Fifty-six steps up a nanobeam cavity: a differentiable open-cavity model, its gradient, and what an independent solver says about it

Author list and affiliations in preparation

Working draft — September 5, 2026

Abstract

We give a complete, self-contained account of one photonic-crystal nanobeam cavity optimization: what the model computes, how its gradient is obtained, how the optimizer uses it, and what an independent solver says afterwards. The cavity resonance is treated as an outgoing complex-frequency pole of a small nonlinear operator $\mathbf{A}(\widetilde{\omega}, \boldsymbol{\theta})\boldsymbol{u} = 0$ built from the guided and radiation-continuum spectrum of the unpatterned beam, so the radiative linewidth is part of the eigenvalue rather than a perturbative afterthought. The gradient of $\log Q$ with respect to all 45 hole coordinates follows from implicit differentiation at the pole and costs one extra linear solve; a full step takes about 701 s on one machine, with no field solver in the loop. Applied to the air-clad Quan–Lončar design, 56 accepted trust-region steps raise the model's quality factor from 7 722.35 to 103 844 while halving the effective mode volume ($1.361 \to 0.6989\,(\lambda/n)^3$); the optimizer achieves this almost entirely by stretching the mirror taper outward by up to 110 nm at fixed hole size. Finite-difference time-domain snapshots show the corresponding mode contracting from roughly $\pm 2\,\mu$m to $\pm 1\,\mu$m along the beam. An independent FDTD checkpoint ran every 3 accepted steps and held a veto: it rejected the last 3 candidates and rolled the design back, ending the campaign on its own. The checkpoints report Q_{FDTD} rising from 13 186.5 to 511 457, but we show that this number is not what the data can support: the ringdown window falls below one photon lifetime after step 9, and the largest gain measured over a full lifetime is 2.124×. We report the analytical trajectory as the result, the FDTD veto as the reason to believe it, and the sub-lifetime fits as the reason the absolute numbers are not yet a claim.

1 Introduction

Most descriptions of a photonic inverse-design campaign report the final device. This paper reports the campaign. Our aim is a single worked example, complete enough that a reader could rebuild it: one cavity, one objective, one gradient, one optimizer, 56 accepted steps, and an independent solver watching from outside the loop.

The example is worth the space because the objective is a hard one. A cavity quality factor is the ratio of a resonant frequency to a linewidth, and for the device here the linewidth is five orders of magnitude smaller than the frequency. Any model that gets the field pattern right but the leakage slightly wrong will report a plausible wavelength and a meaningless Q. Optimizing Q therefore stresses a solver in a way that optimizing a transmission spectrum does not.

Two ingredients make the campaign possible. The first is a forward model in which the leakage is part of the eigenvalue rather than a correction to it: the resonance is an *outgoing* *complex-frequency pole* of a small nonlinear operator assembled from the guided and radiation-continuum spectrum of the unpatterned beam [1, 2, 3]. The operator is a few thousand rows, so a pole solve is minutes, not hours. The second is that this pole can be differentiated exactly. Implicit differentiation at a simple pole gives the derivative of the complex frequency with respect to every geometric coordinate from the left and right null vectors, at the cost of one additional linear solve regardless of how many coordinates there are [4]. Together these turn a 45-dimensional cavity design problem into an ordinary quasi-Newton ascent.

This is the same motivation that produced differentiable guided-mode expansion [5, 6], and the same motivation behind adjoint methods in full-wave solvers [7, 8, 9]. The difference here is the object being differentiated: not a real-frequency band of a periodic cell, and not a field on a discretized domain with an absorbing boundary, but the complex pole of a finite open device.

We are equally interested in the parts of such a campaign that are usually omitted. The optimizer stalled once and required a human decision to continue. The mode-identity gate that protects the pole tracking degraded monotonically, from 0.9278 to 0.759, and had to be relaxed. The independent solver was given a veto, used it three times, and thereby terminated the run. And the quality factors that the independent solver reported at the end were fitted over less than a tenth of a photon lifetime, which bounds what they can be said to mean. Each of these is reported here in the place where it happened.

What this paper is not. It is not a demonstration of a record quality factor; the absolute numbers are explicitly bounded in Section 7. It is not a comparison of solvers; the analytical Q_{ana} and the fitted Q_{FDTD} are treated throughout as two different observables and are never equated. And it is not a survey—a companion study applies the same operator construction to a materially different multilayer device, which is where questions of generality belong.

2 The device and its 45 numbers

The cavity is the canonical deterministic nanobeam design [10, 11]: a silicon beam of cross-section 700 nm×220 nm, refractive index $n = 3.46$, suspended in air, with 15 elliptical through-holes on each side of a mirror plane and a 990 nm unpatterned lead beyond the outermost hole. The design principle is a taper: the holes near the centre are small and closely spaced, and both the hole size and the pitch grow towards the mirror, so that the guided mode is turned around gently rather than abruptly. An abrupt turn scatters into the light cone; a gentle one does not.

The parameter vector. Because the device is mirror symmetric, only the positive-x half is free. Hole j is an elliptical cylinder through the full beam thickness,

$$H_j = \Big\{(x, y, z) : \ \frac{(x - x_j)^2}{r_{x,j}^2} + \frac{y^2}{r_{y,j}^2} \le 1,\ |z| \le \tfrac{T}{2}\Big\}, \tag{1}$$

and the free vector collects the 15 centre-to-centre pitches together with the two semi-axes of each hole,

$$\boldsymbol{\theta} = \big(s_0, \ldots, s_{15-1};\ r_{x,1}, \ldots, r_{x,15};\ r_{y,1}, \ldots, r_{y,15}\big) \in \mathbb{R}^{45}. \tag{2}$$

Centres follow from the pitches by $x_1 = s_0/2$ and $x_j = s_0/2 + \sum_{\ell<j} s_\ell$. Storing pitches rather than absolute positions has one consequence that will matter in Section 6: a small reduction of every pitch accumulates, so the device length is not an independent variable but a sum of 15 of them. The beam width, the thickness, and the lead length are held fixed throughout.

