Supplementary Information

An Adaptive Hybrid Deep Learning Framework for Known and Zero-Day Cyberattack Detection in Next-Generation Smart Grids

Electronic Supplementary Material

This file is the electronic supplementary material of the article above, submitted to *Smart Grids and Sustainable Energy*. It specifies the IEEE 14-bus co-simulation benchmark in full, gives the composition of the fused feature vector and of the evaluation sets, extends the proofs of Propositions 2–6 with numerical illustrations, defines every evaluation metric and closes with a reproducibility checklist. Every value is either restated from the main text or computed exactly from parameters stated there; no additional experiments are reported. Numbers prefixed with S (Section S2, Eq. (S1), Table S1, Fig. S1) refer to this file; all other section, equation, table, figure and algorithm numbers refer to the main text.

S1 Reading guide

Each of the five research gaps of Section 3.1 maps to one or more objectives of Section 3.2, to the PG-ADTA-ID components that address it and to specific results in the main text (Table S1). The last column points to the material in this file that extends each row.

S2 Co-simulation benchmark specification

The benchmark is built on the standard IEEE 14-bus case: 14 buses, 15 lines, 5 transformer branches, a slack generator, one further generator, three synchronous condensers and 11 load buses carrying 259.0 MW and 73.5 Mvar (Tables S2 and S3). The network values below were checked against the MATPOWER case14 data as distributed with pandapower (Zimmerman et al. 2011; Thurner et al. 2018). The photovoltaic plants, measurement layers, attacks and drift settings are those of Section 6.1 and Table 2.

Test-system data. Table S2 gives the bus data and Table S3 the branch data; the notes under Table S3 mark the branches whose switching or outage matters in the experiments.

Table S1 Traceability from the research gaps (Section 3.1) and objectives (Section 3.2) to the components of PG-ADTA-ID and the supporting evidence

Research gap	Objective	Component and where it is defined	Evidence in the main text	Detail in this file
G1 Closed-world detection	1, 2	Three-way open-set decision of PC-OAF (Definition 1; Eqs. (41)–(44); Algorithm 4)	Zero-day AUROC 0.934 against at most 0.804 for eleven baselines; 90.7% of zero-day episodes flagged (Tables 6, 8 and 14; Figs. 12 and 13)	S5.3, S5.4, S7
G2 Incomplete use of cyber-physical structure	1	Kirchhoff, Ohm's-law and AMI–SCADA residuals (Eqs. (10)–(14)); breaker-aware graph embedding (Eqs. (17)–(21)); PGST (Algorithm 2)	Known-attack macro-F1-score 98.74%; without physics guidance the seed-0 macro-F1-score falls from 98.33% to 92.69% and the AUROC from 0.931 to 0.736 (Tables 5, 7 and 9)	S2, S3, S5.1
G3 Novelty scoring unaware of operating conditions	2	Error signature on operating-point-independent groups (Eq. (34)); class-conditional reconstruction and physics surprise (Eqs. (37)–(40))	Pooled class statistics lower the AUROC to 0.895; 140 of the 151 seed-0 false positives (93%) occur during rare benign topology changes (Section 7.3; Table 9)	S3, S5.2, S5.3
G4 Unsafe adaptation under concept drift	2	CDCL: physics-gated trust, KS test with effect-size floor, bounded threshold update, replay fine-tuning (Eqs. (45)–(51); Algorithm 5; Proposition 6)	Drift-period FPR falls from 100% (static detector) to 11.01% (full CDCL) (Table 10; Fig. 15)	S6
G5 Evaluation protocols	3	Withheld attack families, benign disturbances, chronological splits, drift period and LOCO tuning without zero-day data (Sections 5.7.2, 6.1 and 6.2)	Event-level detection rates and delays (Table 6); block-bootstrap confidence intervals (Table 14)	S2, S4, S7, S8

Physical layer and benign disturbances. Table S4 summarises the physical-layer models and the benign disturbances that are labelled normal.

Measurement and communication layers. Table S5 lists the four data sources fused by CP-DAP, and Table S6 the 22 communication-traffic features.

Attack episodes and the drift period. Table S7 gives the parameters of the nine attack families.

During the drift period (days 11–12, Section 6.1.4) five changes act together:

- load grows by 18%;
- PV capacity doubles;
- line 9–10 is permanently out of service;
- PMU noise increases by 50%;
- the communication baseline shifts by +30% traffic, four new hosts and +3 ms round-trip time.

Table S2 Bus data of the IEEE 14-bus test system (bus types and loads of the standard base case). Every bus carries a PMU and an RTU (Fig. 9b)

			Base load			
Bus	Bus type	Generation	P (MW)	Q (Mvar)	PV plant	AMI meter
1	Slack	Generator	0.0	0.0	—	No
2	Voltage-controlled	Generator (40 MW)	21.7	12.7	Yes	Yes
3	Voltage-controlled	Synchronous condenser	94.2	19.0	—	Yes
4	Load	—	47.8	−3.9	—	Yes
5	Load	—	7.6	1.6	—	Yes
6	Voltage-controlled	Synchronous condenser	11.2	7.5	—	Yes
7	Load (no demand)	—	0.0	0.0	—	No
8	Voltage-controlled	Synchronous condenser	0.0	0.0	—	No
9	Load	19 Mvar shunt capacitor	29.5	16.6	—	Yes
10	Load	—	9.0	5.8	—	Yes
11	Load	—	3.5	1.8	—	Yes
12	Load	—	6.1	1.6	Yes	Yes
13	Load	—	13.5	5.8	Yes	Yes
14	Load	—	14.9	5.0	—	Yes
Total		5 units	259.0	73.5	3 plants	11 buses

