A Torus-Structured Generalisation of NTRU: the NTC Assumption, its Cryptanalysis, and a Compact KEM

Sidoine Djimnaibeye[1], Djiby Sow[2], and Mahamat Borgou Hassan[3]

[1] ChadUP, N'Djamena, Chad
d.sidoine@chad-up.com
[2] Department of Mathematics and Computer Science, Université Cheikh Anta Diop (UCAD), Dakar, Senegal
dsow@ucad.edu.sn
[3] Department of Mathematics and Computer Science, ENASTIC, N'Djamena, Chad
mahamatborgou@gmail.com

Abstract. We introduce *Noisy Torus Conjugation* (NTC), a lattice assumption in which a short secret is confined to a non-split maximal torus of $\mathrm{GL}_k(R_q)$ and acts by conjugation on a uniform matrix, the result being masked by a short additive error. NTRU is the $k = 1$ member of the family. Passing to $k \geq 2$ changes the geometry of the underlying lattice in two specific ways. The planted module occupies a fraction $1/(2k)$ of the published lattice's dimension, against NTRU's $1/2$; and the norm-map shortcut that governs the overstretched regime is blocked once the conjugated matrix is required to be uniform over the full matrix algebra instead of the torus.

We develop the structure theory of the assumption: marginal uniformity of each component, invariance along the torus orbit, a rigidity theorem identifying the full set of short solutions, and a reduction from search to decision. On it we build an IND-CCA key encapsulation mechanism whose passive security reduces tightly to NTC together with one isolated decisional assumption. At NIST categories 1, 3 and 5 it reaches public keys within 1.08 to 1.16 times Kyber's and ciphertexts 2.0 to 2.1 times Kyber's.

The new assumption is not load-bearing but purchasable. Widening the key distribution to the smoothing parameter of the key lattice would make the public key statistically uniform and remove it altogether, leaving IND-CPA on module-LWE and hence on a worst-case problem. We price that variant at a factor 3.4 on the public key and 4.0 on the ciphertext. It rests, however, on a regularity statement not established for the completely split rings our transform uses; we isolate that statement as a conjecture and give a modulus class for which it is not needed.

Concrete parameters are selected with an estimator calibrated against the published core-SVP figures of Kyber, and validated against a hybrid meet-in-the-middle model whose single free constant is fitted on Kyber. The fatigue predictions underlying the modulus window are tested further by lattice reduction. We reduce small instances of the published lattice against NTRU controls of identical dimension, determinant and

planted-vector norm, and at every modulus the NTRU plant is discovered as a dense sublattice while the sparser NTC plant is not. The assumption is new and has no worst-case reduction; we state throughout what is proved, what is heuristic, what is measured, and what remains open.

Keywords: Post-quantum cryptography · Lattice-based cryptography · NTRU · Key encapsulation mechanism · Non-split maximal torus · Overstretched NTRU · Dense sublattice discovery

Mathematics Subject Classification (2020). 94A60 · 68P25 · 11T71 · 11H06

1 Introduction

NTRU [17] remains, after a quarter of a century, the most economical lattice construction we have. Its public key is a single ring element, its operations are a handful of polynomial multiplications, and its security rests on a statement of disarming simplicity: given $h = g/f$ in R_q with f, g short, recover (f, g). That economy is also its exposure. The NTRU lattice carries a planted module of rank n inside dimension $2n$ — exactly half the ambient space — and this density is what drives the *overstretched* phenomenon: past a modulus threshold, dense sublattice discovery finds the plant far faster than the generic estimate predicts [21,11]. The threshold has been mapped precisely enough that deployed parameters stay clear of it, but the margin is a constraint on the design rather than a consequence of it, and it forecloses the regime of large q that correctness of an encryption scheme naturally wants.

The module-LWE constructions that were standardised instead — Kyber, now ML-KEM [5,28], and Dilithium, now ML-DSA [10,27] — avoid the phenomenon by having no plant at all: their public key is pseudorandom by a reduction from a worst-case problem [25,23], and the price is a public key and ciphertext that grow with the module rank. The two families thus sit at opposite ends of a trade-off, and the question that motivates this work is whether the gap between them contains anything.

The idea. We keep NTRU's planted structure and change what the plant is planted in. Fix a commutative R_q-algebra

$$\mathcal{A} \;:=\; R_q[y]/(y^k - \alpha), \qquad k \geq 2,$$

with α chosen so that $\mathcal{A}$ is a product of the large fields $\mathbb{F}_{q^k}$ rather than a product of copies of R_q — the *non-split* condition. Left multiplication embeds $\mathcal{A}$ into the matrix algebra $\mathcal{M}_k = M_k(R_q)$ as a maximal commutative subalgebra $M(\mathcal{A})$, a non-split maximal torus of GL_k. An NTC instance is then

$$t \;=\; \big(M_s A + e\big)\, M_s^{-1}, \qquad A \leftarrow\!\!\$\; U(\mathcal{M}_k), \quad s \in \mathcal{A} \text{ short}, \quad e \text{ short}: \tag{1}$$

a uniform matrix conjugated by a short torus element, perturbed additively. At $k = 1$ the torus is the whole ring, M is the identity, and (1) collapses to

$t - A = es^{-1}$, which after the public shift A is decisional NTRU. The family therefore strictly contains NTRU, and any parameter set must respect the known NTRU constraints.

What changes at $k \geq 2$ is the geometry. The short solutions of (1) are not isolated: they form a planted module, as in NTRU. But that module is sparser: inside the full lattice, of dimension $nk(1+k)$, it has density $1/\big(k(k+1)\big)$, and inside the lattice the schemes actually publish, of dimension $2nk$, it has density $1/(2k)$ — one quarter at $k = 2$ — against NTRU's one half. It is the latter figure that governs the attack, and Section 5.6 measures what it buys. Dense sublattice discovery is sensitive to precisely this density, and the sparser plant pushes the fatigue point up. A second effect points the same way. The reason one might hope to reduce (1) to a lower-dimensional NTRU instance is the norm map $N_{\mathcal{A}/R_q} = \det \circ M$, which turns a multiplicative relation in $\mathcal{A}$ into one in R_q and would collapse the dimension from nk to n. That route is available exactly when A is itself conditioned to lie in the torus, in which case the instance degenerates to $t - A = es^{-1} \in \mathcal{A}$ and the norm map applies. Requiring A uniform over *all* of $\mathcal{M}_k$ turns the relation into a difference rather than a quotient, and the shortcut disappears. The non-split condition is what blocks the remaining slotwise attack: because $\mathcal{A}$ has no factor isomorphic to R_q, shortness is a global condition in coefficient space and cannot be attacked one CRT slot at a time.

1.1 Contributions

1. The assumption and its structure theory (Section 3). We give search and decision forms of NTC and establish what can be established unconditionally. Each component of an instance is *exactly* marginally uniform, so all information lies in the joint correlation and no test applied to t alone can gain any advantage. The instance distribution is invariant under translation along the torus orbit, which yields a normal form but — we are explicit about this — a self-reduction of codimension $k^2 - k$ only, not a full random self-reduction. A rigidity theorem, resting on the generic triviality of the centraliser of a random matrix, shows that every short solution is (us, ue) for a short scalar $u \in R_q$: the analogue of "NTRU recovers (f, g) up to a unit". Search reduces to decision; as for NTRU, the converse is open, and we say why the standard rerandomisation machinery is unavailable.

