SEPARATING THE MODEL FROM THE DECISION RULE IN THE EVALUATION OF LONG-TAILED ICD CODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated ICD coding is extreme multi-label classification over thousands of long-tailed codes; most reported progress since 2023 is in macro-F1 and has been credited to new model components. Macro-F1, however, also depends on the decision rule, the thresholds that turn a model's scores into codes. Published systems on the standard MIMIC-IV ICD-10 benchmarks differ in how they set these thresholds or do not state how, so the model's share of any reported gain is unknown. Changed on their own, the thresholds can add as much macro-F1 as nearly half of that progress and alter how systems compare; we therefore propose a reporting standard that separates the model from the rule. The benchmark's reference protocol tunes one threshold on validation for micro-F1; using one per training-frequency bin, tuned for twice micro-F1 plus macro-F1, instead raises the macro-F1 of PLM-ICD, the 2023 state of the art, by $+0.034$, 46% of the progress since. On the reproduced members of the state of the art, GoM-ICD, the same change lifts one member alone above the published three-member ensemble on macro-F1. The gain is positive on all of our more than 60 training runs and all five released checkpoints we scored. The reason is that rare codes keep a high per-label AUC while a threshold shared with common codes sits above their positives: they are rankable but not decidable. Pre-registered tests on legal and protein-function benchmarks find the same, and the gain grows with how well the tail is ranked. Reporting both rules also shows which components improve the model: averaging members that rank codes differently gains under both, while negative focusing in the loss only moves the operating point. Built from the components that gain under both, an average of three long-context clinical encoders sets the state of the art on both standard benchmarks. On the 7,942-code benchmark it improves on GoM-ICD by $+0.057$ macro-F1 at higher micro-F1, of which $+0.048$ comes from the thresholds; on the 26,096-code benchmark it reaches $1.4\times$ the best published macro-F1 at equal micro-F1. The proposed standard reports every result under both rules; a difference that holds under both is a difference between models.

1 INTRODUCTION

Assigning ICD codes to a discharge summary is extreme multi-label classification with a heavy tail. On the most widely used MIMIC-IV ICD-10 benchmark (Edin et al., 2023; Johnson et al., 2023a;b;c), a discharge summary carries a median of 14 of the 7,942 codes, and 45% of the codes have at most 20 training examples. The field reports micro-F1 and macro-F1, and most of its progress over the past three years is in macro-F1. PLM-ICD (Huang et al., 2022) scored 0.585 micro-F1 and 0.211 macro-F1 in the reproduction of Edin et al. (2023). TraceCoder later added multi-source knowledge attention and reached $0.600/0.277$ (Ren et al., 2025), and GoM-ICD, which adds gap-scheme encoders and mixtures of experts (MoE) at three levels, holds the state of the art with the best published micro-F1, at $0.613/0.285$ (Wu et al., 2025a). The state of the art's macro-F1 has thus risen by 0.074 since 2023. Each of these numbers was produced by a model together with a decision rule, the thresholds that convert per-code scores into predicted codes, and papers set these thresholds differently. Of six recent systems on this benchmark, two do not say how, and the other four each use a different procedure (Appendix Table 6).

This matters most for macro-F1. It weights every code equally, so it depends heavily on rare codes, whose predictions are the most sensitive to where the threshold sits. We measure this effect directly:

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.

the model's scores stay the same, and only the thresholds change. The reference protocol of Edin et al. (2023) tunes one global threshold on validation for micro-F1 and reports macro-F1 at that threshold. On the PLM-ICD checkpoint that the same authors released, replacing that threshold with five validation-fit thresholds, one per training-frequency bin (the binned rule), adds $+0.034$ macro-F1, 46% of the macro-F1 gain since 2023. On GoM-ICD's three ensemble members, which we retrained, the same change adds $+0.021$ to $+0.028$, and the best member alone then exceeds the published three-member ensemble's macro-F1. The gain is positive on all 119 checkpoints we trained for this analysis and on all five released checkpoints we scored, and it is larger on the heavier-tailed 26,096-code benchmark of Nguyen et al. (2023).

The size of the gain depends on how well a model ranks rare codes. Rare codes are rankable but not decidable: a model usually scores a rare code's positive notes above its negative ones, but not above a threshold that suits common codes (§3.4). A pre-registered test on the released checkpoints confirms that the gain rises with their mean AUC on rare codes (§3.2). Legal and protein-function benchmarks show the same pattern: across the six label spaces we study, the gain follows how well the tail is ranked rather than the share of rare labels.

Scoring every model both ways also sorts components by whether their gains persist when the thresholds change (§5). Our base model, a long-context clinical encoder (BioClinical-ModernBERT-large; Sounack et al., 2025; reading 8,192-token windows) with a label-attention head, gains over the released PLM-ICD checkpoint under both, and so does averaging three members that rank codes differently. Our system sets the state of the art on both benchmarks (§6). On the 7,942-code benchmark it improves on GoM-ICD by $+0.057$ macro-F1 at higher micro-F1, of which the binned thresholds account for $+0.048$, and it leads on both metrics under the reference protocol's single threshold as well; on the 26,096-code benchmark it reaches $1.4\times$ the best published macro-F1 at equal micro-F1.

Thresholds have been studied in ICD coding before: Edin et al. (2023) introduced the reference protocol in place of untuned thresholds, and Choi et al. (2026) found a macro-F1 boost from frequency-group thresholds on a PLM-ICD model they trained. On released models and on our own, we show that the amount the reference protocol understates macro-F1 grows with how well a model ranks rare codes, and we measure what this is worth against the benchmark's reported progress. We propose that results on long-tailed ICD coding be reported both ways, under the reference protocol and under the binned rule, each fit on validation, with a statement of how macro-F1 is averaged (§8), and provide code that scores both from saved logits.

2 Setting

Benchmarks. We study the two full-label MIMIC-IV ICD-10 benchmarks. The *7,942-code split* of Edin et al. (2023) contains 89,099 training, 13,378 validation, and 19,802 test discharge summaries over 5,803 diagnosis and 2,139 procedure codes, 7,935 of which have at least one test positive. The *26,096-code split* of Nguyen et al. (2023) divides MIMIC-IV ICD-10 admissions by patient into 110,442 training, 4,017 validation, and 7,851 test admissions and keeps every ICD-10 code that occurs in them; we use the 110,398, 4,014, and 7,847 that have both a note and a code (Appendix D). On both splits, micro-F1 pools decisions over all note–code pairs and macro-F1 averages per-label F1 over test-active labels, those with at least one test positive (Appendix E gives other conventions).

The decision rule. Published systems set thresholds in different ways (Appendix Table 6); we compare three rules. **Rule A**, the reference protocol (Edin et al., 2023), selects one global probability threshold on validation to maximize micro-F1, applies it unchanged to test, and reports macro-F1 at that threshold. **Rule B**, the binned rule, gives each of five training-frequency bins its own threshold (codes with no training example, 1–20, 21–100, 101–1,000, and more than 1,000), fit on validation to maximize the micro–macro objective $2\,\mathrm{micro} + \mathrm{macro}$ and then applied unchanged to test; §3.3 explains the weight of 2 on micro-F1. **Rule C** is GoM-ICD's protocol, taken from its released code and its ablations: the best of five thresholds from 0.30 to 0.50, selected on the test set separately for micro-F1 and macro-F1. We use it only on the reproduced GoM-ICD members, to compare with their published numbers. Where noted, we also report the untuned threshold 0.5. All three are applied to the same frozen logits; Appendix A defines them and gives their search grids.

Use of the test split. Everything fitted after training is fit on the validation split only. During development, thresholds are fit on one half of validation and scored on the other, and reported systems are refit on full validation before their test evaluation.

3 What the decision rule is worth

3.1 The released PLM-ICD checkpoint

We scored the PLM-ICD checkpoint released with the reference implementation (Edin et al., 2023), using that implementation's model code. Rule A selects the threshold stored with the checkpoint and reproduces the published micro-F1. Changing only the thresholds raises the checkpoint's macro-F1 from 0.225 to 0.259, a gain of $+0.034$ (95% interval $[+0.030, +0.035]$ from resampling test notes), and lowers its micro-F1 from 0.589 to 0.583. On this checkpoint, every bin of codes seen in training gets a threshold below 0.5, and the rarer the bin, the lower the threshold, from 0.33 for codes with over 1,000 training examples to 0.13 for codes with 1 to 20. This $+0.034$ is 46% of the macro-F1 gain since 2023.

3.2 Across models

Rule B beats rule A on macro-F1 for all 119 checkpoints of Appendix B, from more than 60 training runs, with EMA and epoch variants of a run counted separately, when thresholds are fit on one half of validation and scored on the other. The gain averages $+0.050$ macro-F1 at an average cost of 0.006 micro-F1 and is never below $+0.014$. The nine models we evaluated on test, chosen to span the pool, all gain at least $+0.025$ (Appendix Table 10). Because the gain differs between models, the comparison between them changes too. Among the nine test-evaluated models, the macro-F1 spread shrinks from 0.030 to 0.012, within seed-to-seed variation, and the order changes (Kendall $\tau = 0.44$, Appendix Figure 6): the model ranked first under the reference protocol, trained with $\gamma^- = 6$ and a rare-label boost (§6, Appendix B), gains the least.