Table S3 Branch data on a 100 MVA base. The magnitude $|y^{\mathrm{ft}}| = 1/(|r + jx|\, a)$, with a the off-nominal tap ratio, is the edge weight $|y_k^{\mathrm{ft}}|$ of Eq. (17)

Branch	Type	r (p.u.)	x (p.u.)	b (p.u.)	Tap ratio a	$\lvert y^{\mathrm{ft}}\rvert$ (p.u.)
1–2[a]	Line	0.01938	0.05917	0.0528	1	16.061
1–5	Line	0.05403	0.22304	0.0492	1	4.357
2–3	Line	0.04699	0.19797	0.0438	1	4.915
2–4	Line	0.05811	0.17632	0.0340	1	5.387
2–5	Line	0.05695	0.17388	0.0346	1	5.465
3–4	Line	0.06701	0.17103	0.0128	1	5.444
4–5[b]	Line	0.01335	0.04211	0	1	22.637
4–7	Transformer	0	0.20912	0	0.978	4.890
4–9	Transformer	0	0.55618	0	0.969	1.855
5–6	Transformer	0	0.25202	0	0.932	4.257
6–11	Line	0.09498	0.19890	0	1	4.537
6–12[c]	Line	0.12291	0.25581	0	1	3.524
6–13	Line	0.06615	0.13027	0	1	6.844
7–8	Transformer	0	0.17615	0	1	5.677
7–9	Transformer	0	0.11001	0	1	9.090
9–10[d]	Line	0.03181	0.08450	0	1	11.076
9–14	Line	0.12711	0.27038	0	1	3.347
10–11	Line	0.08205	0.19207	0	1	4.788
12–13[b]	Line	0.22092	0.19988	0	1	3.357
13–14	Line	0.17093	0.34802	0	1	2.579

[a] Most heavily loaded line; 53-min authorised outage on the test days (Section 7.3).

[b] Protection operations on the test days (Section 7.3).

[c] Outage on the test days (Section 7.6.2).

[d] Permanently out of service in the drift period (Section 6.1.4; Fig. 9a).

Table S4 Physical-layer models of the co-simulation (Sections 6.1 and 6.1.1)

Component	Specification
Power flow	Full Newton–Raphson AC power flow every $\Delta t = 10$ s; 8,640 steps per day; 103,680 steps over 12 simulated days
Loads	Each load bus follows its own mixture of residential, commercial and industrial daily profiles, with day-to-day variation, a weekend reduction, first-order autoregressive fluctuations (10-min correlation, 2% standard deviation) and fast noise
Photovoltaics	Three plants at buses 2, 12 and 13 supplying about 13% of the peak load; clear-sky profile modulated by a two-state Markov cloud process with fast transients
Dispatch	Generators follow the net load through participation factors; the slack bus (bus 1) absorbs the losses
Benign disturbances, labelled normal	Authorised line switching 1–2 times per day for 10–60 min; faults with protection tripping and automatic reclosing 1–2 times per day; 3–6 load steps per day; generator redispatch; nightly data backups that generate heavy traffic

Table S5 Data sources fused by CP-DAP (Sections 5.2 and 6.1.2). Sources are aligned to the 10-s PMU time base by sample-and-hold, and 0.3% of measurement packets are lost at random

Source	Quantities	Reporting interval	Measurement noise	Path to the control centre	Block in Eq. (16)
PMU (14 buses)	Voltage magnitude and angle; system frequency	10 s	Standard deviation 0.002 p.u., 0.002 rad and 2 mHz	Phasor data concentrator	$\mathbf{Z}_g$
RTU (14 buses)	Active and reactive bus injections; branch flows at both ends; generator output; breaker status	20 s	$0.01\|x\| + 0.003$ p.u. (injections, flows); $0.005\|x\| + 0.002$ p.u. (generator output)	Substation gateway	$\mathbf{Z}_g$, $\mathbf{Z}_b$
AMI	Aggregated net demand of each of the 11 load buses	60 s	$0.01\|x\| + 0.001$ p.u.	AMI head-end	$\mathbf{Z}_a$
Communication network	22 flow, protocol and timing statistics (Table S6)	Aligned to the 10-s time base	—	—	$\mathbf{Z}_n$

S3 Composition of the fused feature vector

For the IEEE 14-bus system the fused vector $\mathbf{x}_t$ of Eq. (16) has $d = 280$ channels in seven blocks (Table S8). The 79 residual channels are the real and imaginary parts of 14 nodal Kirchhoff residuals and 20 branch Ohm's-law residuals, plus 11 AMI–SCADA residuals; this count follows from Eqs. (10)–(12) and matches Table 4.