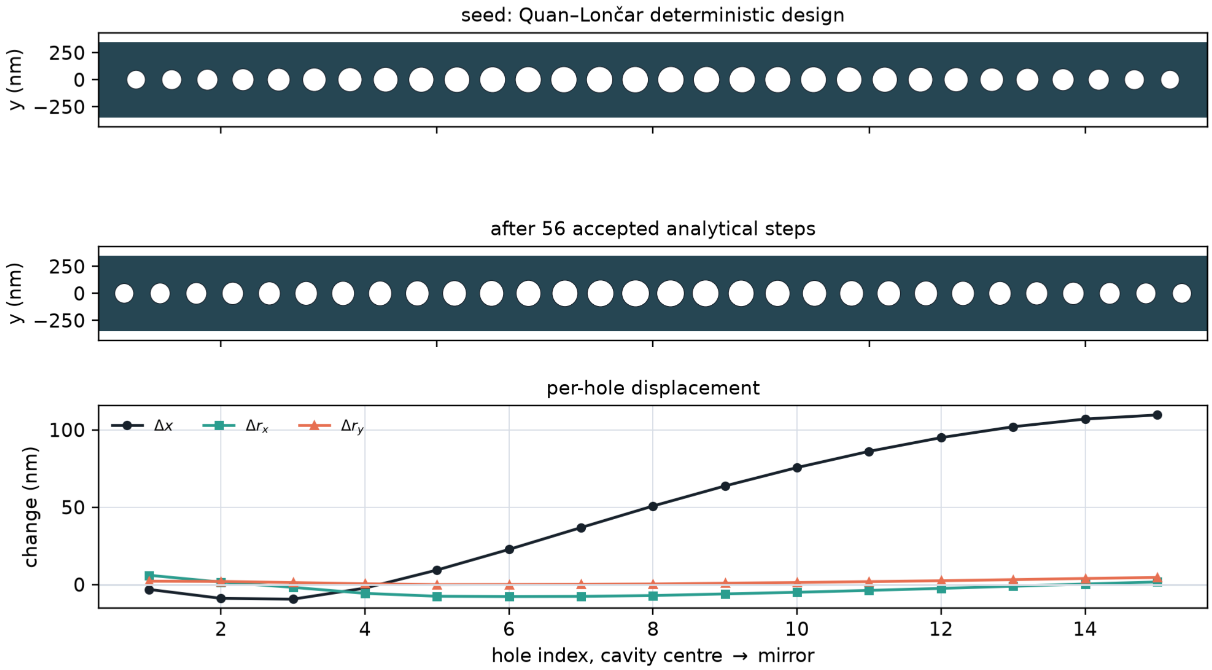

Figure 1: The geometry, before and after. Top: the seed, a Quan–Lončar deterministic design with a constant 330 nm pitch and radii tapering from the cavity centre to the mirror. Middle: the design after 56 accepted analytical steps, on matched axes. Bottom: the per-hole change. Silicon is dark, apertures white. Only the positive-x half is a free parameter; the negative half is generated by mirror symmetry. The two top panels look nearly identical, which is the point: the optimizer moved the outer holes outward by up to 110 nm while changing every radius by less than 7.75 nm.

The objective. We maximize $\log Q$ of the fundamental E_y-polarized cavity mode, where for a pole $\widetilde{\omega} = \omega_r + \mathrm{i}\omega_i$ with $\omega_i < 0$,

$$Q = -\frac{\omega_r}{2\omega_i}, \qquad \delta \log Q = \frac{\delta\omega_r}{\omega_r} - \frac{\delta\omega_i}{\omega_i}. \tag{3}$$

The logarithm makes multiplicative gains additive and puts the two parts of the pole on an equal relative footing. It also exposes the difficulty in one line: since $|\omega_i|/\omega_r = 1/(2Q)$, at the seed value $Q_{\mathrm{ana}} = 7\,722.35$ the loss rate is already only 6.5×10^{-5} of the resonant frequency, and at the end of the campaign it is 4.8×10^{-6}. Everything in Section 3 exists to compute that small number reliably; everything in Section 7 is about the fact that the independent solver eventually could not.

3 The forward model: Q from a pole

This section states the whole forward calculation. It is compact by design; the point is that every ingredient is either closed-form or a one-dimensional quadrature, which is why the model is fast enough to sit inside an optimizer.

3.1 The resonance as a pole

Use $\exp(-\mathrm{i}\omega t)$ and write the patterned permittivity as the unpatterned beam plus a hole contrast,

$$\epsilon_{\boldsymbol{\theta}} = \epsilon_0(y,z) + \Delta\epsilon_{\boldsymbol{\theta}}, \qquad \Delta\epsilon_{\boldsymbol{\theta}} = (\epsilon_{\mathrm{air}} - \epsilon_{\mathrm{Si}}) \sum_j \mathbf{1}_{H_j(\boldsymbol{\theta})}. \tag{4}$$

Let $\mathbf{G}^+_{\mathrm{wg}}(\omega)$ be the *outgoing* Green tensor of the unpatterned beam. The source-free Maxwell problem is then equivalent to the homogeneous Lippmann–Schwinger equation

$$\mathbf{E} = \frac{\omega^2}{c^2}\, \mathbf{G}^+_{\mathrm{wg}}(\omega)\, \Delta\epsilon_{\boldsymbol{\theta}}\, \mathbf{E}. \tag{5}$$

Because $\Delta\epsilon_{\boldsymbol{\theta}}$ vanishes outside the holes, the unknown lives only on 15 small ellipses — not on a computational domain. This is the single structural reason the model is cheap, and it does not become more expensive when the mirror is made longer.

A resonance is a frequency at which Eq. (5) has a nontrivial solution:

$$\mathbf{A}(\widetilde{\omega}, \boldsymbol{\theta})\, \boldsymbol{u} = 0, \qquad \boldsymbol{v}^T \mathbf{A}(\widetilde{\omega}, \boldsymbol{\theta}) = 0, \tag{6}$$

where $\mathbf{A} = I - k_0^2(\epsilon_{\mathrm{air}} - \epsilon_{\mathrm{Si}})\mathbf{G}^+_H$ is Eq. (5) projected onto a basis inside the holes, and $\boldsymbol{v}, \boldsymbol{u}$ are its left and right null vectors. Equation (6) is a *nonlinear* eigenproblem: $\mathbf{A}$ depends on ω through the reference Green tensor and through the radiation branch cuts, not merely through a factor ω^2.

3.2 Three ingredients

(i) The hole basis, and why the geometry is exact. Inside hole j we use normalized coordinates $x - x_j = r_{x,j}\rho\cos\theta$, $y = r_{y,j}\rho\sin\theta$ and expand in Zernike functions on the disk times Legendre functions across the thickness,

$$\boldsymbol{\phi}_{j,nm\sigma\ell s} = \frac{Z^\sigma_{nm}(\rho,\theta)}{\sqrt{r_{x,j} r_{y,j} N_{nm}}} \sqrt{\tfrac{2\ell+1}{T}}\, P_\ell(2z/T)\, \boldsymbol{e}_s, \tag{7}$$

which is complete in the hole as the retained orders grow. Because the holes are through-holes and the vertical functions are shared with the waveguide cross-section basis, the z projection is exact. The production run uses a $P2$ hole basis at one vertical order against a $y3$–$z1$ cross-section basis.