For the other checkpoints the reference implementation released (CNN, Bi-GRU, CAML, MultiResCNN, LAAT), we pre-registered three predictions (Appendix F). The first two, that rule B raises each model's test macro-F1 and that the gain grows with its rare-label AUC (mean AUC on validation over codes with 1 to 20 training examples), hold on PLM-ICD and the four checkpoints that reproduce their stored threshold: the gain is positive on every model and rises with rare-label AUC (Spearman 0.9 across the five), from $+0.003$ for the CNN to $+0.034$ for PLM-ICD. The third, that binned thresholds would order the models more closely by rare-label AUC, did not hold: the order is unchanged, and the gap between the best and worst model widens from 0.141 to 0.173, as it does among the reproduced GoM-ICD members (Table 9). On the models we scored, which differences persist depends on what separates them. Differences in architecture, among the released checkpoints and among GoM-ICD's members, persist and grow. Differences in loss and optimizer settings within one architecture, among our nine, mostly disappear: the $\gamma^- = 6$ rare-boost model leads the base model by $+0.030$ macro-F1 under A but by only $+0.002$ under B.

The reproduced state of the art. GoM-ICD does not release checkpoints, so we retrained its three gap-scheme members from the published code at the paper's stated configuration, one seed each, with label-title embeddings we built because they are not released (Appendix C). Our best member lands 0.010 micro-F1 below its published value, so the comparisons below are, if anything, conservative.

On these models, five binned thresholds add $+0.021$ to $+0.028$ macro-F1 over the reference protocol (Appendix Table 9). With them, the best reproduced member alone reaches 0.297 macro-F1 in the form GoM-ICD reports, the harmonic mean of macro precision and recall (Appendix E), above the published three-member ensemble's 0.285, at lower micro-F1 (0.593 against 0.613). Fit on validation, they also beat the macro-F1 threshold of GoM-ICD's own protocol, chosen on test from its grid, by about $+0.013$ on every member. That threshold is always the lowest on the grid, 0.30, which is still too high for rare codes.

A heavier-tailed benchmark. On the 26,096-code split, where 78% of labels have 1 to 20 training examples against 45% on the 7,942-code split, the binned thresholds are worth more, as predicted.

They add about $+0.06$ macro-F1 to each of our three members (Appendix D), more than to any 7,942-code model, and $+0.083$ to the mean of their logits, at a cost of 0.028 micro-F1 (Appendix Figure 2). On this split the reference protocol of Edin et al. (2023) does worse than no tuning at all: its micro-F1-optimal threshold lies above 0.5 and suppresses the tail, so its macro-F1 is 0.218, against 0.262 at 0.5. Outside MIMIC-IV, tail share alone does not predict the size of the effect (§4).

3.3 Granularity and objective

When labels are partitioned into k equal-count frequency bins, each with a threshold fit on validation, test macro-F1 rises from $k = 1$ to a peak at $k = 5$ to 8 and then declines on each of three models: the base model (§6), the $\gamma^- = 6$ rare-boost model of Appendix Table 10, and the eleven-model per-bin fusion of Appendix B (Appendix Figure 4). Per-label thresholds, each maximizing its own label's F1 on validation, do worse than eight bins on all three models and cut micro-F1 to 0.20 on the base model (Appendix A).

Of rule B's $+0.053$ macro-F1 on the base model, $+0.024$ comes from tuning one global threshold for the micro–macro objective instead of micro-F1, and the remaining $+0.029$ from binning. The same decomposition holds on validation halves for the released PLM-ICD checkpoint and our three-model system (Table 1). The weight on micro-F1 in the objective sits where the tradeoff bends: tuning the bins for macro-F1 alone costs about 0.02 micro-F1 for less than 0.01 more macro-F1, while with a weight of 16 on micro-F1, five bins exceed the micro-tuned threshold on both metrics for the released PLM-ICD checkpoint and the three-model system, and on the base model match its micro-F1 within seed noise while adding $+0.025$ macro-F1 (Appendix Figure 5; seed variation in Appendix B). The five validation-fit thresholds are as good as test-fit ones: refitting them on the test set changes macro-F1 by less than 0.0001 on the three models of the granularity study. A two-parameter bias linear in log-frequency, post-hoc logit adjustment (Menon et al., 2021), matches the five bins' macro-F1 within seed noise (Appendix A), so the gain does not depend on the choice of bins.

3.4 Why rare codes gain

Figure 1 distinguishes two properties of a per-label score: whether it ranks positives above negatives, measured by AUC, and whether its positives lie above a fixed operating point, measured by F1 at 0.5. On the validation split of a single AdamW run ($\gamma^- = 3$, probability shift 0.04), mean AUC stays above 0.94 across all five bins, while F1 at 0.5 falls from 0.623 for the most common to 0.134 for the rarest. Per-label thresholds fit in sample, an upper bound on what thresholds can recover, close much of the gap. The same separation underlies decoupled classifier retraining (Kang et al., 2020), in which representations learned on long-tailed data rank rare classes well and the classifier's boundary favors common ones. For calibrated scores, the F1-optimal threshold of a label is half the F1 it can attain (Lipton et al., 2014), so labels with lower attainable F1, usually the rare ones, need lower thresholds, and one threshold shared across all frequencies sits too high for the tail.

4 Beyond clinical coding

We pre-registered four predictions for two other domains (Appendix H) and tested them on four label spaces: EURLEX57K, 57,000 EU legal acts labeled with 4,271 EuroVoc concepts (Chalkidis et al., 2019), with a linear model on TF-IDF features; and the three Gene Ontology subontologies of the DeepGO-SE protein-function benchmark (Kulmanov et al., 2024), with the authors' MLP on ESM2 protein-language-model embeddings (Lin et al., 2023), retrained with their released code. In each, the thresholds of §2 were fit on validation and applied to test (Table 2).

As predicted, the pattern of §3.4 holds in all four label spaces: labels with 1 to 20 training examples keep a mean AUC above 0.86, yet at a fixed 0.5 their F1 is at most 0.18 and no more than half that of the most frequent labels. A single threshold tuned on validation already raises macro-F1 over a fixed 0.5 in all four label spaces, and five binned thresholds add up to $+0.022$ more, as predicted (Table 2). Neither prediction about tail share was supported: restricting each protein label space to terms with at least 50 training proteins, as in DeepFRI (Gligorijević et al., 2021), increases the gain instead of reducing it (Appendix Table 12); and across the six label spaces in this paper the gain is only weakly ordered by tail share (Spearman 0.14).

Table 1: Threshold structure against tuning objective on validation halves: each rule is fit on the first 6,689 validation notes and scored on the other 6,689, all on the same grid of logit offsets. Cells give micro-F1 / macro-F1, with macro-F1 over labels that have a positive in the scored half.

thresholds	objective	released PLM-ICD	base model	three models
one	micro-F1	0.5919 / 0.2451	0.6219 / 0.2968	0.6300 / 0.2979
one	$2\,\text{micro} + \text{macro}$	0.5880 / 0.2580	0.6187 / 0.3108	0.6265 / 0.3342
five bins	micro-F1	0.5931 / 0.2343	0.6222 / 0.2921	0.6309 / 0.3205
five bins	$2\,\text{micro} + \text{macro}$ (rule B)	0.5873 / 0.2773	0.6147 / 0.3500	0.6241 / 0.3721

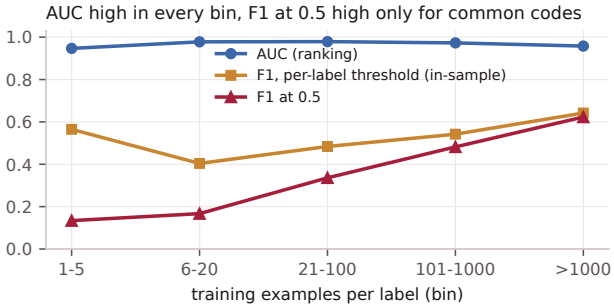

Figure 1: Mean per-label AUC, F1 at a fixed 0.5, and F1 under in-sample per-label thresholds by training-frequency bin, on validation for one AdamW run. Bins divide the 1–20 bin at 5; codes with no training example are not shown.

Consistent with the mechanism of §3.4, the gain tracks how well the tail is ranked, across all six label spaces and on the released checkpoints (§3.2). Where the tail is ranked well (MIMIC-IV, EURLEX57K, and GO cellular component), the validation-fit thresholds rise monotonically from rare to frequent labels, as on the released PLM-ICD checkpoint in §3.1, and the gain is at least $+0.02$ macro-F1. Where it is ranked less well (GO molecular function and biological process), the fitted thresholds differ by at most 0.7 logits without a monotone trend, and a single global threshold is already close to optimal. Lowering the threshold for rare labels raises their F1 only if their positives are ranked above most negatives; thresholds can use only the ranking the model already has.

