EEG Muse báo trước N2 khoảng 90 giây

Trên 540 bản ghi Muse công khai, EEG chỉ chuyển sang trạng thái "giống ngủ" khoảng **90 s trước epoch N2 đầu tiên** (median của bước nhảy sleep-likeness score là -90 s, IQR -140 đến -55 s), và trong 90 s cuối mọi feature đều thay đổi theo dạng ramp tuyến tính chứ không có bước nhảy nào ở mốc 40 s. Hệ quả trực tiếp cho W-bMAE: bin 0-40 s chỉ chứa 5.3 % số cửa sổ nhưng mang **52.9 % trọng số**, trong khi một cửa sổ đơn lẻ chỉ tách được bin này khỏi bin 40-90 s với AUC 0.65-0.70 (tổ hợp tốt nhất có smoothing 30 s đạt 0.75). Hai baseline hiện có (boosted trees và CNN+GRU, khoảng 47 s mỗi bản ghi trong cross-validation so với 62 s của hằng số 35 s) thắng hằng số 19 điểm ở các bin 90-600 s nhưng **trả lại 3.5-4.3 điểm ở các bin 0-90 s**, và dự đoán của chúng không bao giờ vượt quá khoảng 300 s dù 44 % cửa sổ có target từ 450 s trở lên. Ba hướng có bằng chứng mạnh nhất vì thế đều nằm ở phần "đầu ra và dạng input" chứ không ở encoder: một head phân phối (hazard, ordinal hoặc soft-bin) giải mã bằng weighted median kèm lớp cap 600 s tường minh; input causal tương đối theo bản ghi (EMA 30 s trừ running mean, đi cùng mức thô); và huấn luyện bền với trạng thái khởi động của bản ghi (random-start crop), vì elapsed time vừa là predictor đơn lẻ mạnh nhất (rho -0.44) vừa là sản phẩm của cách cắt bộ dữ liệu công khai ở 25 phút. Y văn không có công trình nào dự đoán "số giây còn lại đến N2 đầu tiên": tiền lệ gần nhất (Li et al. 2025) dự đoán tipping point với sai số 0.82 ± 2.14 phút trên PSG phòng lab, nên độ chính xác cỡ phút, không phải cỡ chục giây, là mốc tham chiếu thực tế. Trong ba bộ PSG phụ trợ chỉ PhysioNet/CinC 2018 có khối lượng đáng kể (292 h tín hiệu trước N2 so với 110 h của Muse), và chỉ các feature tương đối (sigma/beta, độ dốc phổ) chuyển được sang Muse. Các foundation model trong neuralbench/braindecode dùng được về mặt kỹ thuật (REVE-base là model duy nhất được cache sẵn trên worker) nhưng bằng chứng độc lập cho thấy chúng chỉ nhỉnh hơn mạng nhỏ trên cửa sổ 5 s với ít kênh, nên nên được xem là một thí nghiệm ưu tiên thấp chứ không phải nền tảng của lời giải. Hai giới hạn cần nhớ khi đọc báo cáo: scorer của tập sealed chưa công khai, và mọi khuyến nghị rút từ y văn đều chưa được kiểm chứng trên chính bài toán này.

Mục lục

Bối cảnh, quy ước và nguồn số liệu 2

Dữ liệu Muse: tín hiệu nằm ở 90-600 s, trọng số nằm ở 90 s cuối 2

540 bản ghi của 203 người, và chỉ 39 % phương sai onset thuộc về "người" 2

Bin 0-40 s chứa 5 % cửa sổ nhưng 53 % trọng số 3

Artifact giảm từ 20-30 % xuống 1-3 % khi tiến gần N2 4

Công suất trên 16 Hz giảm đơn điệu, delta có hình chữ U 5

Hai họ feature: "slow drifters" suốt 10 phút và "late movers" trong 2 phút cuối 6

90 giây cuối là một ramp, không có bước nhảy ở mốc 40 s 9

Không feature causal đơn lẻ nào vượt rho tuyệt đối 0.4, và 35-50 % là "đồng hồ" 10

Hai baseline thắng 19 điểm ở xa onset rồi trả lại 3.5-4.3 điểm ở gần onset 11

Các đêm của một người khác nhau gần bằng những người khác nhau 12

Train và test không tách được, trừ phút đầu tiên sạch hơn 13

Nhóm off-grid bị cắt mất phần khởi động, và không có dấu hiệu nhãn sai 14

Ba bộ PSG phụ trợ: chỉ CinC-2018 có khối lượng, và chỉ feature tương đối chuyển được 14

Y văn: chưa ai dự đoán "số giây đến N2", và N1 là điểm yếu của mọi stager 17

Sinh lý chuyển tiếp: tipping point 1.5-3.5 phút trước N2, với độ phân tán 2-3 phút 17

Nhãn "N2 đầu tiên" tự mang sàn sai số cỡ chục giây 17

Tiền lệ dự đoán: sai số cỡ phút trên PSG, không có con số nào theo giây 17

SOTA sleep staging: MF1 78-83 trên Sleep-EDF, N1 F1 quanh 50 18

Lợi ích của pre-training lớn nhất khi dữ liệu đích hiếm 19

Dữ liệu có thể thêm: BOAS là bộ gần Muse nhất và có giấy phép CC0 19

Phương pháp có bằng chứng: chuẩn hoá causal theo bản ghi, head phân phối, và bền với artifact 19

Tiền xử lý: lọc causal có trạng thái, dừng ở 45 Hz, và không chuẩn hoá mất biên độ . . . 20

Artifact: báo cho model biết thay vì làm sạch . . . 20

Feature: quỹ đạo đã smoothing của feature phổ đơn giản mang phần lớn sức dự đoán . . . 20

Augmentation: lợi ích nhỏ ở cỡ dữ liệu này, ưu tiên các dịch chuyển đã đo được . . . 21

Domain adaptation: chuẩn hoá phổ cho +1 điểm trung bình và +5-6 điểm ở cohort lệch nhất . . . 21

Mô hình hoá: dự đoán một phân phối rồi giải mã bằng weighted median . . . 22

Validation: nhóm theo người, chấm theo bản ghi, nhiều seed . . . 22

Pretrained model trong neuroai: dùng được, nhưng bằng chứng không ủng hộ đặt cược vào chúng . . . 23

Những gì thực sự có sẵn . . . 23

Hiệu năng: nhỉnh hơn khi fine-tune trên epoch 30 s, yếu khi đóng băng, và không hơn mạng nhỏ trên cửa sổ ngắn ít kênh . . . 24

Khuyến nghị: một thí nghiệm có giới hạn, không phải nền móng . . . 24

Các thí nghiệm nên chạy tiếp, theo thứ tự ưu tiên . . . 24

Kết luận . . . 25

Tài liệu tham khảo . . . 26

Bối cảnh, quy ước và nguồn số liệu

Track 03 yêu cầu cứ mỗi 5 s, từ 4 kênh EEG Muse (TP9, AF7, AF8, TP10, 128 Hz, µV thô, causal), dự đoán số giây còn lại đến epoch N2 đầu tiên, cắt trần ở 600 s. Thước đo xếp hạng là W-bMAE: MAE tính riêng trong bốn bin của target thật [0, 40), [40, 90), [90, 300), [300, 600] s với trọng số 10 / 5 / 3 / 1, tính trong từng bản ghi trên các bin không rỗng rồi lấy trung bình qua các bản ghi; điểm sealed là trung bình của nhóm người đã thấy (seen) và chưa thấy (unseen) (competition README; NeuralBench Track 3 guide). Mốc tham chiếu đã có của dự án: trong cross-validation 5-fold seen/unseen, hằng số tốt nhất (35 s) đạt W-bMAE khoảng 62 s, boosted trees trên feature phổ causal khoảng 46 s và CNN+GRU causal khoảng 45 s; trên 40 bản ghi Muse test công khai hai model đạt 52.5 s và 55.8 s. Leaderboard warm-up trên Codabench ngày 2026-10-12 vẫn chấm trên proxy Sleep-EDF bằng bMAE không trọng số (CNN nhanh của nhóm: 114.7, khoảng median của bảng), và trên proxy đó target có thể đọc ra từ vị trí của cửa sổ trong bản ghi, nên các điểm top ở đó nói rất ít về mô hình hoá EEG.

Toàn bộ số liệu EDA Muse dưới đây được tính trên cache cục bộ `data/cache/muse_v1` bằng các script `scripts/muse_*.py` trong thư mục ghi chú; nguồn của từng con số là log tương ứng `scripts/logs/muse_NN.log`, được ghi trong ngoặc ở cuối mỗi đoạn. Quy ước: "cửa sổ" là một ô 5 s không chồng lấn tính từ đầu bản ghi (113,375 cửa sổ, trong đó 80,975 trước N2 là cửa sổ đánh giá); "TP" là trung bình TP9 và TP10, "AF" là trung bình AF7 và AF8; công suất tính bằng dB (10·log10 µV²); "dz" là effect size ghép cặp trong bản ghi; một kênh-cửa sổ là "BAD" khi max |x| > 200 µV hoặc RMS 30-45 Hz > 5 µV. Mọi feature theo cửa sổ đều causal; những phân tích dùng `n2_onset_s`, median cả bản ghi hoặc bộ lọc zero-phase được ghi rõ là "chỉ để phân tích". Các kết quả đo được và các suy luận được tách riêng trong từng mục.

Dữ liệu Muse: tín hiệu nằm ở 90-600 s, trọng số nằm ở 90 s cuối

540 bản ghi của 203 người, và chỉ 39 % phương sai onset thuộc về "người"

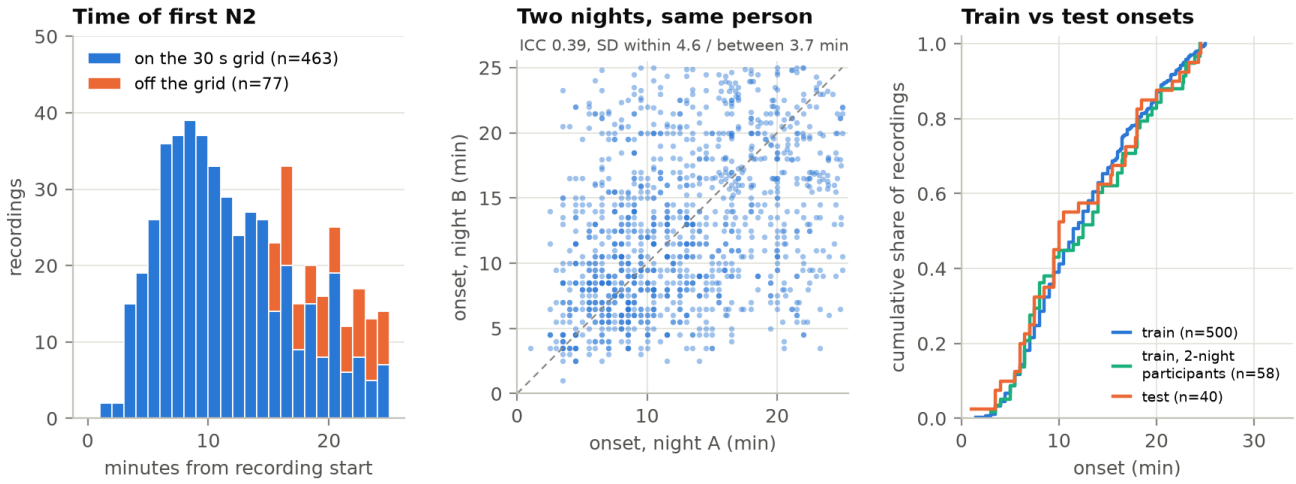
Bộ dữ liệu gồm **540 bản ghi ngắn (6-30 phút) của 203 người**, tổng 157.5 h, mỗi người có 1-5 đêm (1 đêm: 50 người, 2 đêm: 49, 3 đêm: 42, 4 đêm: 44, 5 đêm: 18). Thời điểm N2 đầu tiên tính từ đầu bản ghi có phân vị 5/25/50/75/95 là 268 / 450 / 690 / 1002 / 1379 s (trung bình 750 s, SD 350 s); 39.4 % bản ghi có onset dưới 600 s nên không bao giờ chạm trần, 7.0 % dưới 300 s, và không bản ghi nào có onset dưới 60 s. Độ dài bản ghi luôn bằng đúng onset + 300 s, tức độ dài chính là nhãn và tuyệt đối không được đưa vào

model (`muse_02.log`).

Hai đêm của cùng một người chỉ giống nhau yếu: ICC(1) của onset là **0.39** (bootstrap 95 % CI 0.29-0.48), SD trong người 275 s so với SD giữa người 220 s. Dự đoán onset của một đêm từ các đêm khác của cùng người (leave-one-night-out) giảm MAE từ 299 s (trung bình toàn cục) xuống 261 s (trung bình các đêm khác) và 252 s (ước lượng co theo ICC), tức cá nhân hoá prior chỉ đáng khoảng 13-16 % MAE của onset. Không có xu hướng đơn điệu theo số thứ tự session (Spearman 0.03), dù session 2 có onset muộn hơn session 1 trong dữ liệu ghép cặp (+76 s, Wilcoxon p = 0.012, n = 153).

Có một dấu vết cấu trúc đáng chú ý: 463 trong 540 onset (85.7 %) là bội số của 30 s tính từ đầu bản ghi, còn 77 bản ghi "off-grid" đều có onset từ 913.8 s trở lên và chiếm 41 % của 188 bản ghi có onset ≥ 900 s. Nhóm này được phân tích riêng ở mục cuối của phần EDA. Về train so với test: cả 40 người của tập test đều có đúng hai đêm (đêm test là session 1 ở 17 người, session 2 ở 23 người), onset median 690 s (train) so với 600 s (test), Mann-Whitney p = 0.72. Tập test công khai vì thế chỉ đo "đêm mới của người đã biết"; không thể dựng phép thử cho người chưa thấy từ chính split này, nên giao thức 5-fold nhóm theo người vẫn là công cụ validation chính.

Metadata BIDS để lại nhiều khoảng trống: phần cứng ghi là "Muse S family" nhưng thế hệ, firmware, điện cực reference và ground chưa được xác nhận; việc down-sample từ 256 Hz xuống 128 Hz là giả định của người biên tập; trường `PowerLineFrequency` luôn ghi 60 bất kể điện lưới thật; onset được định nghĩa "so với đầu bản ghi được cung cấp, không nhất thiết là lúc tắt đèn"; phương pháp chấm N2 không được mô tả; không có nhân khẩu học hay ngày giờ. Do đó không khảo sát được ảnh hưởng circadian hay tuổi lên onset.



Onset N2 trải từ 1 đến 25 phút, 77 bản ghi off-grid đều muộn hơn 15 phút, hai đêm của cùng một người chỉ tương quan yếu (ICC 0.39), và phân phối onset của train và test trùng nhau.

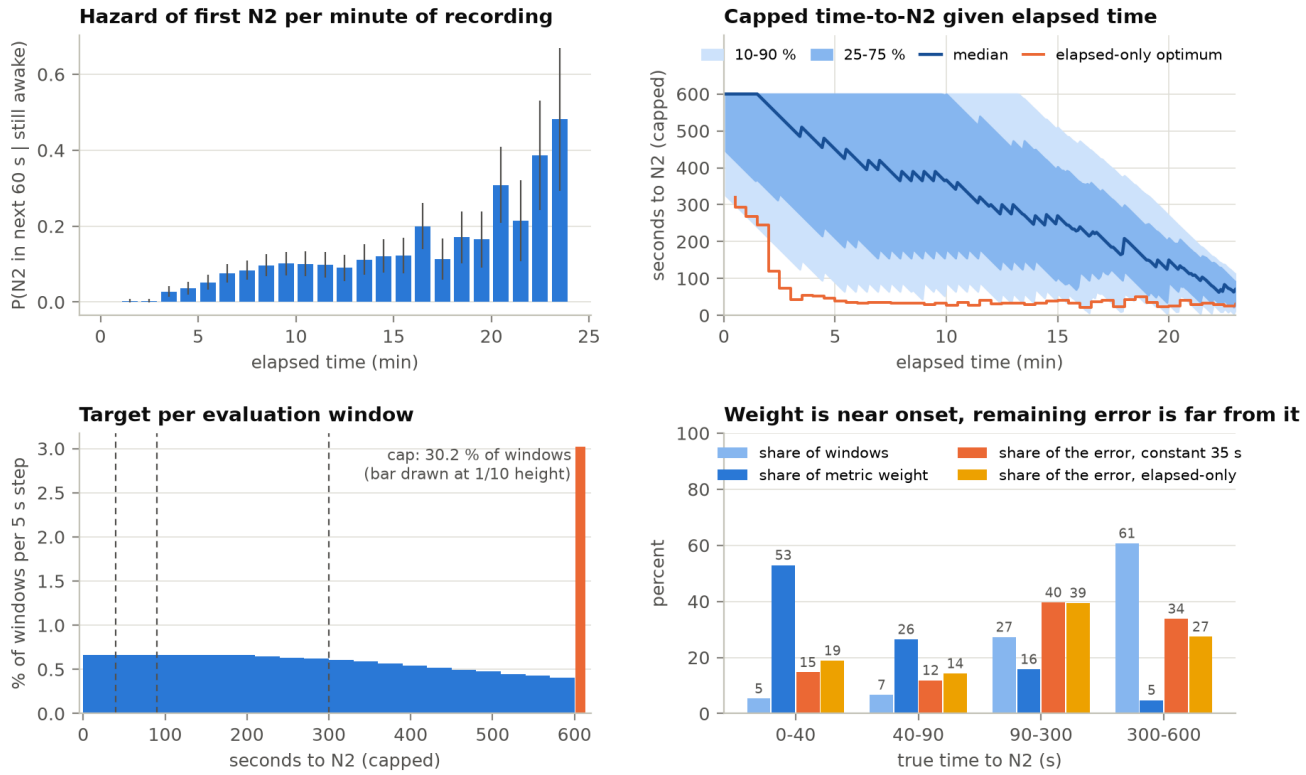
Bin 0-40 s chứa 5 % cửa sổ nhưng 53 % trọng số

Có 80,975 cửa sổ đánh giá, median 138 cửa sổ mỗi bản ghi; **30.2 % cửa sổ nằm đúng ở trần 600 s**. Bảng dưới đặt cạnh nhau tỉ lệ cửa sổ, tỉ lệ trọng số thực tế của W-bMAE (tính trong từng bản ghi rồi trung bình) và phần đóng góp của hằng số tốt nhất (35 s, chính là weighted median, cho W-bMAE 61.99).

Bin target	% cửa sổ	% trọng số	Điểm của hằng số 35 s	MAE của hằng số (s)
0-40 s	5.3	52.9	9.12	17.2
40-90 s	6.7	26.5	7.37	27.9
90-300 s	27.3	15.8	24.56	156.5
300-600 s	60.7	4.8	20.95	483.1

Một cửa sổ của bin 0-40 s nặng xấp xỉ 100 cửa sổ của bin 300-600 s, và các cửa sổ chạm trần chỉ mang 1.5 % trọng số. Tuy vậy, vì hằng số đã "rẻ" ở hai bin gần onset (9.1 + 7.4 điểm), **phần sai số còn lại của hằng số nằm chủ yếu ở 90-300 s và 300-600 s**. Các phép phân rã oracle làm rõ điều này: biết đúng bin và dự đoán weighted median của bin cho 21.75 điểm (5.29 / 3.31 / 8.24 / 4.91); hoàn hảo chỉ trong 90-600 s cho 15.87; hoàn hảo chỉ trong 0-40 s chỉ hạ được xuống 41.08. Predictor chỉ dùng elapsed time (weighted median theo từng 30 s elapsed, 5-fold theo người) đạt 57.50 (`muse_02.log`).

Hazard của N2 đầu tiên tăng theo thời gian: bằng 0 trong phút đầu, 2.8-5.2 % mỗi phút ở phút 4-6, bình nguyên khoảng 10 % mỗi phút từ phút 9 đến 16, rồi lên 20-48 % mỗi phút sau phút 20 vì mọi bản ghi đều kết thúc trước 25 phút. Spearman giữa elapsed time và target là -0.44 trên toàn bộ cửa sổ nhưng chỉ 0.00 / -0.01 / -0.13 / -0.27 bên trong bốn bin. Việc cắt ở 25 phút là đặc tính của cách chọn tập công khai; không biết cohort sealed có cùng kiểu cắt hay không, nên một model dựa nhiều vào elapsed time sau khoảng 15 phút là model có rủi ro. Suy luận quan trọng nhất của mục này: bài toán về bản chất là **phân loại thứ tự bốn trạng thái của time-to-N2, cộng với hồi quy mịn chỉ cần thiết trong 90-300 s**, và chi phí bất đối xứng đòi hỏi một head bảo thủ khi không chắc chắn. Lưu ý rằng mọi trọng số ở trên giả định scorer sealed hoạt động như warm-up (theo bản ghi, rồi trung bình).