The important property is that the geometry never touches a grid. The Fourier factor of a hole is closed-form,

$$\widehat{\mathbf{1}}_{H_j}(\boldsymbol{k}) = e^{-\mathrm{i}k_x x_j}\Big[2\pi r_{x,j} r_{y,j} \tfrac{J_1(\rho_j)}{\rho_j}\Big]\Big[\tfrac{2\sin(k_z T/2)}{k_z}\Big], \quad \rho_j^2 = (k_x r_{x,j})^2 + (k_y r_{y,j})^2, \tag{8}$$

so moving a hole is exactly a phase, $\partial_{x_j} = -\mathrm{i}k_x$, and changing a radius is an exact Bessel derivative. No permittivity is ever sampled, no subpixel smoothing rule enters, and—as Section 5 needs—the geometry produced by a step of size α tends to the previous geometry as $\alpha \to 0$.

(ii) The reference Green tensor, containing both spectra. The unpatterned beam is invariant along x, so

$$\mathbf{G}^+_{\mathrm{wg}}(x - x'; \omega) = \int_{\mathcal{C}_\beta} \frac{\mathrm{d}\beta}{2\pi}\, e^{\mathrm{i}\beta(x-x')}\, \widetilde{\mathbf{G}_{\mathrm{wg}}}(\beta, \omega), \qquad \widetilde{\mathbf{G}_{\mathrm{wg}}} = \big[I - k_0^2 \chi_b \widetilde{\mathbf{G}}_{\mathrm{c}}\big]^{-1} \widetilde{\mathbf{G}}_{\mathrm{c}}, \tag{9}$$

with $\widetilde{\mathbf{G}}_{\mathrm{c}} = P_T/(k^2 - k_c^2) - P_L/k_c^2$ the exact air symbol and χ_b the beam contrast. Two things follow. The guided modes of the beam are not inserted by hand—they are the poles of the same resolvent that carries the radiation continuum, and their residues are normalized by the pole derivative rather than by a box norm. And the longitudinal projector P_L, which is the singular part of the electric dyadic Green tensor in real space, stays algebraic in momentum space, so no principal value or depolarization factor is ever guessed.

(iii) The contour, which is where openness lives. Openness is not imposed by an absorber; it is the choice of which side of the singularities the β integral passes. In the scaled variable $u = \beta/k_c(\omega)$, the radiation branch points sit at $u = \pm 1$ for every frequency, and we integrate along

$$u(t) = t - \mathrm{i}h \tanh(t/d) \exp[-(t/L)^8], \tag{10}$$

which dips below every positive-u singularity, rises above its negative partner, and returns to the real axis in the far tails. Holding $u(t)$ fixed while moving ω off the real axis is an explicit analytic-continuation homotopy; a singularity crossing the frozen path shows up as a jump in $\widetilde{\omega}$ and is rejected rather than silently accepted on the wrong sheet.

3.3 Solving for the pole

The pole search starts on the real axis, where a singular-value scan of $\mathbf{A}(\omega)$ in the declared symmetry sector locates the resonance; the smallest left and right singular vectors then seed the scalar equation $\boldsymbol{v}^T\mathbf{A}(\omega)\boldsymbol{u} = 0$, which is solved by a secant iteration in complex ω with the vectors refreshed each step. Acceptance requires all three of a small scalar residual, a small operator residual $\|\mathbf{A}\boldsymbol{u}\|/\|\boldsymbol{u}\|$, and $\operatorname{Im}\widetilde{\omega} < 0$. Over the whole campaign the worst operator residual was 9.19×10^{-10}, the worst reciprocity error $\|\mathbf{A} - \mathbf{A}^T\|/\|\mathbf{A}\|$ was 9.2×10^{-14}, and the worst null-vector equation error was 3.13×10^{-15}.

3.4 What the model reports

Beyond Q and the wavelength, the solved pole yields two diagnostics used throughout this paper. The effective mode volume follows from the null vector. And the linewidth splits into the part carried away along the beam by the guided channel and the remainder,

$$\frac{1}{Q} = \frac{1}{Q_{\mathrm{end}}} + \frac{1}{Q_{\mathrm{rad}}}, \tag{11}$$

which is how we can say in Section 6 what the optimizer actually suppressed.

4 The gradient, in three lines

Differentiate Eq. (6) with respect to one geometry coordinate θ_j, contract on the left with $\boldsymbol{v}$, and observe that the unknown field derivative multiplies $\boldsymbol{v}^T\mathbf{A}$, which is zero. What remains is

$$\boldsymbol{v}^T(\partial_{\theta_j}\mathbf{A})\boldsymbol{u} + \frac{\partial\widetilde{\omega}}{\partial\theta_j}\,\boldsymbol{v}^T(\partial_\omega\mathbf{A})\boldsymbol{u} = 0, \tag{12}$$

that is,

$$\boxed{\frac{\partial\widetilde{\omega}}{\partial\theta_j} = -\frac{\boldsymbol{v}^T(\partial_{\theta_j}\mathbf{A})\boldsymbol{u}}{\boldsymbol{v}^T(\partial_\omega\mathbf{A})\boldsymbol{u}}} \tag{13}$$

and then, from Eq. (3), the objective gradient

$$\frac{\partial \log Q}{\partial \theta_j} = \frac{1}{\omega_r}\frac{\partial \omega_r}{\partial \theta_j} - \frac{1}{\omega_i}\frac{\partial \omega_i}{\partial \theta_j}. \tag{14}$$

Four remarks are all that is needed to implement this.

The denominator is the normalization. $\boldsymbol{v}^T(\partial_\omega \mathbf{A})\boldsymbol{u}$ is the quasinormal-mode normalization *for this operator*. It is generated by the frequency dependence of the reference Green tensor and the contour, it is nonzero exactly when the pole is simple, and it is invariant under rescaling either null vector. Substituting a closed-cavity energy integral $\int \epsilon|\mathbf{E}|^2$ in its place is the standard way to obtain a believable $\partial\omega_r/\partial\theta$ together with a wrong $\partial\omega_i/\partial\theta$, and by Eq. (14) it is the second that carries Q [3, 1].

The pairing is a transpose, not a conjugate. $\boldsymbol{v}^T$, not $\boldsymbol{v}^\dagger$. On the real-frequency axis with real basis functions they agree, which is why the substitution is easy to make and its consequences are delayed; only the bilinear form is analytic in ω and survives continuation into the lower half plane. Reciprocity then predicts $\mathbf{A} = \mathbf{A}^T$, which we check numerically rather than impose.