5 Components under both rules

The 119 checkpoints of §3.2 share one architecture, a long-context clinical encoder with a label-attention head, and differ in optimizer, loss, training length, and seed (Appendix B; §6 gives the configuration); the AdamW configuration with $\gamma^- = 3$ is the base model. Table 3 compares each component under both, which shows whether its gain persists when the thresholds change.

The base model scores $+0.044$ macro-F1 above the released PLM-ICD checkpoint under A and $+0.064$ under B; averaging three members adds $+0.025$ under A and $+0.020$ under B, and the lift tracks rank diversity between members (Appendix B): the three members of our system are less correlated with one another (mean per-label Spearman correlation 0.60) than two seeds of the base model are (0.74). Negative focusing improves neither the ranking nor the diversity: changing γ^- alone leaves members as correlated as two seeds, and in three otherwise identical runs, moving γ^- from 3 to 5 raises the best threshold from 0.53 to 0.61, yet with the threshold tuned the runs end within 0.001 micro-F1 of one another. Negative focusing thus moves only the operating point, which the thresholds set anyway.

Nor does more elaborate combination help: a GoM-style stacker (Wu et al., 2025a) with top-2 gating (Shazeer et al., 2017) and a learned per-bin fusion, both trained on eleven of our members, score no higher than the plain mean of their logits on test (Appendix Table 11). More members help little:

Table 2: Outside clinical coding: test macro-F1 over test-active labels, one model per label space, thresholds fit on validation (Appendix H). GO-CC, -MF, -BP: the cellular-component, molecular-function, and biological-process subontologies. Tail: share of labels with 1 to 20 training examples; rare AUC: their mean per-label AUC on test.

label space	labels	tail	rare AUC	fixed 0.5	A	B	B − A	95% CI
EURLEX57K (legal)	4,271	0.67	0.956	0.360	0.404	0.426	+0.022	[+0.016, +0.023]
GO-CC (protein)	2,829	0.66	0.929	0.088	0.116	0.136	+0.020	[+0.014, +0.024]
GO-MF (protein)	6,851	0.76	0.866	0.044	0.068	0.073	+0.005	[+0.004, +0.007]
GO-BP (protein)	21,356	0.67	0.866	0.022	0.042	0.043	+0.0003	[−0.0002, +0.0006]

Table 3: Components compared on frozen scores under each rule, 7,942-code split unless noted. Cells give differences in micro-F1 / macro-F1; ranges are macro-F1 across the three members; "obj." marks differences in the micro–macro objective. "val": fit on one validation half and scored on the other, with AUC on full validation. cRT: the classifier-retrained model of Appendix Figure 3b. For scale, the binned thresholds add +0.042 macro-F1 on average across the nine test-evaluated models (§3.2).

component	under rule A	under rule B	split
the model			
base model vs. released PLM-ICD	+0.028 / +0.044	+0.029 / +0.064	test
one → three decorrelated members	+0.009 / +0.025	+0.008 / +0.020	test
EMA vs. raw weights (26,096 codes)	+0.003–0.011	+0.005–0.015	val
classifier re-training (cRT)	−0.002 / +0.012	−0.004 / +0.009	test
the operating point			
negative focusing γ^-, effect on ranking	AUC spread 0.0004, within seed noise		val
combining members, under rule B			
three → eleven members		+0.002 / +0.009	test
learned per-bin fusion vs. mean of logits		obj. −0.0004	test
gated MoE stacker vs. mean of logits		obj. −0.0005	test
distillation vs. matched $\lambda = 0$ control		obj. −0.0016	test

eleven add +0.009 macro-F1 over three at 3.7× the inference cost. Distilling the ensemble into one model (Hinton et al., 2015) has no effect against a matched control (Appendix G).

6 The resulting system

The system combines the components of §5 that gain under both: the base model, a long-context clinical encoder with a label-attention head and EMA weights, and an average of three members that rank codes differently, scored with binned thresholds.

Encoder and head. BioClinical-ModernBERT-large (Sounack et al., 2025; Warner et al., 2025), a 396M-parameter clinical encoder whose pretraining corpus includes MIMIC-IV notes, processes each note in windows of 8,192 tokens with stride 4,096. A label-attention head (Vu et al., 2020) with one query per code aggregates tokens into a code-specific vector scored by one linear output shared across codes (LAAT has one per code), with separate diagnosis and procedure heads; label queries are initialized from the encoder's token embeddings of the code descriptions. PLM-ICD, GoM-ICD, and TraceCoder use the same head family, with encoders that read notes in chunks of at most 512 tokens. Attention scores in each head also receive a section prior, a learned bias for each of 15 note-section types found by matching headings; removing these 30 parameters changes validation F1 by less than 0.002.

Loss and optimization. We train with the asymmetric loss of Ridnik et al. (2021), with negative focusing $\gamma^- = 3$ and probability shift 0.02, and with 8-bit AdamW (Loshchilov & Hutter, 2019; Dettmers et al., 2022), at learning rate 4.5×10^{-5} for the encoder and 9×10^{-4} for the head, for four epochs at effective batch 28 in bfloat16; a single model trained this way is the base model. We

Table 4: MIMIC-IV ICD-10, 7,942 codes, test split. Published rows are taken from their papers, and our thresholds are fit on validation only. All rows train on the benchmark's labeled ICD-10 notes only; with added ICD-9-coded notes, DualLAAT reaches $0.580/0.299$. GoM-ICD's published macro-F1 is in the harmonic form, so the final column is conservative for our rows.

system	members	rule	micro-F1	macro-F1	vs. GoM-ICD
PLM-ICD (2022; 2023)	1	A (val)	0.585	0.211	
TraceCoder (2025)	1	global (dev)	0.600	0.277	
GoM-ICD (2025a)	3 + MoE	unstated	0.613	0.285	
Choi et al. (2026)	1	3 groups, macro (val)	0.568	0.288	
DualLAAT, ICD-10 (2026)	1	unstated	0.575	0.279	
ours, base model	1	A	0.6168	0.2690	$+0.004$ / -0.016
ours, base model (cRT)	1	A	0.6165	0.2858	$+0.004$ / $+0.001$
ours, three models	3	A	**0.6257**	0.2943	$+0.013$ / $+0.009$
ours, base model	1	B	0.6118	0.3225	-0.001 / $+0.037$
ours, three models	3	B	0.6201	0.3422	$+0.007$ / $+0.057$
ours, eleven models	11	B	0.6219	**0.3516**	$+0.009$ / $+0.067$

Table 5: MIMIC-IV ICD-10, 26,096 codes, test split. Macro-F1 averages over all 26,096 labels, as in published work on this split, and micro-F1 counts all 26,096 columns. Published rows are the benchmark's baselines and peer-reviewed systems. Text preprocessing varies across works on this split (Appendix D).

system	rule	micro-F1	macro-F1
CAML / LAAT / JointLAAT / MSMN (2023)	as published	0.527–0.559	0.041–0.057
PLM-ICD (2023)	as published	0.570	0.049
CoRelation (2024)	as published	0.578	0.063
Choi et al. (2026), PLM-ICD with COMIC-S loss	3 groups (val)	0.588	0.076
ours, three models	fixed 0.5	0.5902	0.0919
ours, three models	A	**0.6171**	0.0766
ours, three models	B	0.5890	**0.1057**

evaluate an exponential moving average (EMA) of the weights, with decay 0.999 by default, from the first epoch (Morales-Brotons et al., 2024); on the 26,096-code split, the EMA weights of each model we average outperform its raw weights. An optional variant of classifier re-training (cRT) (Kang et al., 2020), used for one row of Table 4 and in no ensemble, fine-tunes the base model's head with the encoder frozen (Appendix B).

Members and cost. On both splits, our system averages the logits of three members that share the architecture and are trained with settings that favor micro-F1 or macro-F1 under a single threshold; under binned thresholds, what they add is rank diversity (§5). On the 7,942-code split they are two seeds of a Muon-hybrid model with a small probability shift and an AdamW model with $\gamma^- = 6$, a rare-label gradient boost, and longer training (Appendix B). Under rule B the mean of the three exceeds every member on both metrics (Appendix Table 10); §5 shows that a learned fusion does no better and that eleven members add little. The binned thresholds are five numbers fit on validation, at negligible cost. A four-epoch model trains in about nine hours on two A100 GPUs on the 7,942-code split and in about eleven on the 26,096-code split.