2. A key encapsulation mechanism (Section 4). Two features of the assumption dictate the design rather than leaving it free. The trapdoor rescales everything by s, which rules out a Kyber-style $\lceil q/2 \rfloor$ encoding and forces the p-adic route of NTRU; and the encryption randomness must commute with M_s, hence must live in $\mathcal{A}$ itself and not in $\mathcal{M}_k$. IND-CPA security follows tightly from NTC together with one further decisional assumption, $\mathsf{cLWE}_{\mathcal{A}}$, which we isolate and whose truncation is ordinary module-LWE; a Fujisaki–Okamoto transform [14,18] gives IND-CCA in the ROM. Three levers — row truncation, MSB encoding, decoupled noise widths — remove from the transcript what the cryptanalysis of Section 5 shows the attacker never uses, bringing public keys to 1.08–1.16× Kyber's

and ciphertexts to 2.0–2.1× Kyber's at comparable security. Appendix F then argues that the new assumption is not load-bearing but purchasable: widening the key to the smoothing parameter of the key lattice makes pk statistically uniform, so that IND-CPA would follow from module-LWE alone and therefore from a worst-case lattice problem. That statement is conditional on a regularity conjecture (Conjecture 1), and we say where it binds and how to avoid it. The price is a factor 3.4 on pk and 4.0 on ct — markedly milder than the corresponding penalty for NTRU, because the NTC key lattice has dimension $2nk$ rather than $2n$.

3. Concrete parameters and cryptanalysis (Sections 5 and 6). Every security figure is produced by one estimator, calibrated by running it on the module-LWE instances of Kyber and comparing to its published core-SVP numbers; agreement is within 1.2 bits. Comparative claims are stated as parity under this shared estimator rather than as absolute costs. We model the hybrid meet-in-the-middle attack [20] in the Babai/GSA form due to Wunderer [30], calibrate its single free constant against Kyber, and show that the stricter variant of the model is ruled out by that same calibration, since it predicts a 22-bit gain on Kyber-512 that the published analyses exclude. Under a BKZ profile simulator matched to our measured Gram–Schmidt profiles the hybrid binds at no parameter set, so parameters are selected against the primal figure, with the GSA closed form retained as a stress model; Section 5.5 reports how far the stress number moves with the calibration constant.

1.2 Limitations

We state the limitations here rather than leaving them to be discovered.

The assumption is new and has no worst-case reduction *at the efficient parameters*. Appendix F argues it can be dispensed with entirely at a quantified cost, but the regularity lemma that would make that statement rigorous is not established for the completely split rings our transform requires: the available union bound is $2^{-13.8}$ at the parameters proposed there, against the 2^{-127} claimed, so we state the missing step as Conjecture 1 rather than assert the theorem outright. Remark 16 also records that the obstruction is an artefact of constraints this variant does not need, and that a modulus class exists — $q \equiv 3, 5 \pmod 8$ — for which the Stehlé–Steinfeld lemma applies as written. The situation is the same as NTRU's in a further respect: the decision problem must be assumed separately from the search problem, because a key defines a single instance and the torus orbit is too small to rerandomise it. Confidence can come only from dedicated cryptanalysis. We have supplied a first round of it, not a definitive one: primal uSVP with column subselection, the norm and adjugate linearisations, CRT and slot attacks, dense sublattice discovery, hybrid.

Three quantitative caveats recur. The fatigue margins are now measured rather than extrapolated (Section 5.6), but by a *predictor* adapted from [11] and validated against its authors' NTRU figures. Lattice reduction on small NTC instances confirms the ordering the margins depend on, but the model is

used far beyond the dimensions in which either its authors or that experiment checked it, so the size of the margins remains predicted rather than observed. Decryption failure probabilities are computed from the exact distribution of the noise term (Section 4.5); the Gaussian approximation, which under-estimates the tail of a sum of products by 5 to 12 bits at these parameters, is reported only for comparison. What δ bounds is honest-execution failure: the use an adversary could make of failures is among what is not modelled (Section 5.7).

There is no implementation and no benchmark. Two distinct number-theoretic transforms are needed across the security categories, which an implementation covering both ring families must carry.

1.3 Related work

NTC belongs to the NTRU-with-errors family, whose members replace NTRU's exact quotient by a noisy relation: NTWE [15] combines NTRU and LWE over a ring, and MiNTRU [16] works with matrix secrets. Our construction differs from both in confining the secret to a *non-split* maximal torus of GL_k and not to GL_k itself, and in taking the instance to be a noisy conjugation rather than a quotient; these are exactly the two features that block the norm-map and CRT shortcuts. Non-split commutative algebras have also served directly as a platform for key establishment: NAIBI [9] builds reconciliation-based KEMs and ephemeral key agreement over algebras of this kind. NTC departs from it by embedding the algebra as a maximal torus of a matrix algebra, and in the planted, noisy-conjugation shape of the instance.

The closest matrix relative on the *exact*-quotient side is Module-NTRU [6,7], which generalises NTRU's quotient to a secret matrix ranging over all of $\mathrm{GL}_d(R_q)$ and has served as a trapdoor for identity-based encryption [6] and for compact hash-and-sign signatures [7]. The same two axes separate it from NTC. Its relation stays exact, so the published lattice is an NTRU lattice in higher dimension, to which the overstretched analysis applies as it stands; and its secret is unconstrained within the matrix algebra, whereas for us the torus is what denies the adversary the norm map. We inherit from that line the observation that matrix secrets buy parameter flexibility; what NTC adds is the error and the confinement. Variants of NTRU that alter the ring rather than the group — NTRU Prime [4], which removes the cyclotomic structure, and NTTRU [26], which chooses a ring optimised for the NTT — are orthogonal to what we do, and the second family of rings we use, $\mathbb{Z}[x]/(x^N - x^{N/2} + 1)$, is the same one NTTRU popularised.

On the cryptanalytic side we rely on the overstretched-NTRU literature [21,11], on the standard 2016 estimate for primal uSVP [1] with sieving exponents from [3,22], and on the hybrid attack [20,30]. Two adjacent efforts locate the present work at its two ends: on the signature side, ElGamal-like lattice signatures have been explored in the NIST additional-signatures process [19], and on the applications side, lattice-based two-factor mutual authentication and key exchange for telecare medicine information systems [8] illustrates the class of constrained deployments a compact KEM is meant to serve.

1.4 Organisation

Section 2 fixes notation, the two ring families with their expansion factors, and the non-splitness criterion. Section 3 defines NTC and develops its structure theory. Section 4 builds the KEM, taking two design choices on credit from the cryptanalysis that follows. Section 5 assembles that cryptanalysis and the concrete methodology, including the calibration of the estimator and of the hybrid model and the measured fatigue experiment, and discharges the credit. Section 6 instantiates the scheme. Section 8 closes with the status of the assumption and what remains open. Every proof is given where its statement is made; the conjectural worst-case variant, the calibration details, and implementation considerations are collected in the appendices.