The SA-DAE error signature $\mathbf{r}_t$ of Eq. (34) uses $G = 10$ groups whose channels do not scale with the operating point (Section 5.4.4):

Table S6 The 22 communication-traffic features and the cyber effects of Table 2 that name each statistic. All 22 features are compressed with $\log(1 + x)$ before standardisation (Section 5.2.3)

No.	Feature	Table 2 cyber effects that name this statistic
1	Packet rate	DoS (1.8–8× packet rate)
2	Byte rate	—
3	Mean packet size	—
4	Mean inter-arrival time	—
5	Standard deviation of inter-arrival times	—
6	Number of distinct sources	—
7	Number of distinct destination ports	Malware (ports); reconnaissance (20–200 extra ports)
8	SYN ratio	DoS (SYN); reconnaissance (SYN/RST surge)
9	RST ratio	Reconnaissance (SYN/RST surge)
10	Modbus/DNP3 read commands	Reconnaissance (probing reads)
11	Modbus/DNP3 write commands	Command injection (write burst)
12	Unsolicited responses	—
13	Round-trip time	DoS (RTT)
14	Jitter	—
15	Loss	DoS (loss)
16	Retransmission	DoS (retransmissions)
17	Sequence anomalies	FDIA (sequence noise); replay (duplicate sequence numbers)
18	Timestamp skew	Replay (timestamp skew)
19	Authentication failures	Malware; command injection
20	Port entropy	Malware (entropy)
21	Outbound byte ratio	Malware (outbound ratio)
22	New-connection rate	—

1. breaker context;
2. traffic volume statistics;
3. traffic connection statistics;
4. traffic command statistics;
5. traffic quality statistics;
6. data freshness;
7. Kirchhoff residuals;
8. Ohm's-law residuals;
9. AMI–SCADA residuals;
10. physics summaries.

The five remaining groups hold the grid and AMI blocks, whose values scale with the operating point. They enter only the "All 15 error groups" variant of Table 9.

Two further dimensions used later follow from Table 4. Each bus contributes an 8-dimensional node vector to the graph embedding (Section 5.3.1): standardised voltage magnitude and angle, active and reactive injections, the two Kirchhoff residual components, AMI demand and AMI residual. The PC-OAF evidence vectors have dimensions $m_v = 64$ for $\mathbf{z}_t$ (the last and mean tokens of Eq. (27), each of width $d_m = 32$), 10 for $\mathbf{r}_t$ and 6 for $\mathbf{q}_t$; these are the dimensions used in Section S5.2.

Table S7 Attack-episode parameters (Section 6.1.3 and Table 2). Every episode lasts 4–16 min; episode counts refer to the test days (Table 6)

Family	Onset	Magnitude and targets	Episodes
Known families (all splits)			
DoS	Abrupt	PMU and RTU channels of 3–6 buses lose 30–90% of packets (stale data)	12
FDIA	Ramp over 5–20 steps	Bias on 2–6 random measurements: V 1–5%, θ 0.01–0.06 rad, P/Q and flows 5–30% plus 0.02–0.1 p.u.	12
Replay	Abrupt	PMU, RTU and breaker telemetry replaced by records from 24 h earlier (2–24 h on the first day)	12
Malware	Abrupt	No physical effect; command-and-control beaconing and lateral movement	12
Zero-day families (test and drift days only)			
Stealthy FDIA	Ramp over 5–20 steps	AC-consistent falsified state $\mathbf{z}_a = \mathbf{h}_{\mathrm{pf}}(\boldsymbol{\xi} + \Delta\boldsymbol{\xi})$ at 1–3 load buses ($\Delta\theta$ 0.0015–0.007 rad)	10
Time-sync (GPS)	Angle ramp of 0.0005–0.003 rad per step	PMU phase angles at 1–3 buses, up to 0.02–0.15 rad	10
Command injection	Abrupt	Real unauthorised line trip	10
Reconnaissance	Abrupt	No physical effect; 20–200 extra ports, SYN/RST surge, probing reads	10
AMI tampering	Ramp over 5–20 steps	Metered demand of 1–3 buses scaled by 0.6–0.9	10

Table S8 Blocks of the fused feature vector for the IEEE 14-bus system (Eq. (16); Table 4)

Block in Eq. (16)	Channels	Content	Normalisation
Grid, $\mathbf{Z}_g$	141	PMU voltage magnitudes, angles and frequency; SCADA bus injections, branch flows at both ends and generator outputs	z-score with verified-normal training statistics, clipped to $[-10, 10]$ (Eq. (15))
AMI, $\mathbf{Z}_a$	11	Aggregated net demand of each load bus	As for $\mathbf{Z}_g$
Breaker status, $\mathbf{Z}_b$	20	One binary open indicator per branch	None (binary)
Network, $\mathbf{Z}_n$	22	Traffic statistics of Table S6	$\log(1 + x)$, then as for $\mathbf{Z}_g$
Freshness, $\boldsymbol{\phi}$	1	Fraction of physical channels not imputed at the step (Eq. (9))	As for $\mathbf{Z}_g$
Physics residuals, $\psi(\mathbf{R})$	79	Kirchhoff: 14 buses $\times$ 2 parts = 28; Ohm's law: 20 branches $\times$ 2 parts = 40; AMI–SCADA: 11 load buses	Robust median/MAD standardisation (Section 5.2.2), then signed logarithm (Section 5.2.3)
Physics summaries, $\mathbf{Q}$	6	Mean and maximum of each residual family (Eqs. (13)–(14))	Standardised, as stated under Eq. (16)
Total	280		