On the 7,942-code split (Table 4; Appendix Figure 3a), with binned thresholds the three-model system improves on GoM-ICD, the published state of the art, by $+0.007$ micro-F1 and $+0.057$ macro-F1, 77% of the macro-F1 gain since 2023. The base model alone, with the same thresholds, already exceeds GoM-ICD's macro-F1 by $+0.037$. The lead does not need them: under the reference protocol our three models are still ahead on both metrics, and the base model with classifier re-training matches GoM-ICD's macro-F1 at higher micro-F1. Eleven models extend the margin to $+0.067$ macro-F1. Systems aimed at rare codes (Choi et al., 2026; Liu & Nguyen, 2026) trade micro-F1 for macro-F1; our three-model system exceeds each of them on both metrics either way.

On the 26,096-code split (Table 5), we take published micro-F1 to count false positives on all 26,096 columns, as ours does. With binned thresholds our three-model system reaches 0.106 macro-F1, $1.4\times$ the best published value, at equal micro-F1, and under the reference protocol it has the highest reported micro-F1, $+0.029$ over the best published.

Of the three-model system's $+0.057$ macro-F1 over GoM-ICD, $+0.048$ comes from the binned thresholds on our own frozen scores (0.342 against 0.294). The remaining $+0.009$ is the gap between our system under the reference protocol and GoM-ICD's published number, which uses the harmonic form (Appendix E) and an unstated threshold.

7 Related work

ICD coding. Benchmark models range from CAML (Mullenbach et al., 2018) through MultiResCNN (Li & Yu, 2020), LAAT (Vu et al., 2020), and MSMN (Yuan et al., 2022) to PLM-ICD (Huang et al., 2022). Edin et al. (2023) found that under-tuned baselines and inconsistent evaluation, including untuned decision thresholds, had distorted earlier comparisons, and released the 7,942-code split with a validation-tuned global threshold as the reference protocol. In their ablation on MIMIC-III, tuning that threshold instead of fixing it at 0.5 raises PLM-ICD's macro-F1 by 0.036, and their threshold curves place the macro-F1 optimum below the micro-F1 optimum; we measure what that gap is worth on MIMIC-IV. Subsequent systems add knowledge attention (Ren et al., 2025), code relations (Luo et al., 2024), gap-scheme encoders with mixtures of experts (Wu et al., 2025a), semantic embedding rectification with long-tail refinement (Wu et al., 2025b), label-attention pretraining (Saha Roy et al., 2025), multi-version training (Liu & Nguyen, 2026), and concept bottlenecks (Syed & Lu, 2026). Choi et al. (2026) compare global, per-label, and frequency-group thresholds for PLM-ICD and find the grouped ones the most stable. Our design follows audits that re-measure an accepted result under a controlled protocol, in ICD coding (Edin et al., 2023) and elsewhere (Liu et al., 2019; Chen et al., 2019; Musgrave et al., 2020; Ferrari Dacrema et al., 2019), and work on variance in benchmarks (Bouthillier et al., 2021).

Decision rules for F1. Per-label thresholds are the standard way to optimize F1 (Fan & Lin, 2007; Lipton et al., 2014; Ye et al., 2012; Koyejo et al., 2015), but they overfit labels with few positives (Fan & Lin, 2007; Lipton et al., 2014; Jasinska et al., 2016), and a single shared threshold can do better on macro-F1 (Koyejo et al., 2015). On a long-tailed benchmark we find that the best choice lies between these extremes: five to eight frequency bins outperform both one global threshold and one threshold per label (§3.3). Post-hoc logit adjustment, which shifts each class's logit by a multiple of its log prior and is consistent for the balanced error in long-tailed multiclass classification (Menon et al., 2021), also lowers the thresholds of rare classes. Our five binned thresholds can be read as a coarse, F1-tuned form of it; a two-parameter log-prior version performs almost as well (§3.3). For the focal loss (Lin et al., 2017), the risk minimizer is an order-preserving transform of the posterior (Charoenphakdee et al., 2021), which agrees with our finding that γ^- moves the operating point rather than the ranking (§5). Extreme multi-label classification mostly avoids thresholds by evaluating top-k predictions (Jain et al., 2016), although per-label thresholds at extreme scale (Jasinska et al., 2016) and plug-in rules for macro-at-k metrics (Schultheis et al., 2023; 2024) have been studied. Protein function prediction evaluates with Fmax, which selects one global threshold on the evaluation set (Radivojac et al., 2013); §4 finds the same pattern for rare Gene Ontology terms.

8 Discussion

A reporting standard. We propose that work on long-tailed ICD coding report micro-F1 and macro-F1 under two rules, both fit on the validation split only and named next to every number: rule A, the reference protocol, and rule B, five frequency-binned thresholds fit for the micro–macro objective $2\,\mathrm{micro} + \mathrm{macro}$ (Appendix A). Under this standard, macro-F1 averages per-label F1 over labels with at least one test positive, and micro-F1 counts false positives on every label column; results that use another convention should say so. Reporting both shows whether a gain persists when the thresholds change; §5 applies this test to the components of our system. Gains from loss weighting and tuning reported under a single threshold should be checked with binned thresholds before they are credited to the model (§3.2). Five bins sit in the range where macro-F1 peaks (§3.3), and

weighting micro-F1 by 2 avoids most of the micro-F1 cost of tuning for macro-F1 alone (Appendix Figure 5). Both can be scored in one pass over saved logits with the scorer in the supplementary material, which works for any multi-label benchmark.

Limitations. Our models share one backbone family. At the time of writing, none of the systems published since the reference implementation has released the checkpoint behind its reported number (DualLAAT releases only its models trained with added ICD-9 data), so the reproduced GoM-ICD members stand in for them. Our state-of-the-art numbers, and the base model's gain over the released PLM-ICD checkpoint, use an encoder whose pretraining corpus includes MIMIC-IV notes, likely including this benchmark's test notes; the rule's gains do not depend on it, since they appear on the released checkpoints as well. Note-to-code mappings were not used in its pretraining. Outside MIMIC-IV we test one simpler model per label space, and the account of when binning helps (§4) rests on six label spaces and five released models. The bin edges and the objective weight of rule B were chosen on ICD coding; other label spaces may call for different ones.

Conclusion. Changing only the decision rule of the released PLM-ICD checkpoint adds macro-F1 equal to 46% of the 7,942-code benchmark's gain since 2023, and the same change raises macro-F1 on every model we scored, including the reproduced GoM-ICD members. Under both rules, our base model and an average of members that rank codes differently still add macro-F1, and a system built from them sets the state of the art on both benchmarks, with $+0.048$ of its $+0.057$ macro-F1 margin coming from the thresholds. The gain is largest where rare codes are ranked best, in ICD coding and in the legal and protein label spaces alike. Because thresholds change both the level of macro-F1 and the gaps between models, results should be reported under common, stated rules; we propose two and provide code to score them.

Reproducibility statement

Training and evaluation code, the patches applied to the GoM-ICD code, the pre-registered predictions and the selection files frozen before test evaluations, the result files behind every table, the 26,096-code benchmark builder, and a dataset-agnostic scorer for both decision rules are supplied as supplementary material and will be released. MIMIC-IV requires credentialed access through PhysioNet, so per-note scores cannot be distributed; upon acceptance, the trained checkpoints will be released on PhysioNet under the same credentialed access.

Ethics statement

All clinical data are de-identified MIMIC-IV records used under the PhysioNet (Pollard et al., 2026) credentialed data-use agreement; the legal and protein-function datasets of §4 are public. PhysioNet's guidance notes that the agreement prohibits sending credentialed data to third-party APIs and names services that may be used instead, Claude among them (PhysioNet, 2023). MIMIC-IV data were stored and processed only on an access-controlled institutional cluster with two-factor login and in a private cloud workspace in our own account. The only assistant that executed code on these machines was Claude, which itself ran on a separate workstation with no MIMIC-IV data. It acted only through sessions we authenticated and saw code, logs, and aggregate results, never note text or patient-level records. This work evaluates code-assignment models and does not deploy them; any clinical use would require validation on the target institution's data.

AI-use statement

We used large language model assistants for retrieval (reading papers and released code), execution (writing and running data-preparation, training, and evaluation code to our specifications), ideation (discussing hypotheses and interpretations, giving feedback on experimental design, drafting the pre-registration documents, and proposing some of the variants we tested, including two loss variants in Appendix B), and writing (drafting text, tables, and figures, which we revised). No AI tool generated data or labels or produced a reported number: every number reported for our models is computed by code from saved result files. We made all decisions, reviewed all text, and take full responsibility for the content.

REFERENCES

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. doi: 10.1038/75556.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Dmitriy Serdyuk, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, volume 3, pp. 747–769, 2021.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. Large-scale multi-label text classification on EU legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6314–6322, 2019. doi: 10.18653/v1/P19-1636.

Nontawat Charoenphakdee, Jayakorn Vongkulbhisal, Nuttapong Chairatanakul, and Masashi Sugiyama. On focal loss for class-posterior probability estimation: A theoretical perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5198–5207, 2021. doi: 10.1109/CVPR46437.2021.00516.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms. In *Advances in Neural Information Processing Systems*, volume 36, pp. 49205–49233, 2023.

