Noise Covariance as the Geometry of Signal Reconstruction

…and how a reconstruction can tell when its own noise model stops holding

---

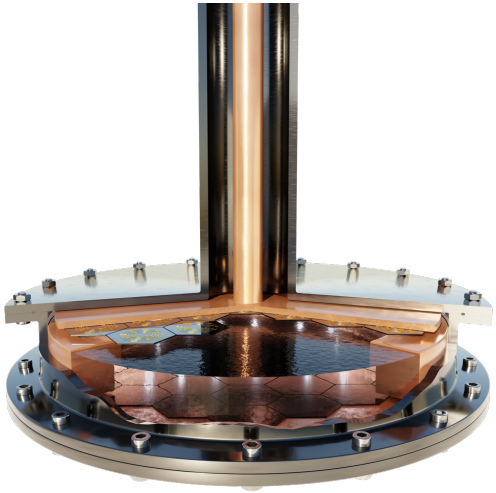
Dowling Wong

57. Herbstschule für Hochenergiephysik · Bad Honnef, September 2026

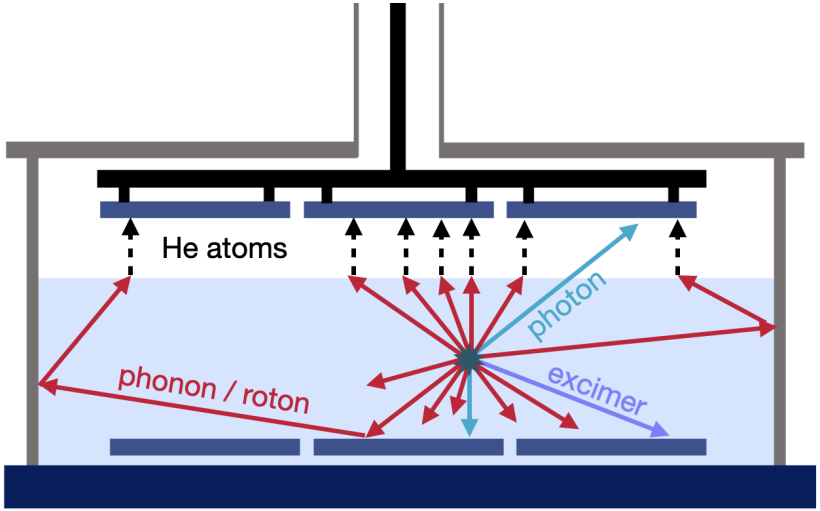
KIT ETP — Institut für Experimentelle Teilchenphysik,
Karlsruher Institut für Technologie

One story, start to finish. If a technical detail does not land, let it go — I would rather you leave with the story than with the algebra

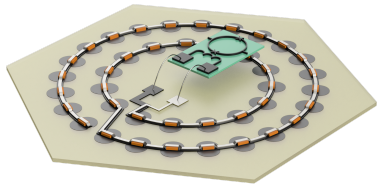
A low-energy event lives or dies in reconstruction



DELight cryostat



quasiparticles, photons and excimers reach the calorimeter array



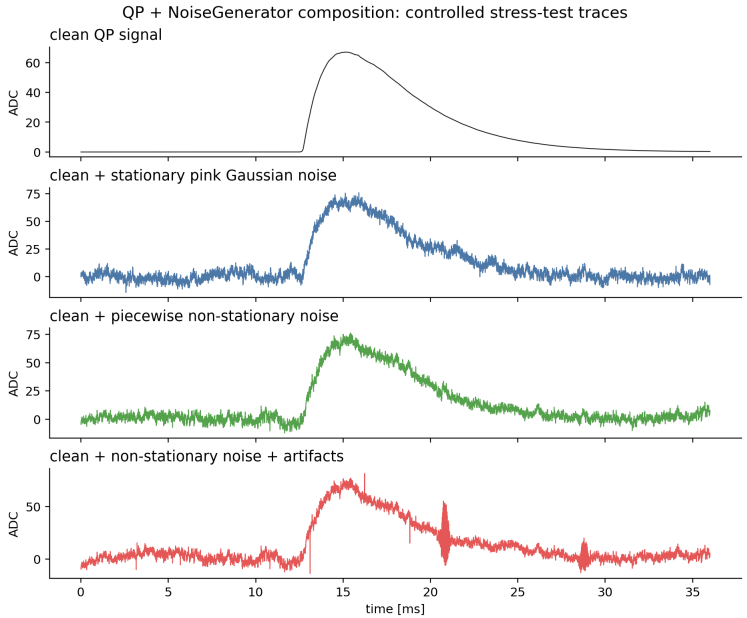
56 MMC wafer calorimeters

DELight searches for light dark matter in superfluid helium, read out by 56 MMC calorimeters.

At the lowest masses the signal is a **quasiparticle pulse**: its energy, position, timing and shape are all carried by the waveform — the model must reconstruct the pulse — and know when its assumptions fail.

Results in this talk come from simulation and public datasets, not from DELight data.