Hazard của N2 tăng theo elapsed time, 30.2 % cửa sổ nằm ở trần 600 s, và bin 0-40 s mang 53 % trọng số của metric với chỉ 5 % số cửa sổ trong khi phần lớn sai số của hằng số và của elapsed-only nằm ở 90-600 s.

Artifact giảm từ 20-30 % xuống 1-3 % khi tiến gần N2

Tín hiệu là EEG điện cực khô điển hình: nền yên tĩnh có SD mỗi cửa sổ 5 s khoảng 5 µV (trán) đến 10 µV (thái dương) với một đuôi artifact nặng. Median SD trước N2 là 9.8 / 5.5 / 5.2 / 10.0 µV cho TP9 / AF7 / AF8 / TP10, phân vị 95 là 162 / 92 / 77 / 148 µV, và tỉ số thái dương trên trán là 1.84. Histogram của |x| có một mode thứ hai ở **806-820 µV** trên mọi kênh và 519 trong 540 bản ghi chạm ≥ 780 µV ở ít nhất một kênh, nên các mẫu trên khoảng 780 µV nên được coi là đã bị clip (suy luận: dải ADC đối xứng khoảng ±840 µV, không xác nhận được từ file). Không có flatline hay dropout: không cửa sổ nào có SD 1 s dưới 1 µV. Trung bình của bản ghi nằm trong ±2.6 µV, tức dữ liệu hành xử như đã qua high-pass (`muse_04.log`).

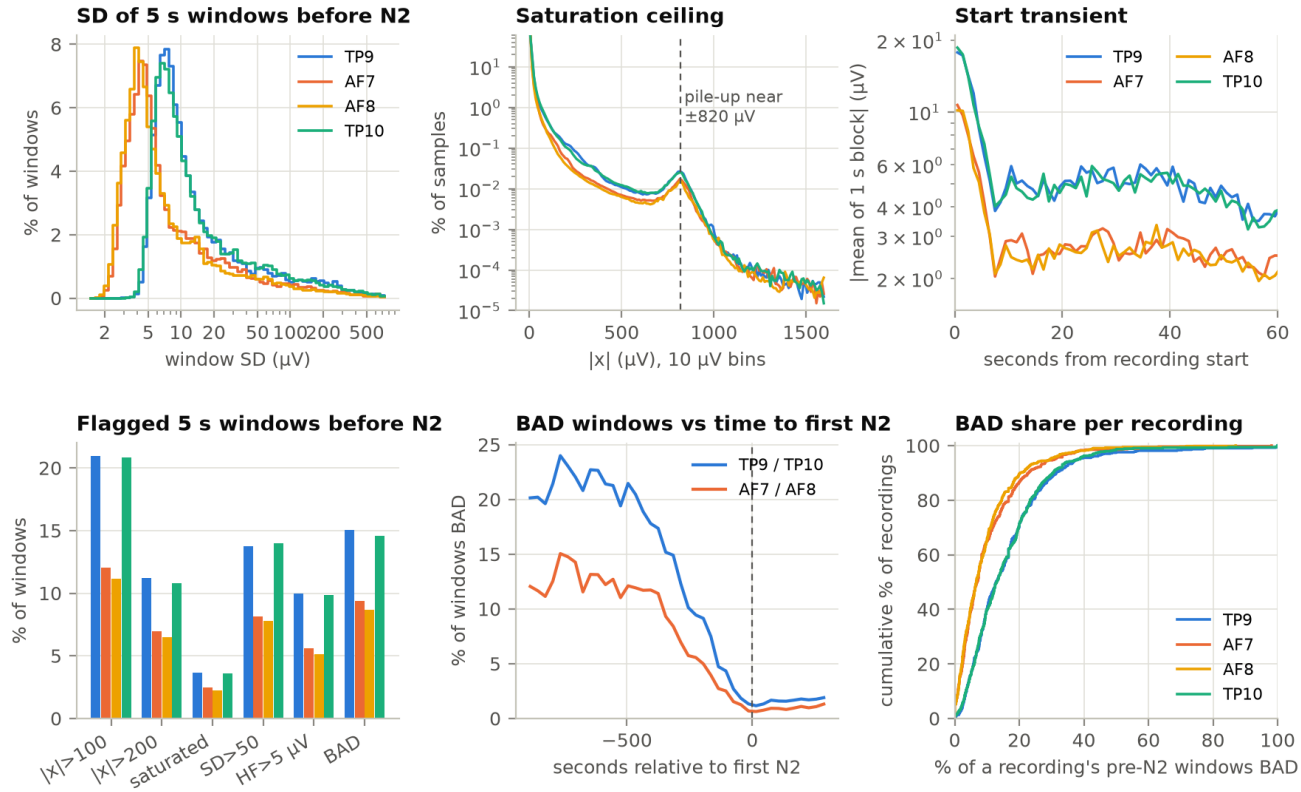
Phần đầu bản ghi khác hẳn phần còn lại. 62 % bản ghi bắt đầu bằng một mẫu bằng 0 trên cả bốn kênh, và |trung bình 1 s| giảm về hai lần mức ổn định trong 5-6 s (phù hợp với một high-pass bậc một gần 0.1 Hz đang ổn định; đây là suy luận từ dữ liệu, không phải từ metadata). Độc lập với transient đó, những phút đầu rất nhiễu: tỉ lệ kênh-cửa sổ BAD là 41 % trong 30 s đầu, 41 % trong 30-60 s và 30 % trong 60-120 s.

Nhiễu điện lưới có mặt ở 96 % bản ghi và chia thành hai nhóm sạch: 200 bản ghi có đỉnh 50 Hz, 319 bản ghi có đỉnh 60 Hz, 1 có cả hai, 22 không có; hai nhóm có cùng phổ băng rộng và cùng onset (median 705 so với 690 s, p = 0.73). Phổ median rơi dốc trên khoảng 52 Hz. Kênh AF7 có thêm các vạch hẹp ở 21.33 Hz (fs/6) và 42.67 Hz (fs/3) trong 92 và 113 bản ghi, gắn với người tham gia (vạch fs/3: luôn có ở 29 người, không bao giờ ở 150 người); chúng vô hại với band power nếu band tránh hai tần số đó, nhưng là một shortcut nhận dạng thiết bị tiềm tàng cho model sâu.

Về tổng thể, 21.6 % cửa sổ trước N2 có ít nhất một kênh BAD nhưng chỉ 5.2 % có cả bốn kênh BAD; 78.4 % cửa sổ có đủ bốn kênh tốt và 88.0 % giữ được ít nhất một kênh thái dương và một kênh trán tốt. Phát hiện có

giá trị mô hình hoá nhất là **tỉ lệ artifact tự nó là một marker của thức**: tỉ lệ cửa sổ có kênh BAD (bất kỳ kênh nào) là 31.3 % khi còn hơn 600 s, 26.4 % ở 300-600 s, 12.9 % ở 90-300 s, 5.0 % ở 40-90 s, 3.4 % ở 0-40 s và 3.9 % sau N2. Chuẩn hoá mạnh tay xoá artifact sẽ xoá luôn tín hiệu này, vì vậy nên giữ các feature artifact tường minh (số kênh BAD, log max |x|, RMS cao tần). Tương quan liên kênh trong các cửa sổ sạch cũng định hướng cách gộp kênh: TP9-TP10 có median r = 0.52, hai cặp cùng bên (TP9-AF7, AF8-TP10) 0.39 và 0.36, còn AF7-AF8 gần 0 (0.04). Lấy trung bình TP9 và TP10 là an toàn; lấy trung bình AF7 và AF8 thì loại bỏ thành phần ngược pha (chuyển động mắt ngang), nên hiệu AF7 - AF8 xứng đáng là một feature riêng.

Những gì chưa biết ở đây: reference, ground, tần số lấy mẫu gốc và các bộ lọc không được ghi lại; không có kênh impedance hay accelerometer, nên "tiếp xúc kém" và chuyển động chỉ suy được từ biên độ và mức nhiễu điện lưới.

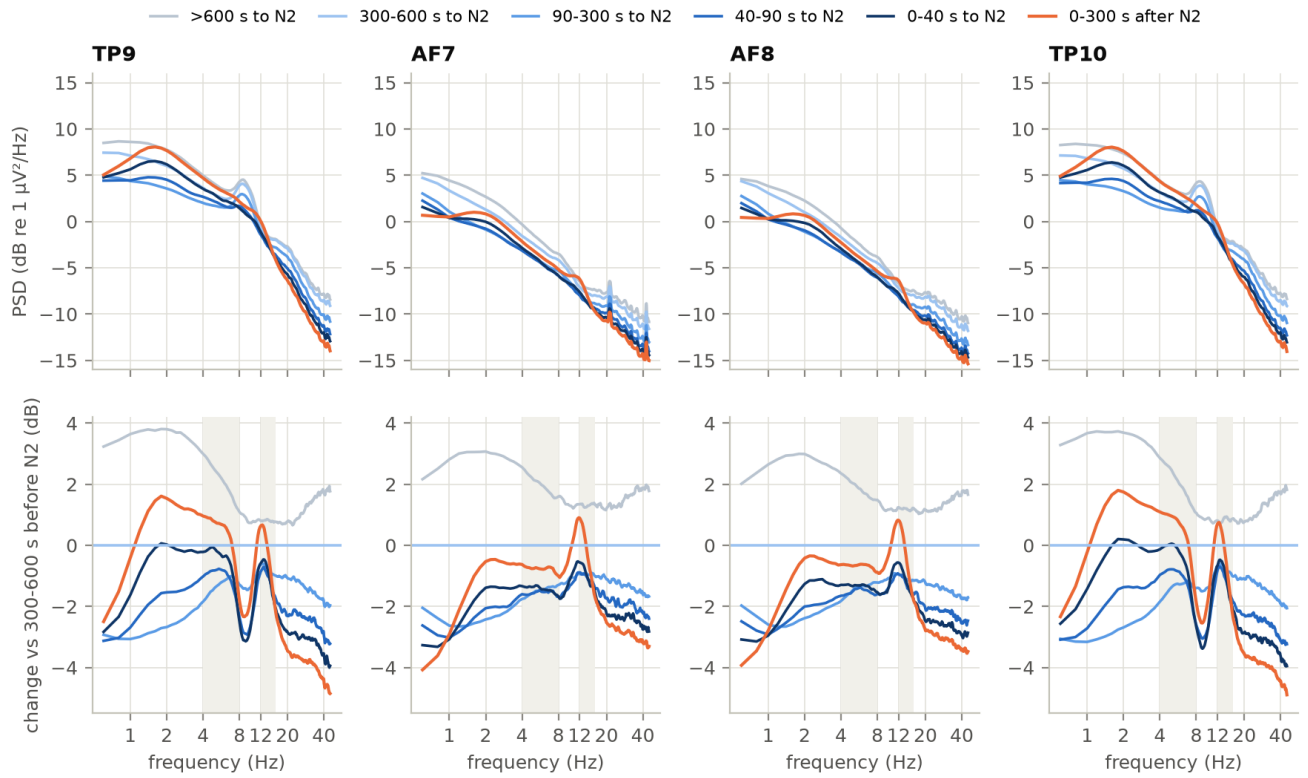


Kênh thái dương có biên độ lớn hơn và nhiễm artifact gấp đôi kênh trán, mẫu dồn ở trần ±820 µV, 5 s đầu có transient ổn định, và tỉ lệ cửa sổ BAD giảm từ khoảng 20 % xuống khoảng 1 % khi tiến gần N2.

Công suất trên 16 Hz giảm đơn điệu, delta có hình chữ U

Khi N2 đến gần, ba điều thay đổi trong phổ (`muse_06.log`). Thứ nhất, công suất trên 16 Hz giảm đơn điệu ở mọi kênh và là hiệu ứng lớn nhất, sớm nhất: giữa 300-600 s và 0-40 s, beta thái dương giảm 2.79 dB (dz -1.56) và gamma 3.16 dB (dz -1.13); beta trán giảm 2.13 dB (dz -0.89). Thứ hai, alpha thái dương sụp khoảng 3 dB (dz -1.10). Thứ ba, công suất tần số thấp **có hình chữ U**: delta thái dương giảm 2.69 dB từ 300-600 s sang 90-300 s (khi artifact thức và chuyển động mắt biến mất), rồi tăng lại 0.76 dB sang 40-90 s và thêm 1.44 dB (dz +0.78) sang 0-40 s. Ở kênh trán, delta và theta tuyệt đối ngay trước N2 lại thấp hơn 5-10 phút trước đó (-1.8 và -1.2 dB), vì công suất chậm ở trán khi thức bị chi phối bởi hoạt động mắt.

Sau N2 xuất hiện một đỉnh riêng ở 12-14 Hz, rõ nhất so với nền ở kênh trán (tăng lớn nhất tại 11.6 Hz, +0.8 đến +1.0 dB so với 300-600 s). Trong phổ hiệu, bướu sau N2 trải khoảng 11-14.5 Hz, tức **vắt ngang ranh giới 12 Hz** giữa "alpha" và "sigma" mà các feature baseline đang dùng. Suy luận thực dụng: nên định nghĩa một band spindle riêng 11-15 Hz (hoặc đo sigma như phần dư trên fit aperiodic) và thu alpha về 8-11 Hz; delta thô là feature kém khi đứng một mình và cần đi kèm công suất cao tần hoặc chỉ báo artifact để giải nhập nhằng. Cần lưu ý nhóm "> 600 s" chỉ gồm các bản ghi onset muộn và những phút đầu của mọi bản ghi, nên phần dư công suất thấp tần của nó trộn lẫn "còn xa giấc ngủ" với "vừa đeo băng đô".



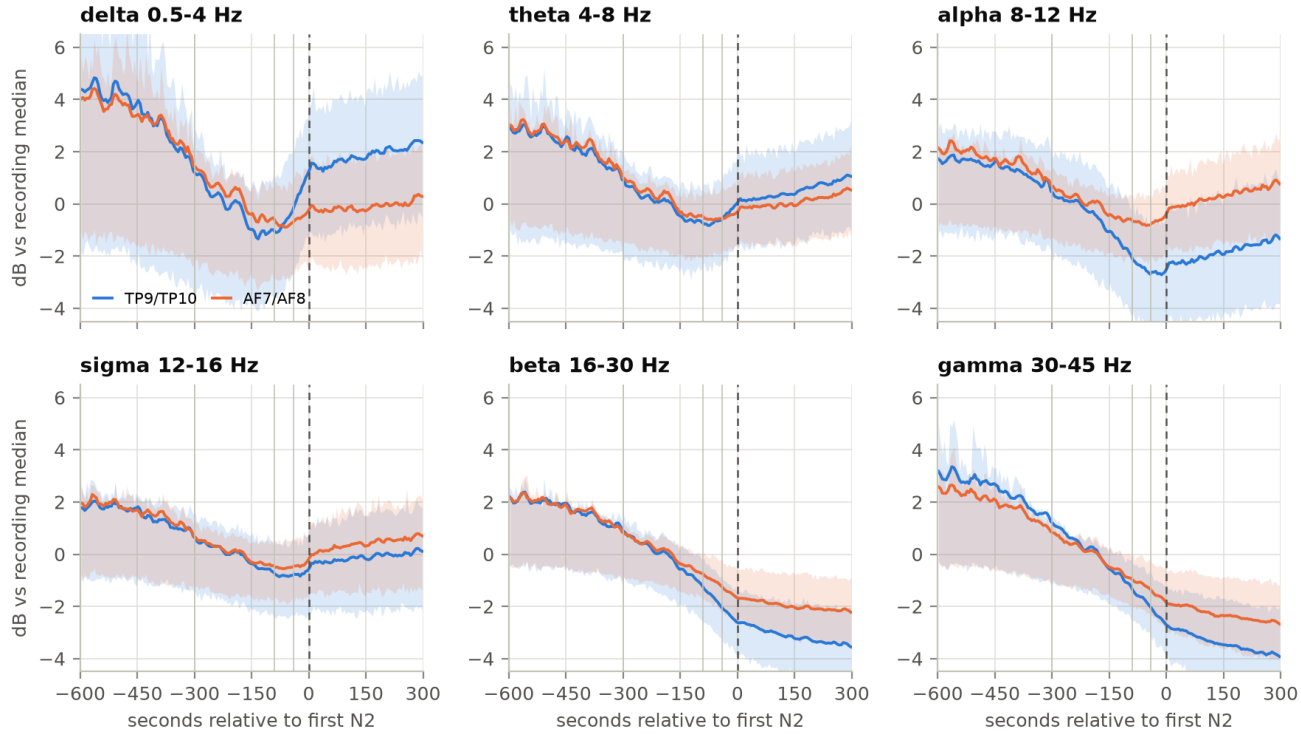
Khi tiến gần N2, công suất trên 16 Hz giảm đơn điệu ở cả bốn kênh, alpha thái dương 8-11 Hz sụp xuống, công suất tần số thấp có hình chữ U, và sau N2 xuất hiện đỉnh spindle 12-14 Hz.

Hai họ feature: "slow drifters" suốt 10 phút và "late movers" trong 2 phút cuối

Căn theo N2 đầu tiên, các feature chia thành hai họ rõ rệt (`muse_07.log`). **Slow drifters** (beta, gamma, biên độ, tỉ lệ artifact, công suất chậm ở trán) giảm suốt mười phút và đã đi được nửa đường 4-6 phút trước N2: block 30 s đầu tiên mà mức thay đổi đạt ít nhất một nửa giá trị cuối là -240 s với beta thái dương, -270 s với gamma và alpha thái dương, -300 s với beta trán, -360 s với SD cửa sổ. **Late movers** (sigma/beta, 1/f exponent, theta/alpha và delta/beta thái dương, delta thái dương) phẳng cho đến khoảng 2 phút trước N2 rồi đổi nhanh qua onset: mốc nửa đường là -90 s với sigma/beta thái dương, -60 s với sigma/beta trán và 1/f exponent thái dương, -30 s với delta/beta thái dương. Về độ lớn (từ -600..-450 s đến 30 s cuối, n = 327 bản ghi có onset ≥ 600 s): alpha thái dương -3.9 dB (dz -1.09), beta -3.8 dB (-1.19), gamma -4.6 dB (-1.13); sigma/beta thái dương +2.0 dB (dz +1.38), 1/f exponent thái dương +0.29 (+0.81), số kênh BAD -0.47 (-0.68), tương quan băng rộng trước-sau cùng bên +0.11 (+0.89).

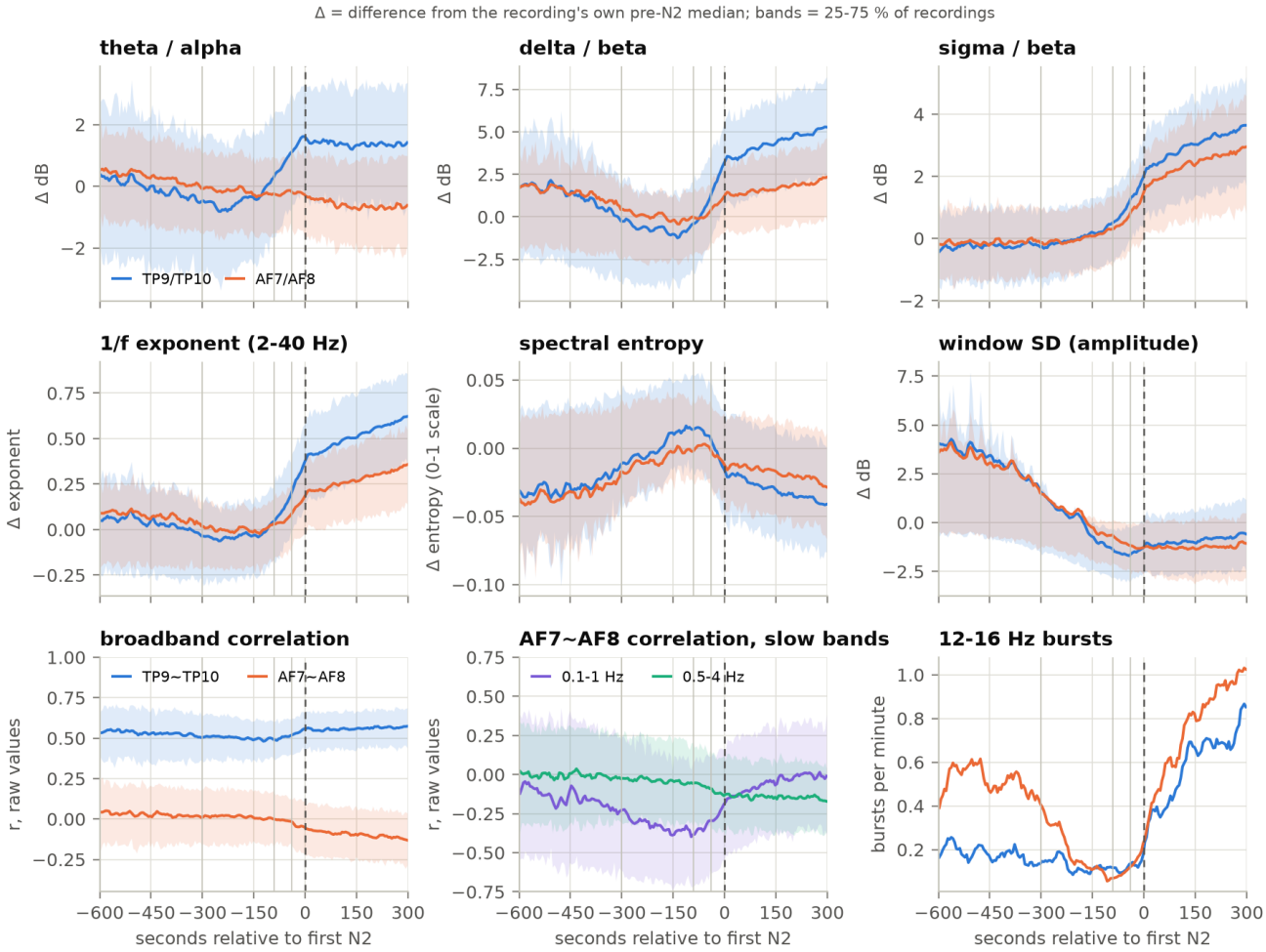
Sự khác nhau giữa hai họ không chỉ ở thời điểm mà còn ở tỉ số tín hiệu trên độ phân tán giữa các bản ghi. SD giữa các bản ghi của mức baseline là 3.4-3.8 dB với alpha, sigma và beta và 7.2 dB với SD cửa sổ, tức **bằng hoặc lớn hơn chính mức thay đổi trong bản ghi (2-5 dB)**; với sigma/beta nó chỉ là 1.3-1.5 dB so với mức thay đổi 2-3.5 dB. Vì vậy slow drifters cần chuẩn hoá causal theo bản ghi hoặc ngữ cảnh dài, còn late movers dùng được gần như không cần chuẩn hoá.