The cost is one solve, not 45. Only the scalar $\boldsymbol{v}^T(\partial_{\theta_j}\mathbf{A})\boldsymbol{u}$ is ever required, never the matrix $\partial_{\theta_j}\mathbf{A}$. Seeding the reverse pass with $\boldsymbol{u}\boldsymbol{v}^T$ and accumulating backwards through the primitives—contour quadrature, Dyson inverse, disk quadrature, Bessel form factors, and the pitch-to-centre accumulation—returns all 45 components in a small constant multiple of one assembly. The pitch parameterization of Eq. (2) contributes a reverse cumulative sum, so the derivative with respect to an inner pitch automatically gathers every hole outboard of it.

Every geometry derivative is analytic. $\partial_{\theta_j}\mathbf{A}$ reduces through Eq. (9) to derivatives of the hole projections, and those are the closed forms of Eq. (8): a phase for a displacement, a Bessel expression for a radius. There is no remeshing and no finite differencing anywhere in the chain, so a finite-difference check of Eq. (13) tests the implementation rather than a step-size choice.

In practice one accepted step—assemble, solve the pole, run the reverse pass, line search—took a median of 701 s.

5 The optimization loop

5.1 Inner loop: trust-region ascent

The inner loop is deliberately ordinary. A BFGS direction is built from the gradient of Eq. (14); the step is scaled so that no single coordinate moves further than the current trust radius; candidate geometries along that direction are solved on *common connectivity*—the same quadrature nodes, basis, and contour—so that the measured change is a property of the geometry and not of a re-discretization. Realized single-coordinate moves ranged from 0.346 nm early in the campaign down to 0.0394 nm at the end, and the trust radius closed to 0.125 nm.

A step is accepted only if all four of the following hold.

(i) *The pole is certified*: passive sign, operator residual at tolerance, complex symmetry within bound.

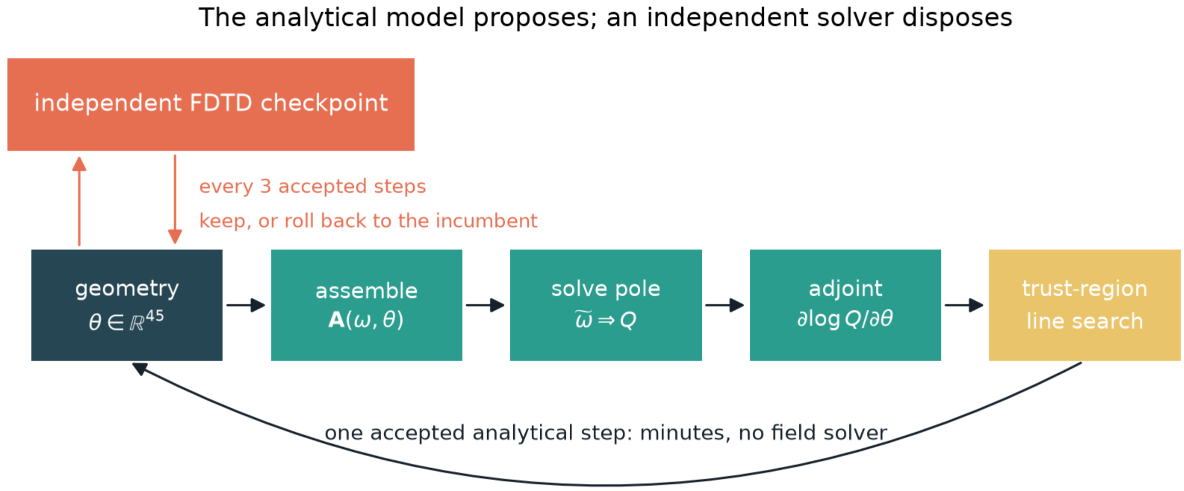

Figure 2: The loop. The inner cycle is analytical and runs on its own; the outer cycle calls an independent FDTD solver every 3 accepted steps and may roll the geometry back. The FDTD result never enters a gradient or a line search.

(ii) *The mode is the same mode.* Candidates are ranked by biorthogonal overlap with the accepted quasinormal mode, then by localization in the cavity region, then by symmetry sector, then by agreement with the first-order prediction from Eq. (13); frequency proximity is a tie-breaker only. This ordering matters because a sub-nanometre step can legitimately move the resonance by many of the previous linewidths, so frequency-nearest tracking is not a safe rule.

(iii) *The geometry is valid*: bounds satisfied, no hole overlap, and the applied displacement equal to the requested one.

(iv) *The measured surrogate gain is positive.* A predicted gain is not enough; the pole is re-solved at the applied geometry.

Failure of any gate shrinks the trust radius and records nothing in the BFGS history, because one poisoned curvature pair persists for many iterations. Of 60 attempted iterations, 56 were accepted and 4 line searches failed outright.

5.2 Outer loop: an independent solver with a veto

Every 3 accepted steps the current geometry was rendered into an independent finite-difference time-domain simulation [12, 13]: 64 ps of ringdown, 12 steps per wavelength, 1 wavelength of padding, with the incumbent geometry re-measured under the identical mesh so that the comparison is matched rather than against a stored number. If the candidate did not beat the incumbent, the geometry was rolled back and the trust radius reduced.

The separation of duties is strict and is what makes the outer loop informative. FDTD supplies no gradient, participates in no line search, and cannot cause a step to be accepted—only to be undone. Had FDTD results been allowed to steer step selection, any subsequent agreement between the two would be partly circular. The whole campaign consumed 3.56 solver credits.

5.3 One intervention, recorded

At accepted step 34 the inner loop stalled: repeated line searches failed, the trust radius had collapsed to 0.0625 nm, the analytical quality factor stood at 63 046.8, and the mode-overlap gate—originally 0.90—was being violated by a mode whose overlap had drifted to 0.8002. The campaign was stopped automatically.

We restarted it by hand, lowering the overlap gate to 0.75 and resetting the trust radius to 1 nm. This is a discretionary choice and it is visible in every figure that follows, marked at its step. It is also the single most questionable decision in the campaign: relaxing a mode-identity gate is exactly how one ends up optimizing a different resonance from the one intended. Section 7 returns to what the subsequent trajectory does and does not therefore support.

6 Results

6.1 The trajectory

Figure 3 is the campaign in one picture. Over 56 accepted steps the model's quality factor rises from 7 722.35 to 103 844, a factor of 13.45, with a peak of 113 613 at step 48 and a slight decline thereafter. The resonance wavelength barely moves, from 1 521.77 nm to 1 521.43 nm: the optimizer improved the linewidth without retuning the cavity, which is the behaviour Eq. (3) was written to encourage.