The obstacle is structured detector noise



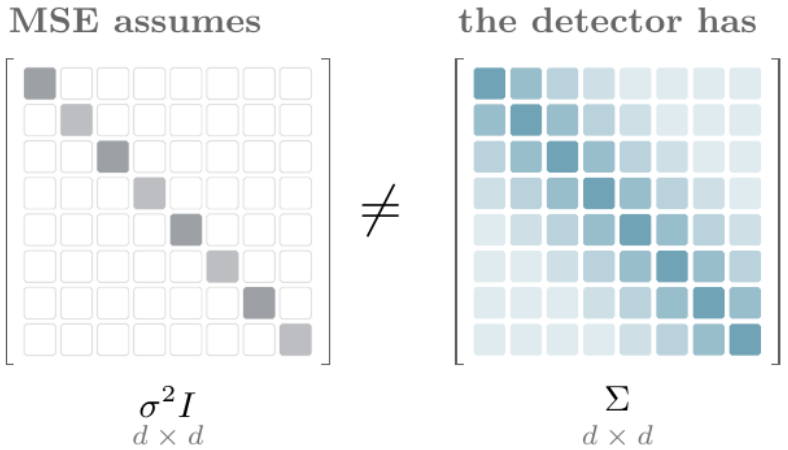
the same pulse under white, colored, non-stationary noise and artifacts

You already accept this when you fit with a full covariance matrix instead of bin-by-bin errors. The talk is what happens when a learned model has to do the same.

x = s(z) + n, n ~ N(0, Σ)

s: the pulse you want. n: everything the instrument adds — and it is almost never white:

colored (1/f, resonances), correlated across channels, drifting with the run.



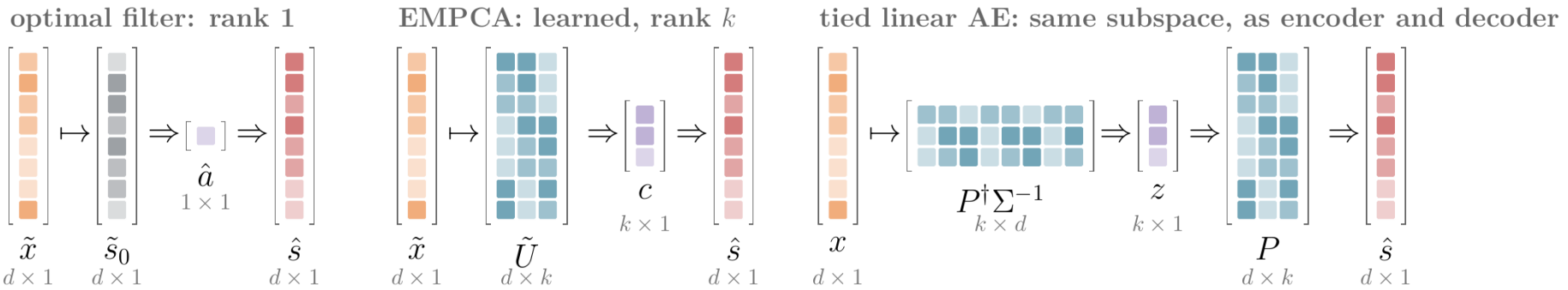
Some directions of x are **well measured**; others are mostly noise. A reconstruction has to know the difference.

Whitening reveals one shared projection

optimal filter: Gatti & Manfredi (1986) · EMPCA: Bailey (2012) · new here: the (Σ^{-1}, F) frame and the tying condition

$$\chi^2(\hat{s}) = (x - \hat{s})^T \Sigma^{-1} (x - \hat{s}) \quad \Leftrightarrow \quad \|\tilde{x} - \tilde{s}\|^2, \quad \tilde{x} = \Sigma^{-1/2} x$$

Σ^{-1} — and nothing else — decides which residuals are expensive. Whitening turns the weighted problem into an ordinary Euclidean projection. Reconstruction is not make the residual small; it is **make it small in the right geometry**.



Same projection, same objective. **Only the admissible column space of P changes.**

fixed template at known arrival time → optimal filter · learned subspace → EMPCA · tied autoencoder: encoder = Σ^{-1}-adjoint of the decoder ($P^T\Sigma^{-1}P = I_k$) — principal-angle cosine > 0.9999 at every rank tested.

The autoencoder adds no accuracy here, and that is the point: tying is what carries a measured metric into a trainable model. Untie it, or train it under MSE, and the subspace is no longer the likelihood's.

Gaussian noise, linear models, matched preprocessing. Non-Gaussian transients — glitches, pileup — are outside the statement.

EMPCA learns the signal in that geometry

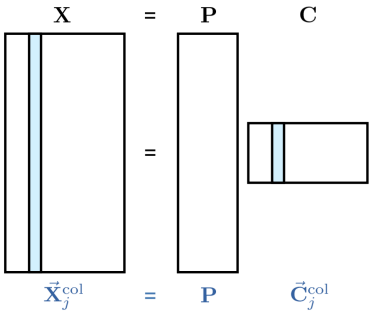
PCA's optimality assumes isotropic noise. Detector noise is **colored** and **correlated across channels**.