Seunghun Choi, Won-Seok Hong, Hyun Won Lee, and Hyun-Soo Choi. Mitigating multi-label imbalance in ICD coding: A comprehensive evaluation of loss functions and threshold optimization. *IEEE Access*, 14:106530–106544, 2026. doi: 10.1109/ACCESS.2026.3708716.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. In *International Conference on Learning Representations*, 2022.

Joakim Edin, Alexander Junge, Jakob D. Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. Automated medical coding on MIMIC-III and MIMIC-IV: A critical review and replicability study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2572–2582, 2023.

Rong-En Fan and Chih-Jen Lin. A study on threshold selection for multi-label classification. Technical report, Department of Computer Science, National Taiwan University, 2007.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pp. 101–109, 2019.

Vladimir Gligorijević, P. Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C. Taylor, Ian M. Fisk, Hera Vlamakis, Ramnik J. Xavier, Rob Knight, Kyunghyun Cho, and Richard Bonneau. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1):3168, 2021. doi: 10.1038/s41467-021-23303-9.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Chao-Wei Huang, Shang-Chi Tsai, and Yun-Nung Chen. PLM-ICD: Automatic ICD coding with pretrained language models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pp. 10–20, 2022.

Himanshu Jain, Yashoteja Prabhu, and Manik Varma. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 935–944, 2016.

Kalina Jasinska, Krzysztof Dembczyński, Róbert Busa-Fekete, Karlson Pfannschmidt, Timo Klerx, and Eyke Hüllermeier. Extreme F-measure maximization using sparse probability estimates. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1435–1444, 2016.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. PhysioNet, version 2.2, 2023a. URL https://doi.org/10.13026/6mm1-ek67.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV-Note: Deidentified free-text clinical notes. PhysioNet, version 2.2, 2023b. URL https://doi.org/10.13026/1n74-ne17.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. PhysioNet, version 3.1, 2024. URL https://doi.org/10.13026/kpb9-mt58.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023c.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks. https://kellerjordan.github.io/posts/muon/, 2024.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Oluwasanmi Koyejo, Nagarajan Natarajan, Pradeep Ravikumar, and Inderjit S. Dhillon. Consistent multilabel classification. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Maxat Kulmanov, Francisco J. Guzmán-Vega, Paula Duek Roggli, Lydie Lane, Stefan T. Arold, and Robert Hoehndorf. Protein function prediction as approximate semantic entailment. *Nature Machine Intelligence*, 6(2):220–228, 2024. doi: 10.1038/s42256-024-00795-w.

Fei Li and Hong Yu. ICD coding from clinical text using multi-filter residual convolutional neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8180–8187, 2020.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2999–3007, 2017. doi: 10.1109/ICCV.2017.324.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Zachary C. Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. Optimal thresholding of classifiers to maximize F1 measure. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, volume 8725 of *Lecture Notes in Computer Science*, pp. 225–239, 2014. doi: 10.1007/978-3-662-44851-9_15.

Jinghui Liu and Anthony Nguyen. Bridging the version gap: Multi-version training improves ICD code prediction, especially for rare codes. In *Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026)*, pp. 372–381, 2026. doi: 10.18653/v1/2026.bionlp-1.29.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *International Conference on Learning Representations*, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Junyu Luo, Xiaochen Wang, Jiaqi Wang, Aofei Chang, Yaqing Wang, and Fenglong Ma. CoRelation: Boosting automatic ICD coding through contextualized code relation learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pp. 3997–4007, 2024.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

Daniel Morales-Brotons, Thijs Vogels, and Hadrien Hendrikx. Exponential moving average of weights in deep learning: Dynamics and benefits. *Transactions on Machine Learning Research*, 2024.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of NAACL-HLT*, pp. 1101–1111, 2018.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *Computer Vision – ECCV 2020*, volume 12370 of *Lecture Notes in Computer Science*, pp. 681–699. Springer, 2020. doi: 10.1007/978-3-030-58595-2_41.

Thanh-Tung Nguyen, Viktor Schlegel, Abhinav Kashyap, Stefan Winkler, Shao-Syuan Huang, Jie-Jyun Liu, and Chih-Jen Lin. MIMIC-IV-ICD: A new benchmark for extreme multilabel classification. *arXiv preprint arXiv:2304.13998*, 2023.

Matteo Pagliardini, Pierre Ablin, and David Grangier. The AdEMAMix optimizer: Better, faster, older. In *International Conference on Learning Representations*, 2025.

Wongi Park, Inhyuk Park, Sungeun Kim, and Jongbin Ryu. Robust asymmetric loss for multi-label long-tailed learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 2703–2712, 2023. doi: 10.1109/ICCVW60793.2023.00286.

PhysioNet. Responsible use of MIMIC data with online services like GPT. https://physionet.org/news/post/gpt-responsible-use, 2023.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, 2026. doi: 10.1038/s44360-026-00096-z.

Predrag Radivojac, Wyatt T. Clark, Tal Ronnen Oron, Alexandra M. Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, et al. A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3):221–227, 2013. doi: 10.1038/nmeth.2340.

Mucheng Ren, He Chen, Yucheng Yan, Danqing Hu, Jun Xu, and Xian Zeng. TraceCoder: Towards traceable ICD coding via multi-source knowledge integration. In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2730–2735, 2025. doi: 10.1109/BIBM66473.2025.11356294.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 82–91, 2021.

Debjyoti Saha Roy, Byron C. Wallace, and Javed A. Aslam. Don't pay attention, PLANT it: Pretraining attention via learning-to-rank. *arXiv preprint arXiv:2410.23066*, 2025.

Erik Schultheis, Marek Wydmuch, Wojciech Kotłowski, Rohit Babbar, and Krzysztof Dembczyński. Generalized test utilities for long-tail performance in extreme multi-label classification. In *Advances in Neural Information Processing Systems*, volume 36, pp. 22269–22303, 2023.

Erik Schultheis, Wojciech Kotłowski, Marek Wydmuch, Rohit Babbar, Strom Borman, and Krzysztof Dembczyński. Consistent algorithms for multi-label classification with macro-at-k metrics. In *International Conference on Learning Representations*, 2024.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Thomas Sounack, Joshua Davis, Brigitte Durieux, Antoine Chaffin, Tom J. Pollard, Eric Lehman, Alistair E. W. Johnson, Matthew McDermott, Tristan Naumann, and Charlotta Lindvall. Bio-Clinical ModernBERT: A state-of-the-art long-context encoder for biomedical and clinical NLP. *arXiv preprint arXiv:2506.10896*, 2025.

Mohammed Sameer Syed and Xuan Lu. ShifaMind: A multiplicative concept bottleneck for interpretable ICD-10 coding. *arXiv preprint arXiv:2605.08482*, 2026.

The Gene Ontology Consortium. The Gene Ontology knowledgebase in 2026. *Nucleic Acids Research*, 54(D1):D1779–D1792, 2026. doi: 10.1093/nar/gkaf1292.

The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 2025. doi: 10.1093/nar/gkae1010.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. A label attention model for ICD coding from clinical text. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3335–3341, 2020.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and stabilizing Shampoo using Adam for language modeling. In *International Conference on Learning Representations*, 2025.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547, 2025. doi: 10.18653/v1/2025.acl-long.127.

Yifan Wu, Weiyan Qiu, Min Zeng, Xi Chen, Min Li, and Hongtao Zhu. GoM-ICD: Automatic ICD coding with gap schemes and mixture of experts. *Big Data Mining and Analytics*, 8(6): 1211–1224, 2025a. doi: 10.26599/BDMA.2025.9020019.

Yuhao Wu, Yifan Wu, Wei Fan, and Min Li. Enhancing ICD classification with semantic embedding rectification and long-tail refinement. *Knowledge-Based Systems*, 330:114530, 2025b. doi: 10.1016/j.knosys.2025.114530.

Nan Ye, Kian Ming A. Chai, Wee Sun Lee, and Hai Leong Chieu. Optimizing F-measures: A tale of two approaches. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. Code synonyms do matter: Multiple synonyms matching network for automatic ICD coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 808–814, 2022.

Wenqiao Zhang, Changshuo Liu, Lingze Zeng, Bengchin Ooi, Siliang Tang, and Yueting Zhuang. Learning in imperfect environment: Multi-label classification with long-tailed distribution and partial labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1423–1432, 2023. doi: 10.1109/ICCV51070.2023.00137.

A Decision rules

All comparisons evaluate frozen per-note, per-label logits $z \in \mathbb{R}^{N \times L}$ with a per-label additive offset vector b: the prediction is $z + b > 0$. The supplementary scorer, `protocol_compare_generic.py`, scores a fixed threshold of 0.5, rule A, and rule B from saved validation and test logits for any label space. Table 6 summarizes the rules used by published systems.

Rule A. The offset is $b = -\mathrm{logit}(\theta)\,\mathbf{1}$, with θ the validation micro-F1-optimal probability threshold from $\mathrm{linspace}(0, 1, 100)$; the classifier re-training results of Table 4 and Figure 3b use a 0.01-step grid instead (0.58 and 0.62).