The ascent is smooth and unremarkable for the first thirty steps, which is the best thing that can be said about an optimizer. The interesting structure is at the two ends. Around step 34 the trust radius collapses and the run stalls; after the intervention it climbs again. Around step 48 the model's own objective turns over, and the independent checkpoints begin to disagree with it.

6.2 What the optimizer actually did

Figure 1 answers this, and the answer is simple enough to state in a sentence: *it stretched the taper*. Hole radii moved by at most 7.75 nm longitudinally and 4.61 nm transversely—about five percent—while the outer holes moved outward by up to 110 nm, and the innermost pitch contracted slightly, from 330 nm to 323.7 nm. The patterned half-length grew from 4 870.7 nm to 4 982.1 nm, an increase of 111 nm.

In other words the optimizer did not invent a new kind of mirror. It took the deterministic design's own principle—turn the mode around gradually—and pushed it further, tightening the centre and lengthening the gentle region on either side. That the answer is recognizable is a point in the method's favour: a 45-dimensional gradient ascent that had latched onto an artefact of the model would not be expected to reproduce the design rule that motivated the device.

6.3 The mode

Figure 4 shows the mode before and after, as computed by the independent solver rather than by the model being optimized. The envelope contracts from roughly $\pm 2\,\mu$m along the beam to $\pm 1\,\mu$m, and the number of visible antinodes falls accordingly. The model agrees quantitatively: its effective mode volume drops from 1.361 to 0.6989 $(\lambda/n)^3$, a factor of 0.514.

This is worth pausing on, because the naive expectation runs the other way. A higher-Q nanobeam cavity is usually a *gentler* one, and gentleness suggests a longer, more delocalized mode. What happened here is both: the mirror region got longer while the mode got shorter.

Table 1: Every FDTD checkpoint. The incumbent column is the previous accepted geometry *re-measured on the same mesh*, so each comparison is matched. $\mathcal{N}$ is the number of amplitude lifetimes the ringdown window actually spans (Section 7); values below 1 mean the fit did not observe a full decay. Source: `results/quan-campaign.json`.

step	Q_{ana}	incumbent Q_{FDTD}	candidate Q_{FDTD}	λ (nm)	$\mathcal{N}$	verdict
3	9 857.8	13 186	19 185	1503.03	2.090	keep
6	12 443	19 185	28 007	1502.50	1.432	keep
9	15 530	28 007	40 167	1501.88	0.999	keep
12	19 168	40 167	55 316	1501.31	0.726	keep
15	23 412	55 316	81 025	1499.60	0.496	keep
18	28 311	81 025	1.0692×10^5	1498.89	0.376	keep
21	33 915	1.0692×10^5	1.3818×10^5	1498.17	0.291	keep
24	40 267	1.3818×10^5	1.6928×10^5	1497.45	0.238	keep
27	47 409	1.6928×10^5	2.0253×10^5	1496.77	0.199	keep
30	55 376	2.0253×10^5	2.3736×10^5	1496.15	0.170	keep
33	62 667	2.3736×10^5	3.3636×10^5	1496.55	0.120	keep
36	69 348	3.3636×10^5	3.9009×10^5	1496.11	0.103	keep
39	79 557	3.9009×10^5	3.9589×10^5	1495.02	0.102	keep
42	90 698	3.9589×10^5	5.0373×10^5	1493.90	0.080	keep
45	1.0279×10^5	5.0373×10^5	5.1146×10^5	1493.24	0.079	keep
48	1.1361×10^5	5.1146×10^5	4.4467×10^5	1492.28	0.091	**roll back**
51	1.0919×10^5	5.1146×10^5	4.9922×10^5	1492.64	0.081	**roll back**
54	1.0597×10^5	5.1146×10^5	4.8477×10^5	1492.80	0.083	**roll back**

The taper stretch pushes the turning point outward, and the field that remains is confined more tightly between the two turning points rather than leaking into them.

6.4 Which loss channel was suppressed

The model separates the linewidth as in Eq. (11), and the separation makes the mechanism explicit. At the seed, 0.98972 of the decay already leaves through the guided channel along the beam. At the end of the campaign that fraction is 0.99639: the residual-loss quality factor reaches 2.875×10^7, some $276\times$ larger than the end-channel quality factor 104 221.

The optimizer therefore suppressed out-of-plane radiation almost to irrelevance, and what remains is essentially the power that propagates out along the beam past a finite mirror. That is a design statement, not a numerical one: it says the device is now mirror-limited, and that further gains require more mirror periods rather than better ones.

6.5 The independent checkpoints

Table 1 lists all 18 checkpoints. 15 were kept and 3 were rejected—at steps 48, 51, 54—so the last geometry the independent solver endorsed is the one at step 45. The three rejections are consecutive and they follow the turnover of the model's own objective at step 48 by three steps. Two readings are available: the model had begun to overstate its gains, or the checkpoint fits had become too poor to resolve the difference. Section 7 argues that the second is at least as likely as the first, and that this is why the outer loop earned its veto either way.

Taken at face value, the checkpoints report Q_{FDTD} rising from 13 186.5 at the seed to 511 457, a factor of 38.79; against the separately recorded 24 ps measurement of the same seed geometry, 15 449.5, the factor is 33.11. We quote both because the difference between them — 17.2 % in the seed value alone, from nothing but a change of ringdown length — is a useful measure of how protocol-dependent these numbers are. Neither factor is the paper's claim; Section 7 explains what is.

7 What the numbers support

7.1 The ringdown bound

A quality factor fitted to a time-domain ringdown is only constrained if the recorded window is long compared with the amplitude lifetime of the mode. With $\gamma = \omega/2Q$ the amplitude decay rate,

$$\tau = \frac{1}{\gamma} = \frac{2Q}{\omega}, \qquad \mathcal{N} = \frac{t_{\text{window}}}{\tau}, \tag{15}$$

and $\mathcal{N}$ counts the lifetimes the fit observed. The checkpoint protocol used a fixed 64 ps window throughout, while τ grew in proportion to the very quantity being measured. The result is Fig. 5, left: $\mathcal{N}$ starts at 2.6, crosses one at step 9, and ends at 0.0789 — under a tenth of a lifetime. 16 of the 19 recorded fits, including every fit above $Q_{\text{FDTD}} = 4.02 \times 10^4$, saw less than a full decay.