Same projection, right metric:

- fixed template → optimal filter
- learned subspace → EMPCA

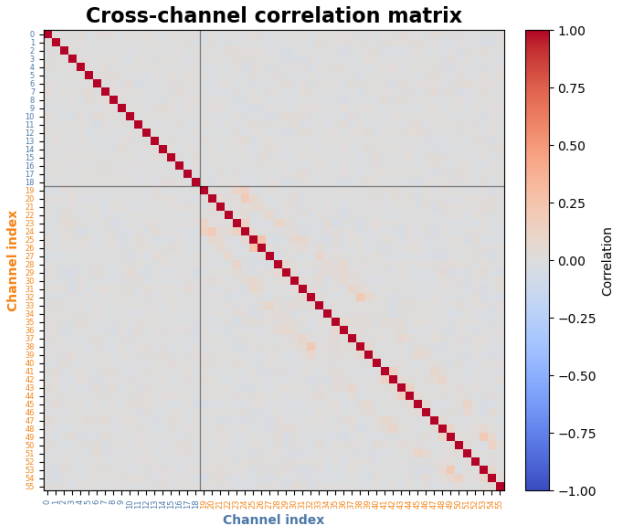
$$\min_{P,C} \sum_i (x_i - Pc_i)^\top \Sigma^{-1} (x_i - Pc_i)$$

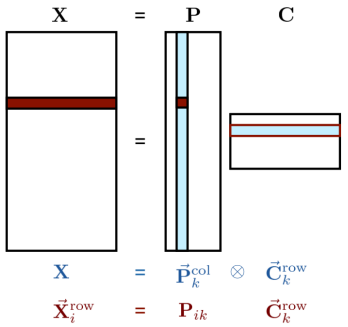
Σ^{-1} = inverse detector covariance



E-step: solve the coefficients

M-step: refine the basis





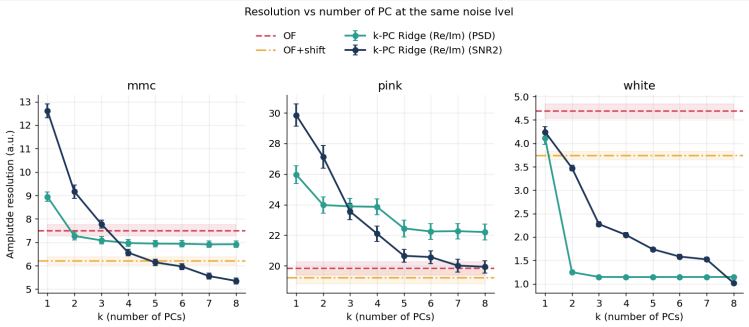
MSE ranks residuals by variance; Σ^{-1} ranks them by information — the same residual spectrum can mean different reconstruction quality.

Only the signal class changes across models

multilinear PCA: Lu et al. (2008) · GLS matrix decompositions: Allen & Tibshirani (2014) · matrix-normal PCA: Zhang et al. (2019) — no new factorization is claimed

Richer signal manifold

→ better resolution



resolution vs number of learned modes, at one noise level

a fixed template is a single point on this curve

exploratory — outside the paper's frozen evidence bundle

Channel × time structure — NFPA

a smaller class, the same metric

$$\widehat{X} = \sum_{i=1}^{k} c_i \left(a_i b_i^T \right)$$

each event is a matrix Z: channel × time — a_i is a channel mode, b_i a time mode

separable noise, Σ = Σ_t ⊗ Σ_c (t = time, c = channel): the whitening factors across the axes

the projector is their Kronecker product — EMPCA restricted to a product of Grassmannians

a full-rank factor does not give EMPCA back: the restriction always binds

62.9° to the likelihood-aligned subspace vs 86.7° isotropic — closer, not aligned: the metric places it, not the prior

The metric is fixed by the likelihood; only the admissible class changes: template → subspace → factored → nonlinear.

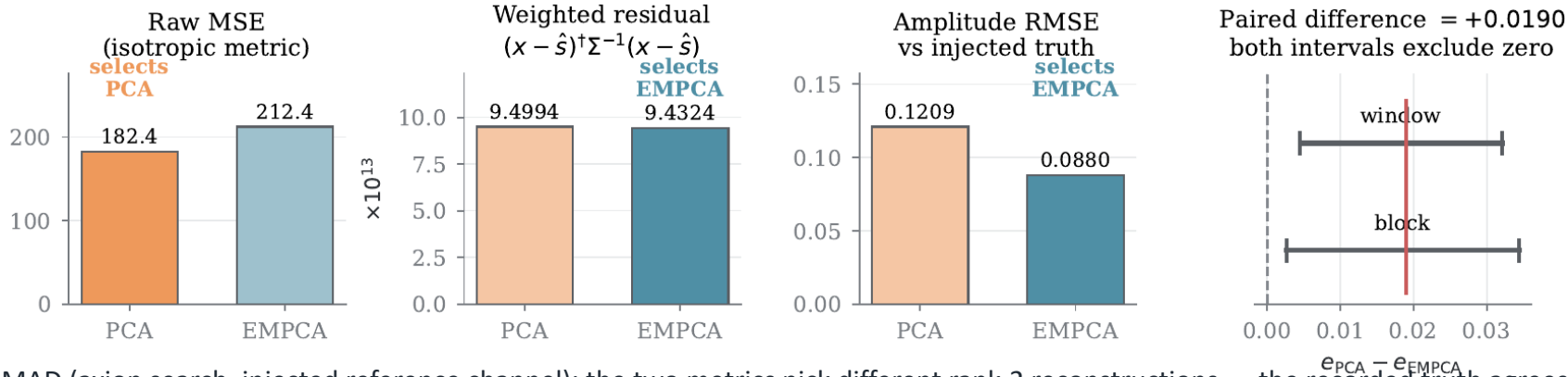
The metric can reverse which reconstruction wins

Isotropic MSE is the Gaussian likelihood **only** when $\Sigma = \sigma^2 I$. Under colored noise it follows raw *variance* instead of *information*: two reconstructions of equal rank can satisfy

MSE(A) < MSE(B) while **$\chi^2(A) > \chi^2(B)$**

Each is optimal — in a different geometry. I call this **metric reversal**.