Hai dấu hiệu sinh lý kinh điển cho kết quả trái chiều. Đếm spindle bằng detector ngưỡng cố định quá thưa để định vị onset trong luồng 5 s: số burst 12-16 Hz mỗi phút (thái dương / trán) là 0.13 / 0.20 ở -300..-90 s và chỉ lên 0.71 / 0.91 ở 120-300 s sau N2; 172 trong 540 bản ghi không có burst trán nào được phát hiện sau N2, và detector còn bắn nhầm vào beta/EMG khi thức. Ngược lại, phần dư sigma trên fit 1/f theo cửa sổ leo đều (thái dương +0.1 dB ở -600..-300 s, +1.5 ở -90..-40 s, +2.7 ở 120-300 s sau N2) và mang cùng thông tin với phương sai nhỏ hơn nhiều. Chuyển động mắt chậm thì hiện ra như tương quan âm giữa AF7 và AF8 trong 0.1-1 Hz: median -0.05 ở -600..-450 s, **-0.29 ở -180..-90 s**, rồi về -0.02 ở 60-300 s sau N2. Đây là một "marker N1" không đơn điệu, có thể giúp tách 40-300 s khỏi cả thức sớm lẫn N2. Hạn chế: cả phân tích spindle lẫn băng chậm đều dùng bộ lọc zero-phase và ngưỡng cả bản ghi; phiên bản causal chưa được đánh giá, và periodogram 5 s thông thường (độ phân giải 0.2 Hz, detrend tuyến tính) che mất tín hiệu này.

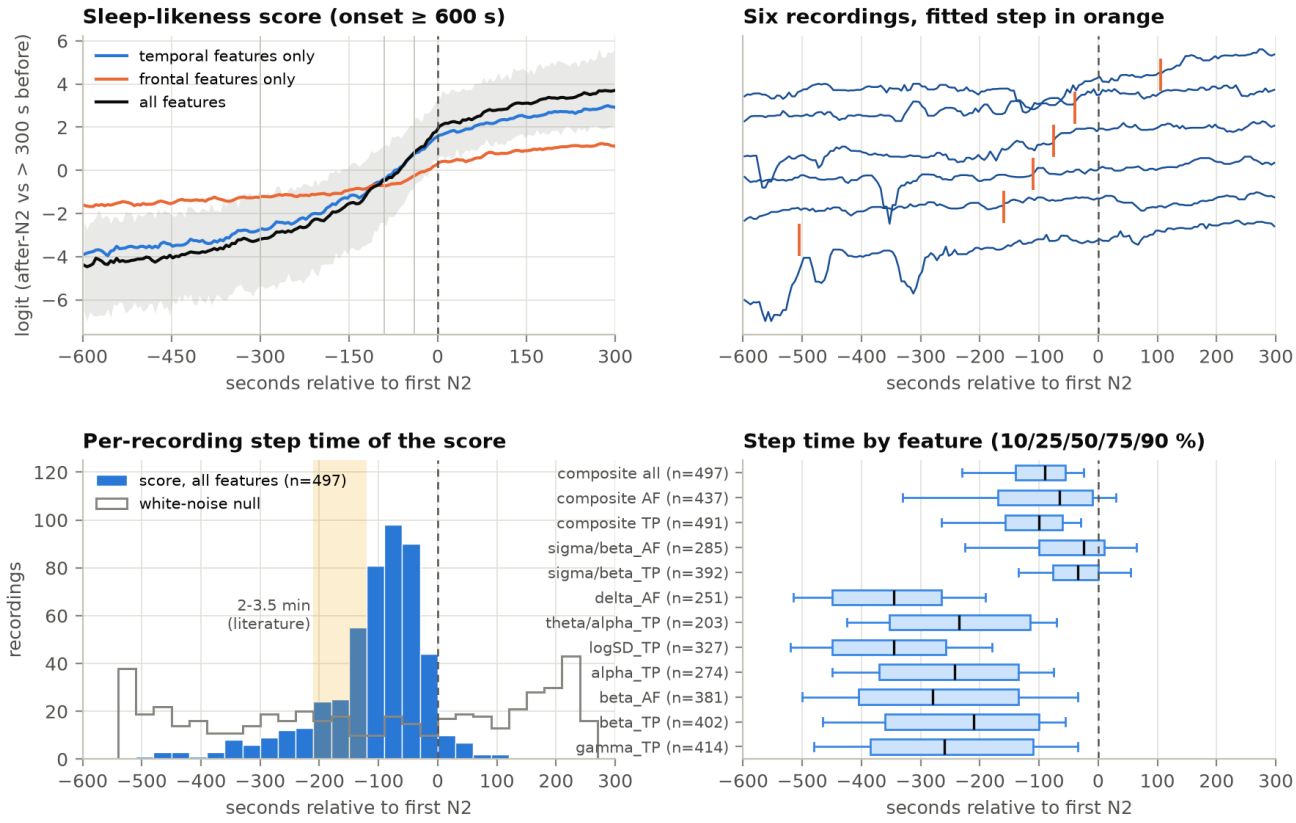


Band power căn theo onset: beta và gamma giảm suốt 10 phút, alpha thái dương giảm nhanh nhất giữa -300 và -60 s, còn delta thái dương quay đầu tăng khoảng 90 s trước N2.

Để gộp các feature, một sleep-likeness score được dựng bằng logistic regression trên cả 78 feature một cửa sổ, huấn luyện để tách cửa sổ sau N2 khỏi cửa sổ cách N2 hơn 300 s, 5-fold theo người: AUC out-of-fold 0.972 (mọi feature), 0.962 (27 feature thái dương), 0.878 (28 feature trán). Khớp một bước nhảy đơn vào score của từng bản ghi (502 bản ghi có onset ≥ 300 s) cho thời điểm bước nhảy có phân vị 10/25/50/75/90 là **-230 / -140 / -90 / -55 / -25 s**; 23 % bản ghi nhảy trong cửa sổ 2-3.5 phút của y văn, 31 % trong vòng 60 s trước N2 và 4 % sau N2. Một null nhiễu trắng trên cùng miền cho phân phối phẳng, nên sự tập trung này là thật. Ở mức nhóm, score đạt nửa đường tại -78 s (CI -88 đến -68 s); chỉ feature thái dương: -82 s; chỉ feature trán: -38 s, tức trán muộn hơn khoảng 45 s. "Trán tip muộn hơn" đúng cho composite chứ không đúng cho công suất cao tần thô: beta/gamma trán giảm sớm hơn beta/gamma thái dương vì phần lớn là hoạt động cơ trán và cơ mắt thả lỏng sớm. Một lưu ý về phương pháp: mô hình bước nhảy đơn gán một ramp từ từ cho một bước ở giữa ramp, nên độ phân tán nêu trên trộn biến thiên thật về thời điểm với biến thiên về hình dạng chuyển tiếp.



Sigma/beta, 1/f exponent, theta/alpha thái dương và delta/beta chỉ tăng mạnh trong 60-120 s cuối; tương quan âm AF7-AF8 ở băng chậm (chuyển động mắt chậm) đạt đỉnh 1-3 phút trước N2 rồi biến mất; burst sigma chỉ tăng sau N2.



Sleep-likeness score đa feature nhảy bậc với median 90 s trước N2 đầu tiên (IQR -140 đến -55 s), muộn hơn cửa sổ 2-3.5 phút của y văn, và các feature trán tip muộn hơn feature thái dương.

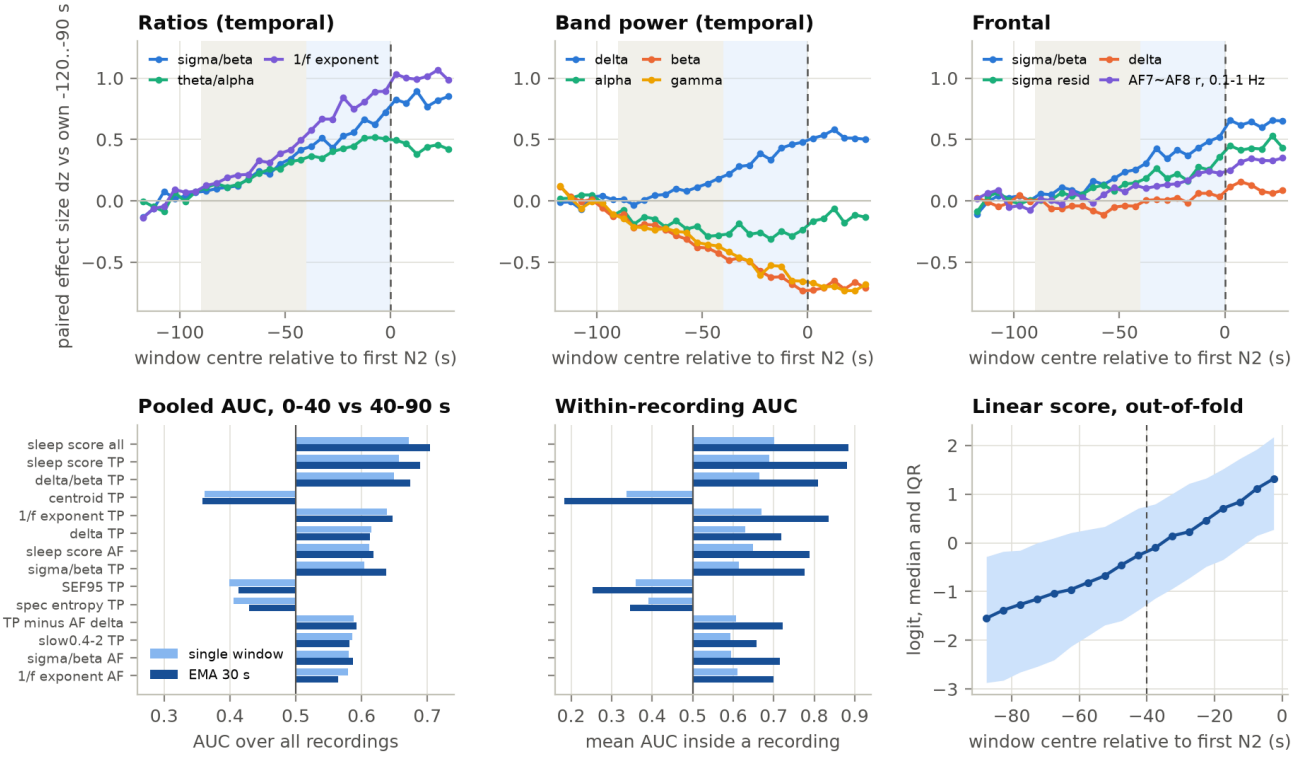
90 giây cuối là một ramp, không có bước nhảy ở mốc 40 s

Đây là mục quyết định kỳ vọng cho hai bin nặng nhất (`muse_08.log`, `muse_08b`, `muse_08c`, `muse_08d`). Ở độ phân giải 5 s, mọi feature có thay đổi trong hai phút cuối đều đi theo một đường gần thẳng bắt đầu khoảng -90 s và tiếp tục sau N2: sigma/beta thái dương có dz so với mức -120..-90 s của chính bản ghi là 0.42 tại -42.5 s, 0.72 tại -2.5 s và 0.82 tại +2.5 s; 1/f exponent thái dương 0.49, 0.89, 1.03; beta thái dương -0.43, -0.73, -0.73. Alpha thái dương, delta trán và SD cửa sổ thì phẳng trong 90 s cuối. **Không có bước nhảy nào ở 40 s.** Ở độ phân giải 1 s chỉ có một bước nhỏ đúng tại onset được gán nhãn (sigma/beta thái dương +0.34 dB, dz +0.25) và chỉ ở các bản ghi on-grid; các bản ghi off-grid thay đổi ít hơn và không có bước (+0.16 dB, dz +0.12).

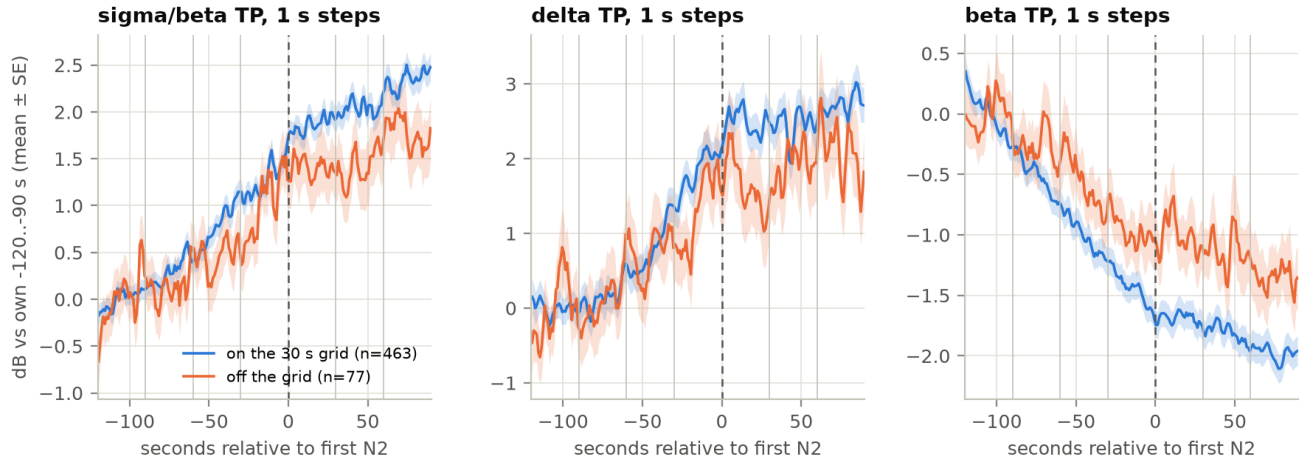
Khả năng tách của từng cửa sổ vì thế bị chặn. Với contrast A (bin 0-40 s so với 40-90 s), AUC gộp của một cửa sổ 5 s là 0.650 cho delta/beta thái dương, 0.638 cho 1/f exponent thái dương, 0.604 cho sigma/beta thái dương và 0.671 cho composite; trong 78 feature đơn không feature nào lệch khỏi 0.5 quá 0.15. Smoothing causal 30 s giúp 0.01-0.03 (composite 0.704), còn EMA 2 phút và 5 phút làm loãng (0.640 và 0.582). Mô hình đa biến tốt nhất (logistic regression trên feature EMA 30 s, 5-fold theo người) đạt AUC gộp **0.754**; score của nó tách 0-20 s khỏi 60-90 s với AUC 0.829 nhưng tách 20-40 s khỏi 40-60 s chỉ với 0.631. Elapsed time đơn lẻ cho AUC 0.536. Với contrast B (40-90 s so với 90-300 s) composite đạt 0.725, và đúng như dự đoán từ mục trước, slow drifters là bộ tách đơn tốt nhất (beta thái dương 0.365, tức 0.635 theo chiều ngược).

Quy ra điểm metric: khi được cho biết N2 còn dưới 90 s, một score EEG (10 decile ánh xạ sang weighted median của target) đạt 13.54 điểm từ hai bin gần so với 15.87 của hằng số trong 90 s cuối và 8.60 nếu biết đúng bin; tức **EEG thu hồi được 2.3 trong 7.3 điểm khả dĩ** (`muse_09.log`). Suy luận: nên kỳ vọng model dự đoán một "time to N2" trơn, đúng trong khoảng ±20-30 s ở 90 s cuối, chứ không đặt được một cạnh sắc ở 40 s; phần lợi lớn hơn của EEG nằm ở 90-600 s.

Một phép kiểm soát giúp tránh hiểu nhầm AUC trong bản ghi: khi dịch cùng phép so sánh hai block ra xa N2 thêm 0 / 60 / 120 / 180 s, AUC trong bản ghi của delta/beta thái dương (EMA 30 s) giảm 0.810 / 0.671 / 0.536 / 0.469, tức late movers có tín hiệu đặc hiệu cho 90 s cuối; còn beta thái dương gần như không đổi theo độ dịch, tức AUC rất cao của beta, gamma và biên độ đã smoothing là drift chung chứ không phải đặc hiệu onset. Các bản ghi off-grid khó hơn một chút ở contrast A (composite 0.633 so với 0.678, chênh +0.045 với CI 0.017-0.072) và không khác ở contrast B.



Trong hai phút cuối mọi feature đều đi theo một ramp bắt đầu khoảng 90 s trước N2, không có bước nhảy ở 40 s; một cửa sổ đơn lẻ tách bin 0-40 s khỏi 40-90 s với AUC tối đa 0.65-0.70, và score tuyến tính out-of-fold tăng trơn từ logit khoảng -1.6 tại -87 s lên +1.3 tại -2 s.



Ở độ phân giải 1 s, sigma/beta, delta và beta thái dương thay đổi dần trong 90 s cuối; chỉ các bản ghi on-grid có một bước nhỏ tại onset được gán nhãn, còn bản ghi off-grid thay đổi ít hơn và không có bước.

Không feature causal đơn lẻ nào vượt rho tuyệt đối 0.4, và 35-50 % là "đồng hồ"

Trên 80,975 cửa sổ đánh giá, elapsed time có Spearman -0.436 với target đã cắt trần, tốt hơn bất kỳ feature EEG đơn lẻ nào: feature một cửa sổ tốt nhất là gamma thái dương (|rho| 0.35), feature smoothing 30 s tốt nhất là 0.38, composite học được là 0.386 ở dạng thô (`muse_09.log`). Hai phép biến đổi causal nâng mức liên hệ hữu ích. Smoothing: median |rho| qua 81 feature là 0.125 (thô), 0.186 (EMA 30 s), 0.190 (EMA 2 phút), 0.141 (EMA 5 phút). Chuẩn hoá theo chính bản ghi: biến thể "EMA 30 s trừ running mean của bản ghi tính đến hiện tại" đưa composite từ 0.386 lên **0.476**, alpha thái dương từ 0.280 lên 0.358 và beta thái dương từ 0.315 lên 0.355; riêng running mean không kèm smoothing lại làm giảm median (0.125 xuống 0.083) vì running mean trễ và xoá cả drift thật.