2 Preliminaries

Let n be a power of two and q a prime with $q \equiv 1 \pmod{2n}$, so that $R_q = \mathbb{Z}_q[x]/(x^n+1)$ admits an NTT and splits as $R_q \cong \prod_{i=1}^{n} \mathbb{F}_q$. Elements are identified with their coefficient vectors, with representatives in $(-q/2, q/2]$. Fix $\alpha \in R_q$ and an integer $k \geq 2$, and define the commutative R_q-algebra

$$\mathcal{A} := R_q[y]/(y^k - \alpha),$$

a free R_q-module of rank k. We write $\alpha_i \in \mathbb{F}_q$ for the i-th NTT evaluation of α.

Definition 1 (Non-split parameter). *α is* non-split of order k *if for every $i = 1, \ldots, n$ the polynomial $y^k - \alpha_i$ is irreducible in $\mathbb{F}_q[y]$.*

For $k = 2$ this says that every α_i is a non-square in $\mathbb{F}_q$; taking α to be the smallest non-square constant $c \in \{2, 3, \ldots\}$ satisfies the condition in every slot at once.

Lemma 1 (CRT structure). *If α is non-split of order k, then $\mathcal{A} \cong \prod_{i=1}^{n} \mathbb{F}_{q^k}$ as rings. In particular $\mathcal{A}$ is a product of finite fields and has no factor isomorphic to R_q.*

This is what blocks the CRT attack: since $\mathcal{A}$ does not decompose through R_q, shortness is a global condition in coefficient space rather than a slotwise one.

Lemma 2 (Invertibility). *Let α be non-split of order k. An element $a \in \mathcal{A}$ is invertible if and only if its evaluation in $\mathbb{F}_{q^k}$ is nonzero in every CRT slot. For a drawn from a centred binomial distribution with $\eta \geq 1$, the probability of non-invertibility is at most $n\, q^{-k}$.*

Finally, every $a \in \mathcal{A}$ acts on the basis $\{1, y, \ldots, y^{k-1}\}$ by left multiplication, yielding a matrix $M_a \in \mathcal{M}_k := M_k(R_q)$. The map $M : a \mapsto M_a$ is an injective R_q-algebra homomorphism, the *regular representation*; it carries $\mathcal{A}^\times$ into $\mathrm{GL}_k(R_q)$ and its image $M(\mathcal{A}) = R_q[Y]$, $Y := M_y$, is a maximal commutative subalgebra of $\mathcal{M}_k$, non-split by Lemma 1. For $k = 2$,

$$M_{(u,v)} = \begin{pmatrix} u & \alpha v \\ v & u \end{pmatrix}.$$

2.1 Rings and expansion factors

Two families are considered. The first is the power-of-two cyclotomic $R = \mathbb{Z}[x]/(x^n + 1)$ with $n \in \{256, 512\}$. The second is

$$R = \mathbb{Z}[x]/\big(x^N - x^{N/2} + 1\big) = \mathbb{Z}[x]/\Phi_{3N}(x), \qquad N = 2^a \cdot 3,$$

the identity holding because $x^N - x^{N/2} + 1 = \Phi_6(x^{N/2})$ and $\Phi_m(x^r) = \Phi_{mr}(x)$ whenever every prime dividing r divides m; the degree is $\varphi(3N) = N$. For $N = 384$ this is Φ_{1152}, the half-size analogue of the NTTRU ring.

Unlike the power-of-two case, reduction in the second family is not a signed permutation, and products expand. Define

$$N_k^{(s)} := \#\big\{(i,j) \in [0,N)^2 \ : \ x^{\,s+i+j} \bmod f \text{ has a nonzero coefficient at } x^k\big\},$$

$$\mathsf{E}_R := \max_k N_k^{(0)}, \qquad \mathsf{E}_\alpha := \max_k \big(N_k^{(0)} + N_k^{(1)}\big).$$

Then a product of two independent random elements with coefficient variance σ_1^2, σ_2^2 has worst-coordinate variance $\mathsf{E}_R\,\sigma_1^2\sigma_2^2$, and the worst row of M_s with $\alpha = x$ contributes $\mathsf{E}_\alpha\,\sigma_1^2\sigma_2^2$. Computing $N_k^{(s)}$ from the reduction rules $x^{N+r} = x^{N/2+r} - x^r$ and $x^{3N/2+r} = -x^r$ gives:

Table 1. Expansion factors of the two ring families, and the relative cost of the twist $\alpha = x$.

Ring	E_R	E_α	$\mathsf{E}_\alpha/\mathsf{E}_R$
$\mathbb{Z}[x]/(x^n + 1)$	n	$2n$	2
$\mathbb{Z}[x]/(x^N - x^{N/2} + 1)$	$3N/2$	$3N$	2

Two readings. The second family costs 50% more variance per product at equal degree. But in both families $\mathsf{E}_\alpha = k\,\mathsf{E}_R$ for $k = 2$: relative to a plain product, the twist $\alpha = x$ costs a factor k in either ring, which is what makes it vastly preferable to a constant α (factor $1 + (k-1)\alpha^2$, i.e. 10 for $\alpha = 3$ and 26 for $\alpha = 5$).

2.2 Non-splitness of $\alpha = x$

Both families admit the same criterion, which we state once.

Lemma 3 (Non-splitness of the twist $\alpha = x$). *Let $R = \mathbb{Z}[x]/f(x)$ be one of the two families above, and let M be the multiplicative order of x modulo f, so $M = 2n$ in the first family and $M = 3N$ in the second. Assume $M \mid q - 1$, so that f splits completely modulo q and R_q admits a full NTT. Then $\alpha = x$ is non-split of order 2 in the sense of Definition 1 if and only if*

$$q \ \equiv \ M + 1 \pmod{2M}.$$

Proof. In every CRT slot, x evaluates to a root ζ of f, that is, to an element of exact order M; M is even in both families. Write $q-1 = Mt$. Then $\zeta^{(q-1)/2} = \left(\zeta^{M/2}\right)^t = (-1)^t$, since $\zeta^{M/2}$ has order 2. Hence ζ is a square in $\mathbb{F}_q$ if and only if t is even, i.e. if and only if $2M \mid q-1$. The value $(-1)^t$ does not depend on the slot, so either all slots are non-split or none is, and non-splitness holds exactly when t is odd, i.e. $q \equiv M+1 \pmod{2M}$. □

Specialising: $q \equiv 2n+1 \pmod{4n}$ in the power-of-two family, and $q \equiv 1153 \pmod{2304}$ for Φ_{1152}. Both conditions refine, and are compatible with, the full-NTT condition $M \mid q-1$.

Remark 1 (Exclusion of constant twists and of $x^{N/2}$). In the power-of-two family, $M = 2n \geq 512$ forces $q \equiv 1 \pmod 8$, so 2 is always a square and the smallest admissible constant is $\alpha = 3$. In Φ_{1152} one has $1152 = 2^7 3^2$, so $q \equiv 1 \pmod 8$ and $q \equiv 1 \pmod 3$: both 2 and 3 are squares and the smallest admissible constant is $\alpha = 5$, with expansion 26. The tempting choice $\alpha = x^{N/2}$, a primitive sixth root of unity, always *fails*: its order is 6 and $12 \mid 3N \mid q-1$, so it is a square in every slot.