S4 Windows, splits and evaluation sets

No window crosses a split boundary. With L steps in a split, window length $T = 20$ and stride s, the number of windows is given by Eq. (S1), which reproduces every window count of Table 3 (Table S9):

$$N = \left\lfloor \frac{L - T}{s} \right\rfloor + 1, \qquad L = 8640\, n_{\text{days}}. \tag{S1}$$

Table S9 Window counts per split ($T = 20$ steps of $\Delta t = 10$ s; a window is labelled by its last step)

Split	Days	Steps L	Stride s	N from Eq. (S1)	Windows in Table 3
Training	1–5	43,200	2	21,591	21,591
Validation	6	8,640	1	8,621	8,621
Calibration	7–8	17,280	1	17,261	17,261
Test	9–10	17,280	1	17,261	17,261
Drift	11–12	17,280	1	17,261	17,261

Of the 17,261 test windows, 1,862 are transition windows: labelled normal, but their 200-s span still overlaps an attack that has just ended. They are excluded from all window-level metrics (Section 6.2), which leaves 15,399 evaluated windows (Table S10).

Table S10 Composition of the test days (days 9–10)

Subset	Windows	Attack episodes
Normal, excluding transition windows	9,521	—
Known attacks: DoS 719, FDIA 762, replay 624, malware 911	3,016	48 (12 per family)
Zero-day attacks: five withheld families	2,862	50 (10 per family)
Transition windows (excluded)	1,862	—
Total	17,261	98

Closed-set metrics use the 12,537 known-class windows (normal plus known attacks; Table 5). The AUROC and AUPR contrast the 2,862 zero-day windows with these 12,537 windows, the FPR uses the 9,521 normal windows, and the open-set macro-F1-score uses all 15,399 windows. The 151 false-positive windows of the seed-0 model correspond to its FPR of $151/9{,}521 = 1.59\%$ (Table 9).

S5 Extended theoretical results

This section extends the proofs of Section 5.9 and evaluates them at the parameter values selected in the main text. It gives a shorter positive-definiteness argument behind Proposition 2, the scale constants of Proposition 3, the decision geometry implied by Proposition 4 and the conformal rank behind Proposition 5.

S5.1 Positive definiteness behind Proposition 2(iii)

Lemma S1 *Let* $\mathbf{B}$ *be the DC nodal susceptance matrix of a connected network with positive branch susceptances* b_{kl}*, and let* $\mathcal{A}$ *be a non-empty proper subset of its buses. Then the principal submatrix* $\mathbf{B}_{\mathcal{A}\mathcal{A}}$ *is symmetric positive definite.*

Proof Take $\mathbf{x}$ on the buses of $\mathcal{A}$ and let $\tilde{\mathbf{x}}$ be $\mathbf{x}$ extended by zeros to all buses. Then

$$\mathbf{x}^{\top}\mathbf{B}_{\mathcal{A}\mathcal{A}}\,\mathbf{x} = \tilde{\mathbf{x}}^{\top}\mathbf{B}\,\tilde{\mathbf{x}} = \sum_{(k,l)} b_{kl}\,(\tilde{x}_k - \tilde{x}_l)^2 \geq 0. \tag{S2}$$

The sum vanishes only if $\tilde{\mathbf{x}}$ takes the same value at both ends of every branch, so $\tilde{\mathbf{x}}$ is constant on the connected network. Because $\tilde{\mathbf{x}}$ is zero on at least one bus outside $\mathcal{A}$, the constant is zero and $\mathbf{x} = \mathbf{0}$. □

The lemma replaces the diagonal-dominance step in the proof of Proposition 2(iii) and yields a quantitative bound on the change of the metered injections:

$$\|\Delta\mathbf{P}_{\mathcal{A}}\|_2 \geq \lambda_{\min}(\mathbf{B}_{\mathcal{A}\mathcal{A}})\,\|\Delta\boldsymbol{\theta}_{\mathcal{A}}\|_2. \tag{S3}$$

Table S11 evaluates the bound for falsified states confined to one load bus of the IEEE 14-bus system, with branch susceptances $b = 1/(x\,a)$ from Table S3 and the stealthy-FDIA angle range of Table S7. For sets of two or three load buses, the smallest eigenvalue over all admissible sets is 5.73 p.u. (buses 13 and 14) and 3.13 p.u. (buses 6, 12 and 13). When every falsified angle is at least 0.0015 rad, the injection changes therefore have a Euclidean norm of at least 1.22 MW and 0.81 MW. Whether such a change is detected depends on the noise of the AMI–SCADA residual, which the robust standardisation of Section 5.2.2 estimates from normal data.

Table S11 Change of the injection at a load bus whose falsified angle alone is shifted (DC approximation, 100 MVA base)

		ΔP_i (MW)		
Load bus i	B_{ii} (p.u.)	$\Delta\theta_i = 0.0015$ rad	$\Delta\theta_i = 0.007$ rad	Base load (MW)
2	33.374	5.01	23.36	21.7
3	10.898	1.63	7.63	94.2
4	42.011	6.30	29.41	47.8
5	38.239	5.74	26.77	7.6
6	20.871	3.13	14.61	11.2
9	26.478	3.97	18.53	29.5
10	17.041	2.56	11.93	9.0
11	10.234	1.54	7.16	3.5
12	8.912	1.34	6.24	6.1
13	15.553	2.33	10.89	13.5
14	6.572	0.99	4.60	14.9

S5.2 Scale constants of the PC-OAF indicators (Proposition 3)

With the evidence dimensions of Section S3, the scales of Proposition 3 are $\kappa_z = 134.48$ for the latent term of u_t, $\kappa_r = 33.48$ for a_t and $\kappa_q = 24.25$ for c_t (Table S12). The common scale holds at the anchor $s = 1/2$, where every indicator exceeds s with probability exactly 0.01 under Gaussian class statistics. Away from the anchor the tails depend on the dimension, as Eq. (56) predicts.