This is a bound, not a correction. A sub-lifetime exponential fit is ill-conditioned in the decay rate and, empirically in this data set, biased towards larger Q; but we cannot say by how much without rerunning. So we state the consequence directly:

> *The largest quality-factor improvement this campaign measured over a full photon lifetime is* 2.124×*, at checkpoint step 6, where* $Q_{\text{FDTD}} = 28\,006.7$ *with* $\mathcal{N} = 1.43$*. Every larger number quoted here, including the* 38.79× *endpoint ratio, rests on fits that did not observe a full decay.*

Figure 5, right, shows why this matters for interpreting the disagreement between the two models rather than merely for the absolute values. The two observables track each other closely at the start, where the fit is adequate, and separate progressively as $\mathcal{N}$ falls. That correlation does not prove the separation is an artefact of the fit — the model may genuinely be under-predicting — but it does mean the data cannot distinguish the two explanations.

7.2 The three rejections, reread

The rejections at steps 48, 51, 54 were the outer loop working as designed, and they ended the campaign. But by then $\mathcal{N} \approx 0.0789$, so the checkpoints were comparing two poorly conditioned fits against each other. The honest reading is that the outer loop stopped the run for the right reason — the model's own objective had already turned over at step 48 — while the specific evidence it used was weak. A protocol that scaled the window with the incumbent lifetime would have made those three decisions trustworthy at a cost of a few more solver credits, and that is the first thing we would change.

7.3 The relaxed gate

The intervention at step 34 lowered the mode-overlap threshold from 0.90 to 0.75. Figure 3, middle panel, shows the consequence: the overlap continued to decay, ending at 0.759, so roughly a quarter of the accepted mode no longer projects onto the template that defined the target resonance at step zero.

Two things are nonetheless true. The *local* branch overlap between consecutive accepted states never fell below 0.999795, so no single step jumped to a different pole; the drift is cumulative, not discontinuous. And the FDTD checkpoints, which know nothing about the template, continued to report improvement past the intervention. What we cannot claim is that the mode at step 56 is the same mode as at step 0 in any stronger sense than continuity of the tracking. A campaign of this length should carry an absolute mode-identity criterion, not only a relative one.

7.4 Two observables, never merged

Throughout this paper Q_{ana} and Q_{FDTD} are treated as distinct. They are computed by different means from different approximations of the same device, and the model is not calibrated against the solver at any point. The ratio between them is reported as a measure of disagreement, never as an error in either. In particular, the 13.45× analytical gain and the 38.79× checkpoint gain are not two measurements of one quantity, and their difference is not a residual to be minimized.

8 Reproducing this

Everything needed to check the claims above, short of rerunning the simulations, is a small set of records.

Inputs. The seed geometry is a single JSON file listing 15 hole centres and radii plus the beam dimensions. The final geometry is the 45-entry vector stored in the campaign state file. Those two vectors, plus Eq. (1), define both structures completely; every geometry number in this paper is derived from them by the same script that draws Fig. 1.

The campaign log. The optimizer wrote one record per event: for each of the 56 accepted steps, the solved pole, the quality factor before and after, the trust radius, the largest coordinate move, the full 45-component direction, and every operator diagnostic quoted in Section 3; for each of the 4 failed line searches, the attempts and the reduced radius; for each of the 18 checkpoints, the candidate and matched-incumbent quality factors, the solver task identifier, the realized cost, and the accept/reject decision; and for the intervention, the previous and new gate values. Table 1 and every panel of Fig. 3 are generated from that log.

No hand-typed numbers. Every numeric value printed in this paper is defined once, by a script that reads those records and emits LaTeX macros. Changing a record and rebuilding changes the paper; there is no path by which prose and evidence can drift apart. The same applies to Table 1, which is emitted rather than typed.

Field data. The three panels of Fig. 4 are the retained plots from the independent solver's own field export, with their original titles and colour bars intact rather than recoloured, so that the normalization statement in the caption is verifiable. The underlying field arrays are large and are kept outside version control, indexed by absolute path, size, and SHA-256.

Cost. The analytical campaign is a laptop-scale computation: a median of 701 s per accepted step, 56 steps. The entire independent-solver budget for 18 checkpoints was 3.56 credits on a commercial FDTD service.

9 Conclusion

We set out to describe one optimization completely. The parts worth carrying away are these.

A finite open cavity can be written as a pole of a small nonlinear operator whose ingredients are the guided and radiation-continuum spectrum of the unpatterned waveguide and a complete basis inside the etched holes. Because openness is imposed by the choice of integration contour rather than by an absorbing boundary, the radiative linewidth is part of the eigenvalue, and because the geometry enters through closed-form form factors, the model is exact in its parameters rather than in a discretization of them.

That pole differentiates cleanly. One line of implicit differentiation, Eq. (13), gives every geometric derivative from the left and right null vectors at the cost of one extra solve, and the normalization it requires is generated by the operator itself rather than assumed. A 45-dimensional cavity problem then becomes an ordinary trust-region ascent, at about 701 s per step with no field solver in the loop.

Run on the Quan–Lončar design, that ascent produced a recognizable answer: 56 accepted steps, a 13.45× improvement in the model's quality factor, an effective mode volume halved, and a geometry change that is almost entirely a stretch of the existing mirror taper. The loss decomposition says what was gained — out-of-plane radiation was suppressed until the device became mirror-limited, with a residual-loss quality factor 276× above the end-channel one — and the independent field snapshots agree that the mode contracted.

Placing an independent solver outside the loop, with a veto and no vote, was the right structure. It rejected the last 3 candidates and stopped the campaign, and because it never fed a gradient or a line search, that verdict is not circular.

Finally, the discipline that made this account worth writing is the same one that bounds it. A fixed 64 ps ringdown window measuring a lifetime that grows with the quantity being measured stops being a measurement somewhere around step 9; past that point the campaign's quality factors are indicative, not established. The improvement this campaign certified over a full photon lifetime is 2.124×. Extending that certificate to the endpoint requires nothing more interesting than a longer window, and that is the experiment we would run next.

References

[1] Philippe Lalanne, Wei Yan, Kevin Vynck, Christophe Sauvan, and Jean-Paul Hugonin. Light interaction with photonic and plasmonic resonances. *Laser & Photonics Reviews*, 12(5):1700113, 2018.

[2] Philip Trost Kristensen, Kathrin Herrmann, Francesco Intravaia, and Kurt Busch. Modeling electromagnetic resonators using quasinormal modes. *Advances in Optics and Photonics*, 12(3):612–708, 2020.

[3] Christophe Sauvan, Jean-Paul Hugonin, Ivan S. Maksymov, and Philippe Lalanne. Theory of the spontaneous optical emission of nanosize photonic and plasmon resonators. *Physical Review Letters*, 110:237401, 2013.