3 The NTC Assumption

We isolate the hardness assumption underlying the scheme and state it in both a search and a decision form. We call it NTC, for *Noisy Torus Conjugation*: the public key is a conjugate of a uniform matrix by an element of a non-split maximal torus, masked by an additive error.

3.1 Parameters and instance distribution

Throughout, the public parameters are

$$\mathsf{pp} = (n,\, q,\, k,\, \alpha,\, \chi_s,\, \chi_e),$$

where n is a power of two, $q \equiv 1 \pmod{2n}$ is prime, $R_q = \mathbb{Z}_q[x]/(x^n+1)$, $k \geq 2$, and $\alpha \in R_q$ is *non-split of order k* in the sense of Definition 1, so that

$$\mathcal{A} := R_q[y]/(y^k - \alpha) \;\cong\; \prod_{i=1}^{n} \mathbb{F}_{q^k} \qquad \text{(Lemma 1)}.$$

We write $M : \mathcal{A} \hookrightarrow \mathcal{M}_k = M_k(R_q)$ for the regular representation $a \mapsto M_a$; it is an injective ring homomorphism carrying $\mathcal{A}^\times$ into $\mathrm{GL}_k(R_q)$. We identify $a \in \mathcal{A}$ with its coefficient vector in R_q^k, and equip $\mathcal{A}$ (resp. $\mathcal{M}_k$) with the norm $\|\cdot\|_\infty$ taken over its nk (resp. nk^2) integer coefficients, with representatives in $(-q/2, q/2]$.

The distribution χ_s over $\mathcal{A}$ is supported on $\|s\|_\infty \leq \eta_s$ and χ_e over $\mathcal{M}_k$ on $\|e\|_\infty \leq \eta_e$ (e.g. coefficientwise centred binomial $\mathcal{B}_\eta$). We write $\chi_s^\times$ for χ_s conditioned on $s \in \mathcal{A}^\times$.

Definition 2 (Instance distribution). $\mathcal{D}_{\mathsf{pp}}$ *is the distribution over* $\mathcal{M}_k \times \mathcal{M}_k$ *obtained by*

$$A \leftarrow\!\!\$\, U(\mathcal{M}_k), \quad s \leftarrow\!\!\$\, \chi_s^{\times}, \quad e \leftarrow\!\!\$\, \chi_e, \quad t := (M_s A + e)\, M_s^{-1},$$

and outputting (A, t)*. We call* (s, e) *the* planted witness*.*

The conditioning is benign: by Lemma 2, $\Pr[s \notin \mathcal{A}^{\times}] \le n\, q^{-k}$, so $\chi_s^{\times}$ and χ_s are statistically indistinguishable.

3.2 Search problem

Definition 3 (NTC-Search). *Let* $\beta = (\beta_s, \beta_e)$ *with* $\beta_s \ge \eta_s$ *and* $\beta_e \ge \eta_e$*. Given* $(A, t) \leftarrow\!\!\$\, \mathcal{D}_{\mathsf{pp}}$*, find* $(s', e') \in \mathcal{A} \times \mathcal{M}_k$ *such that*

$$t\, M_{s'} = M_{s'} A + e', \qquad s' \neq 0, \qquad \|s'\|_\infty \le \beta_s, \quad \|e'\|_\infty \le \beta_e.$$

For an adversary $\mathcal{B}$ *we set* $\mathbf{Adv}^{\mathsf{search}}_{\mathsf{pp},\beta}(\mathcal{B}) = \Pr[\,(s', e')\ \textit{valid}]$*.*

The planted witness is always a solution, so the problem is non-vacuous as soon as $\beta \ge (\eta_s, \eta_e)$. Note that e' is *determined* by s' once (A, t) are fixed; the search is effectively over s' alone, subject to $tM_{s'} - M_{s'}A$ being short as well.

Remark 2 (The constraint $s' \in \mathcal{A}$ *is constitutive).* Requiring $s' \in \mathcal{A}$ rather than $s' \in \mathcal{M}_k$ is what makes the problem non-trivial: without it, any short approximate kernel vector of the Sylvester operator $\mathcal{L}_{t,A} : X \mapsto tX - XA$ would qualify, and such vectors are abundant. The assumption is precisely that the secret is confined to the commutant $R_q[Y] = M(\mathcal{A})$, a non-split maximal torus of GL_k.

3.3 Decision problem

Definition 4 (NTC-Dec). *Distinguish* $(A, t) \leftarrow\!\!\$\, \mathcal{D}_{\mathsf{pp}}$ *from* $(A, u) \leftarrow\!\!\$\, U(\mathcal{M}_k) \times U(\mathcal{M}_k)$*, with advantage*

$$\mathbf{Adv}^{\mathsf{dec}}_{\mathsf{pp}}(\mathcal{B}) = \big|\Pr[\mathcal{B}(\mathcal{D}_{\mathsf{pp}}) = 1] - \Pr[\mathcal{B}(U \times U) = 1]\big|.$$

Assumption 5 (NTC)**.** *For every admissible parameter family* $\mathsf{pp} = \mathsf{pp}(\lambda)$ *and every PPT adversary* $\mathcal{B}$*, the quantity* $\mathbf{Adv}^{\mathsf{search}}_{\mathsf{pp},\beta}(\mathcal{B})$ *(resp.* $\mathbf{Adv}^{\mathsf{dec}}_{\mathsf{pp}}(\mathcal{B})$*) is negligible in* λ*.*

3.4 Basic structure

Lemma 4 (Marginal uniformity). *For any fixed* (s, e) *with* $s \in \mathcal{A}^{\times}$*, if* A *is uniform on* $\mathcal{M}_k$ *then* $t = (M_s A + e)M_s^{-1}$ *is* exactly *uniform on* $\mathcal{M}_k$*.*

Proof. $A \mapsto M_s A M_s^{-1}$ is a bijection of $\mathcal{M}_k$, hence $M_s A M_s^{-1}$ is uniform; t is the translate of a uniform variable by the fixed matrix eM_s^{-1}, hence uniform. □

Consequently each component of an instance is individually uniform, and all the information sits in the *joint* correlation of (A, t): no statistical test applied to t alone (spectrum, determinant, coefficient distribution) can achieve any advantage. In this respect NTC-Dec behaves like decisional NTRU.

Lemma 5 (Torus translation invariance). *For every $c \in \mathcal{A}$ (arbitrary, not necessarily short), the map $(A, t) \mapsto (A + M_c,\ t + M_c)$ maps $\mathcal{D}_{\mathsf{pp}}$ to $\mathcal{D}_{\mathsf{pp}}$* with the same witness (s, e), *and $U \times U$ to $U \times U$.*

Proof. $(t + M_c)M_s - M_s(A + M_c) = tM_s - M_sA + (M_cM_s - M_sM_c) = e$, the commutator vanishing because $c, s \in \mathcal{A}$ and $\mathcal{A}$ is commutative. □

Remark 3 (Normal form, and the limits of self-reduction). Lemma 5 yields a rerandomisation along the torus orbit and a normal form: fixing an R_q-linear projection $\pi : \mathcal{M}_k \to M(\mathcal{A})$, one may assume w.l.o.g. $\pi(A) = 0$, so that the effective instance lives in $(\mathcal{M}_k/M(\mathcal{A})) \times \mathcal{M}_k$. We stress that the orbit has codimension $k^2 - k$: this is *not* a full random self-reduction, and no worst-case to average-case reduction is known for NTC.