Table S12 Scale constants $\kappa = \chi^2_{0.99}(m_v)/\ln 2$ and tail probabilities $\Pr(I > s)$ of Eq. (56) under the idealised Gaussian assumption

Indicator (scale; definition)	Evidence vector	m_v	$\chi^2_{0.99}(m_v)$	κ	$\Pr(I > 0.25)$	$\Pr(I > 0.50)$	$\Pr(I > 0.75)$
Latent term of u_t (κ_z; Eq. (35))	$\mathbf{z}_t$	64	93.217	134.484	0.995	0.010	6.7×10^{-14}
Reconstruction surprise a_t (κ_r; Eq. (37))	$\mathbf{r}_t$	10	23.209	33.484	0.473	0.010	1.2×10^{-6}
Physics surprise c_t (κ_q; Eq. (39))	$\mathbf{q}_t$	6	16.812	24.254	0.323	0.010	8.0×10^{-6}

Shrinkage, non-Gaussian tails and, for $\mathbf{z}_t$, the layer normalisation of Eq. (27), which removes at least one degree of freedom, make these probabilities approximate; this is why the threshold of Eq. (43) is calibrated empirically.

S5.3 Decision geometry of the fused score (Proposition 4)

With the selected weights and the seed-0 threshold $\tau_0 = 0.609$, a window is flagged as a suspected zero-day attack when $0.23u_t + 0.63a_t + 0.14c_t > 0.609$ (data freshness $\bar{\phi}_t = 1$). Fig. S1 draws this boundary in the plane of reconstruction and physics surprise.

Corollary S1 (Freshness modulation) *For every $\bar{\phi}_t \in [0, 1]$, the weights of Eq. (42) satisfy $\alpha_t \le 0.23/0.86 \approx 0.267$ and $\gamma_t \le 0.140$, both below τ_0. Hence only the reconstruction surprise can raise a zero-day decision on its own, once $a_t > \tau_0/\beta_t$. No window with a_t below 0.379 ($\bar{\phi}_t = 1$) or 0.466 ($\bar{\phi}_t = 0$) can be flagged, whatever its u_t and c_t (Table S13).*

Proof By Eq. (42), $\alpha_t = \alpha/(\alpha+\beta+\gamma\bar{\phi}_t)$ is largest at $\bar{\phi}_t = 0$, where it equals $0.23/0.86 \approx 0.267$, and γ_t is largest at $\bar{\phi}_t = 1$, where it equals 0.140. The trigger condition is Proposition 4. Since $u_t, c_t < 1$, $S_t < \alpha_t + \beta_t a_t + \gamma_t$, so $S_t > \tau_0$ requires $a_t > (\tau_0 - \alpha_t - \gamma_t)/\beta_t$. □

With $\bar{\phi}_t = 1$ the score is linear in the indicators, so the mean score of a family equals the score of its mean indicators. Table S14 applies this to the three families whose mean indicators are stated in Section 7.10. For AMI tampering, even the largest possible physics surprise ($c_t = 1$) gives $S_t = 0.515 < \tau_0$ at the family's mean u_t and

Table S13 Effective fusion weights and trigger levels as functions of the window freshness ($\tau_0 = 0.609$)

Mean freshness $\bar{\phi}_t$	α_t	β_t	γ_t	Single-indicator trigger τ_0/β_t	Smallest a_t that can be flagged
1.00	0.230	0.630	0.140	0.967	0.379
0.75	0.238	0.653	0.109	0.933	0.401
0.50	0.247	0.677	0.075	0.899	0.423
0.25	0.257	0.704	0.039	0.865	0.445
0.00	0.267	0.733	0.000	0.831	0.466

a_t, so physics evidence alone cannot flag such a window. It needs $a_t > 0.719$ at the mean u_t and c_t, or an unusually large uncertainty ($u_t > 0.916$ at the mean a_t and c_t), which is why most AMI-tampering windows stay below the threshold (Section 7.10).

Table S14 Fused score of the mean indicator vectors of Section 7.10 ($\bar{\phi}_t = 1$, $\tau_0 = 0.609$)

Family	u_t	a_t	c_t	S_t	$S_t - \tau_0$	a_t needed for $S_t = \tau_0$
Normal	0.19	0.24	0.17	0.219	−0.390	0.860
AMI tampering	0.29	0.49	0.64	0.465	−0.144	0.719
Reconnaissance	0.73	0.93	0.14	0.773	+0.164	0.669

S5.4 Rank of the split-conformal threshold (Proposition 5)

For n calibration scores, the split-conformal threshold is the k-th smallest score, and its exact false-alarm level follows from the rank argument of Proposition 5:

$$k = \lceil (n+1)(1-\delta) \rceil, \qquad \Pr(S_{n+1} > \tau_0) \le \frac{n+1-k}{n+1} \le \delta. \tag{S4}$$

Table S15 Rank of the split-conformal threshold for $n = 17{,}261$, the number of calibration windows in Table 3