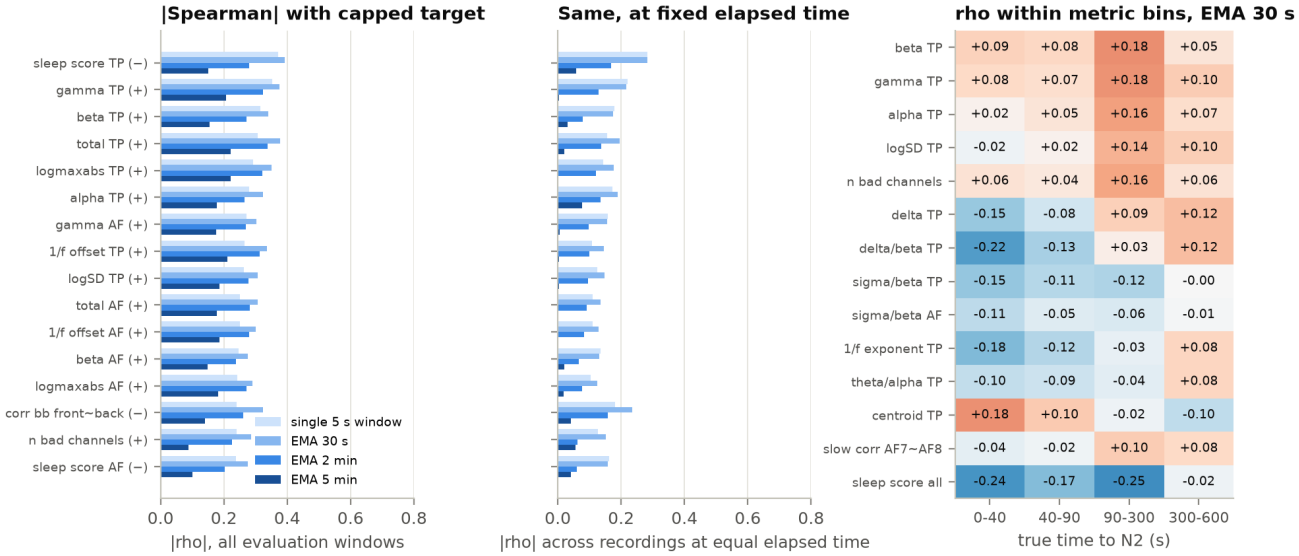
Kiểm soát elapsed time cho thấy một phần lớn mối liên hệ là drift chung của bản ghi (ổn định, thư giãn) chứ không phải tiến gần giấc ngủ: partial rho so với rho thô là 0.225 so với 0.352 cho gamma thái dương, 0.182 so với 0.315 cho beta thái dương, 0.165 so với 0.307 cho tổng công suất thái dương, tức **khoảng 35-50 % mối liên hệ của biên độ và cao tần là chung với elapsed time**. Ở elapsed time cố định, median |rho| của bảy biến thể là 0.063 (thô), 0.085 (EMA 30 s), 0.062 (EMA 2 phút), 0.024 (EMA 5 phút), 0.079 (runmean), 0.141 (EMA 30 s trừ runmean), 0.113 (EMA 30 s trừ EMA 5 phút): EMA 5 phút chủ yếu tái tạo đồng hồ, còn các biến thể chuẩn hoá là những biến thể duy nhất giữ được liên hệ lớn. Theo pha của bản ghi, rho của composite (EMA 2 phút) ở elapsed time cố định là 0.039 trong 5 phút đầu, -0.206 ở 5-10 phút và -0.435 về sau: **trong 5 phút đầu EEG gần như không nói gì thêm ngoài đồng hồ.** Bên trong hai bin gần onset (79 % trọng số) không feature nào vượt |rho| 0.25.

Bảng sau trích 8 trong 15 feature causal nhiều thông tin nhất (xếp theo |rho| ở elapsed time cố định của biến thể tốt nhất; đã bỏ composite và các feature tương quan hạng trên 0.9 với một feature tốt hơn).

Feature (vùng)	Biến thể causal tốt nhất	rho mọi cửa sổ	rho ở elapsed cố định	rho trong bin 0-40 s	rho trong bin 300-600 s
beta 16-30 Hz (TP)	EMA 30 s - runmean	0.355	0.313	0.09	0.17
alpha 8-12 Hz (TP)	EMA 30 s - runmean	0.358	0.286	0.01	0.13
log line length (TP)	EMA 30 s - runmean	0.331	0.266	0.05	0.19
gamma 30-45 Hz (AF)	EMA 30 s - runmean	0.298	0.264	0.04	0.18
sigma/beta (TP)	EMA 30 s - EMA 5 phút	-0.228	-0.244	-0.15	-0.04
tương quan băng rộng trước-sau	EMA 30 s	-0.323	-0.235	-0.06	-0.08
sigma residual trên fit 1/f (AF)	EMA 30 s - runmean	-0.280	-0.229	-0.05	-0.14
số kênh BAD	EMA 30 s - runmean	0.242	0.217	0.02	0.16

Bảng cho thấy lại sự phân công: biên độ và beta có cùng dấu trong mọi bin và mạnh nhất ở xa onset, còn sigma/beta là feature hiếm hoi có liên hệ trong bin 0-40 s. Về kênh, TP9/TP10 mang thông tin beta, alpha và

biên độ (rho 0.27-0.31 ở elapsed cố định so với 0.19-0.26 ở trán), trong khi sigma/beta, 1/f offset và gamma nhiều thông tin gần như ngang nhau ở cả bốn kênh; trung bình hai kênh thái dương (0.313) hầu như không hơn từng kênh (0.304 / 0.305). Tỉ lệ phương sai giữa các bản ghi (median qua 81 feature) giảm từ 0.229 ở dạng thô xuống 0.099 sau khi trừ running mean. Dạng input hợp lý vì thế là feature smoothing causal, tương đối theo bản ghi, **đi kèm mức thô** (running mean là tham chiếu tối trong những phút đầu, và số artifact cùng biên độ tốt hơn ở dạng thô). Hạn chế: Spearman chỉ đo phụ thuộc đơn điệu; các composite dùng nhãn khi huấn luyện và chưa có ước lượng nested-CV.



Feature causal có |Spearman| tối đa 0.4 với target đã cắt trần; smoothing 30 s đến 2 phút giúp ích, phần lớn mối liên hệ của biên độ và beta biến mất ở elapsed time cố định, và bên trong các bin của metric không feature nào vượt 0.25.

Hai baseline thắng 19 điểm ở xa onset rồi trả lại 3.5-4.3 điểm ở gần onset

Phân tích lỗi chỉ dùng dự đoán cross-validation (fold 0-4), nhóm "unseen" gồm 500 bản ghi train mỗi bản được dự đoán bởi model chưa từng thấy người đó (`muse_11.log`). Bảng dưới tách W-bMAE theo bản ghi thành điểm của từng bin.

Model	0-40 s	40-90 s	90-300 s	300-600 s	Tổng
Hằng số 35 s	9.12	7.37	24.61	20.99	62.09
Elapsed-only	10.51	7.71	22.56	15.86	56.65
EEG-only	11.82	8.85	14.22	12.50	47.39
Boosted trees (EEG + elapsed)	11.87	8.91	14.24	12.21	47.22
CNN+GRU	9.86	10.10	15.37	11.37	46.69

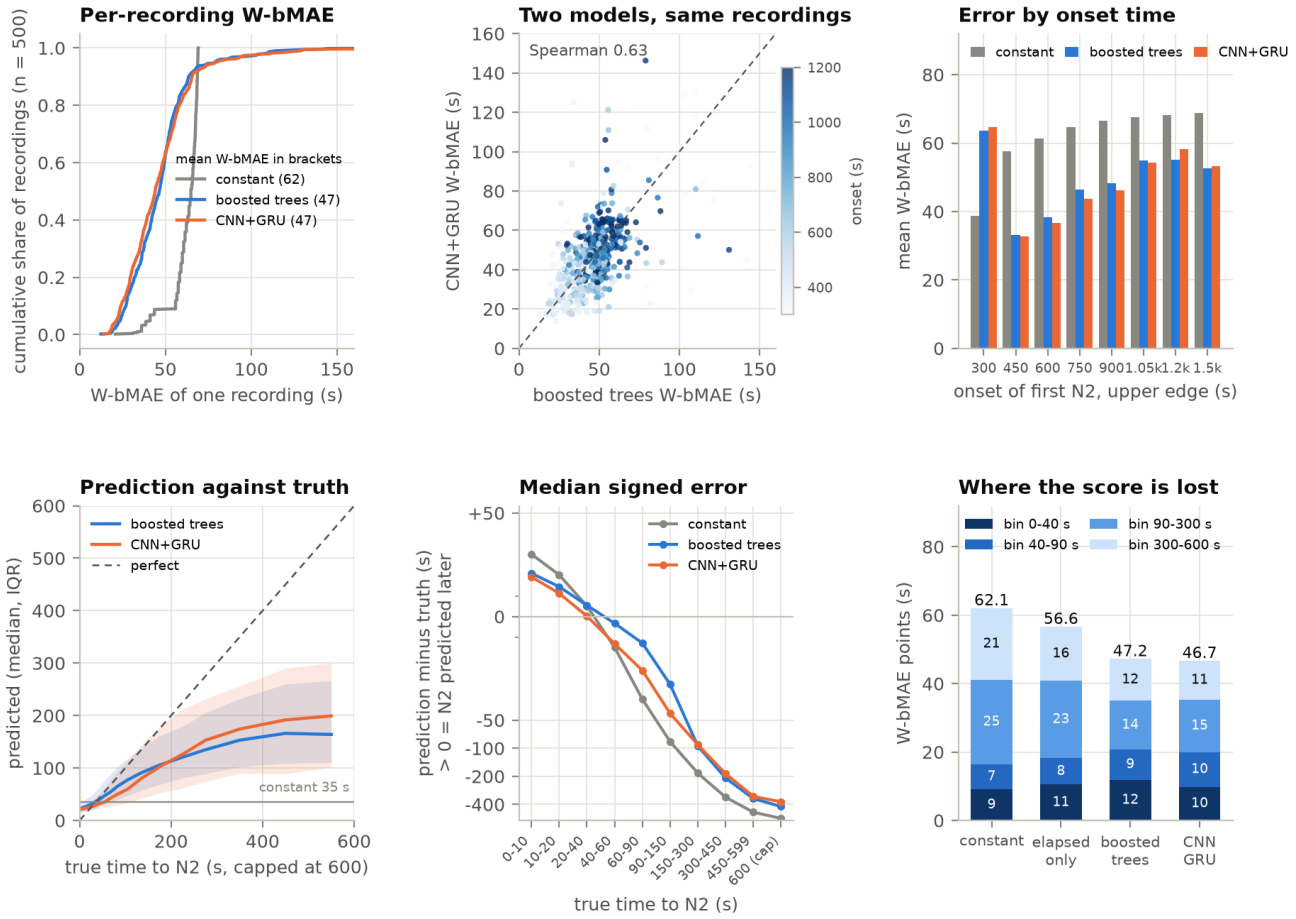
So với hằng số, trees được -2.75 / -1.54 / +10.38 / +8.78 điểm (ròng +14.87) và CNN+GRU -0.74 / -2.73 / +9.25 / +9.62 (ròng +15.40); hai bin gần onset vẫn giữ 20.8 điểm (44 %) của trees và 20.0 điểm (43 %) của CNN. MAE gộp trong 40 s cuối là 17.2 s (hằng số), 22.2 s (trees), 18.5 s (CNN+GRU); trên 40 bản ghi test công khai khoảng cách còn rộng hơn (28.1 và 26.7 s so với 17.1 s), với W-bMAE 60.7 (hằng số), 52.5 (trees) và 55.8 (CNN+GRU). Con số tiêu đề của dự án (trung bình của hai nhóm seen và unseen: 45.97 và 45.32, khớp với 45.84 và 45.27 trong `results.json`) thấp hơn một chút so với trung bình theo bản ghi của nhóm unseen ở bảng trên.

Dự đoán hồi quy mạnh về trung bình. Median dự đoán (trees / CNN) là 24 / 22 s khi thật sự còn 0-10 s, 84 / 69 s ở 90-150 s, 157 / 179 s ở 300-450 s và chỉ **179 / 226 s ở trần 600 s**; dải IQR của cả hai chỉ lên tới khoảng 300 s trong khi 44 % cửa sổ (32,848 trong 75,106) có target từ 450 s. Sai số có dấu vì thế dương (dự đoán muộn hơn thật) chỉ trong 10-20 s cuối, gần 0 ở 20-60 s, rồi âm dần tới khoảng -100 đến -420 s sau 150 s. Về hành vi phát hiện, 14.8 % (trees) và 21.8 % (CNN) cửa sổ còn cách N2 ít nhất 300 s bị dự đoán dưới 90 s. Trong 40 s cuối, trees tệ hơn hằng số ở mọi bin elapsed time dưới 20 phút, và đặc biệt tệ trước phút thứ 6 (MAE 28-44 s so với 16-18 s).

Lỗi tập trung ở hai đầu của dải onset. Với onset ≤ 300 s (n = 44), W-bMAE trung bình là 38.8 (hằng số) so với **63.7 (trees) và 64.6 (CNN+GRU)**; với onset 300-450 s là 57.6 so với 33.1 và 32.7; với onset 900-1050 s là

67.6 so với 54.9 và 54.3. 10 % bản ghi tệ nhất gồm 34 % onset ≤ 300 s (so với 8.8 % trong toàn bộ) và 42 % onset > 900 s. Chất lượng tín hiệu giải thích rất ít: hồi quy lỗi theo hạng của onset, tỉ lệ BAD, biên độ thái dương và cờ off-grid cho R^2 0.051 (trees) và 0.048 (CNN); liên hệ đơn lẻ lớn nhất là công suất 30-45 Hz thái dương (AUC 0.66 cho việc rơi vào 10 % tệ nhất). Hai model thất bại trên cùng các bản ghi (Spearman 0.63, partial 0.53 khi kiểm soát onset; 20 trong 50 bản ghi tệ nhất là chung, so với 5.0 kỳ vọng ngẫu nhiên), nên **ensemble đơn thuần của hai model tương tự sẽ được ít**. Cuối cùng, trên cùng 121 bản ghi, việc đã thấy người đó khi huấn luyện không giúp phát hiện được: seen trừ unseen là -0.65 s cho trees (CI -2.02 đến +0.73) và -0.99 s cho CNN+GRU (CI -3.62 đến +1.51).

Ba chỗ mất điểm rút ra từ đây: (i) các bin 90-600 s vẫn giữ 26 trong 47 điểm dù đã tốt hơn hằng số 19 điểm; (ii) các bin 0-90 s giữ 20-21 điểm và tệ hơn hằng số 2-4 điểm; (iii) một đuôi dài bản ghi onset rất sớm hoặc rất muộn. Suy luận: một head bảo thủ (co về hằng số khi model không chắc) sẽ gỡ phần lớn (ii) và phần onset sớm của (iii), còn một lớp "chạm trần" tường minh sẽ dùng 22,666 cửa sổ ở 600 s tốt hơn một bộ hồi quy L1 thuần. Chưa phân tích được: run `experiments/neural_randstart` có trong repo nhưng chưa được đưa vào phân tích này, và các dự đoán đã cache không có độ bất định theo cửa sổ nên chưa xét được calibration.



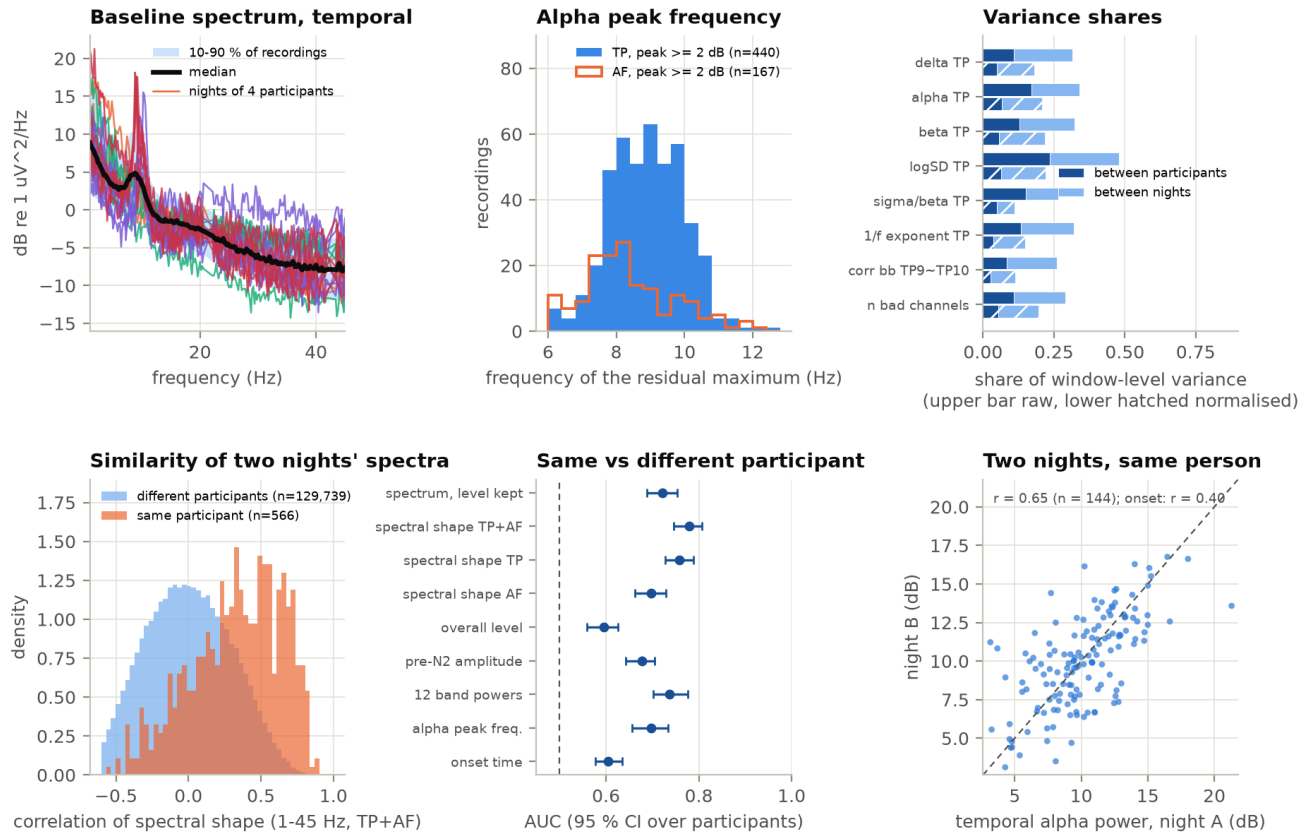
Boosted trees và CNN+GRU đều đạt khoảng 47 s mỗi bản ghi so với 62 s của hằng số và thất bại trên cùng các bản ghi (tương quan hạng 0.63); dự đoán hồi quy về trung bình và không vượt quá khoảng 300 s, và 19 điểm thắng ở các bin 90-600 s bị trả lại một phần ở các bin 0-90 s.

Các đêm của một người khác nhau gần bằng những người khác nhau

Người tham gia khác nhau nhiều ngang với chính các thay đổi trong bản ghi mà model phải phát hiện: SD giữa người của alpha thái dương baseline là 2.46 dB, của beta 1.52 dB, của biên độ cửa sổ 5.17 dB (`muse_12.log`). Đỉnh alpha thái dương rõ (≥ 2 dB trên fit) ở 440 trong 517 bản ghi (85 %), với tần số 9.02 ± 1.10 Hz; ở trán chỉ 32 % bản ghi có đỉnh rõ. Nhưng một người không giống chính mình qua các đêm nhiều như ta tưởng: với các feature mức cửa sổ, 8-25 % phương sai là giữa người, 11-25 % giữa các đêm của một người và 50-78 % trong một đêm; ICC người-so-với-đêm của các mức baseline là 0.44-0.69 cho alpha, sigma, beta và biên độ nhưng chỉ 0.32-0.43 cho delta, theta, gamma và 1/f exponent.

Hai kết quả định hướng chiến lược cá nhân hoá. Thứ nhất, **trừ running mean của chính bản ghi cắt khoảng hai phần ba phần phương sai giữa người** (median 0.150 xuống 0.052; SD giữa người của alpha thái dương từ 2.41 xuống 1.30 dB) mà không cần danh tính hay đêm trước đó; phần còn lại chủ yếu là

phương sai mức đêm và trong đêm, thứ mà không mô hình theo người nào sửa được. Thứ hai, nhận dạng người từ một baseline 2.5 phút đầu là yếu: AUC "cùng người so với khác người" là 0.70-0.78 nhưng top-1 chỉ 21.3 % và top-5 36.3 %, và cả onset (ICC 0.395) lẫn thời điểm chuyển tiếp so với N2 (ICC 0.234) đều chỉ phụ thuộc người ở mức yếu. Kết hợp với việc baseline không hưởng lợi từ người đã thấy, kỳ vọng đơn giản là hai nửa seen và unseen của điểm sealed sẽ gần nhau. Chưa thử: adaptation có giám sát theo người (ví dụ fine-tune trên một đêm trước); ở đây chỉ đo cấu trúc tương đồng và phương sai.