Rule B. The offset is $b_\ell = \beta_{\mathrm{bin}(\ell)}$, with five bins by training-positive count, edges $\{1, 21, 101, 1001\}$, i.e., unseen, 1–20, 21–100, 101–1,000, and over 1,000; each β is chosen from $\{-2.0, -1.9, \ldots, 2.0\}$ by four rounds of coordinate descent on the micro–macro objective, $2\,\mathrm{micro} + \mathrm{macro}$, over validation. On the 7,942-code split the bins hold 3, 3,575, 2,822, 1,291, and 251 labels; on the 26,096-code split, 881, 20,261, 3,128, 1,519, and 307. The bin edges were chosen before the granularity study of §3.3 and not changed afterwards; that study replaces them with k equal-count frequency quantiles.

Rule C. The threshold is swept over $\{0.30, 0.35, 0.40, 0.45, 0.50\}$ on the test split, as the released GoM-ICD code does, and the best one is kept separately for each metric, as the paper describes for its ablation table.

Log-frequency bias. A two-parameter alternative, $b_\ell = c - \tau \log(1 + n_\ell)$, fit on validation on the same objective, scores within 0.002 macro-F1 of the five bins on the base model (0.6094/0.3221 against 0.6118/0.3225) and on the three-model mean (0.6170/0.3441 against 0.6201/0.3422).

Validation halves. Development fits on the first 6,689 validation notes and scores on the remaining 6,689; a system that wins on the scoring half, with a ranking consistent across halves, is refit on full validation before its test evaluation.

Validation or test. Whether a threshold is selected on validation or on test matters little: choosing the global threshold on test changes micro-F1 by at most +0.0001 across nine models, so the advantage of per-metric thresholds over rule A on the best GoM-ICD member (+0.016 macro-F1, Table 9) comes mainly from selecting the threshold for macro-F1. Per-label thresholds fit on the test labels themselves reach only 0.31 micro-F1 on the base model, because when a label's positives are hard to separate, the threshold that maximizes its F1 predicts it on many notes.

Table 6: Decision rules of recent ICD coding papers.

work	decision rule
Edin et al. (2023), reference	one global threshold tuned on validation for micro-F1; macro-F1 reported at that threshold
Ren et al. (2025)	one global threshold tuned on the development split
Wu et al. (2025a), state of the art	headline threshold not stated; released code selects a global threshold from $\{0.30, \ldots, 0.50\}$ on test for micro-F1; ablations report each metric at its own test-selected threshold
Choi et al. (2026)	three frequency-group thresholds tuned on validation for macro-F1
Syed & Lu (2026)	fixed 0.5
Liu & Nguyen (2026)	not stated in the paper

B The trained models

The 119 checkpoints come from more than 60 training runs; EMA decays, raw weights, and evaluation epochs of one run are counted as separate checkpoints. They use the encoder and head of §6 and

vary: optimizer (8-bit AdamW, Muon-hybrid (Jordan et al., 2024; Liu et al., 2025), AdEMAMix-hybrid (Pagliardini et al., 2025), Lion (Chen et al., 2023), SOAP-hybrid (Vyas et al., 2025)), negative focusing $\gamma^- \in \{3, 4, 5, 6\}$, the asymmetric loss probability shift $\in \{0.004, 0.005, 0.02, 0.04\}$, two loss variants, one adapted from COMIC (Zhang et al., 2023) through its simplification COMIC-S (Choi et al., 2026) and one from the robust asymmetric loss (Park et al., 2023), rare-label gradient boosts, epochs (4 to 7) and evaluation epoch, EMA decay $\in \{0.999, 0.995, 0.9997\}$ and raw weights, seeds, classifier re-training, low-rank adaptation (Hu et al., 2022), and distillation students; variants of one run, such as different evaluation epochs or EMA decays, count separately. Every run multiplies the gradient of each label's query row by a factor that grows with rarity, $m(n) = 1 + s\,(\min\{1.25,\ 1 + 0.275/(1 + (n/30)^{1.22})\} - 1)$ for a label with n training examples ($m = 1$ for $n = 0$), ramped in linearly over the first 714 optimizer steps by default. The base configuration uses $s = 1$, a factor of at most 1.25; runs with a rare-label boost use $s = 4$, at most $2\times$, or in a few cases $s = 8$. Muon-hybrid and SOAP-hybrid update only the encoder's attention and feed-forward weight matrices, AdEMAMix-hybrid the whole encoder; all other parameters stay on AdamW. Three seeds of the base configuration, at probability shift 0.04 rather than 0.02, give a seed standard deviation of 0.002 micro-F1 and 0.012 macro-F1 at a fixed threshold of 0.5 on validation.

The members of the three-model systems of Tables 4 and 5 are, on the 7,942-code split, two seeds of a Muon-hybrid configuration with probability shift 0.005 and an AdamW model with a rare-label boost, seven epochs, and $\gamma^- = 6$; on the 26,096-code split, the base configuration, a Muon-hybrid model with $\gamma^- = 4$, and an AdamW model with the same rare-label boost, seven epochs, and $\gamma^- = 6$. The fifteen-member ensemble combines five optimizers, each with three members trained to favor micro-F1, macro-F1, or both; the eleven-member ensemble is the subset of these fifteen that scored best on validation. A learned per-bin fusion fits member weights, temperatures, and a bias within each frequency bin of rule B on validation. Under rule B it matches the plain mean on the three 7,942-code members ($0.6200/0.3417$ against $0.6201/0.3422$) and trails it on eleven (Table 11). Under rule A it scores higher on macro-F1 ($0.6239/0.3117$ against $0.6257/0.2943$) because its per-bin biases already act as bin-wise thresholds; to keep the rule separate from the model, we report the mean.

Rank diversity and ensemble lift. We compared four three-member ensembles, each averaged and thresholded by rule B fit on one validation half and scored on the other: the base model with two models trained at $\gamma^- = 4$ and 5; the base model with a probability-shift variant and a second seed; three models trained to favor macro-F1; and our three members. In that order, the mean per-label Spearman correlation between members falls from 0.75, about the level of two seeds of the base model (0.74), to 0.60, and the gain on the micro–macro objective over the best member rises from $+0.022$ to $+0.042$.

Classifier re-training. As in cRT (Kang et al., 2020) the encoder is frozen, but the head is fine-tuned from the base model rather than re-initialized, with $\gamma^- = 5$ and the rare-label gradient boost. This yields $+0.017$ macro-F1 at equal micro-F1 under rule A (Table 4).

C The GoM-ICD reproduction

The members follow the configuration in Table 2 and Section 4.3 of Wu et al. (2025a), with the repository's default encoder and the hidden size of 384 that its code requires. As the repository directs, we used PLM-ICD's evaluator and the reference implementation's preprocessed MIMIC-IV data, and we built the label-title embeddings, which are not released, from the ICD dictionaries. The changes we made to run the released code are included as patches in the supplementary material. Our members rank in the published order and fall below the values in Table 5 of Wu et al. (2025a) by 0.010, 0.034, and 0.071 micro-F1, from the best member to the weakest.

D The 26,096-code benchmark

The admission-identifier splits released with Nguyen et al. (2023) were matched to the hospital tables of MIMIC-IV version 3.1 (Johnson et al., 2024), with notes from MIMIC-IV-Note version 2.2; after uppercasing and removing dots, the split admissions carry 26,096 distinct ICD-10 codes (16,153 diagnosis, 9,943 procedure), matching the published label count. Note text follows the

preprocessing of Edin et al. (2023) rather than that of the benchmark paper; both apply CAML-style normalization but differ in tokenization details. Thirteen admissions with no ICD-10 code and 38 split identifiers absent from the note table were removed, leaving 110,398 / 4,014 / 7,847 notes.

Three members were trained as in §6, varying by optimizer and loss (Appendix B). The EMA variant of each member and both rules' parameters were chosen on full validation and frozen before the single test pass; the pre-registered prediction was that the rule would gain more than on the 7,942-code split. Table 7 lists the members.

Table 7: The three 26,096-code members on the test split under both rules, full-matrix micro-F1 and macro-F1 over the 9,158 test-active labels. The last row is the three-model system of Table 5, where its macro-F1 averages over all 26,096 labels.

member	A micro	A macro	B micro	B macro
AdamW, γ^-=3, shift 0.02	0.6087	0.2188	0.5833	0.2791
Muon-hybrid, γ^-=4	0.6091	0.2286	0.5760	0.2902
AdamW, γ^-=6, rare boost, 7 epochs	0.6034	0.2199	0.5738	0.2776
mean of logits	0.6171	0.2182	0.5890	0.3011

E METRIC CONVENTIONS

Macro-F1 form. The reference implementation and this paper average per-label F1; the CAML-lineage evaluator that GoM-ICD inherits computes $2PR/(P+R)$ from macro-averaged precision and recall. On the reproduced members the two forms differ by $+0.014$ to $+0.021$ across the rules of Table 9. The published $0.211 \rightarrow 0.285$ trajectory thus mixes forms, which makes the 46% share of §3.1 conservative.

