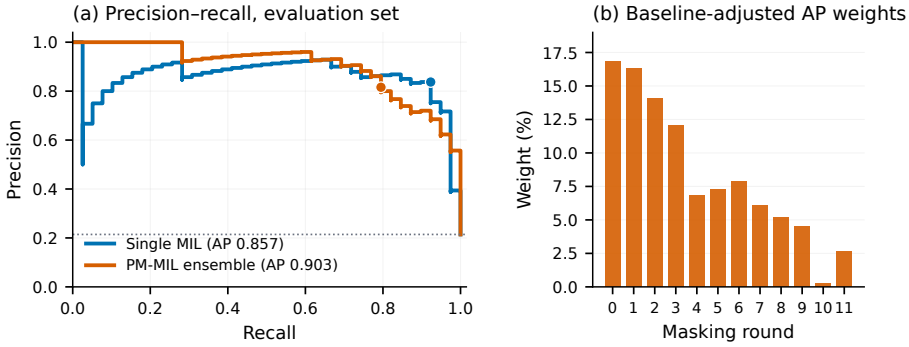



Figure 5: Observed next-event transitions under the shared vocabulary for (a) Single MIL and (b) PM-MIL. Rows are normalized by outgoing visits, and the DIS/ND columns are next-step terminals appended from the observed labels rather than eventual absorption. Single-MIL chains are essentially one event followed by the terminal, whereas PM-MIL chains contain 143 family-to-family transitions.




182 evaluation discharges (39 DIS / 143 ND). Markers are the operating points at 7/143 ND false positives.

Figure 6: (a) Full-record precision–recall curves on the 182 evaluation discharges. Ensemble weights and threshold are fitted on this set, and the cross-fitted AUPRC is 0.896. (b) The twelve weights of Eq. 4, with no special treatment of round 0.

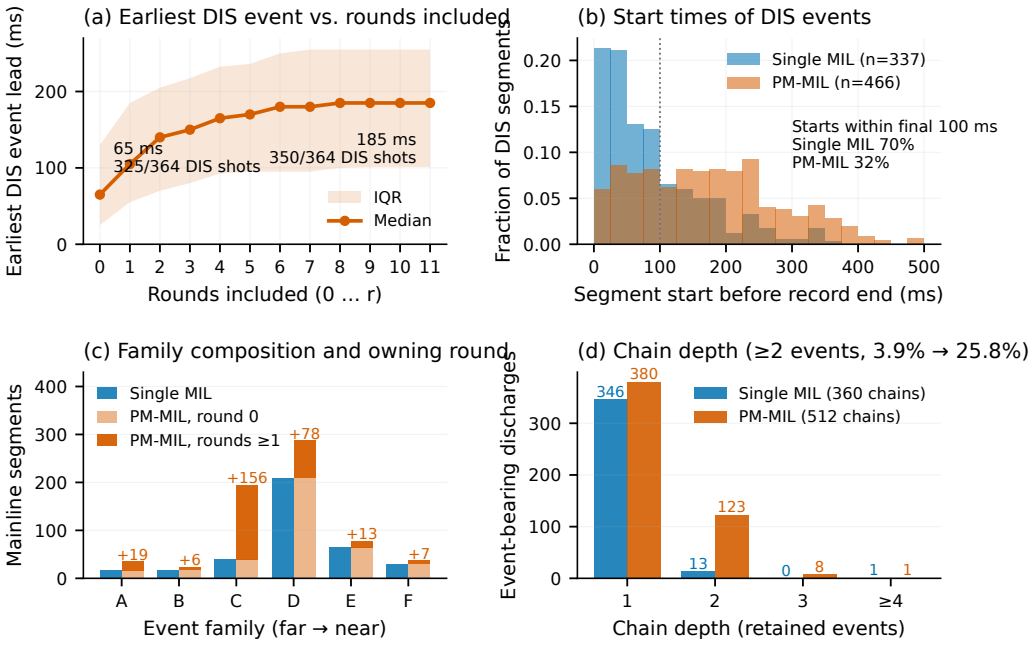



Figure 7: Single MIL against the PM-MIL reference run under the shared vocabulary and consolidation rule (development pool). (a) Median and IQR of the earliest DIS event lead as rounds accumulate. (b) Start times of DIS mainline events. (c) Mainline segments per family, where the dark part of each PM-MIL bar is owned by rounds $\geq 1$, which supply 156 of the 195 segments of the early family C. (d) Chain depth, where 3.9% of Single-MIL chains contain two or more events against 25.8% of PM-MIL chains.