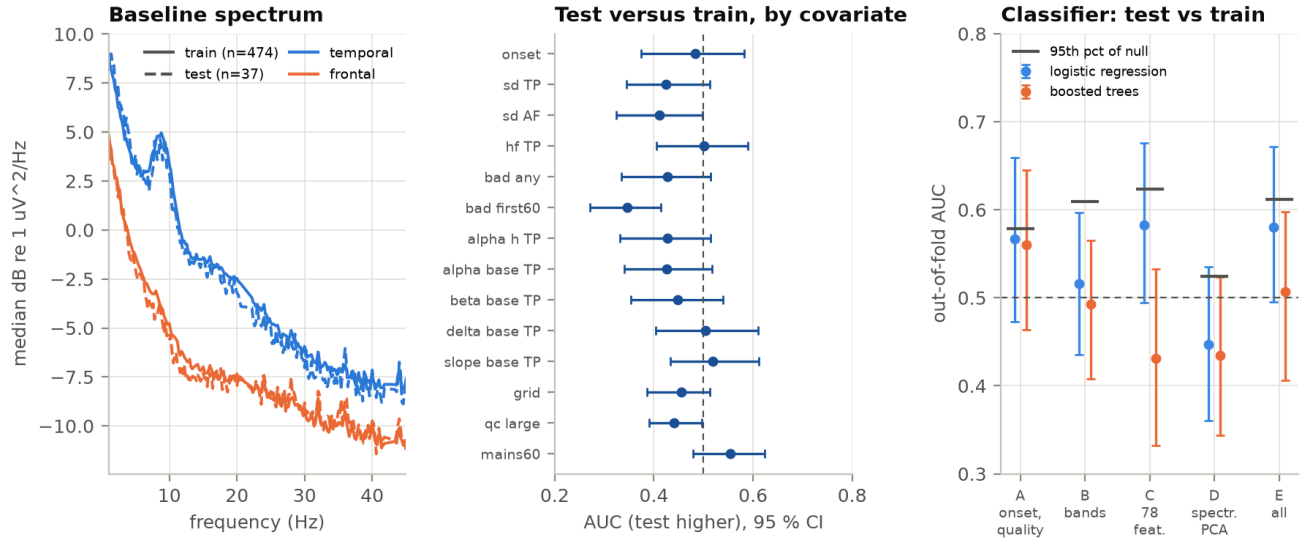


Phổ baseline và đỉnh alpha mang tính cá nhân (9.0 ± 1.1 Hz) nhưng các đêm của một người khác nhau gần bằng những người khác nhau; AUC cùng người so với khác người là 0.70-0.78, và chuẩn hoá bằng running mean loại bỏ khoảng hai phần ba phần phương sai giữa người.

Train và test không tách được, trừ phút đầu tiên sạch hơn

Không classifier mức bản ghi nào tách được 40 bản ghi test khỏi 500 bản ghi train: AUC out-of-fold 0.43-0.58 trên năm bộ feature (onset và chất lượng, band power baseline, median của 78 feature cửa sổ, PCA của phổ, và tất cả), đều nằm trong dải null của phép hoán vị nhãn (`muse_14.log`). Onset, biên độ, tỉ lệ artifact, tỉ lệ điện lưới, đỉnh alpha và phổ đều khớp trong sai số lấy mẫu; thống kê toàn phổ có p = 0.61. Khác biệt vững duy nhất là **phút đầu tiên của test sạch hơn**: tỉ lệ BAD median 0.198 so với 0.417 (Mann-Whitney p = 0.001, Holm 0.038), vẫn còn khi chỉ xét bản ghi on-grid (0.29 so với 0.48, p = 0.003). Ngoài ra có một xu hướng nhất quán nhưng mang tính thăm dò: đêm test có công suất baseline thấp hơn 0.7-1.9 dB so với đêm train của chính người đó; nó không lặp lại được ở so sánh không ghép cặp, không phải hiệu ứng thứ tự session, và là một trong nhiều phép so sánh, nên chỉ là giả thuyết (nếu thật thì chuẩn hoá theo bản ghi sẽ trung hoà nó).

Giới hạn của kết luận này là cỡ mẫu: với 40 bản ghi, không loại trừ được các khác biệt yếu hơn khoảng AUC 0.58 hay một dịch chuyển onset dưới khoảng 2 phút; SEM của W-bMAE theo bản ghi trên 40 bản ghi vào khoảng 1.7-7 s, nên tập test công khai không xếp hạng được các cải tiến nhỏ. Tập sealed, kể cả nửa người chưa thấy, hoàn toàn không so sánh được với dữ liệu train ở đây.



Phổ baseline và gần như mọi covariate mức bản ghi của 40 bản ghi test trùng với 500 bản ghi train, ngoại lệ là phút đầu sạch hơn, và không classifier nào tách được hai tập vượt quá null hoán vị nhãn (AUC 0.43-0.58).

Nhóm off-grid bị cắt mất phần khởi động, và không có dấu hiệu nhãn sai

Bất thường cấu trúc duy nhất là nhóm off-grid (`muse_13.log`, `muse_13c.log`). 77 bản ghi này không chỉ muộn: số mẫu của chúng không phải bội số của 640 (5 s) ở cả 77 bản ghi, trong khi mọi bản ghi on-grid đều là bội số; không bản ghi nào bắt đầu bằng mẫu 0 (so với 72.8 % ở on-grid); và phút đầu của chúng sạch (tỉ lệ BAD median 0.021 so với 0.438 ở các bản ghi on-grid muộn; SD cửa sổ phẳng quanh 8 µV ngay từ cửa sổ đầu so với 26-34 µV). Một logistic regression trên 16 feature mức bản ghi tách hai nhóm với AUC 0.960, nhưng khi bỏ phút đầu thì chỉ còn 0.590. Thời điểm chuyển tiếp của composite so với N2 thì giống nhau (-90 so với -70 s, p = 0.23). Suy luận hợp lý nhất: phần đầu của các bản ghi này đã bị cắt đi một số giây không nguyên, loại bỏ transient khởi động và 2-5 phút đầu nhiều artifact; **nhãn N2 không lệch, nhưng "elapsed time kể từ đầu" mang nghĩa khác** ở chúng. Việc cắt này là suy ra từ thống kê chứ không được ghi trong metadata, nên cũng có thể là một giao thức ghi hay firmware khác. Nhóm này chiếm 13.6 % train và 22.5 % test (Fisher p khoảng 0.15), cùng hướng với phát hiện phút đầu của test sạch hơn: cả hai đều là dịch chuyển train/test ở đúng phần của luồng mà các cửa sổ 300-600 s và chạm trần nhìn thấy.

Các phép kiểm tra còn lại đều âm tính. Không có đoạn trùng lặp chính xác nào (không chuỗi 0.25 s nào của TP9 hay AF7 xuất hiện hai lần trong 540 bản ghi) và không có near-duplicate; không có đường nối ở ranh giới 30 s. Không bản ghi nào trông như nhãn sai ở mức cả bản ghi: AUC theo bản ghi của composite giữa cửa sổ sau N2 và cửa sổ trước N2 có median 0.975, chỉ một bản ghi dưới 0.7 (sub-142 ses-001, 69 % cửa sổ BAD), và không bản ghi nào giống ngủ ngay từ đầu. Lệch thời điểm cỡ vài chục giây thì không loại trừ được (9 bản ghi có bước nhảy 35-105 s sau N2, 25 bản ghi hơn 300 s trước N2), và phép sàng lọc này phụ thuộc vào một composite huấn luyện trên chính các nhãn đó, nên chỉ loại được lỗi nhãn thô. Về phần cứng, có 33 bản ghi (21 người) với tín hiệu common-mode mạnh trên các kênh (median SD cửa sổ 57 µV ở thái dương so với 9.4 µV chung), 8,022 bước nhảy lớn cô lập (median 418 µV) và 11 bản ghi có hơn 80 % cửa sổ BAD: model cần xử lý bền theo từng kênh (clip, mask, feature đếm artifact) thay vì giả định bốn kênh sạch.

Ba bộ PSG phụ trợ: chỉ CinC-2018 có khối lượng, và chỉ feature tương đối chuyển được

Phần này tổng hợp các phân tích đã cache trên Sleep-EDF (sleep-cassette), HMC và phần train của PhysioNet/CinC 2018 (các script `psg_*.py`, `cinc_*.py`; số liệu trong `psg_out/psg_summary.json` và `cache_cinc2018/summary_hypno.json`). Chưa có thí nghiệm transfer nào (pre-train PSG, fine-tune Muse) được chạy, nên mọi khuyến nghị ánh xạ kênh bên dưới chỉ dựa trên độ tương đồng phổ.

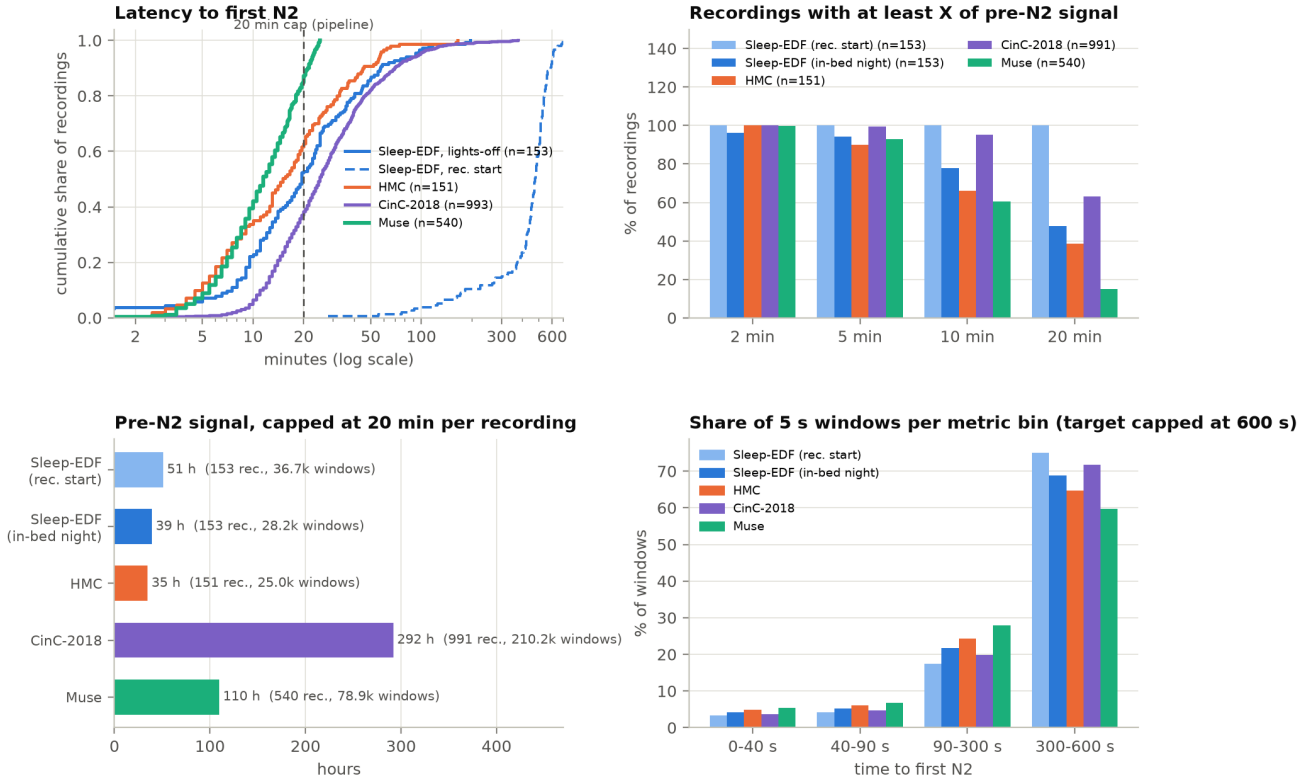
	Sleep-EDF SC	HMC	CinC-2018 train	Muse
Bản ghi / người	153 / 78	151 / 151	994 / 994	540 / 203
Quần thể	khoẻ mạnh, 25-101 tuổi, tại nhà	phòng khám giấc ngủ, 54 ± 15 tuổi	bệnh nhân lab giấc ngủ, median 56 tuổi	không có metadata

	Sleep-EDF SC	HMC	CinC-2018 train	Muse
Tần số, kênh EEG	100 Hz; Fpz-Cz, Pz-Oz	256 Hz; F4-M1, C4-M1, O2-M1, C3-M2	200 Hz; 6 EEG tham chiếu mastoid	128 Hz; TP9, AF7, AF8, TP10
N2 đầu tiên (median)	7.9 h từ đầu file; 19.5 phút từ lights-off	15.0 phút	25.0 phút (13.5 phút từ lúc bắt đầu chấm)	11.5 phút
Giờ trước N2 (trần 20 phút)	39.2 (đêm); 51.0 (pipeline)	34.8	291.9	109.6
Cửa sổ 5 s	28.2k (36.7k)	25.0k	210.2k	80,975

Ba điểm đáng nhớ. Thứ nhất, **CinC-2018 là nguồn PSG duy nhất có khối lượng thật**: 991 đối tượng dùng được, 292 h tín hiệu trước N2 ở trần 20 phút (Muse: 110 h), 210,150 cửa sổ (gấp 2.6 lần Muse), với tỉ lệ cửa sổ theo bin (3.8 / 4.7 / 19.8 / 71.8 %) gần với Muse (5.3 / 6.7 / 27.3 / 60.7 %); Sleep-EDF (đêm) và HMC chỉ thêm 28k và 25k cửa sổ. Thứ hai, **Sleep-EDF theo pipeline chính thức chủ yếu là thức ban ngày**: N2 đầu tiên nằm ở median 7.9 h sau đầu bản ghi (bản ghi bắt đầu khoảng 15:40), và 21 % bản ghi có N2 trước lights-off (ngủ trưa); chỉ sau khi cắt tại lights-off các con số mới giống một bài toán sleep onset (median 19.5 phút). Điều này giải thích vì sao proxy Sleep-EDF của warm-up ít giá trị. Thứ ba, CinC-2018 bắt đầu bằng một đoạn không được chấm (median 510 s) và là quần thể bệnh lý (chỉ số sự kiện hô hấp median 17.5 mỗi giờ; sự kiện hô hấp dồn về onset, 0.5 sự kiện mỗi phút tại N2 đầu tiên), nên pseudo-recording phải bắt đầu ở cuối đoạn chưa chấm.

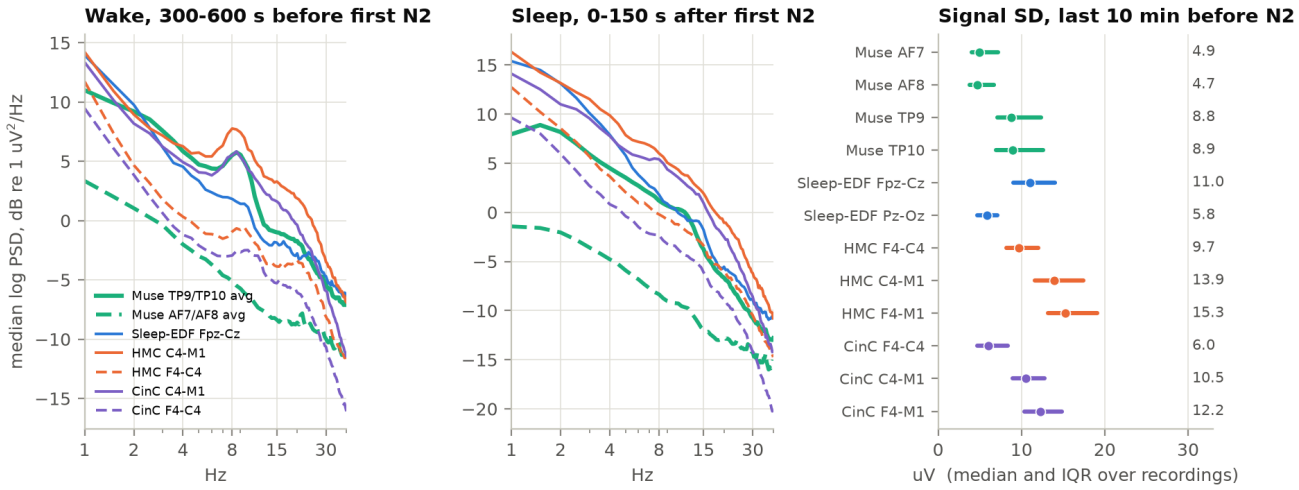
Cấu trúc hypnogram quanh onset giống nhau ở cả ba bộ và bổ sung điều Muse không cho thấy (Muse chỉ lưu thời điểm N2 đầu tiên): epoch ngay trước N2 đầu tiên là N1 ở 97-99 % bản ghi, chuỗi N1 liền trước dài median 120-210 s, và giấc ngủ không ổn định sau đó (có ít nhất một epoch thức trong 5 phút sau N2 đầu tiên ở 22-32 % bản ghi, trong 20 phút ở khoảng một nửa). Target của các cửa sổ trước N2 không bị ảnh hưởng, nhưng nhãn sau N2 thì có.

Về domain gap, biên độ là khác biệt đầu tiên: AF7/AF8 của Muse có SD nhỏ hơn khoảng 2.5 lần (4.9 / 4.7 µV) so với mọi đạo trình trán-mastoid của PSG (10-15 µV) và thấp hơn 5-9 dB về PSD, trong khi TP9/TP10 (8.8 / 8.9 µV) gần với các đạo trình tham chiếu mastoid (+0.5 đến +3 dB). Về hình dạng phổ, các đạo trình mastoid của HMC và CinC cách TP9/TP10 chỉ 0.8-0.9 dB RMS, còn không đạo trình nào gần AF7/AF8 (tốt nhất là lưỡng cực trán-trung tâm F4-C4: 1.2 dB ở HMC, 1.5 dB ở CinC). Muse còn thiếu phần tăng sóng chậm sau onset (công suất tại 1 Hz là 8.0 dB ở TP trung bình so với 14-17 dB ở PSG); nguyên nhân chưa được điều tra.



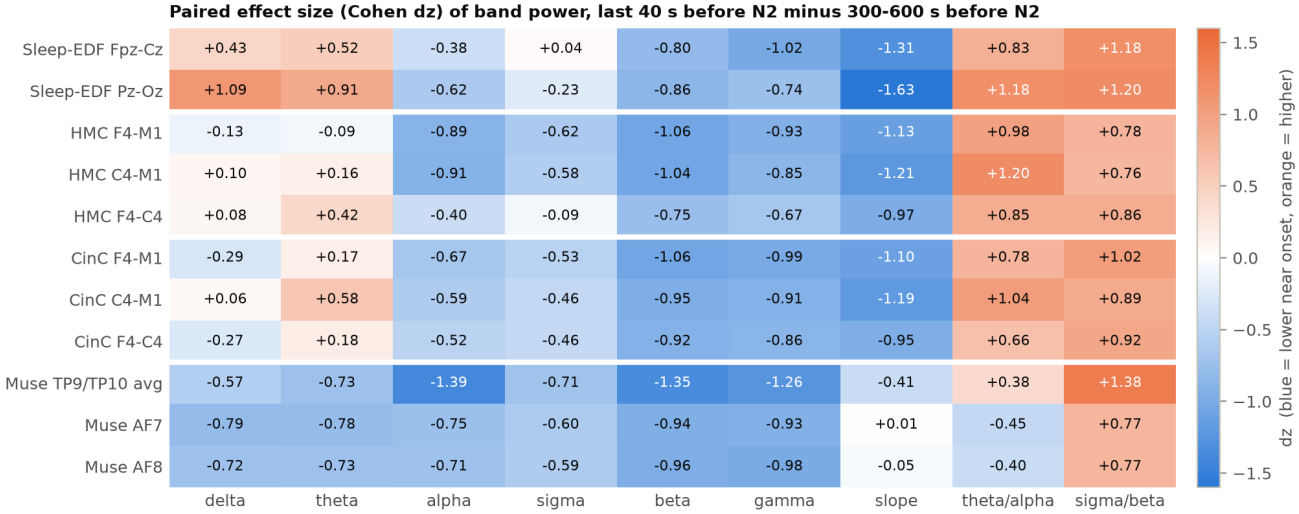
Sleep-EDF (trong giường), HMC và CinC ngủ muộn hơn Muse (median 19.5, 15.0, 25.0 so với 11.5 phút), chỉ CinC cho nhiều tín hiệu trước N2 hơn Muse ở trần 20 phút (292 h so với 110 h), và mốc đầu bản ghi của Sleep-EDF vô nghĩa cho bài toán onset.

Về sinh lý căn theo onset (dz của 40 s cuối trừ 300-600 s trước N2), alpha, beta, gamma giảm và sigma/beta tăng ở cả bốn nguồn với cùng dấu và độ lớn tương tự: beta -0.75 đến -1.07 ở PSG so với -0.94 đến -1.35 ở Muse; sigma/beta +0.76 đến +1.20 ở PSG so với +0.77 (AF) và +1.38 (TP) ở Muse. Ngược lại, delta và theta tuyệt đối tăng ở Sleep-EDF và một số đạo trình trung tâm nhưng **giảm ở mọi kênh Muse** (-0.57 đến -0.79), theta/alpha tăng ở PSG (+0.66 đến +1.20) nhưng giảm ở AF của Muse, và Muse có một đợt giảm công suất băng rộng 2.5-6 dB trong 10 phút cuối mà PSG không có (PSG thay đổi tối đa khoảng 2.5 dB). Kết luận thực dụng: **feature tương đối (sigma/beta, độ dốc phổ) chuyển được; band power tuyệt đối thì không**. Giả thuyết (chưa kiểm chứng) cho khác biệt này là pha thức của Muse chứa artifact chuyển động và tiếp xúc đang tắt dần, còn thức của PSG là thức yên tĩnh trên giường.