Macro-F1 label set. On the 26,096-code split, published macro-F1 averages over all labels (Choi et al., 2026), so the 16,938 labels without a test positive, whose F1 is zero under any rule, enter the mean. This multiplies every macro-F1 on the split, and every difference between rules, by $9{,}158/26{,}096 = 0.35$. We use this convention only in Table 5, to compare with published results. Everywhere else we average over test-active labels, so that the rule effects of §3 are comparable across the two splits. On the 7,942-code split the choice hardly matters: averaging over all 7,942 columns instead of the 7,935 test-active labels changes macro-F1 by at most 0.0003.

Micro-F1 denominator. Micro-F1 can count false positives on every label column, the full matrix, or only on labels with a test positive. On the 7,942-code split, where 7,935 labels are test-active, the two agree within 0.001. On the 26,096-code split, 16,938 labels have no test positive; counting false positives on them, the standard convention, reduces rule B's micro-F1 from 0.6063 to 0.5890 and rule A's from 0.6181 to 0.6171. Table 5 uses the full-matrix convention.

F THE RELEASED CHECKPOINTS

The PLM-ICD checkpoint for MIMIC-IV ICD-10 released by Edin et al. (2023) was evaluated with the model code and 128-token chunking of the reference implementation, the checkpoint's saved label index, and the benchmark's released split files. On validation, rule A selects the stored threshold, 0.3636 on the implementation's 100-point grid. The checkpoint's macro-F1 under rule A is 0.2248, against a published 0.211; the published figure is a ten-seed mean with a standard deviation of 0.023, and a single released seed lies within one standard deviation of it. The bootstrap of §3.1 draws 2,000 resamples of test notes; its distribution lies slightly below the point estimate because resampling notes thins the support of rare labels.

The other released checkpoints. The same authors released CNN, Bi-GRU, CAML, MultiResCNN, and LAAT checkpoints for this benchmark. We scored them with the reference implementation's model code and each checkpoint's own vocabulary, and pre-registered three predictions:

R1, rule B exceeds rule A on test macro-F1 for every model; R2, across the models the gain increases with rare-label AUC on validation; R3, test macro-F1 is more closely ordered by rare-label AUC under rule B than under rule A. A model entered the comparison only if rule A on validation selected exactly the threshold stored in its checkpoint. Four did; MultiResCNN selected 0.444 against a stored 0.434 and was excluded before any test scores were computed. The rules were fit on validation and frozen before one test pass. R1 and R2 hold (Table 8; Spearman 0.9 across the five models); R3 does not, because the five models keep the same order under both rules.

Table 8: Checkpoints released by Edin et al. (2023) on the 7,942-code test split, with both rules fit on validation and macro-F1 averaged over test-active labels. Rare-label AUC is the mean per-label AUC on validation over labels with 1 to 20 training examples. The PLM-ICD row is the evaluation of §3.1.

model	rare-label AUC	A micro	A macro	B micro	B macro	B − A macro
CNN	0.863	0.4763	0.0836	0.4753	0.0863	+0.0027
Bi-GRU	0.914	0.4981	0.1116	0.4954	0.1246	+0.0131
CAML	0.875	0.5537	0.1628	0.5503	0.1800	+0.0172
LAAT	0.950	0.5799	0.2047	0.5741	0.2292	+0.0245
PLM-ICD	0.962	0.5890	0.2248	0.5832	0.2588	+0.0340

G ADDITIONAL RESULTS

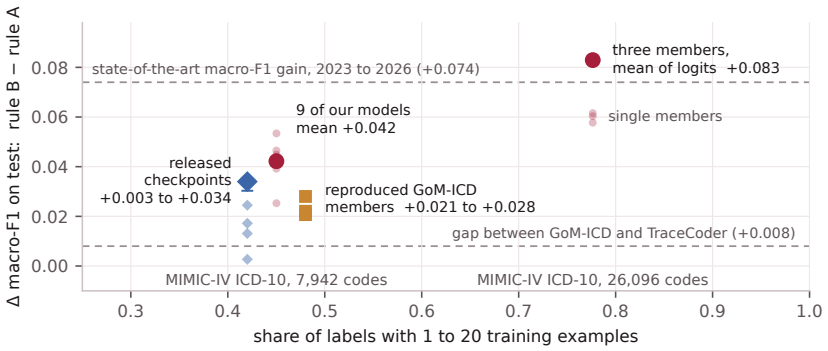

Figure 2: Gain in test macro-F1 from replacing rule A with rule B on the two MIMIC-IV label spaces, against the share of labels with 1 to 20 training examples. On the 7,942-code split, the five released checkpoints, nine of our models, and the reproduced GoM-ICD members all have $x = 0.45$ and are offset by ± 0.03 for readability; on the 26,096-code split, the points are the three members and the mean of their logits. The error bar is PLM-ICD's 95% bootstrap interval. The dashed lines give two published differences on the 7,942-code split for scale: GoM-ICD over TraceCoder, and the macro-F1 gain since 2023. The label spaces of §4 are left out because outside MIMIC-IV tail share does not predict the effect.

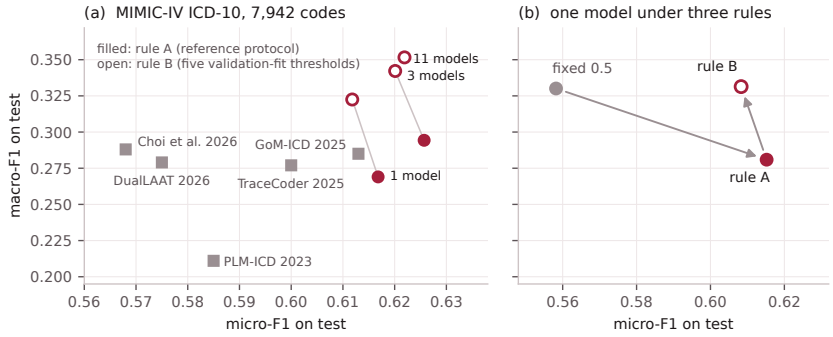

Figure 3: (a) The 7,942-code leaderboard in the (micro, macro) plane. Gray squares are published systems at their reported numbers; filled circles are ours under rule A, open circles under rule B; the eleven-model system is shown under rule B only. (b) A classifier-retrained model, whose head ($\gamma^- = 5$) already favors the tail at 0.5: test macro-F1 is 0.330 at a fixed 0.5, 0.281 under rule A, and 0.331 under rule B.

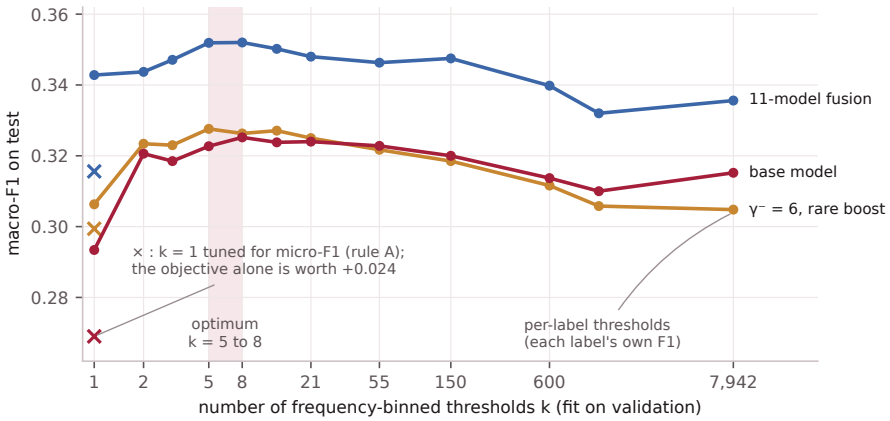

Figure 4: Test macro-F1 against the number k of equal-count frequency bins, each with its own threshold fit on validation for the micro–macro objective, for the base model, the $\gamma^- = 6$ rare-boost model of Table 10, and the eleven-model per-bin fusion of Appendix B. All three curves peak between $k = 5$ and 8. The last point ($k = 7{,}942$) instead uses per-label thresholds that each maximize their own label's F1, and falls below $k = 8$ on every model. Crosses mark $k = 1$ tuned for micro-F1, which is rule A.