Budget δ	Rank k	Calibration scores above the threshold	Exact level $(n+1-k)/(n+1)$
0.0035	17,202	59	0.003476
0.005	17,176	85	0.004982
0.01	17,090	171	0.009964
0.02	16,917	344	0.019986
0.03	16,745	516	0.029950

The exact level stays below δ in every row. Each rank changes the level by $1/(n+1) = 5.8 \times 10^{-5}$, so an empirical quantile one rank lower changes it by less

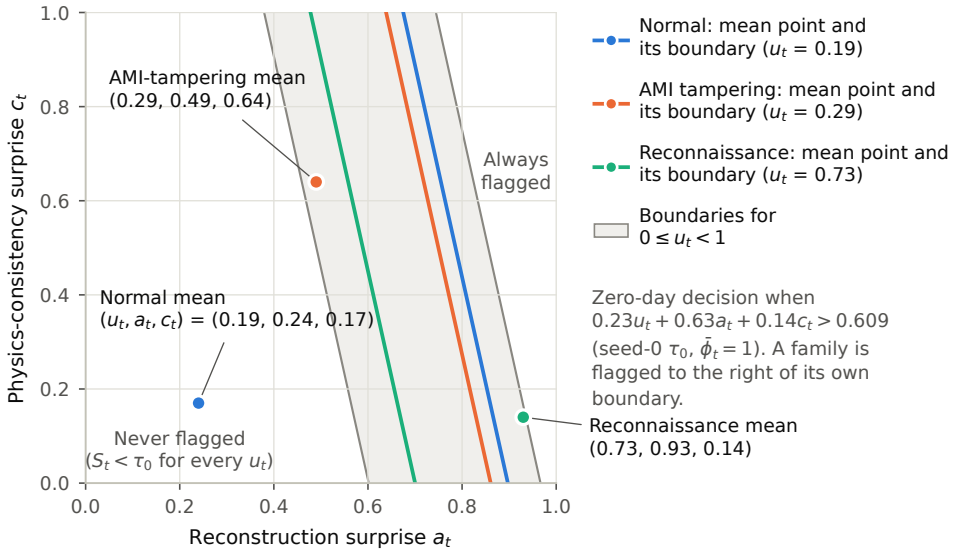

Fig. S1 Zero-day decision geometry of PC-OAF in the plane of reconstruction surprise a_t and physics surprise c_t (seed-0 $\tau_0 = 0.609$, $\bar{\phi}_t = 1$). Each coloured line is the boundary $S_t = \tau_0$ at the mean uncertainty of one family, and its marker is that family's mean indicator vector (Section 7.10). The shaded band spans the boundaries for $0 \le u_t < 1$: windows to its left are never flagged, windows to its right always are

than 10^{-4} (Section 5.9.2); for example, rank 17,201 at $\delta = 0.0035$ has level 0.003534. The guarantee still requires PC-OAF statistics fitted on other data and test windows exchangeable with the calibration windows (Section 7.8.2).

S6 Safeguards of the continual-learning stage

At the CDCL settings of Table 4 and the effect-size floor $D_{\min} = 0.58$ of Section 5.9.3, an adversary needs more than 40% of the trusted buffer to move the threshold when the clean scores lie at a KS distance of 0.30 from the reference. The floor is 3.7–6.8 times the KS critical value at the buffer sizes used, and each update of Eq. (47) moves the threshold by at most 0.05 (Fig. S2).

S6.1 Poisoning barrier (Proposition 6(ii) and (iii))

By Eq. (58), drift, and hence any movement of the threshold, requires an adversarial fraction $\nu > (D_{\min} - D_{\text{clean}})/(1 - D_{\text{clean}})$ of the trusted buffer. With $D_{\min} = 0.58$ and a 360-window buffer, this means at least 209 windows when the clean scores match the reference ($D_{\text{clean}} = 0$), more than 144 at $D_{\text{clean}} = 0.30$ and at least 58 at $D_{\text{clean}} = 0.50$ (Fig. S2a). Every adversarial window must also pass the trust gate of Eq. (45).

Proposition 6(iii) bounds the effect of fewer adversarial scores. With $\delta = 0.0035$, the $(1-\delta)$ quantile of a 360-window buffer sits near its second-largest score ($\delta \times 360 = 1.26$). A single adversarial score ($\nu = 1/360 < \delta$) can only move that quantile between the clean-score quantile levels 0.99649 and 0.99928.

S6.2 The effect-size floor decides the drift test

The asymptotic critical value of the two-sample Kolmogorov–Smirnov test for a buffer of n scores and a reference of m scores is (Smirnov 1948)

$$D_{\text{crit}} = c(\alpha_{\text{KS}})\sqrt{\frac{n+m}{n\,m}}, \qquad c(\alpha) = \sqrt{-\tfrac{1}{2}\ln(\alpha/2)}. \tag{S5}$$

At $\alpha_{\text{KS}} = 0.01$, $c = 1.628$. For a reference much larger than the buffer, D_{crit} is 0.149 at $n = B_{\text{min}} = 120$, 0.121 at $n = B_{\text{ft}} = 180$ and 0.086 at $n = 360$; with $m = 10n$ it is 4.9% larger. The floor $D_{\text{min}} = 0.58$ therefore exceeds the critical value by a factor of 3.7–6.8 for any reference of at least $10n$ scores (Fig. S2b), and a distance of 0.58 has an asymptotic p-value below 10^{-31} even at $n = 120$ and $m = 10n$. Once the buffer holds 120 windows, drift is decided by the floor rather than by the p-value, which is the behaviour Section 5.6.2 requires. These p-values assume independent scores; consecutive windows overlap and are positively correlated, which makes the test fire more readily and strengthens the case for the floor.