Read the panels left to right: the two metrics disagree about the winner, and the third is the referee — it is the only one that uses recorded truth.



TIDMAD (axion search, injected reference channel): the two metrics pick different rank-2 reconstructions — the recorded truth agrees with the weighted one (27 % lower amplitude RMSE; 150 held-out windows, both bootstrap intervals exclude zero).

One file, one interval. On the same release the analytic rank-2 prediction was not confirmed in three attempts, and this linear estimator loses the release's own denoising benchmark.

Choosing MSE is not a neutral default. It is an assertion that your detector's noise is white.

One recipe transfers across detectors

How to apply

1. **Measure the noise**
   noise-only data → $\hat{\Sigma}$, whose error propagates
2. **Whiten**
   apply $\Sigma^{-1/2}$
3. **Choose the signal class**
   template · subspace · channel × time
4. **Reconstruct**
   project after whitening
5. **Check**
   held-out noise; recorded truth

DELight simulation

controlled traces

noise and truth both known

CRESST

cryogenic baselines

a noise model per channel, within-channel only

GW150914 strain

off-source strain

PSD supplies the metric

TIDMAD (axion search)

reference channel

recorded truth to check against

Requirements, not architecture: measurable noise · a declared signal class · an independent check.

Three ways a frozen model fails

proposed framing

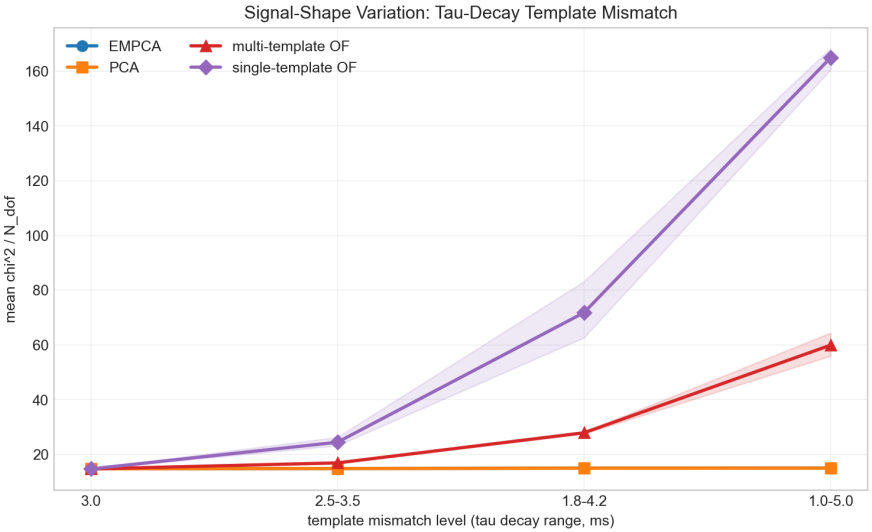
contract	· what moved	
N	**acquisition / noise**	The run conditions drifted away from the calibration: covariance change, jitter, lines, dead channels, module loss. The measurement broke its declared acquisition condition.
S	**training support**	A corner of phase space the training never populated: clean, physically valid events the model has simply never seen. Nothing is broken.
E	**evaluation pipeline**	A bug in the analysis chain: wrong units, coordinates, PSD convention, metric. Caught by deterministic gates, not by a class.

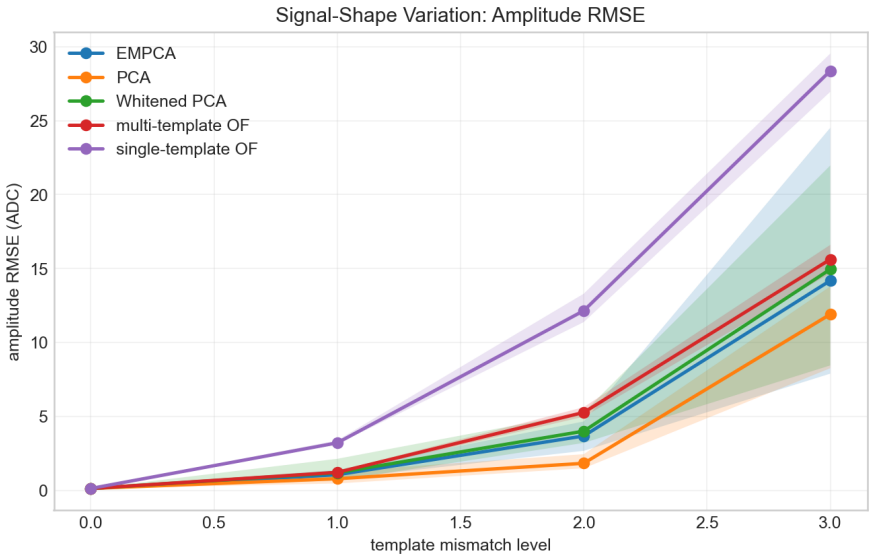
Both N and S move the input distribution, so generic shift detection cannot separate them.