TP9/TP10 của Muse khớp với PSG tham chiếu mastoid về mức và hình dạng phổ, trong khi AF7/AF8 thấp hơn 5-9 dB (SD 4.7-4.9 µV so với 10-15 µV) và Muse có công suất 1-2 Hz sau onset yếu hơn.

Cách dùng từng bộ suy ra từ đó. CinC-2018 là nguồn chính: TP9 từ F3-M2 hoặc C3-M2, TP10 từ F4-M1 hoặc C4-M1, AF7/AF8 từ F3-C3 và F4-C4, bắt đầu pseudo-recording ở cuối đoạn chưa chấm và đánh dấu 131 đối tượng có đoạn chưa chấm dài ít nhất 20 phút. HMC: TP từ C3-M2/C4-M1 (gain khoảng 0.7), AF từ F4-C4, chấp nhận low-pass phần cứng 35 Hz (không dùng gamma). Sleep-EDF: chỉ dùng đêm trong giường (cắt tại lights-off theo `SC-subjects.xls`), và vì chỉ có một cặp kênh nên phù hợp nhất làm regulariser. Cả ba cần resample polyphase về 128 Hz, và vì không đạo trình PSG nào tái tạo được cả hình dạng lẫn mức của AF7 và TP9, nên dùng channel dropout hoặc gán kênh ngẫu nhiên. Những gì chưa tính: thống kê tín hiệu CinC chỉ dựa trên mẫu ngẫu nhiên 150 đối tượng; effect size không có khoảng tin cậy và bỏ qua các đêm lặp lại; không có dữ liệu Muse-PSG đồng thời; sai lệch R&K so với AASM về thời điểm N2 chưa được đo.



Effect size của alpha, beta, gamma và sigma/beta cùng dấu và cùng cỡ ở mọi bộ dữ liệu, trong khi delta, theta và theta/alpha khác dấu giữa PSG và Muse.

Y văn: chưa ai dự đoán "số giây đến N2", và N1 là điểm yếu của mọi stager

Sinh lý chuyển tiếp: tipping point 1.5-3.5 phút trước N2, với độ phân tán 2-3 phút

Y văn mô tả việc đi vào giấc ngủ như một quá trình nhiều phút, không đồng bộ theo vùng: alpha sau giảm và dịch ra trước, rồi theta và hoạt động chậm cùng vertex sharp wave, rồi spindle và K-complex; Ogilvie kết luận rằng không nên dùng một chỉ số đơn lẻ nào (Ogilvie 2001). Nghiên cứu căn theo epoch stage 2 đầu tiên cho thấy alpha trung tâm-sau và beta giảm dần trước onset, delta và theta trán-trung tâm tăng trước onset, còn **sigma chỉ tăng rõ sau onset** (Gorgoni et al. 2019). Nguồn định lượng nhất là Li et al. 2025: trên 1,011 PSG tại nhà (MESA) và 267 đêm lab của 36 người trẻ, một "sleep distance" đa feature (47-50 feature mỗi epoch 6 s) ổn định trong phần lớn thời gian rồi sụp đột ngột; tipping point ở mức cá nhân nằm ở **-2.15 ± 2.56 phút** (cohort 1) và **-3.40 ± 1.94 phút** (cohort 2) trước onset, vùng chẩm tip sớm hơn vùng trán, và từng feature đơn lẻ khớp mô hình kém (peak beta frequency $R^2 = 0.43 \pm 0.28$) với các feature chủ đạo khác nhau giữa người (Li et al. 2025). Cần chú ý định nghĩa: Li et al. lấy onset là phút N2 đầu tiên (hai epoch N2 liên tiếp), còn target của cuộc thi là "epoch đầu tiên được chấm N2, không phải khởi đầu của một chuỗi N2 liên tiếp" (NeuralBench Track 3 guide), một nhãn nhiễu hơn.

EDA Muse tái hiện bức tranh này về hướng nhưng đặt mọi thứ muộn hơn và hẹp hơn: bước nhảy của composite ở median -90 s (IQR khoảng 85 s), trán muộn hơn thái dương khoảng 45 s, phù hợp với "sau tip trước, trước tip sau" của Li et al. Hai khác biệt về định nghĩa giải thích một phần khoảng cách: mốc ở đây là epoch N2 đầu tiên chứ không phải N2 ổn định, và score được neo vào trạng thái sau N2. Ba điểm mà Muse không khớp với y văn PSG đáng được ghi nhớ khi chọn feature: delta/theta trán không tăng trước N2 trên Muse (bị hoạt động mắt che); sigma tuyệt đối không tăng nhưng sigma/beta và sigma residual bắt đầu leo 1-2 phút trước N2; và chuyển động mắt chậm, vốn được báo cáo là đi trước alpha blocking 10-20 phút trên EOG (Magosso et al.), trên Muse hiện ra như tương quan âm AF7-AF8 đạt đỉnh 1.5-3 phút trước N2. Y văn không có nghiên cứu nào mô tả 60-120 s cuối trước epoch N2 đầu tiên theo từng giây, cũng không có nghiên cứu nào đo khả năng phát hiện spindle hay vertex wave trên montage Muse; phần EDA ở trên vì thế là nguồn duy nhất cho câu hỏi đó.

Nhãn "N2 đầu tiên" tự mang sàn sai số cỡ chục giây

Theo quy tắc AASM, N2 bắt đầu ở epoch mà một K-complex không kèm arousal hoặc một spindle xuất hiện trong nửa đầu của epoch đó hoặc nửa sau của epoch trước, và epoch dài 30 s (tóm tắt quy tắc AASM). Sự kiện định nghĩa vì thế nằm đâu đó trong dải ±15 s quanh đầu epoch được gán nhãn; nếu nó phân bố đều trong dải này thì ngay cả một model đoán trúng sự kiện vẫn mang sai số tuyệt đối trung bình khoảng 7.5 s, và trong khoảng một nửa trường hợp sự kiện định nghĩa còn chưa xảy ra khi target về 0 (đây là suy luận số học, không phải kết quả đo). Độ đồng thuận giữa người chấm cũng chỉ ở mức vừa: meta-analysis 11 nghiên cứu cho kappa 0.57 với N2 và 0.24 với N1 (Lee et al. 2022), và chương trình inter-scorer của AASM ghi nhận người chấm đặc biệt khó với epoch W cuối cùng trước sleep onset (Rosenberg & Van Hout 2013). Một nhiễu nhãn cỡ một epoch (30 s) là giả định làm việc hợp lý cho dữ liệu chấm tay.

Với Muse, cách tạo nhãn không được ghi lại: trang NEMAR chỉ nói mỗi bản ghi có một annotation N2-onset (NEMAR nm000287). Nghiên cứu kiểm định độc lập duy nhất về Muse-S (56 người, một đêm lab, stager tự động v2 của InteraXon so với một kỹ thuật viên PSG) cho kappa tổng 0.76, N2 0.75, N1 0.41, loại 9/56 bản ghi (16 %) vì tín hiệu kém, và sleep onset latency lệch -3.4 phút với SD của hiệu 11.4 phút (Lanthier et al. 2025). Nếu nhãn cuộc thi đến từ cùng họ stager (một suy luận, chưa có tài liệu xác nhận), nhãn là một hàm tất định của tín hiệu mà model về nguyên tắc học được, kể cả lưới epoch 30 s; bước nhảy nhỏ tại onset chỉ có ở bản ghi on-grid trong EDA phù hợp với một scorer căn theo epoch. Kết hợp sàn lượng tử hoá, nhiễu người chấm và AUC 0.63 của hai cửa sổ kề mốc 40 s, kết luận nhất quán là: trong bin 0-40 s phần cải thiện khả dĩ so với hằng số vốn nhỏ, và điều quan trọng là **nhận ra bản ghi đang ở trong 40-90 s cuối**.

Tiền lệ dự đoán: sai số cỡ phút trên PSG, không có con số nào theo giây

Dự đoán thật sự (trước thời điểm) thời gian đến giấc ngủ từ EEG rất hiếm. Tiền lệ gần nhất là phần pseudo-real-time của Li et al. 2025: tâm "đã ngủ" và ngưỡng tới hạn lấy từ một đêm huấn luyện của chính người đó, feature z-score theo đêm huấn luyện, s(t) được làm trơn bằng moving median causal; sai số dự đoán tipping point là **0.82 ± 2.14 phút** (n = 944) và 13 % trường hợp cho dự đoán không hợp lệ; con số "trên 0.95" trong abstract là cosine similarity giữa hai đường cong chứ không phải độ chính xác phân loại (Li

et al. 2025). Nhóm Korea University mô hình hoá sleep onset latency như một phân phối: MEDi-SOL so sánh các phân phối đầu ra Normal, log-Normal, Weibull, log-Logistic với một mô hình survival và một mô hình hồi quy trên SHHS, và ensemble log-Logistic với log-Normal cho C-index và MAE tốt nhất (số MAE cụ thể không truy cập được) (MEDi-SOL). Mô hình state-space của Prerau et al. theo dõi "wake probability" và cho thấy epoch N2 đầu tiên là mô tả một-thời-điểm tệ nhất của sleep onset hành vi (Prerau et al. 2014). Phần còn lại của y văn, kể cả thiết bị đeo, là phát hiện sau sự kiện với độ chính xác 80-90 % hoặc độ lệch latency cỡ phút.

Không tìm thấy bài nào báo cáo sai số theo giây cho time-to-first-N2 ở các horizon cố định, hay AUC cho "N2 trong X giây tới". Kết quả tốt nhất được ghi nhận có SD thời điểm khoảng 2 phút cho một mốc chuyển trạng thái vốn đi trước N2 một khoảng biến thiên 1-4 phút, trên PSG sạch và có một đêm cá nhân để huấn luyện. Công thức của Li et al. vẫn chuyển thẳng được sang model causal (vector đa feature, chuẩn hoá causal, khoảng cách đến tâm "đã ngủ", làm trơn median causal), và 300 s sau N2 trong mỗi bản ghi train cho đúng đoạn cần để định nghĩa tâm. Phần "tâm cá nhân cho người đã thấy" thì cần thận trọng trên dữ liệu này: hợp đồng streaming không cho mang gì từ bản ghi này sang bản ghi khác, EDA chỉ nhận dạng đúng người 21 % (top-1) từ baseline 2.5 phút, và baseline không hưởng lợi đo được từ người đã thấy.

SOTA sleep staging: MF1 78-83 trên Sleep-EDF, N1 F1 quanh 50

Bảng dưới gom các kết quả so sánh được, đọc từ bảng của bài gốc. Sleep-EDF-20 là 20 người / 39 đêm (20-fold), Sleep-EDF-78 là 153 đêm (10-fold), một kênh Fpz-Cz; PhysioNet 2018 dùng 994 bản ghi có nhãn.

Bộ dữ liệu	Model	Ghi chú giao thức	Acc	MF1	N1 F1
Sleep-EDF-20	L-SeqSleepNet	pre-train SHHS, L = 200 epoch	88.6	82.9	53.3
Sleep-EDF-20	L-SeqSleepNet	train từ đầu	86.3	79.3	45.3
Sleep-EDF-20	XSleepNet2	chỉ EEG	86.3	80.6	51.8
Sleep-EDF-78	SleePyCo	9 epoch quá khứ, causal	84.6	79.0	50.4
Sleep-EDF-78	XSleepNet2	L = 20	84.0	77.9	49.9
PhysioNet 2018	SleePyCo	một kênh C3-A2	80.9	78.9	59.3
PhysioNet 2018	XSleepNet2	chỉ EEG, 5-fold	80.3	78.6	không có
HMC (domain chưa thấy)	SleepDG	leave-one-domain-out, EEG + EOG	73.85	71.16	không có

Nguồn: L-SeqSleepNet, XSleepNet, SleePyCo, SleepDG. Với HMC không tìm thấy leaderboard in-domain sạch: bộ này chủ yếu xuất hiện như domain ngoài, nơi các model một-site rơi từ kappa khoảng 0.80 xuống khoảng 0.54 (Alvarez-Estevez & Rijsman 2021, đọc ở mức abstract), các model không có domain generalisation đạt Acc 64.0-65.9, và foundation model fine-tune đầy đủ dồn quanh bAcc 70-74 (REVE-Base 0.7401; trên AdaBrain-Bench Conformer 73.84 và ST-Tran 73.73 thắng mọi foundation model) (REVE; AdaBrain-Bench). Các con số HMC vì thế không so sánh được với cross-validation của Sleep-EDF hay PhysioNet 2018. Các model đa bộ dữ liệu, bất khả tri kênh, đạt mức tương đương mà không cần thấy bộ đích: AnySleep (hơn 20,000 bản ghi, 28 bộ) đạt MF1 0.81 trên Sleep-EDF-SC và 0.79 trên PhysioNet 2018, và 0.726-0.760 với hai kênh bất kỳ (AnySleep); ensemble SLEEPYLAND huấn luyện trên NSRR đạt MF1 trung bình 76.2 ngoài phân phối với N1 F1 trung bình 46.4 (SLEEPYLAND).

Điều liên quan trực tiếp đến Track 03 là **N1 là stage yếu nhất ở mọi model và mọi bộ dữ liệu** (F1 khoảng 33-60) và lỗi dồn ở các chuyển tiếp: trong ma trận nhầm lẫn của SleePyCo, recall của N1 là 45.9 % trên Sleep-EDF (30.7 % thành N2, 15.2 % thành Wake) và 54.6 % trên PhysioNet 2018 (SleePyCo); U-Sleep trung bình N1 F1 0.53 trên 21 bộ (U-Sleep). N2 đầu tiên được quyết định bởi đúng những ranh giới yếu nhất này (Wake, N1, N2). Về ngữ cảnh, SleePyCo thấy hơn 10 epoch đầu vào không cải thiện đáng kể và chỉ dùng epoch quá khứ; ngữ cảnh cả chu kỳ (L = 200) chỉ có lợi với kiến trúc thiết kế riêng và pre-training lớn (L-SeqSleepNet). Không bài nào đánh giá "time to first N2" trực tiếp: mọi stager được chấm theo epoch.

Trên thiết bị đeo, các con số thấp hơn PSG vài điểm và N1 vẫn là chỗ yếu: Dreem đạt 83.5 % so với 86.4 % của năm chuyên gia (Arnal et al. 2020); băng đô Bitbrain Ikon (AF7/AF8) đạt 86.4 % có giám sát trên BOAS (Estevan et al. 2025); ear-EEG đạt MF1 84.1 và cEEGrid sau tai chỉ đạt N1 F1 25.8 dù đã pre-train trên SHHS (L-SeqSleepNet); trên 200 đêm Muse S có nhãn chuyên gia, DeepSleepNet một kênh đạt 85.2 % và "Muse Net" đa kênh 86.3 % (Abou Jaoude et al. 2023). Không tìm thấy model sâu nào huấn luyện trên PSG công khai rồi kiểm tra trên Muse với metric theo lớp.

Với chính bài toán của track, số liệu công bố chỉ có trên Sleep-EDF (task batch `sleep_onset` của NeuralBench, bMAE không trọng số): Chance 205.42 ± 0.01 s, EEGNet 143.30 ± 0.40 s, REVE frozen probe 134.89 ± 2.02

s, được ghi rõ là không so sánh được với điểm Muse streaming; **không có baseline Muse nào được công bố** (competition README). Các baseline nội bộ của nhóm trên giao thức 5-fold seen/unseen vì thế là mốc tham chiếu có nghĩa duy nhất.

Lợi ích của pre-training lớn nhất khi dữ liệu đích hiếm

Pre-train trên PSG lớn rồi fine-tune trên thiết bị đích được nhiều kết quả độc lập ủng hộ, nhưng mức lợi phụ thuộc mạnh vào lượng nhãn đích. SHHS khởi tạo cho SleepTransformer nâng Acc trên Sleep-EDF-78 từ 81.4 lên 84.9 (SleepTransformer); L-SeqSleepNet pre-train SHHS nâng MF1 của ear-EEG từ 79.4 lên 84.1 và của cEEGrid (12 bản ghi) từ 57.3 lên 67.9 (L-SeqSleepNet); một băng đô một kênh trán với 52 bản ghi đạt kappa 0.778 khi fine-tune so với 0.733 khi train từ đầu (TU/e). RobustSleepNet đạt khoảng 97 % F1 in-domain khi chuyển trực tiếp sang bộ chưa thấy và fine-tune đạt mức train-từ-đầu chỉ với khoảng 10-40 bản ghi đích (RobustSleepNet). Trên băng đô AF7/AF8, SSL in-domain (SimCLR, Barlow Twins) thêm 7-10 điểm accuracy khi chỉ có ≤ 7.5 % nhãn nhưng chỉ khoảng 1 điểm khi có đủ nhãn (86.43 so với 87.21), còn LaBraM, CBraMod và SleepFM không vượt baseline có giám sát (Estevan et al. 2025). Với 540 bản ghi Muse mà mọi cửa sổ đều có nhãn, nhóm đang ở cột "100 % nhãn": kỳ vọng thực tế từ pre-training là nhỏ, và nó phải được chứng minh bằng thí nghiệm chứ không giả định.

Dữ liệu có thể thêm: BOAS là bộ gần Muse nhất và có giấy phép CC0

Rule 04 cho phép pre-train trên bất kỳ bộ dữ liệu "công khai, phân phối lại được", cấm tập sealed và "closed clinical datasets", giữ nguyên hiệu lực giấy phép gốc và yêu cầu khai báo mọi corpus ngoài cùng ước lượng compute (rules).

Bộ dữ liệu	Quy mô	Kênh / thiết bị	Giấy phép	Nhận định
BOAS (OpenNeuro ds005555)	128 đêm, 100 người	PSG + băng đô Ikon AF7/AF8 đồng thời, 256 Hz	CC0	Gần Muse nhất; ưu tiên số 1
EESM19 / EESM23	320 đêm	ear-EEG khô tại nhà, một phần có scalp EEG	CC0 trên OpenNeuro	Domain đeo sau/trong tai
DOD-H / DOD-O	25 khoẻ mạnh / 55 OSA	có đạo trình trán (F3-F4, F3-M2, F4-M1), 5 người chấm	MIT theo trang Zenodo	Nhãn đồng thuận nhiều người chấm
CAP Sleep Database	108 PSG	ít nhất 3 EEG	ODC-By 1.0	An toàn, PSG lâm sàng
ISRUC, DCSM	100+ và 255 PSG	PSG lâm sàng	không tìm thấy văn bản giấy phép	Cần kiểm tra trước khi dùng
NSRR (SHHS, MESA…), MASS	hàng chục nghìn PSG	chủ yếu C3/C4	thoả thuận sử dụng, không phân phối lại	Rủi ro với Rule 04; hỏi ban tổ chức

Nguồn: OpenNeuro ds005555, EESM, DOD, CAP, số lượng NSRR theo SLEEPYLAND. BOAS là bộ mở duy nhất có EEG trán AF7/AF8 cấp tiêu dùng đồng thời với nhãn PSG, đúng vị trí điện cực trán của Muse. Không tìm thấy bộ Muse qua đêm công khai nào khác. Hai lưu ý trung thực: giấy phép của DOD và ANPHY-Sleep (29 người, EEG mật độ cao, CC0 theo tóm tắt tìm kiếm) chưa được đọc trực tiếp từ văn bản gốc, và đợt rà soát các bộ phát hành 2024-2026 không có hệ thống, nên có thể còn bộ băng đô công khai khác.

Phương pháp có bằng chứng: chuẩn hoá causal theo bản ghi, head phân phối, và bền với artifact

Phần này ghép hai loại bằng chứng: kết quả đo trên Muse ở phần EDA và kết quả đã công bố trên dữ liệu giấc ngủ khác. Cần nói thẳng ngay từ đầu: không có mô hình hay benchmark nào đã công bố cho việc dự đoán một countdown time-to-N2 có trần từ EEG, nên mọi khuyến nghị về loss, head và hậu xử lý là suy luận từ ML tổng quát và các bài toán lân cận, và phải được quyết định bằng thực nghiệm.