3.5 Rigidity of short solutions

The next lemma, together with Heuristic 6 below, is what makes Definition 3 well posed: it identifies the full set of short solutions.

Lemma 6 (Generic commutant). *Let $A \leftarrow_{\$} U(\mathcal{M}_k)$ and let $\mathcal{C}(A_i)$ denote the centraliser of the i-th NTT slot $A_i \in M_k(\mathbb{F}_q)$. With probability at least $1 - n\,\omega(k)\,q^{-k^2/2}$, where $\omega(k)$ is the number of distinct prime divisors of k, for every slot i,*

$$\mathcal{C}(A_i) \cap M(\mathbb{F}_{q^k}) = \mathbb{F}_q \cdot I_k.$$

At the deployed $k = 2$ the bound reads $1 - n\,q^{-2}$.

Proof. Fix a slot and write $T := M(\mathbb{F}_{q^k}) \subset M_k(\mathbb{F}_q)$, a maximal subfield. The intersection $C := \mathcal{C}(A_i) \cap T$ is an $\mathbb{F}_q$-subalgebra of the field T; a finite-dimensional subalgebra of a field is an integral domain, and a finite integral domain is a field, so $C \cong \mathbb{F}_{q^d}$ for some $d \mid k$ — and it is *the* copy of $\mathbb{F}_{q^d}$ inside T, which is unique since a finite field has exactly one subfield of each admissible order. Consequently $C \neq \mathbb{F}_q \cdot I_k$ if and only if the subfield $\mathbb{F}_{q^p} \subset T$ lies in $\mathcal{C}(A_i)$ for some prime $p \mid k$, that is, if and only if A_i lies in the centraliser of that fixed subfield. Viewing $\mathbb{F}_q^k$ as a vector space of dimension k/p over $\mathbb{F}_{q^p}$, that centraliser is $\mathrm{End}_{\mathbb{F}_{q^p}}(\mathbb{F}_q^k) \cong M_{k/p}(\mathbb{F}_{q^p})$, of $\mathbb{F}_q$-dimension k^2/p; since A_i is uniform on $M_k(\mathbb{F}_q)$,

$$\Pr\big[\,\mathbb{F}_{q^p} \subseteq \mathcal{C}(A_i)\,\big] \;=\; q^{\,k^2/p - k^2} \;\le\; q^{-k^2/2}.$$

A union bound over the at most $\omega(k)$ primes dividing k and over the n slots gives the claim. No regularity assumption on A_i is needed: the argument bounds the intersection directly, whether or not $\mathcal{C}(A_i)$ equals $\mathbb{F}_q[A_i]$, and in particular it covers the derogatory slots, whose centralisers are strictly larger. □

Corollary 1 (Parametrisation of solutions). *Write any solution as $s' = su$ with $u := s^{-1}s' \in \mathcal{A}$. Then*

$$e' \;=\; tM_{su} - M_{su}A \;=\; M_s\,[A, M_u] \;+\; e\,M_u.$$

If $u \in R_q$ then $[A, M_u] = 0$ and $(s', e') = (us,\, ue)$. If $u \notin R_q$, Lemma 6 gives $[A, M_u] \neq 0$, and $M_s[A, M_u]$ is heuristically of size $\Theta(q)$, which is incompatible with $\|e'\|_\infty \leq \beta_e \ll q$.

Thus the only structurally guaranteed short solutions form the *planted rank-one module*

$$\mathcal{P} := \{\,(us,\, ue) \;:\; u \in R_q\,\},$$

of rank n over $\mathbb{Z}$ inside a lattice of dimension $nk(1+k)$. This is the analogue of the trivial NTRU rotations $(x^j f, x^j g)$, except that here the planted sublattice occupies a fraction $1/\big(k(1+k)\big)$ of the dimension instead of $1/2$ — which is favourable with respect to dense sublattice discovery (see Section 5.6).

Heuristic 6 (Uniqueness of short solutions)**.** *Let*

$$\Lambda_{A,t} := \big\{(a, E) \in \mathcal{A} \times \mathcal{M}_k \;:\; tM_a \equiv M_a A + E \pmod q\big\},$$

a lattice of dimension $d = nk(1+k)$ and volume q^{nk^2} (for each a, the matrix E is determined mod q). Under the Gaussian heuristic its first minimum outside $\mathcal{P}$ is

$$\lambda_1(\Lambda_{A,t} \setminus \mathcal{P}) \;\approx\; \sqrt{\tfrac{d}{2\pi e}}\; q^{\,k/(k+1)}.$$

We assume β is chosen so that $\sqrt{d}\,\max(\beta_s, \beta_e)$ is well below this bound; then every *solution in the sense of Definition 3 lies in $\mathcal{P}$.*

Under Heuristic 6, NTC-Search is exactly witness recovery up to a short scalar of R_q, mirroring "NTRU recovers (f, g) up to a unit".

3.6 Relation between the two problems

Proposition 1 (Search reduces to Decision). *Assume Heuristic 6, and let $\mathcal{B}$ be an adversary solving $\mathsf{NTC\text{-}Search}_{\mathsf{pp},\beta}$ with advantage ε. Then there is a distinguisher $\mathcal{B}'$ against $\mathsf{NTC\text{-}Dec}_{\mathsf{pp}}$ whose advantage is at least $\varepsilon - \mathrm{negl}(\lambda)$.*

Proof. $\mathcal{B}'$ runs $\mathcal{B}$ on its input and outputs 1 iff the returned pair satisfies the verification of Definition 3. On $\mathcal{D}_{\mathsf{pp}}$ this happens with probability ε. On a uniform instance (A, u), the lattice $\Lambda_{A,u}$ carries no plant, so by the Gaussian heuristic its minimum is $\approx \sqrt{d/2\pi e}\, q^{k/(k+1)}$, which by hypothesis exceeds $\sqrt{d}\max(\beta_s, \beta_e)$; hence no valid solution exists and $\mathcal{B}'$ outputs 1 with negligible probability. □

Remark 4 (The converse is open). No reduction from NTC-Search to NTC-Dec is known. The standard search-to-decision machinery for (Ring-)LWE relies on rerandomising freshly drawn samples, which is unavailable here: a key defines a *single* instance, and Lemma 5 only rerandomises along a codimension-$(k^2 - k)$ orbit. The situation is the same as for NTRU, where the decision variant must be assumed separately.

3.7 The multi-instance variant

Definition 7 (Multi-instance $\mathsf{NTC}^{(\ell)}$). *Sample $s \leftarrow_s \chi_s^\times$ once, then for $j = 1, \ldots, \ell$ sample $A_j \leftarrow_s U(\mathcal{M}_k)$, $e_j \leftarrow_s \chi_e$ and set $t_j := (M_s A_j + e_j) M_s^{-1}$. The adversary receives $(A_j, t_j)_{j \le \ell}$ and must recover a short witness.*

Monotonicity is immediate: hardness of $\mathsf{NTC}^{(1)}$ is implied by hardness of $\mathsf{NTC}^{(\ell)}$ for $\ell \ge 1$, and the KEM uses only $\ell = 1$, the most conservative form. Any variant exposing several t_j under a common s must be re-estimated: the NTRU analogue (several quotients sharing f) is strictly weaker.