Operationally the difference is everything: rejecting S throws away good physics; rejecting N removes measurements that broke the acquisition contract.

Mechanism: N changes the residual within the span the training excited; S moves the representation along the complement, where the model carries no trained sensitivity.

But no metric can replace missing signal coverage





τ-decay template mismatch: a single template fails first; learned modes survive as far as the calibration data reach

The likelihood answers **how residuals should be weighted**. It does not decide **coverage**: did the training data excite the latent factors the physics will use?

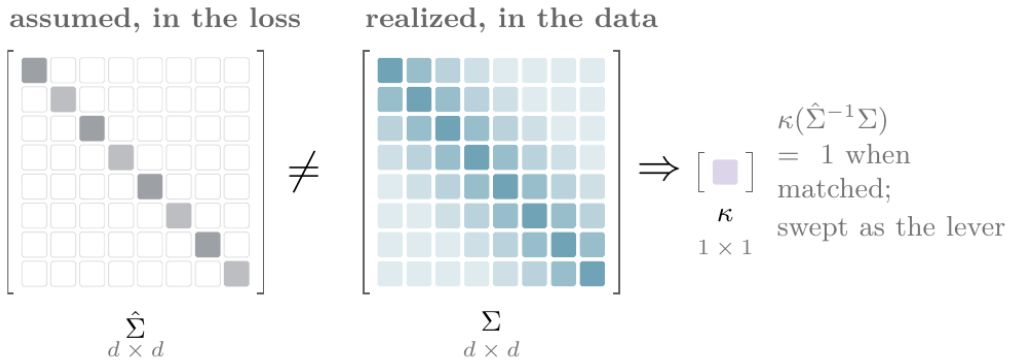
More calibration data → better manifold approximation. **Coverage is a data problem, not a loss problem** — and it is the first way a deployed model fails.

At deployment, the metric is a testable belief

pivot to proposed work

training: $\min \|x - \hat{s}\|^2_{\hat{\Sigma}^{-1}}$ deployment: $n \sim N(0, \Sigma)$, $\Sigma \neq \hat{\Sigma}$

the weighted objective makes the assumed covariance a literal property of the loss



Detectors drift, channels die, lines appear. So the question for a *deployed* model becomes:

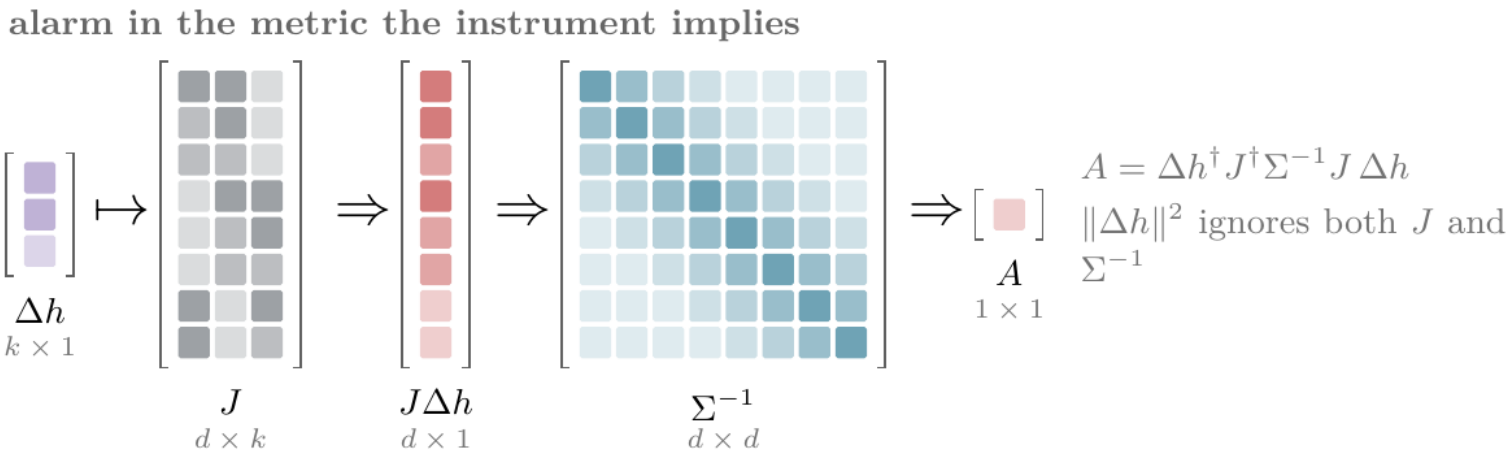
Can a frozen model notice that $\hat{\Sigma} \neq \Sigma$ — and say whether it matters?

A useful alarm measures consequence, not movement

proposed method

In plain terms: a frozen network's internal vector moved — only the part of that motion which can change the physics output should raise an alarm.

Frozen model with representation h and output y = g(h); local Jacobian J = ∂y/∂h. A shift moves the representation by Δh.
Which number should raise the alarm?