Tiền xử lý: lọc causal có trạng thái, dừng ở 45 Hz, và không chuẩn hoá mất biên độ

Các pipeline giấc ngủ đã công bố hội tụ về high-pass khoảng 0.2-0.5 Hz, low-pass 30-45 Hz, lấy mẫu 60-128 Hz và một thang đo bền theo bản ghi. Pipeline gần cuộc thi nhất là của băng đô Bitbrain Ikon (AF7/AF8): down-sample 256 xuống 128 Hz, band-pass 0.5-45 Hz, rồi z-score (Estevan et al. 2025). Dreem dùng 0.4-35 Hz và mỗi epoch chọn "virtual channel" có chất lượng tốt nhất (Arnal et al. 2020); RobustSleepNet dùng IIR 0.2-30 Hz, chuẩn hoá median/IQR theo bản ghi và clip ở 20 lần IQR (RobustSleepNet); YASA dùng 0.40-30 Hz và không loại artifact (Vallat & Walker 2021). Không bài nào đo cái giá của việc làm cho bộ lọc causal; cái giá đã đo được là của việc mất ngữ cảnh tương lai: staging thời gian thực một kênh trán đạt 78.5 % so với 80.65 % khi thêm một epoch 30 s tương lai (Esparza-Iaizzo et al. 2024). Với hop 5 s và sự kiện cách tới 600 s, đòn bẩy thực tế là ngữ cảnh quá khứ dài chứ không phải pha của bộ lọc.

Ghép với EDA, cấu hình mặc định có cơ sở là: band-pass IIR causal có trạng thái (`scipy.signal.sosfilt` mang `zi` qua các cửa sổ), dừng ở 45 Hz vì phổ rơi dốc trên 52 Hz và điện lưới 50/60 Hz chỉ là biến nhiễu không liên quan onset; tránh hoặc lấy median qua tần số quanh 21.3 và 42.7 Hz; clip các mẫu trên khoảng 780 µV; mask hoặc giảm trọng số 5-10 s đầu. Bộ lọc dùng lúc huấn luyện phải đúng là bộ lọc chạy lúc suy luận, kể cả transient khởi động: lọc zero-phase khi train và causal khi test là một dịch chuyển train/test âm thầm. Với đạo trình AF7 - AF8, high-pass không nên cao hơn khoảng 0.3 Hz (hoặc giữ thêm một bản sao lọc thấp), vì băng 0.1-1 Hz chứa dấu hiệu chuyển động mắt chậm.

Về chuẩn hoá, thống kê cả bản ghi là không causal; các thay thế causal là median/IQR cửa sổ mở rộng, EMA mean/variance (`braindecode.preprocessing.exponential_moving_standardize` có sẵn trong image), hoặc InstanceNorm theo cửa sổ. Bằng chứng công bố duy nhất ở đây là gián tiếp: InstanceNorm thắng BatchNorm trên các bộ dữ liệu chưa thấy (78.97 so với 78.38 BACC) (PSDNorm). Bằng chứng trên Muse thì trực tiếp hơn: trừ running mean của bản ghi cắt hai phần ba phương sai giữa người và nâng composite từ 0.39 lên 0.48, nhưng biên độ tuyệt đối và số artifact lại là tín hiệu (số kênh BAD nằm trong 15 feature đầu; biên độ tuyệt đối cũng là feature hàng đầu của YASA). Vì vậy nếu chuẩn hoá dạng sóng thì phải đưa log-RMS và các chỉ số chất lượng vào như feature phụ. Các đạo trình lưỡng cực AF7 - AF8, TP9 - TP10 và trán trừ thái dương là tuyến tính, causal và triệt tiêu reference chung; chưa có so sánh đo được nào giữa các lựa chọn này cho giấc ngủ, nên đây là suy luận từ hình học điện cực và từ tương quan liên kênh trong EDA.

Artifact: báo cho model biết thay vì làm sạch

Bằng chứng tốt nhất trên chính dữ liệu Muse giấc ngủ là Dynamic Spatial Filtering (DSF): một module attention trước lớp đầu, nhận thống kê không gian theo cửa sổ (log-variance hoặc log-covariance của kênh) và xuất ma trận tái trọng số kênh. Huấn luyện trên các đêm Muse S sạch nhất và kiểm tra trên 17 đêm nhiễu nhất, DSF kèm corruption augmentation đạt median balanced accuracy **65.0 % so với 58.4 %** của mạng thường; trong cùng thí nghiệm, corruption augmentation đứng một mình làm mạng thường tệ đi, Autoreject không giúp, và pipeline Riemannian tangent-space hoạt động kém (Banville et al. 2022). ASR không có kết quả định lượng nào cho giấc ngủ hay ít kênh (Miyakoshi 2023), và với 4 kênh thì không gian con để tái tạo tối đa chỉ có 3 chiều. Không nên loại EOG/EMG bằng ICA hay hồi quy: trên Muse, hoạt động mắt ở trán và công suất 30-45 Hz chính là các marker của thức.

Điều này khớp với EDA: 16 % cửa sổ có một đến ba kênh hỏng, artifact dùng chung giữa các kênh (P(TP10 BAD | TP9 BAD) = 0.72), và tỉ lệ artifact tự mang thông tin. Thiết kế có cơ sở là: đưa các scalar chất lượng theo kênh (log-variance, max |x|, tỉ lệ mẫu ở trần, RMS 30-45 Hz) vào cùng tín hiệu; thêm một cổng kiểu DSF; huấn luyện với corruption tổng hợp bắt chước những gì EDA thấy (thay cả kênh bằng nhiễu, bước nhảy lớn vài trăm µV, common-mode); và với model hồi quy, dùng điểm chất lượng để giảm trọng số cửa sổ thay vì bỏ hẳn, giữ cho chuỗi liền mạch.

Feature: quỹ đạo đã smoothing của feature phổ đơn giản mang phần lớn sức dự đoán

YASA cho thấy 9 trong 20 feature quan trọng nhất là các phiên bản đã smoothing hoặc chuẩn hoá theo thời gian, rằng trung bình trượt 2 phút quá khứ (causal) nằm trong số đó, và rằng thời gian đã trôi cũng nằm trong top 20; EEG-only đạt 85.12 % so với 87.46 % khi thêm EOG, EMG và nhân khẩu học, với N1 vẫn là stage yếu nhất (F1 0.432) (Vallat & Walker 2021). Độ dốc aperiodic tách thức khỏi NREM (Lendner et al. 2020), và nghiên cứu augmentation kết luận rằng "bảo toàn tỉ số công suất giữa các tần số là quan trọng" cho staging (Rommel et al. 2022). Trên Muse giấc ngủ nhiễu, feature thủ công với random forest cạnh tranh được với một CNN thường (Banville et al. 2022). Không bài nào xếp hạng feature cho việc dự đoán time-to-first-N2; bảng xếp hạng trong phần EDA là bằng chứng trực tiếp duy nhất.

Họ feature	Bằng chứng trên Muse (EDA)	Bằng chứng y văn	Dùng cho
beta, gamma, alpha TP, biên độ, line length	rho 0.27-0.31 ở elapsed cố định; nửa đường ở -240 đến -360 s	band power là lõi của Dreem, YASA	bin 90-600 s
sigma/beta, sigma residual, 1/f exponent, delta/beta, centroid	AUC 0.60-0.65 cho 0-40 so với 40-90 s; dz đến +1.38	slope: Lendner 2020; sigma sau onset: Gorgoni 2019	90-120 s cuối
EMA 30 s và EMA 30 s trừ runmean	median rho 0.125 lên 0.19; composite 0.39 lên 0.48	9/20 feature đầu của YASA là smoothing	mọi bin
số kênh BAD, log max, RMS 30-45 Hz	BAD 31 % xuống 3 %; top 15 feature	DSF trên Muse: +6.6 BACC	300-600 s và trần
AF7 - AF8 băng 0.1-1 Hz	tương quan -0.29 ở -180..-90 s (zero-phase)	SEM đi trước onset (EOG)	40-300 s; chưa thử causal
elapsed time	rho -0.44; hazard tăng theo thời gian	top 20 của YASA	prior; rủi ro sau 15 phút

Hai chi tiết triển khai rút từ EDA: định nghĩa band spindle là khoảng 11-15 Hz và alpha là 8-11 Hz; và giữ các detector phụ thuộc ngưỡng cả bản ghi (đếm burst spindle) ra khỏi bộ feature causal, vì chúng vừa thưa vừa không causal.

Augmentation: lợi ích nhỏ ở cỡ dữ liệu này, ưu tiên các dịch chuyển đã đo được

Trên Sleep Physionet (78 người, hai kênh), TimeReverse cho tới +13 % tương đối và FTSurrogate tới +12 % tương đối balanced accuracy ở chế độ ít dữ liệu (tương đương gấp 4 lần dữ liệu), FrequencyShift nhỏ (tối đa 0.3 Hz) cho lợi ích nhỏ hơn; GaussianNoise, BandstopFilter và SmoothTimeMask không có tác dụng, và SmoothTimeMask còn có hại cho NREM vì xoá K-complex và spindle; không augmentation không gian nào giúp (với lưu ý là chỉ có hai kênh đường giữa), ChannelsDropout tốt nhất ở p = 0.4 nhưng không có ý nghĩa thống kê (Rommel et al. 2022). Chính sách theo lớp học tự động chưa đem lại mức tăng kỳ vọng (CADDA). Tất cả đều là augmentation lúc huấn luyện nên không vướng causality, với hai ngoại lệ cho bài toán này: không bao giờ đảo thứ tự các cửa sổ của một chuỗi (phá hướng thời gian mà target phụ thuộc), và FTSurrogate phá đúng hình thái spindle/K-complex định nghĩa N2, nên nếu dùng thì chỉ nên áp cho các cửa sổ xa onset.

Với khoảng 81k cửa sổ từ 203 người, dữ liệu nằm giữa chế độ "nhỏ" và "đầy đủ" của Rommel, nên nên kỳ vọng mức tăng tương đối một chữ số thấp từ augmentation dạng sóng. Ưu tiên hợp lý là các augmentation nhắm vào dịch chuyển **đã đo được** trong dữ liệu này: cắt ngẫu nhiên phần đầu bản ghi (nhóm off-grid và phút đầu sạch hơn của test), corruption theo kênh đi cùng DSF, gain ngẫu nhiên theo bản ghi và theo kênh (SD giữa bản ghi của biên độ là 6-7 dB), dịch tần số nhỏ theo bản ghi (alpha peak 9.0 ± 1.1 Hz), và hoán đổi bán cầu AF7 với AF8, TP9 với TP10. Hai mục cuối chưa có kết quả đo trên giấc ngủ.

Domain adaptation: chuẩn hoá phổ cho +1 điểm trung bình và +5-6 điểm ở cohort lệch nhất

PSDNorm, một lớp chuẩn hoá ánh xạ PSD của feature map về một barycenter chạy, đạt BACC trung bình 79.51 so với 78.38 của BatchNorm trong leave-one-dataset-out trên 10 bộ dữ liệu, và 70.57 so với 64.72 trên bộ lệch nhất (CHAT, trẻ em), với mức tăng lớn nhất ở các đối tượng khó nhất; trên MASS nó lại kém BatchNorm (PSDNorm). Phiên bản tiền xử lý TMA/CMMN cho +0.4 trung bình và +5.0 trên CHAT (CMMN). SleepDG thêm 6.3 MF1 so với baseline không DG khi kiểm tra trên domain chưa thấy, và các baseline DG đơn giản (CORAL, MMD, IRM, DANN) thêm 2.7-4.2 MF1 (SleepDG). Huấn luyện đa bộ dữ liệu với lấy mẫu kênh ngẫu nhiên là đòn bẩy còn lại được ủng hộ nhất quán (RobustSleepNet; AnySleep). Test-time adaptation có cập nhật trọng số (kiểu Tent) chưa có bằng chứng thuyết phục cho giấc ngủ: StableSleep không báo cáo số liệu của baseline đóng băng (StableSleep).

Trong hợp đồng của cuộc thi, adaptation "hợp lệ" là bất cứ thứ gì chỉ dùng mẫu quá khứ của bản ghi hiện tại và reset ở ranh giới bản ghi. Dạng an toàn nhất là dựa trên thống kê chứ không dựa trên gradient: running mean/variance, running PSD, feature tương đối theo baseline phút đầu. Các phiên bản causal của TMA và PSDNorm chưa từng được đo (bản công bố dùng cả bản ghi hoặc cửa sổ hai chiều 17.5 phút), và có một cảnh báo riêng cho bài toán này: một ánh xạ PSD với EMA chậm sẽ chuẩn hoá mất một phần chính sự thay đổi phổ từ thức sang ngủ; một ánh xạ đóng băng sau 1-2 phút đầu tránh được điều đó nhưng lại dựa vào đúng đoạn nhiễu nhất và lệch nhất giữa train và test. Kết quả EDA rằng running mean loại hai phần ba phương sai giữa người là phiên bản đơn giản, đã đo được, của cùng ý tưởng.

Mô hình hoá: dự đoán một phân phối rồi giải mã bằng weighted median

Y văn survival và prognostics nhất quán ưu tiên dự đoán một phân phối hay hazard rời rạc theo thời gian rồi đọc ước lượng điểm từ đó: tham số hoá hazard (logistic hazard) nhỉnh hơn tham số hoá PMF, và rời rạc hoá theo phân vị tốt hơn lưới đều với dữ liệu nhỏ (Kvamme & Borgan 2021); hồi quy thứ tự nhất quán hạng cho MAE 2.64 so với 3.34 của phân loại cross-entropy trên ước lượng tuổi (CORAL); thay MSE bằng cross-entropy trên các bin với target làm trơn Gauss (HL-Gauss) cải thiện hồi quy giá trị (Farebrother et al. 2024; Imani & White 2018); và dữ kiện riêng cho giấc ngủ là MEDi-SOL, nơi ensemble phân phối thắng hồi quy điểm (MEDi-SOL). Bằng chứng đều từ lĩnh vực khác; chưa có so sánh trực tiếp hazard so với hồi quy trên một bài toán time-to-event EEG với metric MAE có trọng số theo bin.

Ba đặc điểm của metric quyết định cách dùng các head này. Một, trọng số hiệu dụng theo cửa sổ là trọng số bin chia cho số cửa sổ của bản ghi trong bin đó: khoảng 1.25 mỗi cửa sổ ở [0, 40), 0.5 ở [40, 90), 0.071 ở [90, 300) và không quá 0.016 ở [300, 600]; loss huấn luyện nên tái tạo đúng điều này thay vì nhân 10/5/3/1 theo mẫu. Hai, vì trọng số phụ thuộc target thật, dự đoán điểm tối ưu Bayes là **weighted median có điều kiện** chứ không phải median: mọi model xác suất đều có thể được giải mã như vậy sau khi huấn luyện bằng cách tái trọng số phân phối dự đoán theo w(y) rồi lấy median, kéo dự đoán xuống mỗi khi có xác suất đáng kể là đang gần onset (đây là kết quả chuẩn được suy ra trong ghi chú, không phải trích dẫn). Ba, trần 600 s là một điểm khối lượng (lớp "≥ 600") chứ không phải censoring, nên head rời rạc cần một bin kết thúc riêng; đây đúng là thuốc cho bệnh đã đo ở baseline, vốn không bao giờ dự đoán quá khoảng 300 s. Thứ tự ưu tiên các công thức theo trọng lượng bằng chứng (chưa cái nào được thử ở đây): logistic hazard rời rạc với encoder hồi quy và giải mã weighted median, xấp xỉ ngang head thứ tự kiểu CORN; rồi phân loại soft-bin; rồi hồi quy quantile; rồi L1 với trọng số khớp metric, là đối chứng tự nhiên cần đánh bại. Cấu hình mặc định của neuralbench cho task này (MSE với trọng số 1e-4 và `RegressionBinSampler`) nằm ở cuối thứ tự đó.

Về kiến trúc chuỗi, các hệ causal đã triển khai dùng mô hình hồi quy trên 2.5-15 phút ngữ cảnh quá khứ: Dreem dùng LSTM 2 lớp trên 30 epoch gần nhất (Arnal et al. 2020), và các nghiên cứu offline cho thấy ngữ cảnh quá vài chục phút thêm không quá khoảng 1 điểm (L-SeqSleepNet). Một GRU một chiều khớp với API có trạng thái, từng cửa sổ một. Lời giải hạng nhất của cuộc thi Kaggle phát hiện trạng thái ngủ (accelerometer cổ tay, độ phân giải 5 s) cho bốn bài học có thể chuyển giao: target one-hot học kém nên họ dùng bướu Gauss thu hẹp dần; một model tầng hai xếp chồng trên dự đoán tầng một; hậu xử lý thiết kế quanh metric là mức tăng lớn nhất (model đơn tốt nhất 0.829 lên 0.852 trên private LB); và điểm một fold quá nhiễu nên mọi thí nghiệm được chấm trên tất cả các fold, tin CV hơn leaderboard nhỏ (lời giải hạng nhất).

Hai chỉnh sửa so với ghi chú gốc cần nêu rõ. Ghi chú về mô hình hoá gợi ý head phụ "5 lớp sleep stage của cửa sổ hiện tại" như thể có sẵn trong nhãn Muse; theo EDA và metadata, Muse chỉ lưu thời điểm N2 đầu tiên, nên nhãn stage chỉ có ở các corpus PSG. Trên Muse, mục tiêu phụ khả dụng là "sau N2 so với trước N2" (dùng 300 s sau N2 lúc huấn luyện, chính là cách dựng sleep-likeness score với AUC 0.972) và các head nhị phân đa horizon "N2 trong k giây". Tương tự, không bản ghi Muse công khai nào thiếu N2, nên censoring thật chỉ nảy sinh khi dùng PSG ngoài.

Validation: nhóm theo người, chấm theo bản ghi, nhiều seed

Đơn vị độc lập là 540 bản ghi của 203 người chứ không phải 81k cửa sổ. Các kết quả giấc ngủ đã công bố có SD theo seed hoặc fold 0.1-1.5 điểm với dữ liệu đầy đủ và 3-6 điểm ở chế độ ít dữ liệu (PSDNorm; Estevan et al. 2025), và chỉ thay đổi seed còn đánh giá thấp phương sai thật (Bouthillier et al. 2021). Giao thức có cơ sở: fold cố định nhóm theo người, 5 fold với ít nhất 3 seed, so sánh cấu hình bằng hiệu ghép cặp theo bản ghi của chính W-bMAE với bootstrap theo người, và đo SD của baseline (khoảng 10 seed) trước khi tin bất kỳ ablation nào. Tập test công khai 40 bản ghi (SEM khoảng 6 s) chỉ là phép kiểm tra tỉnh táo. Kiểm tra rò rỉ riêng cho bài toán: thống kê cả bản ghi, bộ lọc zero-phase, cửa sổ trượt căn giữa và lớp hai chiều đều rò mẫu tương lai và thổi phồng CV so với leaderboard streaming; cách xác minh là chạy đúng vòng suy luận streaming trên bản ghi held-out và so với điểm CV theo batch. Hướng dẫn chính thức cũng cảnh báo điều tương tự về baseline REVE của NeuralBench, vốn fit scaler trên cả bản ghi (NeuralBench Track 3 guide).

Phương pháp	Mức lợi đã đo	Đo trên dữ liệu nào	Đã thử trên Track 03?	Khuyến nghị
EMA 30 s trừ running mean	composite rho 0.39 lên 0.48	Muse (EDA)	chỉ ở mức feature	dùng, kèm mức thô
Head phân phối + weighted median	MAE 2.64 so với 3.34 (CORAL, tuổi)	ngoài giấc ngủ	chưa	thử trước tiên

Phương pháp	Mức lợi đã đo	Đo trên dữ liệu nào	Đã thử trên Track 03?	Khuyến nghị
DSF + corruption augmentation	+6.6 BACC	17 đêm Muse nhiễu nhất	chưa	thử
FTSurrogate, TimeReverse	đến +12-13 % tương đối, ít dữ liệu	Sleep Physionet	chưa	ưu tiên thấp
PSDNorm / TMA	+1.1 trung bình, +5.9 trên CHAT	10 bộ PSG	chưa; bản causal chưa tồn tại	sau chuẩn hoá đơn giản
Pre-train PSG rồi fine-tune	+0.045 kappa (52 đêm băng đô)	băng đô một kênh	chưa	thử với CinC-2018, BOAS

Pretrained model trong neuroai: dùng được, nhưng bằng chứng không ủng hộ đặt cược vào chúng

Những gì thực sự có sẵn

Kiểm tra code đã cài cho thấy neuralbench 0.4.0 nối sẵn đúng sáu foundation model EEG (REVE-base, LaBraM-base, CBraMod, LUNA-large, BIOT, BENDR), mỗi model được bọc bởi `DownstreamWrapperModel` với các preset frozen probe (linear hoặc attentive), LoRA (r = 4 hoặc 32), fine-tune toàn phần và LP-FT. braindecode 1.8.1 có thêm loader cho EEGPT, CodeBrain, SignalJEPA, EEGDINO, ST-EEGFormer và ZUNA. Không gói nào kèm model pretrained chuyên cho giấc ngủ; các kiến trúc `USleep`, `DeepSleepNet`, `AttnSleep`, `SleepStagerChambon2018` có trong braindecode nhưng không có weights. Số tham số, giấy phép và kích thước file dưới đây đọc từ code cục bộ và HF API ngày 2026-10-12; chưa tải weights nào, nên **chưa có phép đo nào về chất lượng embedding hay độ trễ trên weights thật**.