3.8 Degeneracies and position in the landscape

Three structural facts locate the assumption, and are developed — with the reductions available to an adversary and their quantitative consequences — in Section 5.1; we only name them here. At $k = 1$ the relation collapses to decisional NTRU, so the family strictly contains NTRU and every parameter set must satisfy the known NTRU constraints. Conditioning A into the torus $M(\mathcal{A})$ degenerates the instance to a commutative one on which the norm map $N_{\mathcal{A}/R_q} = \det \circ M$ applies, collapsing the dimension from nk to n; the assumption therefore *requires* A uniform over all of $\mathcal{M}_k$, which turns the relation into a difference and removes the shortcut. Finally, NTC belongs to the NTRU-with-errors / matrix-NTRU family (NTWE, MiNTRU), from which it differs by confining the secret to a *non-split* maximal torus of GL_k and by the conjugation shape of the instance; no member of this family has a worst-case reduction, and confidence can only come from cryptanalysis (Section 5) and from time.

3.9 Summary of dependencies

Table 2. Assumption dependencies of the two constructions.

Target	Assumption	Additional ingredients
Encryption / KEM	NTC-Dec	no reduction known from search

Finally, the norm bounds (β_s, β_e) appearing in Assumption 5 must be calibrated on what the security *proof* extracts — inflated by the extraction slack, typically a factor of order γ — and not on the honest widths (η_s, η_e). This gap is where provable and concrete security part ways.

4 A key encapsulation mechanism

We build an IND-CCA key encapsulation mechanism on NTC. The construction proceeds in two steps: a passively secure public key encryption scheme PKE

whose IND-CPA security rests on NTC-Dec (Definition 4) together with one additional decisional assumption isolated in Section 4.7, and a Fujisaki–Okamoto transform yielding an IND-CCA KEM in the ROM.

The shape of the scheme is not free. Section E records what the algebra forces: a p-adic encoding, and randomness confined to $\mathcal{A}$. Section 4.1 records what the cryptanalysis permits us to stop publishing — that cryptanalysis comes after the scheme, in Section 5, and is taken here on credit. Together the two determine the scheme of Section 4.3; a reader wanting the intermediate, untruncated design will find it in Remark 11.

Constraints forced by the algebra, in brief. Correctness needs the noisy conjugation to cancel exactly on the row the scheme publishes, which forces the secret to the torus, the twist to be non-split, and the modulus to the class of Lemma 3; the full accounting is in Appendix E.

4.1 Simplifications permitted by the cryptanalysis

The parameter analysis of Section 5 establishes one fact with direct consequences for size, used here and proved there: at every parameter set the optimal column subselection is $m^* = 1$ (Section 5.1). The attacker uses one column; the other $k^2 - k$ error coordinates carry no fresh entropy. A first temptation is to sample them as zero and save the corresponding noise budget. This does not work, as Remark 6 records. What one can do is stop *publishing* them.

Lever 1: row truncation. Take $s \equiv 1 \pmod{p}$ (Remark 7), and restrict the message to the first coordinate,

$$m_0 \in R_p, \qquad m := m_0 \ \in \ \mathcal{A}_p.$$

Then $\mathrm{coef}(sm) = (s_0 m_0, s_1 m_0, \ldots, s_{k-1} m_0) \equiv (m_0, 0, \ldots, 0) \pmod{p}$, so Dec only reads the first coordinate of D. Tracing this back through the decryption identity of Lemma 7, D_1 depends on c_2 only through its first row and on t only through its first row. The receiver therefore needs $t[1,:]$ and $c_2[1,:]$ alone:

$$\mathsf{pk} = (\rho,\, t[1,:]) : nk \text{ coefficients}, \qquad \mathsf{ct} = (c_1,\, c_2[1,:]) : 2nk \text{ coefficients},$$

instead of nk^2 and $nk(k+1)$ respectively. The truncated attack lattice matches exactly the $m^* = 1$ regime of the full analysis, so no security is lost.

Lever 2: MSB encoding, for compression only. For q odd, multiplication by the unit $\Delta := (q+1)/2$ is an isomorphism of R_q that sends the p-adic encoding $c_1 = A\mathrm{coef}(r) + pf_1 + \mathrm{coef}(m)$ to the MSB encoding $c_1 = A\mathrm{coef}(r) + pf_1 + \Delta\mathrm{coef}(m)$. The noise budget is identical; the gain lies elsewhere. Under MSB encoding the rounding error introduced by a compression function $\mathrm{Compress}_d : R_q \to \{0, \ldots, 2^d - 1\}$ enters D_1 additively as ordinary noise, whereas under LSB encoding it would destroy the residue mod p. Two limits are worth naming. The public $t[1,:]$ is *not* compressible: its rounding error multiplies $\mathrm{coef}(rs)$, a product

of two secrets, and even one bit of compression breaks correctness. And $c_1, c_2[1,:]$ tolerate only moderate compression ($d = 11$–12 bits) because, unlike Kyber's v, they multiply a secret rather than a hash.

Lever 3: decoupled distributions. A single noise width η used everywhere, as the untruncated design of Remark 11 does, is suboptimal. Wider χ_r hardens $\mathsf{cLWE}_{\mathcal{A}}$ at only *linear* cost in the noise budget, whereas χ_f enters quadratically. Decoupling $(\chi_f, \chi_r, \chi_{s'})$ into three widths $(\eta_f, \eta_r, \eta_{s'})$ recovers the category-3 slot, in which the parameter search would otherwise fail.

4.2 Parameters and notation

In addition to $\mathsf{pp} = (n, q, k, \alpha)$ of Section 3.1, the scheme fixes a small prime p with $p \nmid q$, three noise widths $(\eta_f, \eta_r, \eta_{s'})$ — decoupled per Lever 3 above — and a compression parameter d. The distributions are $\chi_{s'}$ over $\mathcal{A}$ for the secret, χ_r over $\mathcal{A}$ for the encryption randomness, and χ_f over R_q for the errors, each centred binomial of the corresponding width. The message space is R_2: a single ring element of binary coefficients, carried in the high bits by the encoding $\Delta := (q+1)/2$ of Lever 2, and we take $p = 2$ throughout, so that $\Delta p \equiv 1 \pmod{q}$. This makes R_2 supply n bits of message entropy, ample for a 256-bit KEM key at every category.

For $a \in \mathcal{A}$ we write $\mathrm{coef}(a) \in R_q^k$ for its coefficient vector, so that $\mathrm{coef}(ab) = M_a \mathrm{coef}(b) = M_b \mathrm{coef}(a)$, and $\mathrm{coef}_1(\cdot)$ for the first of those k coordinates; $\mathrm{row}_1(\cdot)$ denotes the first row of a matrix. We also write

$$\mathcal{A}_p := \mathcal{A}/p\mathcal{A} \;=\; R_p[y]/(y^k - \alpha),$$

which appears only in the untruncated reference design of (2), where the message space is $\mathcal{A}_p$ and the p-adic encoding is used in place of Δ.