S6.3 Bounded threshold response (Eq. (47))

By Proposition 6(i), m updates of Eq. (47) move τ_t by at most $m\Delta_\tau$; the re-anchoring that follows a fine-tuning round (Algorithm 5, line 16) is not bounded in this way. Raising τ_t from 0.609 to 0.823, the level reached by the threshold-only mode of Section 7.6.2, which uses Eq. (47) alone, therefore takes at least five declared blocks (50 min), and reaching 1.0 takes at least eight (80 min). Fig. S2c compares Eq. (47) with the same update without the clip when the buffer quantile jumps from 0.609 to 0.90. The clip limits the first three updates to 0.05, where the unclipped rule would propose 0.087, 0.072 and 0.057 from the same thresholds; afterwards both forms approach the new quantile geometrically with ratio $1 - \rho = 0.7$.

S7 Evaluation metrics and statistical procedures

On the test days, window-level metrics are computed on the 15,399 non-transition windows of Table S10 (Tables 5–9, Table 14 and the stationary rows of Table 10). The definitions below apply wherever these metrics are reported. Macro-averaged precision, recall and F1-score follow Eq. (52), and the zero-day detection rate (ZDR) and false-positive rate (FPR) follow Eq. (53).

S7.1 Closed-set metrics

Closed-set metrics compare the PGST class decision with the true class over the 12,537 known-class windows, before the open-set decision of Eq. (44). With N windows, N_{c} of them correct, t_k windows of true class k and p_k windows assigned to class k, the multiclass Matthews correlation coefficient (Gorodkin 2004) and Cohen's κ (Cohen

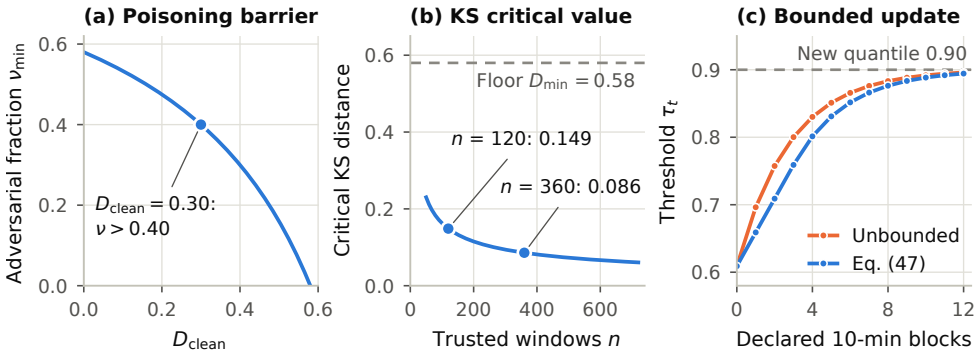

Fig. S2 Safeguards of CDCL at the settings of Table 4 and Section 5.9.3. (a) Fraction $\nu_{\min}$ of the trusted buffer an adversary must control before drift can be declared, as a function of the KS distance D_{clean} of the clean scores (Eq. (58), $D_{\min} = 0.58$). (b) Asymptotic two-sample KS critical value at $\alpha_{\text{KS}} = 0.01$ for a reference much larger than the buffer, against the effect-size floor. (c) Threshold after a jump of the buffer quantile from 0.609 to 0.90: the bounded update of Eq. (47) ($\rho = 0.3$, $\Delta_\tau = 0.05$) and the same update without the clip

1960) are

$$\text{MCC} = \frac{N_c N - \sum_k p_k t_k}{\sqrt{(N^2 - \sum_k p_k^2)\,(N^2 - \sum_k t_k^2)}}, \qquad \kappa = \frac{p_o - p_e}{1 - p_e}, \qquad p_o = \frac{N_c}{N}, \qquad p_e = \frac{\sum_k p_k t_k}{N^2}. \tag{S6}$$

Accuracy is N_c/N.

S7.2 Open-set metrics

- **Zero-day alarm rate:** fraction of zero-day windows that raise any alert, that is, any decision other than normal.
- **Open-set macro-F1-score:** Eq. (52) over $C+1 = 6$ classes, with the five withheld families merged into one unknown class; ZD and UNC decisions count as predictions of the unknown class, as for TP_Z in Eq. (53).
- **AUROC and AUPR:** zero-day windows are the positive class and the 12,537 known-class windows the negative class. Windows are ranked by S_t for PG-ADTA-ID, by one minus the maximum class probability (temperature-scaled for the neural networks) for the supervised baselines and by the anomaly score for the unsupervised baselines.