Model	Tham số	Hợp với 4 kênh Muse	Weights, giấy phép	Cache trên worker	Bằng chứng trên sleep
REVE-base	69.2M	có: position bank chứa TP9, AF7, AF8, TP10; cần 200 Hz	277 MB; REVE Responsible Use License v1.0	có	Sleep-EDF onset 134.89 so với EEGNet 143.30; HMC fine-tune 0.740
LaBraM-base	5.8M	có: từ vựng 128 tên kênh chứa cả 4	23 MB; BSD-3	không	ISRUC 0.763; frozen trên HMC 33.83
CBraMod	4.9M	linh hoạt số kênh; cảnh báo suy giảm dưới 4 kênh	19.7 MB; BSD-3	không	Sleep-EDF fine-tune 74.58; frozen yếu
LUNA-large	40.4M	có: bất biến topology, dùng toạ độ 3-D	CC-BY-ND-4.0 (rủi ro với fine-tune)	không	top-3 NeuralBench; không có số sleep onset
SignalJEPA	3.46M	128 Hz khớp Muse; channel embedding học lại	13.8 MB; MIT	không	không có số sleep
BIOT, BENDR	3.2M, 157M	cần adapter kênh học từ đầu	BSD-3	không	yếu trong NeuralBench

Về mặt vận hành, REVE-base là con đường duy nhất không có rủi ro hạ tầng về weights vì nó nằm trong `staged_hf_models.txt`; các checkpoint nhỏ (CBraMod, LaBraM, SignalJEPA) nhét vừa ZIP, và việc tải từ HF lúc suy luận không được đảm bảo và tính vào giới hạn một giờ. Có bốn cái bẫy cụ thể. Một, hàm dựng `REVE(...)` luôn tạo `RevePositionBank()` và sẽ tải `positions.json` nếu không có cache; file này không nằm trong danh sách staged, nên phải ship kèm và đặt `REVE_POSITIONS_PATH`, hoặc xác minh bằng một debug submission. Hai, image worker (theo pip freeze đã chụp) chạy neuralbench 0.3.1 trong khi máy cục bộ là 0.4.0, nên submission chỉ nên phụ thuộc braindecode và torch, không phụ thuộc nội bộ neuralbench. Ba, giấy phép REVE cho phép nghiên cứu và phân phối lại bản fine-tune nhưng yêu cầu kèm văn bản giấy phép và có các hạn chế sử dụng (REVE license); LUNA là CC-BY-ND, và fine-tune có thể bị coi là tác phẩm phái sinh. Bốn, baseline chính thức `reve_probe.py` resample 128 lên 200 Hz bằng nội suy tuyến tính, z-score từng cửa sổ từng kênh, lấy trung bình feature lớp cuối qua kênh và qua thời gian rồi đưa vào Ridge: nó không lọc, không có ngữ cảnh ngoài cửa sổ 5 s, và phép z-score theo cửa sổ **xoá đúng thông tin biên độ và artifact mà EDA xác định là tín hiệu**.

Hiệu năng: nhỉnh hơn khi fine-tune trên epoch 30 s, yếu khi đóng băng, và không hơn mạng nhỏ trên cửa sổ ngắn ít kênh

Con số duy nhất cho đúng loại bài toán này là trên Sleep-EDF: REVE frozen probe 134.89 so với EEGNet 143.30 s bMAE, tức tốt hơn khoảng 6 %, với Chance ở 205.42; hướng dẫn chính thức tự lưu ý rằng điểm REVE của NeuralBench là lạc quan cho submission streaming (competition README). Bài báo NeuralBench (36 task EEG, 14 kiến trúc) kết luận rằng các foundation model EEG hiện tại "chỉ nhỉnh hơn một chút" so với baseline chuyên biệt, với REVE, LaBraM và LUNA dẫn đầu về hạng trung bình, BENDR và BIOT kém, và sleep staging nằm trong nhóm task "gần bão hoà"; bài báo chỉ fine-tune end-to-end, không có kết quả linear probe (Banville et al. 2026).

Các benchmark độc lập cho một bức tranh nhất quán. Khi fine-tune toàn phần cho staging 30 s, foundation model ngang hoặc hơn mạng có giám sát vài điểm: trên Sleep-EDF (hai kênh) CBraMod 74.58 và LaBraM 72.86 so với 70.20-71.61 của các mạng có giám sát; nhưng cùng nghiên cứu đó thấy trên các task cửa sổ ngắn (≤ 5 s) mạng có giám sát ngang hoặc tốt hơn, và khi bị giới hạn kênh foundation model ngang hoặc hơi kém EEGNet lớn, với kết luận rằng chúng "giúp nhiều nhất ở task ngữ cảnh dài" và lợi ích lớn nhất ở chế độ ít dữ liệu (Kommineni et al. 2026). Khi đóng băng, kết quả trên sleep thường yếu: trên HMC, backbone frozen cho bAcc 20.00 (BENDR), 33.83 (LaBraM), 22.29 (CBraMod) so với 69.67-72.63 sau fine-tune (EEG-FM-Bench); linear probe của REVE-Base đạt 0.604-0.647 trên HMC so với 0.7401 khi fine-tune (REVE). Trên băng đô trán AF7/AF8, LaBraM, CBraMod và SleepFM ngang hoặc dưới một CNN nhỏ có giám sát ở mọi tỉ lệ nhãn từ 15 % trở lên (Estevan et al. 2025). Khi giảm kênh, REVE-Base tụt từ 0.824 (64 kênh) xuống 0.728 (4 kênh) trên motor imagery (REVE). Một bài stress-test còn ghi nhận encoder khởi tạo ngẫu nhiên vượt REVE pretrained trên một task lâm sàng (0.667 so với 0.570 macro-AUROC) (arXiv 2607.24519). Điểm sáng duy nhất cho thiết bị đeo là các model pretrain trên PSG quy mô lớn: SleepFM-2 đóng băng nâng macro-F1 trên EEG băng đô BOAS từ 0.514 lên 0.721 so với cùng kiến trúc khởi tạo ngẫu nhiên, nhưng việc phát hành weights và giấy phép chưa được xác minh (SleepFM-2). Không benchmark nào đánh giá foundation model trên dữ liệu Muse, trên montage 4 kênh trán-mastoid, trên target hồi quy kiểu latency, hay trong suy luận streaming causal.

Khuyến nghị: một thí nghiệm có giới hạn, không phải nền móng

Không nên xây lời giải trên foundation model, vì bốn lý do độc lập cùng chỉ một hướng. Chế độ của track (cửa sổ 5 s, 4 kênh, 157 h dữ liệu có nhãn đầy đủ) đúng là chế độ mà bằng chứng nói foundation model ít lợi nhất. Mức lợi duy nhất đo được trên bài toán tương tự là khoảng 6 % và được chính ban tổ chức gọi là lạc quan. EDA cho thấy thông tin nằm ở quỹ đạo cỡ phút của các feature tương đối theo bản ghi cùng biên độ và artifact, tức ở phần trạng thái hồi quy và chuẩn hoá chứ không ở encoder theo cửa sổ. Và lỗi của baseline nằm ở head (hồi quy về trung bình, không bảo thủ gần onset), thứ mà encoder nào cũng thừa hưởng.

Vẫn đáng chạy một thí nghiệm có giới hạn vì nó rẻ và REVE đã được cache: trích embedding REVE-base theo cửa sổ (đóng băng trước, LoRA trên QKVO sau, với feature theo kênh thay vì trung bình qua kênh) và ghép chúng như input bổ sung vào cùng GRU causal, cùng head, cùng feature biên độ và chất lượng; so sánh ghép cặp trên 5 fold. Nếu mức tăng không vượt nhiễu seed thì bỏ. Thay thế có cùng chi phí và bằng chứng tốt hơn cho thiết bị đeo là pre-train có giám sát chính encoder nhỏ của nhóm trên CinC-2018 và BOAS.

Các thí nghiệm nên chạy tiếp, theo thứ tự ưu tiên

1. **Đo nhiễu của thước đo trước khi so sánh bất cứ thứ gì.** Chạy khoảng 10 seed cho CNN+GRU hiện tại trên fold cố định, báo cáo SD của W-bMAE sealed-style và dựng công cụ so sánh ghép cặp theo bản ghi với bootstrap theo người; xác minh vòng suy luận streaming cho đúng điểm CV. Căn cứ: SEM 1.7-7 s trên 40 bản ghi test; hai baseline chỉ cách nhau 0.53 s với SD của hiệu theo bản ghi là 18.1 s; SD seed/fold 0.1-1.5 điểm trong y văn giấc ngủ; bài học "tin CV trên mọi fold" của lời giải Kaggle hạng nhất.
2. **Loss khớp metric và giải mã bảo thủ trên model hiện có.** Trọng số theo cửa sổ bằng trọng số bin chia số cửa sổ của bản ghi trong bin; sau đó co dự đoán về hằng số khi model không chắc, đặc biệt trong 6 phút đầu. Căn cứ: baseline thua hằng số 3.5-4.3 điểm ở 0-90 s; trees thua hằng số trong 40 s cuối ở mọi bin elapsed dưới 20 phút; 14.8-21.8 % cửa sổ cách N2 hơn 300 s bị dự đoán dưới 90 s. Đây là đối chứng mà mọi head mới phải đánh bại.
3. **Head phân phối có lớp trần tường minh.** Logistic hazard rời rạc hoặc head thứ tự trên các bin của [0, 600) cộng một bin "≥ 600", giải mã bằng weighted median theo trọng số metric; so với L1 có trọng số ở mục 2. Căn cứ: median dự đoán chỉ 179-226 s ở trần trong khi 30.2 % cửa sổ ở trần và 44 % có target ≥

450 s; "biết đúng bin" đáng 62 xuống 22 điểm; Kvamme & Borgan 2021, CORAL, MEDi-SOL.

4. **Input causal tương đối theo bản ghi, hai họ feature.** Thêm vào cả trees lẫn GRU: EMA 30 s, EMA 30 s trừ running mean và mức thô; band sigma 11-15 Hz và alpha 8-11 Hz; sigma residual, 1/f exponent, delta/beta, spectral centroid; số kênh BAD, log max |x|, RMS 30-45 Hz; tương quan trước-sau. Căn cứ: median |rho| 0.125 lên 0.19 và composite 0.39 lên 0.48; AUC đa biến 0-40 so với 40-90 s tăng từ 0.689 lên 0.754 với feature EMA 30 s; running mean loại hai phần ba phương sai giữa người.
5. **Bền với trạng thái khởi động và kiểm toán elapsed time.** Phân tích run `neural_randstart` đã có; huấn luyện với cắt ngẫu nhiên 0-5 phút đầu, mask 5-10 s đầu; báo cáo điểm riêng cho bản ghi off-grid, cho bin 300-600 s và cho model có/không có elapsed time. Căn cứ: off-grid chiếm 13.6 % train và 22.5 % test và thiếu pha ổn định; phút đầu của test sạch hơn (0.20 so với 0.42, Holm p 0.04); 35-50 % mối liên hệ của biên độ là đồng hồ; hazard cuối bản ghi là sản phẩm của việc cắt ở 25 phút; các task PSG streaming của neuralbench đã dùng random start cho val/test, và EEG-only gần như ngang EEG + elapsed (47.39 so với 47.22).
6. **Xử lý riêng các bản ghi onset sớm.** Phân tích 44 bản ghi onset ≤ 300 s và thử prior hazard theo elapsed cùng mức co mạnh hơn trong những phút đầu. Căn cứ: ở nhóm này hằng số đạt 38.8 so với 63.7-64.6 của baseline, và chúng chiếm 34 % của 10 % bản ghi tệ nhất; EEG gần như không mang thông tin ngoài đồng hồ trong 5 phút đầu (rho 0.039).
7. **Front-end bền theo kênh.** Clip ở khoảng 780 µV, scalar chất lượng theo kênh, cổng kiểu DSF và corruption augmentation bắt chước artifact thật (thay kênh bằng nhiễu, bước nhảy, common-mode). Căn cứ: 16 % cửa sổ có 1-3 kênh hỏng, 33 bản ghi common-mode, 8,022 bước nhảy lớn; DSF +6.6 BACC trên các đêm Muse nhiễu nhất trong khi augmentation đứng một mình có hại (Banville et al. 2022).
8. **Feature chuyển động mắt chậm ở dạng causal.** Band-pass 0.1-1 Hz có trạng thái trên AF7 - AF8, tương quan và tỉ số công suất trên cửa sổ 20-30 s; đo mức tăng AUC cho 40-90 so với 90-300 s. Căn cứ: tương quan AF7-AF8 chạm -0.29 ở -180..-90 s và biến mất sau N2, nhưng mới chỉ đo bằng bộ lọc zero-phase; periodogram 5 s che mất nó.
9. **Pre-train có giám sát trên CinC-2018 (rồi BOAS), fine-tune trên Muse.** Đa nhiệm sleep stage và time-to-N2, kênh ánh xạ theo phần PSG, input đã chuẩn hoá hoặc feature tương đối, channel dropout; so sánh ghép cặp với train từ đầu. Căn cứ: CinC có 2.6 lần số cửa sổ của Muse với phân phối bin tương tự; beta, alpha và sigma/beta cùng dấu và cùng cỡ giữa PSG và Muse còn band power tuyệt đối thì không; fine-tune băng đô +0.045 kappa với 52 đêm. Kỳ vọng khiêm tốn vì Muse không khan nhãn.
10. **Ensemble đa dạng trong không gian phân phối.** Trung bình xác suất bin hoặc hazard rồi lấy weighted median, với thành viên khác nhau về bản chất (trees trên feature, GRU trên feature, CNN trên dạng sóng, model chỉ-thái-dương và chỉ-trán). Căn cứ: hai baseline hiện tại chia sẻ lỗi (Spearman 0.63, 20/50 bản ghi tệ nhất chung) nên trung bình chúng được ít; ensemble thắng model đơn trong 94.9 % trường hợp ở SLEEPYLAND.
11. **Bộ theo dõi causal ở tầng hai.** Một tracker ràng buộc tính đơn điệu (dự đoán giảm 5 s mỗi bước trừ khi bằng chứng mới bất đồng mạnh) hoặc một model tầng hai huấn luyện trên dự đoán out-of-fold. Căn cứ: trong một bản ghi mọi cửa sổ chia sẻ một thời điểm onset ẩn; hậu xử lý theo metric là mức tăng lớn nhất của lời giải Kaggle hạng nhất; chưa có bằng chứng đo trên giấc ngủ.
12. **Thí nghiệm REVE có giới hạn.** Embedding REVE-base theo kênh làm input bổ sung cho cùng GRU và cùng head, đóng băng rồi LoRA; ship `positions.json` và kiểm tra bằng debug submission. Căn cứ: mức lợi 6 % trên Sleep-EDF; foundation model không hơn mạng nhỏ trên cửa sổ ≤ 5 s và ít kênh (Kommineni et al. 2026; Estevan et al. 2025).
13. **Hỏi ban tổ chức năm điều.** Nguồn gốc nhãn N2 của Muse (người chấm hay thuật toán, lưới epoch); scorer sealed có tính theo bản ghi và cohort sealed có bị cắt ở 25 phút hay bắt đầu ngẫu nhiên không; `reve_positions.json` có được cache không; phiên bản neuralbench trên worker; NSRR và MASS có hợp lệ theo Rule 04 không.

Kết luận

Điều thay đổi sau đợt phân tích này là vị trí của nút thắt. Trước đó, cách nghĩ tự nhiên là "cần một encoder EEG tốt hơn"; dữ liệu cho thấy nút thắt nằm ở hai đầu còn lại của pipeline. Ở đầu vào, tín hiệu hữu ích là độ lệch cỡ phút của một bản ghi so với chính nó, bị che bởi khác biệt giữa người và giữa đêm lớn ngang với hiệu ứng cần đo, và bị trộn với một chiếc đồng hồ mà cách cắt bộ dữ liệu công khai làm cho trông thông minh hơn thực tế. Ở đầu ra, metric thưởng cho sự bảo thủ có tính toán: một model biết mình đang ở bin nào và nói ra weighted median của bin đó sẽ đạt khoảng 22 điểm, trong khi các model hiện tại ở mức 47 vì chúng vừa

không dám nói "còn xa" vừa nói "sắp rồi" quá sớm ở những bản ghi onset sớm. Sàn sai số của bin 0-40 s (ramp 90 s, lượng tử hoá 30 s của nhãn, AUC 0.63 quanh mốc 40 s) nghĩa là cuộc đua sẽ không được quyết định ở đó, mà ở khả năng tách đúng 40-90 s, 90-300 s và trần.

Hệ quả cho chiến lược là một thứ tự công việc khác với thói quen: đo nhiễu của CV trước, sửa loss và cách giải mã trước khi sửa kiến trúc, coi trạng thái khởi động của bản ghi như một domain shift hạng nhất, và chỉ sau đó mới chi compute cho pre-training PSG hay foundation model, những thứ mà y văn cho thấy đem lại ít nhất đúng trong chế độ nhiều nhãn, cửa sổ ngắn, ít kênh của track này. Rủi ro lớn nhất chưa giải được nằm ngoài dữ liệu: nếu cohort sealed không bị cắt ở 25 phút hoặc bắt đầu ở điểm khác so với lúc đeo băng đô, mọi thứ model học từ elapsed time và từ artifact đầu bản ghi sẽ phản tác dụng, nên mỗi model nộp nên có một phiên bản đã được chứng minh là bền với random start.

Tài liệu tham khảo

Nguồn nội bộ (thư mục `research_notes/EDA và SOTA sleep onset/`): `eda_muse.md` và các log `scripts/logs/muse_02.log` đến `muse_14b.log`; `eda_psg_datasets_brief.md` với `scripts/psg_out/psg_summary.json` và `cache_cinc2018/summary_hypno.json`; `sleep_onset_prediction_literature.md`; `sleep_staging_sota_and_datasets.md`; `foundation_models_and_neuroai.md`; `preprocessing_features_augmentation.md`; `README.md` của workspace.

Cuộc thi và dữ liệu:

- Competition README, Track 03
- NeuralBench Track 3 guide
- Trang tracks và rules của challenge
- Codabench Track 03
- NEMAR nm000287, Muse Sleep-Onset EEG
- Sleep-EDF Expanded, HMC Sleep Staging, PhysioNet/CinC Challenge 2018
- BOAS, OpenNeuro ds005555; EESM19/EESM23, Sci Data 2025; Dreem Open Datasets, Guillot et al. 2019; CAP Sleep Database

Sinh lý, nhãn và dự đoán sleep onset:

- Ogilvie 2001, Sleep Medicine Reviews
- Gorgoni et al. 2019, Frontiers in Neuroscience
- Prerau et al. 2014, PLoS Computational Biology
- Li et al. 2025, Nature Neuroscience
- Magosso et al., slow eye movements
- Lendner et al. 2020, eLife
- Tóm tắt quy tắc AASM, Sleep Review
- Lee et al. 2022, J Clin Sleep Med; Rosenberg & Van Hout 2013
- Lanthier et al. 2025, Muse-S so với PSG
- MEDi-SOL, Oh et al. 2024

Sleep staging và thiết bị đeo:

- L-SeqSleepNet, Phan et al. 2023; XSleepNet; SleePyCo; SleepTransformer
- SleepDG; SLEEPYLAND; AnySleep; U-Sleep; RobustSleepNet
- Alvarez-Estevez & Rijsman 2021
- Arnal et al. 2020, Dreem; Estevan et al. 2025, SSL trên băng đô; Abou Jaoude et al. 2023, Muse S
- Esparza-Iaizzo et al. 2024, staging thời gian thực; TU/e, transfer sang băng đô một kênh
- Vallat & Walker 2021, YASA

Tiền xử lý, augmentation, domain adaptation và mô hình hoá:

- Banville et al. 2022, Dynamic Spatial Filtering; Miyakoshi 2023, ASR
- Rommel et al. 2022, augmentation; CADDA
- PSDNorm; CMMN; StableSleep

- Kvamme & Borgan 2021; CORAL; Farebrother et al. 2024; Imani & White 2018
- Lời giải hạng nhất, Kaggle Detect Sleep States; Bouthillier et al. 2021

Foundation model:

- Banville et al. 2026, NeuralBench
- REVE, El Ouahidi et al. 2025; REVE Responsible Use License
- EEG-FM-Bench; AdaBrain-Bench; Kommineni et al. 2026
- Stress-test foundation model, arXiv 2607.24519; SleepFM-2