4.3 The scheme

We write $\mathrm{row}_1(\cdot)$ for the first row of a matrix and $\mathrm{Compress}_d$ for coefficientwise truncation to d bits.

Algorithm 1 $\mathsf{KeyGen}()$

Output: $\mathsf{pk} = (\rho,\, t[1,:])$, $\mathsf{sk} = s'$ (equivalently s)

$\rho \leftarrow\!\$\ \{0,1\}^{256}$; $\quad A := \mathrm{Expand}(\rho) \in \mathcal{M}_k$
repeat
  $s' \leftarrow\!\$\ \chi_{s'}$ in $\mathcal{A}$; $\quad s := 1 + p\, s'$
until $s \in \mathcal{A}^{\times}$
$e_1 \leftarrow\!\$\ \chi_f^k$ {one row of the ambient error}
$t[1,:] := \big(\mathrm{row}_1(M_s)\, A + e_1\big)\, M_s^{-1} \in R_q^k$
return $\mathsf{pk} = (\rho,\, t[1,:])$, $\mathsf{sk} = s'$

Algorithm 2 $\mathsf{Enc}(\mathsf{pk}, m_0)$

Input: $\mathsf{pk} = (\rho,\, t[1,:])$, message $m_0 \in R_2$
Output: ciphertext ct

$\Delta := (q+1)/2$; $\quad A := \mathrm{Expand}(\rho)$
$r \leftarrow\!\!\$\ \chi_r$; $\quad f_1 \leftarrow\!\!\$\ \chi_f^k$; $\quad g_1 \leftarrow\!\!\$\ \chi_f^k$ {one row}
$c_1 := A\,\mathrm{coef}(r) + f_1 + \Delta\,\mathrm{coef}(m_0) \in R_q^k$, with $\mathrm{coef}(m_0) = (m_0, 0, \ldots, 0)$
$c_2[1,:] := t[1,:]\,M_r + g_1 \in R_q^k$
return $\mathsf{ct} := \big(\mathrm{Compress}_d(c_1),\ \mathrm{Compress}_d(c_2[1,:])\big)$

Algorithm 3 $\mathsf{Dec}(\mathsf{sk}, \mathsf{ct})$

Input: $\mathsf{sk} = s'$, ciphertext ct
Output: message $m_0 \in R_2$

decompress ct to $\tilde{c}_1, \tilde{c}_2$
$D_1 := \tilde{c}_2\,\mathrm{coef}(s) - \mathrm{row}_1(M_s)\,\tilde{c}_1 \in R_q$
return $m_0 := \lfloor 2D_1/q \rceil \bmod 2$

Remark 5 (Where p survives, and where it does not). The reference design of (2) scales its errors by p so that every parasitic term vanishes modulo p and the message can be read as $D \bmod p$. Under the MSB encoding of Lever 2 the message is recovered by rounding, not by reduction modulo p, so that scaling buys nothing and is dropped above: the errors enter as e_1, f_1, g_1 rather than pe_1, pf_1, pg_1. This is not cosmetic. Retaining the scaling would multiply the first bracket of (5) by p^2, inflating σ_{D_1} by a factor 1.6 to 1.8 at the parameter sets of Section 6.1 and raising the failure probability from 2^{-156} to roughly 2^{-56} — which would void the $q_G\delta$ term of Theorem 2. What p is still needed for is the secret: $s = 1 + p\,s'$ is what makes $\mathrm{coef}(sm) \equiv (m_0, 0, \ldots, 0) \pmod p$ and hence what makes row truncation possible, and $p = 2$ is what makes $\Delta p \equiv 1 \pmod q$ so that the residual wrap term is short. One consequence is that $t[1,:]$ is now *literally* a truncated NTC instance rather than a p-rescaling of one, so Lemma 8 is needed only for the reference design.

Remark 6 (No noise-free coordinates). It is tempting to drop the coordinates of e_1 or g_1 that "the attack does not use". This is unsafe. The public relation $t[1,:]\,M_s - \mathrm{row}_1(M_s)\,A = e_1$ is column-by-column a linear equation in $\mathrm{coef}(s)$; setting any coordinate of e_1 to zero makes the corresponding column an exact linear constraint on $\mathrm{coef}(s)$, which combined with shortness of s recovers the key. Every published equation must carry noise; the levers above concern what is published, not how it is sampled.

4.4 Correctness

The decryption identity is a computation in $\mathcal{A}$, and it is cleanest to state it once for the untruncated design of Remark 11, from which the identity for the scheme

of Section 4.3 follows by restriction to the first coordinate. That reference design is the one with

$$t := \big(M_s A + p\,e'\big)M_s^{-1}, \qquad c_1 := A\mathrm{coef}(r) + p\,\mathbf{f}_1 + \mathrm{coef}(m), \qquad c_2 := t\,M_r + p\,f_2, \tag{2}$$

all k rows published, message $m \in \mathcal{A}_p$, and $D := c_2\mathrm{coef}(s) - M_s c_1$.

Lemma 7 (Decryption identity). *For any key and ciphertext generated as in* (2)*,*

$$D \;=\; p\,\big(e'\mathrm{coef}(r) + f_2\mathrm{coef}(s) - M_s\mathbf{f}_1\big) \;-\; \mathrm{coef}(sm) \pmod q.$$

Proof. Since $r, s \in \mathcal{A}$ and $\mathcal{A}$ is commutative, $M_r\mathrm{coef}(s) = \mathrm{coef}(rs) = M_s\mathrm{coef}(r)$. Hence

$$c_2\mathrm{coef}(s) = tM_r\mathrm{coef}(s) + pf_2\mathrm{coef}(s) = tM_s\mathrm{coef}(r) + pf_2\mathrm{coef}(s),$$

while $M_s c_1 = M_s A\mathrm{coef}(r) + pM_s\mathbf{f}_1 + \mathrm{coef}(sm)$. Substituting $tM_s = M_s A + p\,e'$ from (2), the terms $M_s A\mathrm{coef}(r)$ cancel and the claim follows. □

Proposition 2 (Correctness). *If*

$$p\big(\|e'\mathrm{coef}(r)\|_\infty + \|f_2\mathrm{coef}(s)\|_\infty + \|M_s\mathbf{f}_1\|_\infty\big) + \|\mathrm{coef}(sm)\|_\infty \;<\; q/2, \tag{3}$$

then D does not wrap around modulo q, so $D \equiv -\mathrm{coef}(sm) \pmod p$ and Dec *returns m. We write δ for the probability, over* KeyGen *and* Enc*, that* (3) *fails.*

Heuristically each of the three noise terms has magnitude $O(\eta^2\sqrt{nk})$, inflated by the expansion factors E_R and E_α of the ring (Section 2.1). Those factors are what make the twist $\alpha = x$ preferable to any constant: at $k = 2$ the twist costs a factor $\mathsf{E}_\alpha/\mathsf{E}_R = 2$ relative to a plain product, against $1 + (k-1)\alpha^2$, i.e. 10 for $\alpha = 3$ and 26 for $\alpha = 5$, for the smallest admissible constants. Non-splitness and noise growth therefore pull in the same direction rather than against each other.