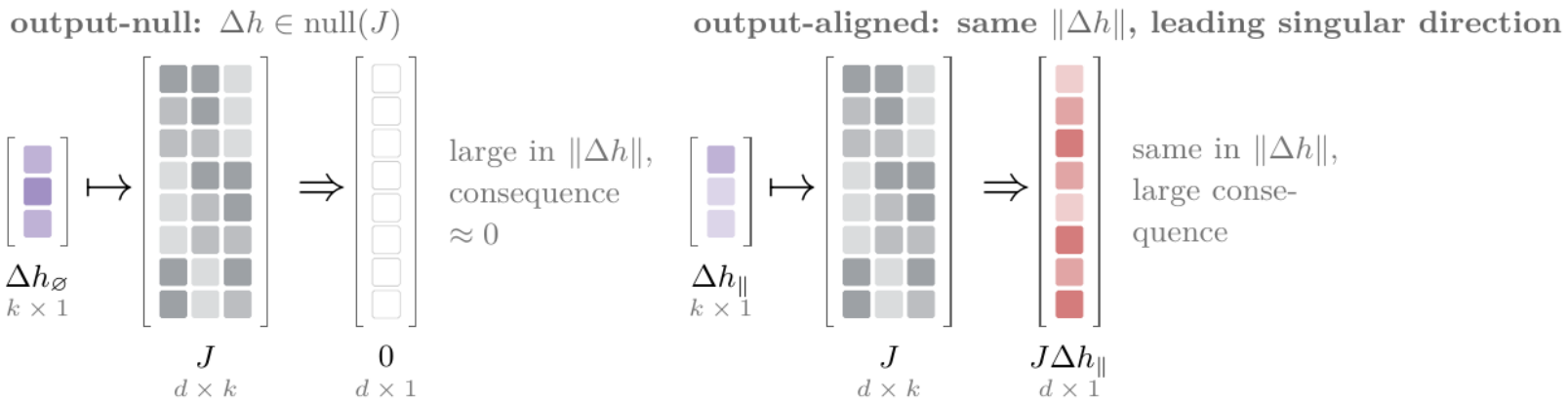
Displacement pulled through J and measured in the output-space precision Σ_y⁻¹: the same construction, one level up.
$J^T\Sigma\_y^{-1}J$ is positive semi-definite — the null space is the point, not a defect. The usual monitor, $\|\Delta h\|^2$, uses neither J nor Σ_y.

Designed test: push h along the null space of J (large ||Δh||, zero consequence) and the same distance along its leading direction (large consequence) — any monitor that ranks by ||Δh|| gets this pair backwards.

…and a test that magnitude-only monitors cannot pass

designed test

Alarm and consequence are both monotone in severity, so a pooled correlation proves nothing. Two perturbation families decouple them by construction:



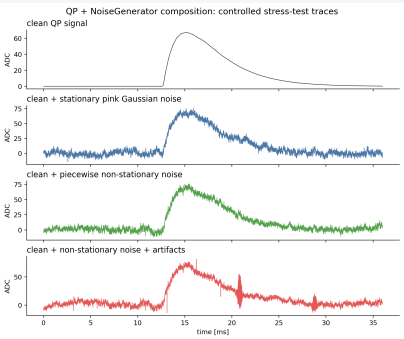
Any monitor that ranks by ‖Δh‖ gets this pair backwards; a J- and Σ_y-aware alarm does not.

Null space of J: large displacement, zero consequence. Leading direction: the same distance, large consequence.

This turns the endpoint into a designed test with a predicted sign, not a correlation.

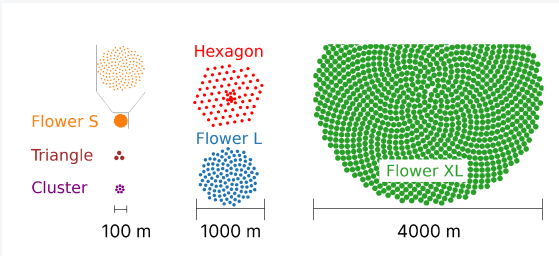
Test the alarm from simulation to real data

designed, not yet run



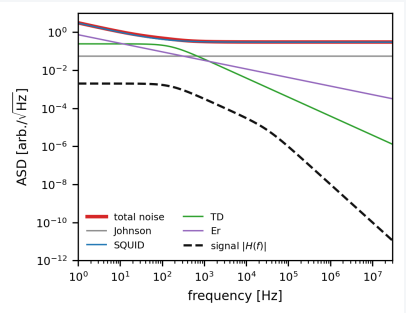
Tier 1 — controlled waveforms

own generator: $\hat{\Sigma}$ and Σ both known; every theoretical claim is decided here



Tier 2 — neutrino telescopes

one injection, two geometries, a frozen public GNN; no $\hat{\Sigma}$ exists — on purpose



Tier 3 — TIDMAD, real data

single stationary channel, so Σ is circulant and the PSD is the whole covariance: $\Sigma^{-1} = F^{\dagger}\mathrm{diag}(1/J_k)F$. One transformer trained twice, MSE vs Σ^{-1} — two $\hat{\Sigma}$, one real Σ

How to apply, on any frozen model

1. hook the stages
2. replay clean and perturbed windows
3. score in the pullback form $J^{T}\Sigma\_y^{-1}J$
4. fix a false-alert budget on clean data
5. attribute N / S, or abstain
6. report the consequence in physical units

One question across all three tiers: at matched severity, does a strong alarm separate the damaging cases from the harmless ones?

Predictions are frozen on Tier 1 before the confirmatory run on Tier 2; generic baselines (input tests, MMD, C2ST, embedding distance, uncertainty) get the same windows and false-alert budget.