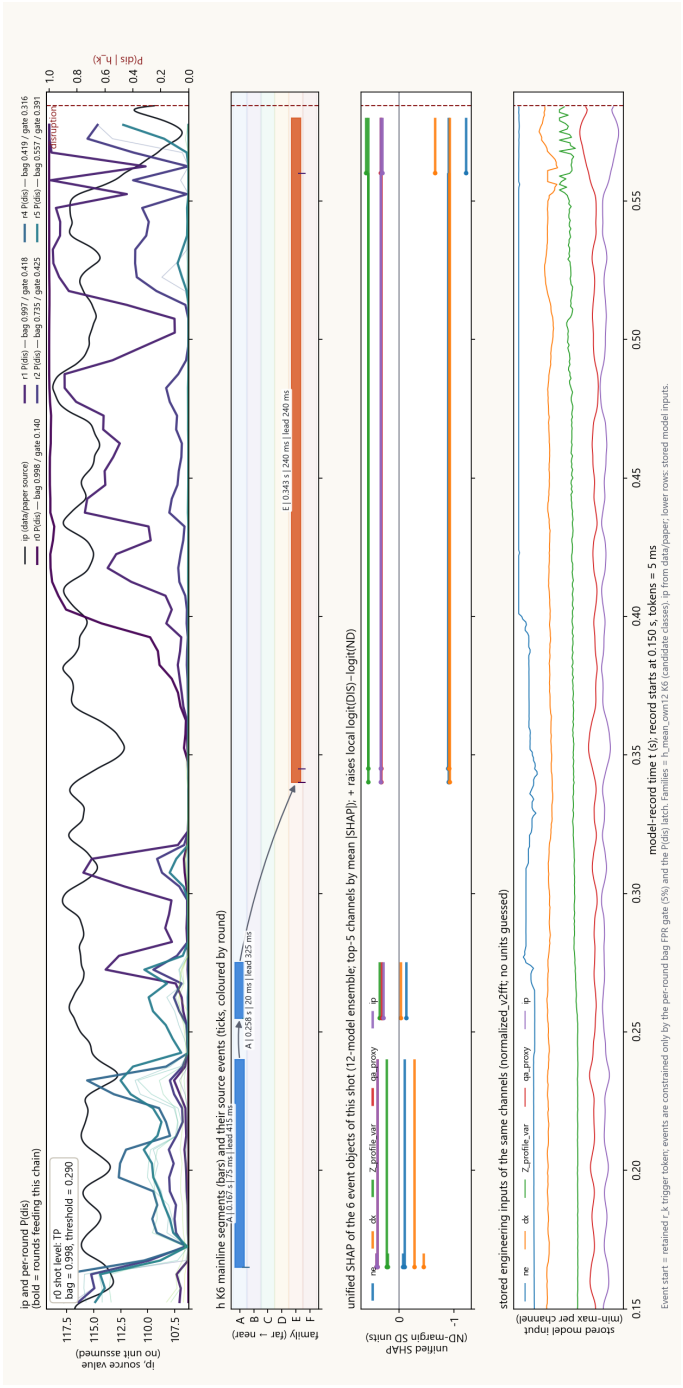



Figure 8: One disruptive development discharge (#1074397, chain A→A→E, classified correctly by round 0 and by the ensemble, $S = 0.653$). The top row shows the plasma current and the per-round local risk $P(\mathrm{DIS} \mid h_k)$, with bold curves for rounds whose events feed the chain. The second row shows the mainline segments with start time, duration and lead. The two A segments (lead 415 and 325 ms) come from rounds 1–5, and the E segment (lead 240 ms) runs into the disruption. The third row shows unified local-margin attributions of the top-5 channels on event tokens, and the bottom row the stored inputs of the same channels. A candidate reading is a displacement excursion followed by a sustained vertical-profile change.

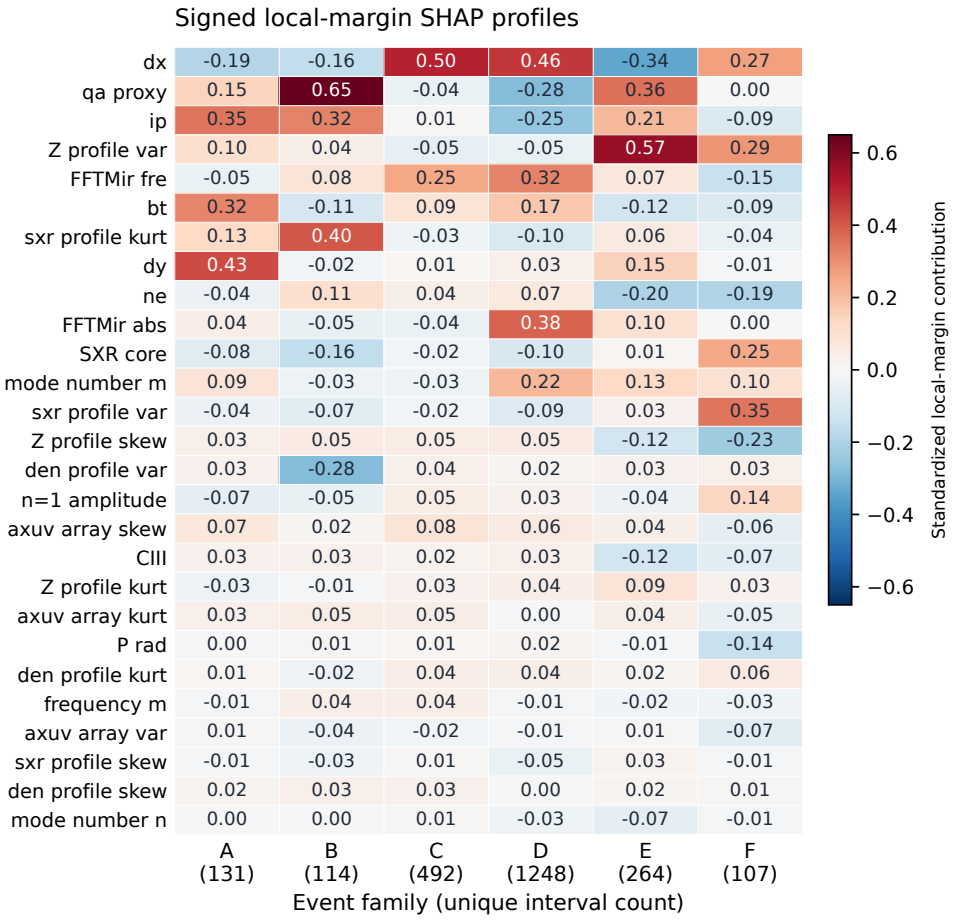



Figure 9: Shared local-margin attribution profiles for all 27 engineering features. Values are equal-shot-weighted class means of the twelve-model standardized local-margin explanation, not attributions of the final ensemble score. Row order uses mean absolute class-profile magnitude with equal family weight, and object counts refer to the 2,356 unique intervals before consolidation.