The AUROC equals the probability that a randomly drawn zero-day window outranks a randomly drawn known-class window, with ties counted as one half. With $\mathcal{W}_Z$ and $\mathcal{W}_K$ the sets of zero-day and known-class test windows,

$$\text{AUROC} = \frac{1}{|\mathcal{W}_Z|\,|\mathcal{W}_K|} \sum_{i \in \mathcal{W}_Z} \sum_{j \in \mathcal{W}_K} \left(\mathbb{I}[S_i > S_j] + \tfrac{1}{2}\,\mathbb{I}[S_i = S_j] \right). \tag{S7}$$

A random ranking has an expected AUROC of 0.5 and an expected AUPR equal to the share of positives, 2,862/15,399 = 0.186; these are the chance levels for the AUROC and AUPR columns of Table 8.

S7.3 Event-level metrics

- **Episode detection:** an attack episode is detected when at least one window ending inside it raises the correct alert, that is, the correct class label for a known family and a ZD or UNC decision for a zero-day family.
- **Detection delay:** time from the onset of the episode to that first correct alert, reported as a median; Table 6 pools the delays of the three seeds.
- **False-alarm events:** clusters of consecutive false alerts on normal windows, reported per day.

S7.4 Statistical procedures

- **Seed variability:** mean ± population standard deviation over three training seeds ($n = 3$; Section 6.2).
- **Paired block bootstrap (Table 14):** the 15,399 windows are cut into 43 non-overlapping blocks of consecutive windows, about one hour (360 windows) each. Each of 2,000 replicates resamples whole blocks with replacement, with the same blocks for PG-ADTA-ID and the baseline, and records the difference Δ of the metric. The 95% percentile interval of Δ is reported, and a difference is significant when the interval excludes zero (Carlstein 1986).
- **Welch's t-test on per-seed metrics** (Welch 1947), with sample means $\bar{x}_1, \bar{x}_2$, sample variances s_1^2, s_2^2 and $n_1 = n_2 = 3$ seeds:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}, \qquad \mathrm{df} = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}. \tag{S8}$$

S8 Reproducibility checklist

Table S16 lists the main experimental settings and where each is reported.

Table S16 Reproducibility checklist

Item	Setting	Where reported
Test systems	IEEE 14-bus (main experiments); IEEE 118-bus, 8 simulated days (scalability)	Sections 6.1 and 7.9.2; Tables S2–S3
Simulation	Newton–Raphson AC power flow every 10 s; 12 days; 103,680 steps	Section 6.1; Table S4
Measurement and communication	PMU every 10 s, RTU every 20 s, AMI every 60 s; 22 traffic features; 0.3% random packet loss	Section 6.1.2; Tables S5–S6
Attack catalogue	Four known and five withheld families; episodes of 4–16 min	Section 6.1.3; Table 2; Table S7
Splits	Chronological: training days 1–5, validation 6, calibration 7–8, test 9–10, drift 11–12	Table 3; Table S9
Windows	$T = 20$ steps (200 s); stride 2 in training, 1 elsewhere; labelled by the last step; 1,862 transition windows excluded from the test metrics	Sections 6.1.4 and 6.2; Table S10
Features	$d = 280$ (14-bus), $d = 2{,}291$ (118-bus)	Table 4; Section 7.9.2; Table S8
Hyperparameters	Network, fusion and CDCL settings; trust limit, fine-tuning set and effect-size floor	Table 4; Sections 5.6.1, 5.6.4 and 5.9.3
Network tuning (Stage A)	Bayesian optimisation with a TPE surrogate; 14 PGST configurations on the validation day	Sections 5.7.2 and 7.7.1
Decision tuning (Stage B)	LOCO pseudo-open-set; BO, PSO and GA with 120 evaluations × 5 runs each; no zero-day data	Section 7.7.2; Table 11
Seeds	Three for PG-ADTA-ID and six baselines; one for the other five baselines; single-model analyses on seed 0	Section 6.2
Hardware and software	PyTorch on a two-core CPU with 7 GB RAM and no GPU	Section 6.3
Computational cost	Training 403 s; 7.23 ms per complete decision; 3,134 windows/s in batches of 256	Table 12
Statistics	Paired block bootstrap with 2,000 replicates; Welch's t-test on per-seed metrics	Section 7.11; Section S7.4
Code and data	Co-simulation generator, CP-DAP, PG-ADTA-ID implementation, evaluation scripts and simulated datasets: available from the corresponding author on reasonable request	Statements and Declarations

References

Carlstein E (1986) The use of subseries values for estimating the variance of a general statistic from a stationary sequence. Ann Stat 14(3):1171–1179. https://doi.org/10.1214/aos/1176350057

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20:37–46

Gorodkin J (2004) Comparing two K-category assignments by a K-category correlation coefficient. Comput Biol Chem 28(5):367–374

Smirnov N (1948) Table for estimating the goodness of fit of empirical distributions. Ann Math Stat 19(2):279–281. https://doi.org/10.1214/aoms/1177730256

Thurner L, Scheidler A, Schäfer F, et al (2018) pandapower—an open source Python tool for convenient modeling, analysis and optimization of electric power systems. IEEE Trans Power Syst 33(6):6510–6521. https://doi.org/10.1109/TPWRS.2018.2829021

Welch BL (1947) The generalization of "Student's" problem when several different population variances are involved. Biometrika 34(1–2):28–35. https://doi.org/10.1093/biomet/34.1-2.28

Zimmerman RD, Murillo-Sánchez CE, Thomas RJ (2011) MATPOWER: steady-state operations, planning, and analysis tools for power systems research and education. IEEE Trans Power Syst 26(1):12–19. https://doi.org/10.1109/TPWRS.2010.2051168