[4] Stefan Güttel and Françoise Tisseur. The nonlinear eigenvalue problem. *Acta Numerica*, 26:1–94, 2017.

[5] Momchil Minkov, Ian A. D. Williamson, Lucio C. Andreani, Dario Gerace, Beicheng Lou, Alex Y. Song, Tyler W. Hughes, and Shanhui Fan. Inverse design of photonic crystals through automatic differentiation. *ACS Photonics*, 7(7):1729–1741, 2020.

[6] Lucio Claudio Andreani and Dario Gerace. Photonic-crystal slabs with a triangular lattice of triangular holes investigated using a guided-mode expansion method. *Physical Review B*, 73:235114, 2006.

[7] Jakob S. Jensen and Ole Sigmund. Topology optimization for nano-photonics. *Laser & Photonics Reviews*, 5(2):308–321, 2011.

[8] Sean Molesky, Zin Lin, Alexander Y. Piggott, Weiliang Jin, Jelena Vuckovic, and Alejandro W. Rodriguez. Inverse design in nanophotonics. *Nature Photonics*, 12:659–670, 2018.

[9] Tyler W. Hughes, Momchil Minkov, Ian A. D. Williamson, and Shanhui Fan. Adjoint method and inverse design for nonlinear nanophotonic devices. *ACS Photonics*, 5(12):4781–4787, 2018.

[10] Qimin Quan and Marko Lončar. Deterministic design of wavelength scale, ultra-high Q photonic crystal nanobeam cavities. *Optics Express*, 19(19):18529–18542, 2011.

[11] Parag B. Deotare, Murray W. McCutcheon, Ian W. Frank, Mughees Khan, and Marko Lončar. High quality factor photonic crystal nanobeam cavities. *Applied Physics Letters*, 94:121106, 2009.

[12] Allen Taflove and Susan C. Hagness. *Computational Electrodynamics: The Finite-Difference Time-Domain Method*. Artech House, 3rd edition, 2005.

[13] Flexcompute. Tidy3D: GPU-accelerated finite-difference time-domain solver. https://www.flexcompute.com/tidy3d/, 2024.

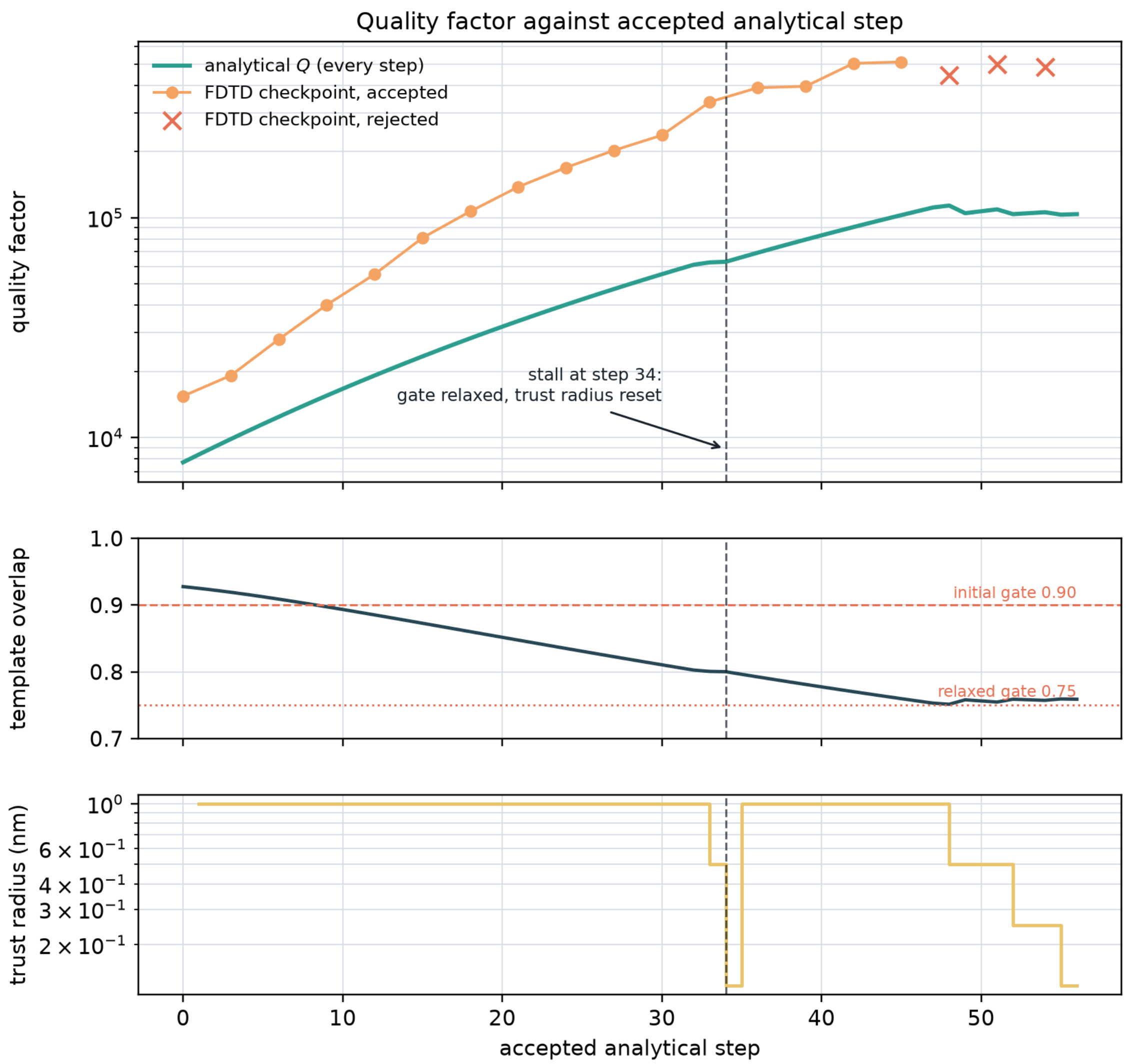

Figure 3: The campaign. Top: the model's quality factor at every accepted step (line) together with the independent FDTD checkpoints (circles when the candidate was kept, crosses when it was rejected and the geometry rolled back). The two are different observables plotted on one axis for trend, not for value. Middle: the mode-overlap gate that protects pole tracking, with the original and relaxed thresholds. Bottom: the trust radius. The dashed vertical line is the manual intervention of Section 5.