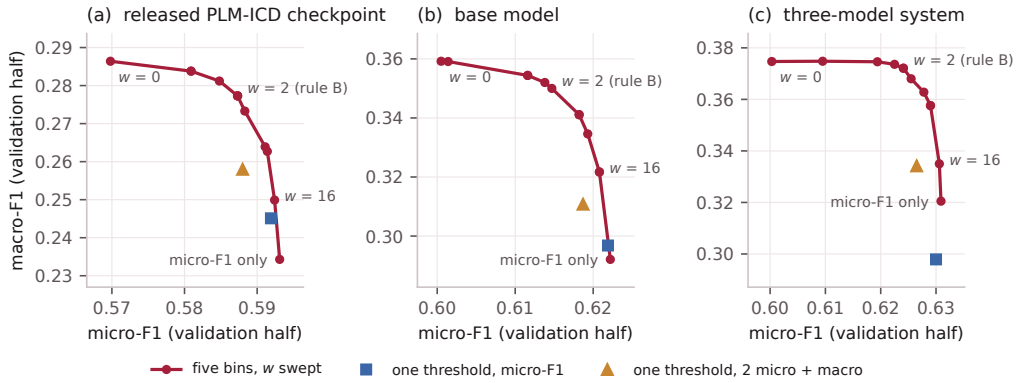

Figure 5: Micro-F1 against macro-F1 on validation halves as the weight w on micro-F1 in the micro–macro objective $w\,\mathrm{micro} + \mathrm{macro}$ is swept, for five binned thresholds (curve) and for one global threshold tuned for micro-F1 or for the micro–macro objective with $w = 2$ (markers), all fit on the first half of validation and scored on the second. $w = 2$ is rule B.

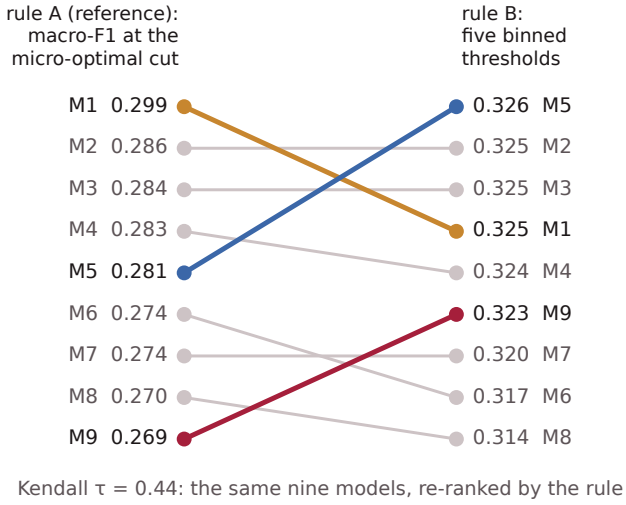

Figure 6: Nine single models ranked by test macro-F1 under rule A (left) and rule B (right); Kendall $\tau = 0.44$. M1–M9 are the rows of Table 10 in order of rule A macro-F1.

Table 9: The reproduced GoM-ICD members on the 7,942-code test split under rules A, B, and C, and for the gap-0 member also at a fixed 0.5. A and B are fit on validation. C is selected on test separately for each metric (§2): micro-F1 at 0.35 and macro-F1 at 0.30. Macro-F1 is given as the mean of per-label F1 over test-active labels and in the harmonic form GoM-ICD reports. Members are named by gap scheme; the gap-0 member is the best of the three.

model	rule	micro-F1	macro-F1 (mean of F1)	macro-F1 (harmonic)
gap-0 member	fixed 0.5	0.5865	0.2211	0.2402
gap-0 member	A	0.5958	0.2483	0.2675
gap-0 member	C, test-selected	0.5957	0.2641	0.2821
gap-0 member	B	0.5927	**0.2763**	**0.2972**
gap-2 member	A	0.5644	0.2074	0.2227
gap-2 member	C, test-selected	0.5644	0.2162	0.2314
gap-2 member	B	0.5613	0.2294	0.2473
gap-1 member	A	0.5241	0.1620	0.1763
gap-1 member	C, test-selected	0.5239	0.1684	0.1828
gap-1 member	B	0.5195	0.1827	0.1992

Table 10: Nine single models on the 7,942-code test split under both rules, fit on validation. The first row is the base model, and the last two are the distillation control and student.

model	A micro	A macro	B micro	B macro
AdamW, γ^-=3, shift 0.02	0.6168	0.2690	0.6118	0.3225
same, seed 3	0.6170	0.2735	0.6117	0.3200
Muon-hybrid, shift 0.005, seed 3	0.6147	0.2814	0.6088	0.3262
Muon-hybrid, shift 0.005, seed 1	0.6137	0.2698	0.6094	0.3138
AdEMAMix-hybrid, γ^-=5, rare boost	0.6156	0.2736	0.6099	0.3173
AdamW, γ^-=5, rare boost, epoch 5	0.6124	0.2859	0.6083	0.3251
AdamW, γ^-=6, rare boost, epoch 7	0.6045	0.2994	0.5989	0.3247
distillation control, λ=0	0.6148	0.2836	0.6106	0.3251
distilled student	0.6144	0.2830	0.6102	0.3242
mean Δ macro	+0.0422 (min +0.0253, max +0.0534)			

Table 11: Ways of combining the eleven members on the 7,942-code test split, all under rule B. The two fusions and the stacker are fit on validation. Objective: the micro–macro objective, $2\,\text{micro} + \text{macro}$.

pooling	micro-F1	macro-F1	objective
mean of member logits, no fitting	0.6219	0.3516	1.5955
linear per-bin fusion, coordinate descent on the objective	0.6238	0.3475	1.5950
gated mixture-of-experts stacker, 10 experts, top-2, cross-entropy	0.6242	0.3466	1.5950
linear per-bin fusion, cross-entropy-trained	0.6204	0.3484	1.5893

Distillation. Students are initialized from the base model and trained for one more epoch with a per-label distillation term toward the fifteen-member ensemble of Appendix B, added to the training loss with weight λ. On validation, the micro–macro objective and macro-F1 both fall as λ increases through three values, and the $\lambda = 0$ control, trained for the same extra epoch, is best on both. On test under rule B, the student scores 1.5447 on the micro–macro objective, against 1.5462 for the control and 1.5460 for the base model it started from (Table 10).

H Outside-domain label spaces

Pre-registration. Before opening any of the data below, we committed the protocol and four predictions to version control. H1: rule B exceeds rule A on test macro-F1 in every label space. H2:

across all label spaces in this paper, the gain increases with tail share (positive Spearman correlation). H3: on each protein subontology, the gain is larger on all terms than on the terms with at least 50 training proteins. H4: in every label space, the bin of labels with 1 to 20 training examples keeps a mean per-label AUC of at least 0.85 while its F1 at a fixed 0.5 is at most half that of the most frequent bin. The rule parameters for each label space were frozen before its single test pass. The intervals of Table 2 are paired bootstraps over test examples, with the rules and the active-label set fixed. H1 and H4 hold in all four label spaces; H2 and H3 do not (§4, Table 12).

EURLEX57K. We use the standard split of 45,000 / 6,000 / 6,000 documents (Chalkidis et al., 2019), with the title and recitals as text. The label set is the 4,271 EuroVoc concepts that occur in any split, 163 of which have no training example. Features are TF-IDF over word unigrams and bigrams (50,000 features, sublinear term frequency, fit on training text), and a single linear layer is trained on them with binary cross-entropy and AdamW. We compared four learning-rate and weight-decay settings by validation micro-F1 at 0.5 with early stopping; the selected model (learning rate 0.03, no weight decay, epoch 31) reaches validation micro-F1 0.730.

Protein function. We use the UniProtKB/Swiss-Prot (The UniProt Consortium, 2025) benchmark released with DeepGO-SE (Kulmanov et al., 2024) and its sequence-similarity split: 38,533 / 1,901 / 2,845 proteins over 6,851 terms for molecular function, 52,584 / 2,870 / 3,275 over 21,356 for biological process, and 52,072 / 2,964 / 4,221 over 2,829 for cellular component. Labels are the authors' propagated experimental annotations, restricted to their term lists, with Gene Ontology terms (Ashburner et al., 2000; The Gene Ontology Consortium, 2026) from the release distributed with the benchmark (2021-11-16). The model is their MLP baseline on 2,560-dimensional ESM2 embeddings (Lin et al., 2023), built with their model and data-loading code: a 1,024-unit hidden block followed by a residual block of the same width, trained with Adam (Kingma & Ba, 2015) at learning rate 10^{-3} and batch size 37 for 256 epochs, keeping the checkpoint with the lowest validation loss. Rules A and B act on the unpropagated logits. To check the reproduction, we compute protein-centric Fmax (Radivojac et al., 2013) with their evaluation procedure, which propagates predictions and selects the threshold on test: 0.532, 0.417, and 0.719 for molecular function, biological process, and cellular component, against published values of 0.517, 0.423, and 0.717.

Table 12: Gain of rule B over rule A in test macro-F1, on all terms and on terms with at least 50 training proteins, with both rules refit on validation for each set. Restricting to well-supported terms increases the gain on all three subontologies, the opposite of what H3 predicted.

	all terms			terms with ≥ 50 training proteins			
subontology	A	B	B − A	labels	A	B	B − A
molecular function	0.068	0.073	+0.005	685	0.160	0.179	+0.019
biological process	0.042	0.043	+0.0003	4,131	0.075	0.076	+0.0013
cellular component	0.116	0.136	+0.020	606	0.158	0.193	+0.034