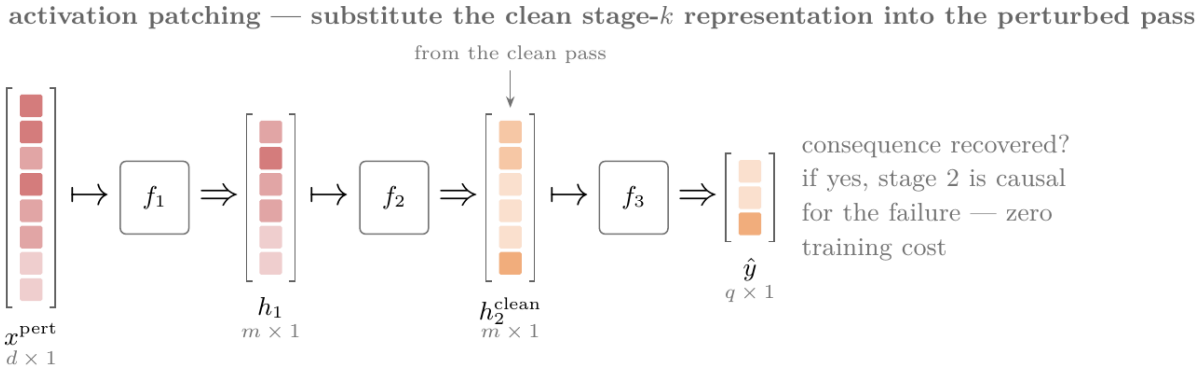
Diagnose first, then repair the failing stage

designed, not yet run

activation patching as a causal check · low-rank adaptation: Hu et al. (2021)

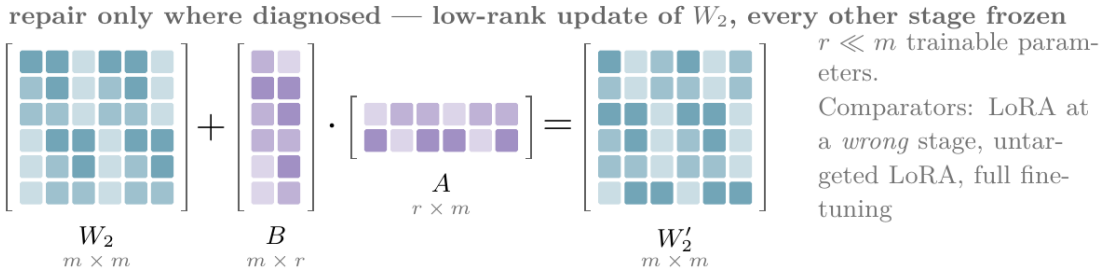
1 Diagnose

swap the clean stage-ℓ representation into the perturbed pass. If the consequence recovers, that stage is causal — at zero training cost.



2 Repair

a low-rank update at the diagnosed stage only, on a separate split. The wrong-stage arm is the comparator that matters.



Order matters: diagnose while frozen, adapt afterwards on a separate split — patch before adapting.

Falsifiable: if diagnosed-stage LoRA does not beat wrong-stage LoRA across seeds, the claim is dropped, not softened.

One object ties the story together: Σ^{-1}

Σ^{-1} is the only weighting the noise allows

filter, EMPCA, tied AE — one projection

the data decide what can be represented

at deployment the trained metric may stop matching the detector

score a shift where it changes physics

Isotropic MSE is a physics assumption, not a default.

The metric is measured; only the signal class is chosen.

And because the metric is a belief, a frozen model can be asked whether that belief still holds — and whether the answer matters.

For a low-threshold experiment like DELight this is the difference between a reconstruction that is *optimised* and one that can say **when it should not be trusted**.

Questions?

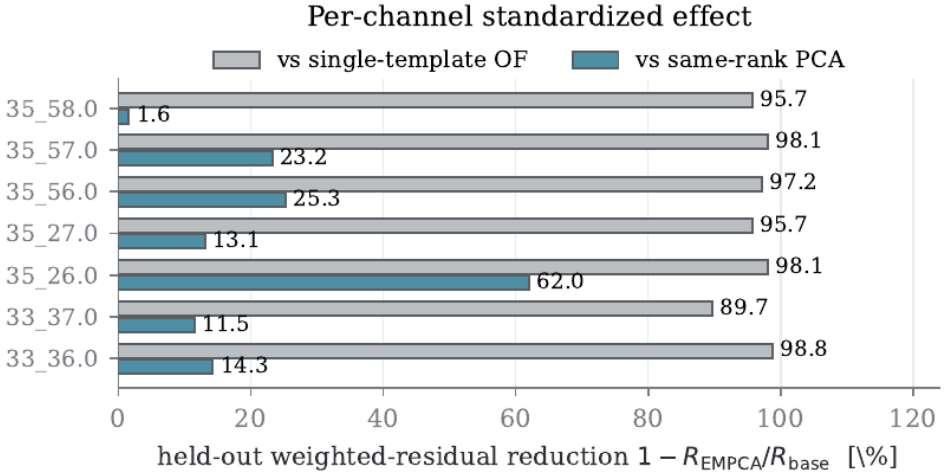
method X+dataset Y→performance gain.