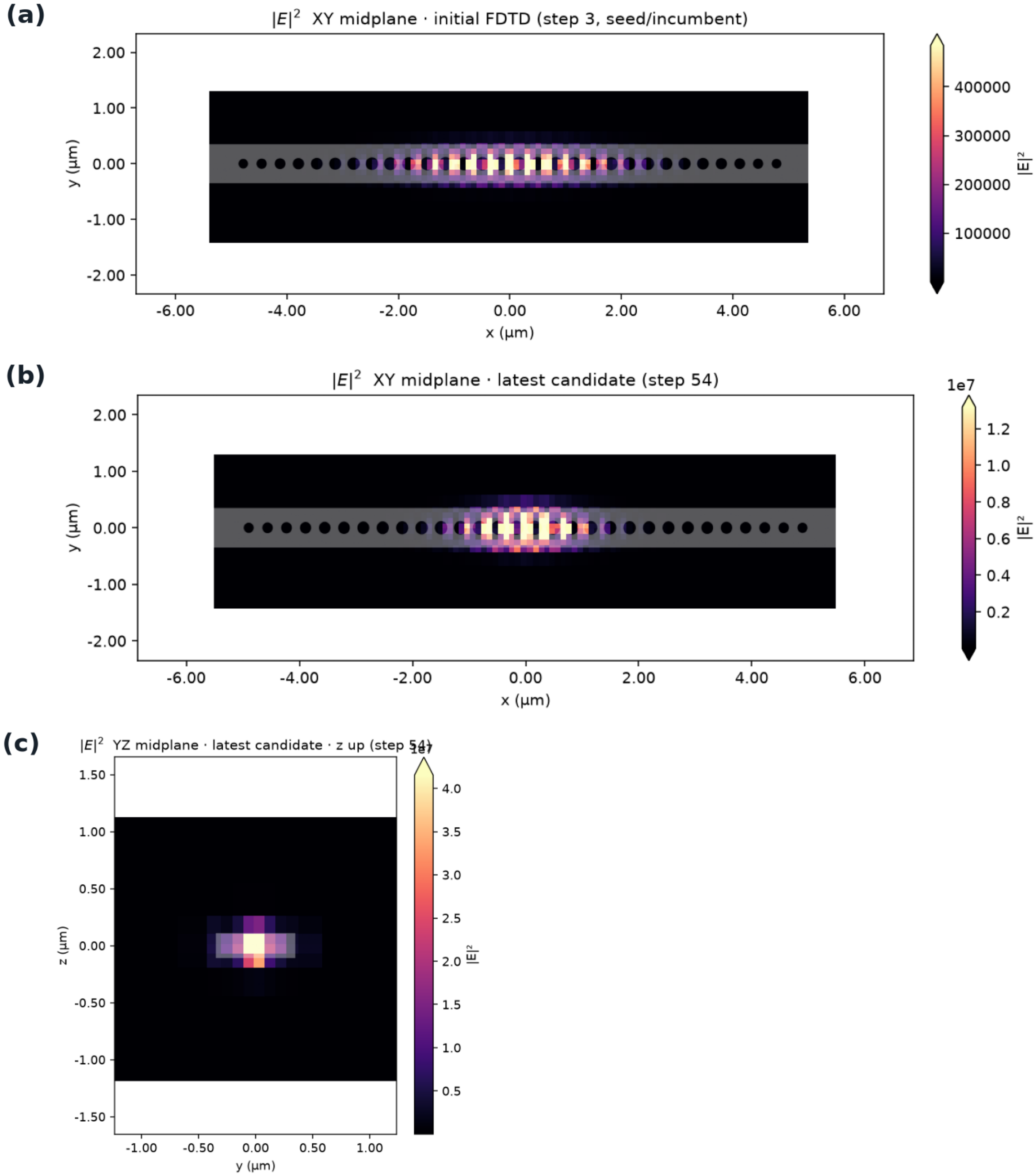

Figure 4: Independent FDTD $|E|^2$ in the beam midplane, with the structure overlaid. (a) the seed geometry; (b) the same quantity after 54 accepted steps; (c) the transverse cut of the same optimized mode. Panels carry *independent* colour scales—these are ringdown snapshots and their absolute amplitudes are not comparable—so the quantity to read is the spatial extent, not the peak value.

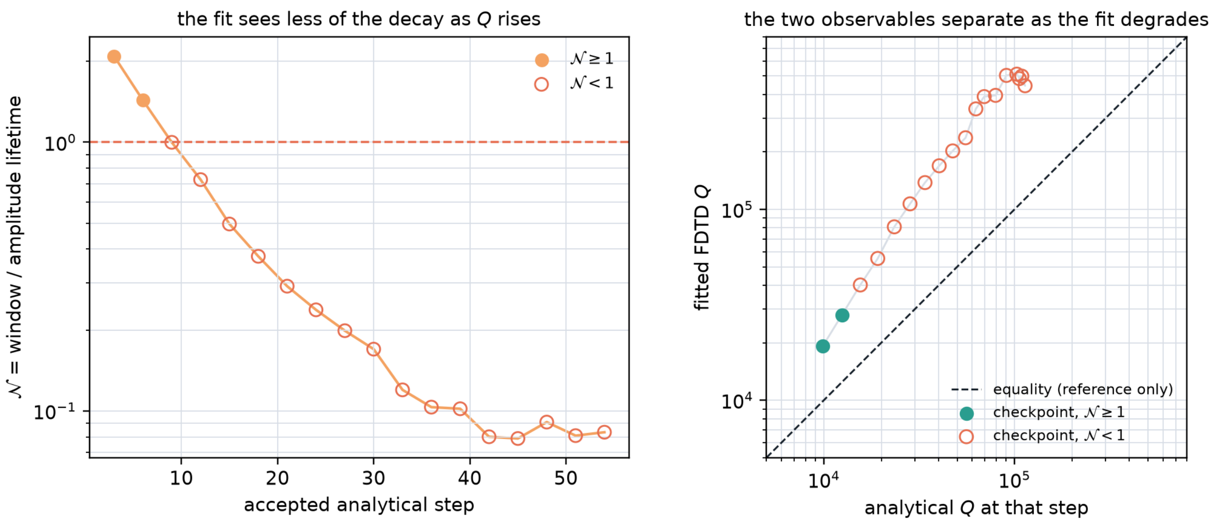

Figure 5: Left: the number of amplitude lifetimes each FDTD checkpoint actually observed, $\mathcal{N}$. It falls below one at step 9 and reaches 0.0789 by the end. Right: the fitted Q_{FDTD} against the model's Q_{ana} at the same step, on matched logarithmic axes; the diagonal is a reference line, not a model. The two observables separate precisely over the range in which the fit has stopped seeing a full decay.