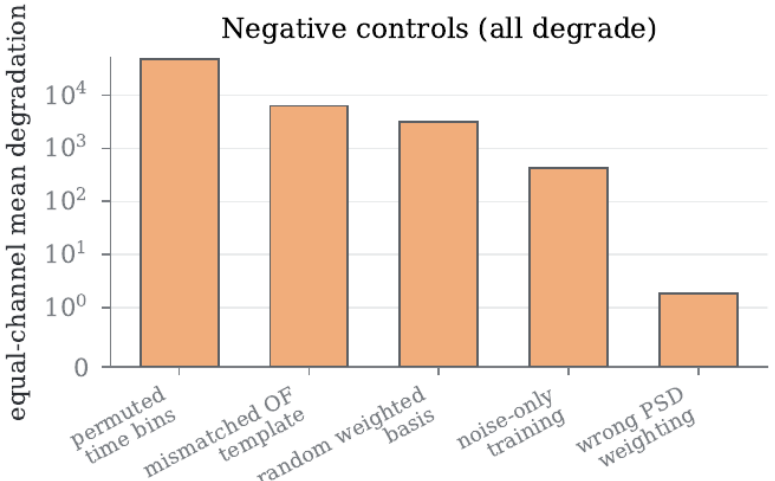
identify limitation→understand why→derive structural requirement→new method→controlled verification→general principle.

dowling.wong@kit.edu

Karlsruher Institut für Technologie

Backup: CRESST — positive, and honestly narrow





Left: the rank is held fixed, so the only difference between the bars is the metric. Right: five predeclared negative controls.

Every channel improves, 1.6 % to 62.0 %, equal-channel mean 21.6 %.

Seven channels from two runs, within-channel only, cluster-level intervals. No cross-channel claim and no detector-sensitivity claim.

Right: all five predeclared negative controls degrade — the pipeline can fail when it should.

Backup: what “tied” means, and what I did not show

The autoencoder recovers the EMPCA subspace when its encoder is the Σ^{-1}-adjoint of its decoder, $P^T\Sigma^{-1}P = I_k$.
Exact only under matched preprocessing, weighting and normalization, and only for linear models. Nonlinear autoencoders, non-Gaussian transients and a misspecified Σ are outside the statement.

· No cross-channel transfer and no detector-sensitivity claim.
· TIDMAD: one frozen operating point. The rank-2 structure prediction was not confirmed on any real file, and the linear estimator does not compete on the release's own denoising benchmark.
· Self-supervised denoisers learn a denoiser; the question here is the residual metric at fixed capacity — context, not comparators.
· Σ is estimated, not given: with N noise traces of length d the estimate is singular for N < d, and whitening amplifies its error. Davis–Kahan bounds the subspace error by ‖E‖ over the spectral gap.
· TIDMAD negative controls: of five predeclared, three degraded the primary metric with intervals excluding zero; two were uninformative rather than passing — reported both ways.

Backup: Tier-1 development run — what the mechanism testbed showed

Development scale, nothing pre-registered; the confirmatory run has not happened. Each line below was a written prediction that passed or failed.

· Three signatures, not two. Covariance-type N gives residual variance in the excited span, as predicted. Signal-deforming N — time jitter, channel loss — gives a mean shift, like S. S along excited coordinates shifts the mean in the span; along unexcited ones, in the complement.
· Abstention is feature novelty, not classifier margin. Split-conformal on the attribution margin failed on the unseen family (AUROC 0.25); a feature-space Mahalanobis score reached 0.71.
· Consequence must be whitened reconstruction error. At SNR 10 the amplitude readout is noise-dominated and cannot rank damage; the reconstruction consequence can.
· Two endpoints as written do not work. Alarm–consequence within strata is uninformative, so the endpoint became pooled ranking plus the designed pair; 10 % event sampling loses far more than five power points. The failures change the pre-registration, not the claims.

Backup: what the diagnostics do not claim

· Identifiability. Diagonalizing a covariance yields decorrelated statistical modes, not labelled physical sources: $\Sigma = \Sigma_1 + \Sigma_2$ has infinitely many PSD splits. Attribution is over declared intervention families under a stated simulator; unknown shifts are routed to abstention.
· Claims C1–C3. Detection at 1 % false alerts within five windows; N/S attribution against an all-generic baseline, where a strong generic baseline may win — that is a meaningful negative; monitoring cost with 10 % sampling and sketches.
· Consequence. TIDMAD's own denoising score is a development metric whose authors disclaim its scientific meaning. The confirmatory consequence there is a relative exclusion-limit comparison.
· Matched cells. Module loss lowers observed multiplicity, so every N cell is matched by clean events at the same multiplicity, charge and time spread — otherwise attribution reduces to recognizing the corruption operator's fingerprint.

Backup: does the alarm rank damage — and does the machinery work at all?

backup

Alarm A vs scientific loss K

	$K < \kappa$	$K \geq \kappa$
low alarm	quiet / benign	**missed**
strong alarm	harmless alarm	**detected**

Primary endpoint: among strong-alarm cells, does A separate $K \geq \kappa$ — at matched severity, and on the designed null / aligned pair? Plus the rate of the top-right cell — the one that costs physics.

Feasibility pilot — frozen NuBench DynEdge, module dropout 0 → 50 %

median angular error

standardised embedding displacement

0.00 → 0.15

10-nearest-neighbour retention

1.00 → 0.39

Monotone: perturbation and hook work. Feasibility only: 256 enriched events, and re-inference does not reproduce the released predictions (0.94° offset) — these are our numbers, not NuBench's. Confirmatory runs stay gated.