Remark 7 (The convention $s \equiv 1 \bmod p$). Two conventions are standard in the NTRU literature: leave χ_s unchanged and invert s in $\mathcal{A}_p$ at decryption, or impose $s \equiv 1 \pmod p$ and avoid the inversion. The scheme of Section 4.3 adopts the second, writing $s := 1 + p\,s'$ with $s' \leftarrow\!\!\$\ \chi_{s'}$: this is what makes $\mathrm{coef}(sm) \equiv (m_0, 0, \ldots, 0) \pmod p$ and hence what makes row truncation possible at all (Section 4.1). The price is that the lattice secret is s rather than s', so its variance is p^2 times that of $\chi_{s'}$ — a point that matters for the estimator and is treated in Section 5.4.

Specialisation to the scheme of Section 4.3.

Let u_1, u_2 be the rounding errors of $\mathrm{Compress}_d$ applied to $c_1, c_2[1,:]$, uniform on $\big(-q/2^{d+1},\, q/2^{d+1}\big]$ per coefficient with variance $V := (q/2^d)^2/12$. The derivation of Lemma 7, adapted to $s = 1 + ps'$, MSB encoding and row truncation, gives

$$\begin{aligned} D_1 \;=\; &-\Delta\,\mathrm{coef}_1(sm) \;+\; \big(e_1\mathrm{coef}(r) + g_1\mathrm{coef}(s) - \mathrm{row}_1(M_s)\,f_1\big) \\ &-\; u_2\mathrm{coef}(s) + \mathrm{row}_1(M_s)\,u_1 \pmod q. \end{aligned} \tag{4}$$

Since $s \equiv 1 \pmod{p}$, $\mathrm{coef}_1(sm) = m_0 + p \cdot (s'm_0)_0$; the low bit of $\Delta\mathrm{coef}_1(sm)$ contributes an additional wrap term whose variance is $\mathsf{E}_\alpha\, \eta_{s'}/8$ (heuristic). Coordinate by coordinate,

$$\begin{aligned}\sigma_{D_1}^2 \;&= k\,\mathsf{E}_R\,\sigma_f^2\sigma_r^2 \;+\; \sigma_f^2 + p^2 k\,\mathsf{E}_R\,\sigma_f^2\sigma_{s'}^2 \;+\; \sigma_f^2 + p^2\,\mathsf{E}_\alpha\,\sigma_f^2\sigma_{s'}^2 \\ &\quad + V\big(1 + p^2 k\,\mathsf{E}_R\,\sigma_{s'}^2\big) + V\big(1 + p^2\,\mathsf{E}_\alpha\,\sigma_{s'}^2\big) + \mathsf{E}_\alpha\,\sigma_{s'}^2/4,\end{aligned} \tag{5}$$

where the factors p^2 come from $s = 1 + ps'$, whose coefficient variance is $p^2\sigma_{s'}^2$. Decryption succeeds when $|D_1 + \Delta m_0| < q/4$; a Gaussian tail bound gives $\delta \le n \cdot 2\bar{\Phi}\big((q/4)/\sigma_{D_1}\big)$, and we require $\delta \le 2^{-140}$.

4.5 The failure probability without the Gaussian approximation

The Gaussian bound of the previous subsection under-estimates the tail of a sum of *products* of bounded variables. Computing the exact distribution of the noise term by convolution — with the rounding term enumerated rather than approximated — costs 2 to 9 bits on the variance and a further 2 to 3 on the tail, so the Gaussian δ is optimistic by 5 to 12 bits at these parameters. Every deployed set is selected against the *exact* δ and clears 2^{-140}; the margin at category 1 falls from 16 bits to 4, which is what makes the correctness constraint bind in the tuning of Section 6.1. The computation, and the comparison table against the Gaussian figures, are in Appendix D.

4.6 The p-adic key is free

The key produced by KeyGen is not literally an NTC instance: the error is $p\,e'$ rather than e'. This costs nothing.

Lemma 8 (Rescaling). *The map $(A, t) \mapsto (p^{-1}A,\, p^{-1}t)$, with p^{-1} taken modulo q, is a bijection of $\mathcal{M}_k \times \mathcal{M}_k$ carrying the output distribution of KeyGen exactly onto $\mathcal{D}_{\mathsf{pp}}$, and the uniform distribution onto itself.*

Proof. p is invertible mod q since $p \nmid q$, so the map is a bijection and preserves uniformity. From $t = (M_s A + pe')M_s^{-1}$ we get $p^{-1}t = \big(M_s(p^{-1}A) + e'\big)M_s^{-1}$, and $p^{-1}A$ is uniform whenever A is; the witness (s, e') is unchanged. □

Consequently the distinguishing advantage against the KeyGen distribution equals that against $\mathcal{D}_{\mathsf{pp}}$, with a tight (indeed advantage preserving) reduction to $\mathsf{NTC\text{-}Dec}$. The same rescaling applies to ciphertexts, so the factor p may be ignored in the security analysis.

4.7 The second assumption

Pseudorandomness of the public key does not by itself hide the message: once (A, t) has been replaced by a uniform pair, one still needs the ciphertext to be pseudorandom. This is the exact analogue of the situation for NTRU-based encryption, where decisional NTRU is combined with an LWE-style assumption over the now-uniform quotient. We isolate it as follows.

Definition 8 (Commutant LWE, $\mathsf{cLWE}_{\mathcal{A}}$). *For $\ell \geq 1$, distinguish*

$$\big(B_i,\ B_i M_r + F_i\big)_{i\leq \ell} \qquad \textit{from} \qquad \big(B_i,\ U_i\big)_{i\leq \ell},$$

*where $B_i, U_i \leftarrow_{\$} U(\mathcal{M}_k)$, $r \leftarrow_{\$} \chi_r$ in $\mathcal{A}$ (*common *to all i) and $F_i \leftarrow_{\$} \chi_f^{k\times k}$. We write $\mathbf{Adv}^{\mathsf{cLWE}}_{\mathsf{pp},\ell}(\mathcal{B})$ for the advantage, and allow individual samples to be truncated to a subset of their columns or rows.*

The deployed scheme uses $\ell = 2$ with the first sample truncated to its first *column* and the second to its first *row*: M_r has first column $\mathrm{coef}(r)$, so $B_1\mathrm{coef}(r)$ is exactly the first column of $B_1 M_r$, and $\mathrm{row}_1(B_2 M_r) = \mathrm{row}_1(B_2)\, M_r$. Throughout, $\mathbf{Adv}^{\mathsf{cLWE}}_{\mathsf{pp},2}$ denotes this truncated form; no statement in this paper invokes the untruncated one, which the definition retains only for the reference design of Remark 11.

Remark 8 (Plausibility). Both deployed samples are standard objects. The column-truncated sample $B_1\mathrm{coef}(r)+\mathbf{f}_1$ is *precisely* module-LWE of rank k over R_q with short secret $\mathrm{coef}(r)$. The row-truncated sample is a single LWE sample over the ring $\mathcal{A}$: under the public R_q-linear identification of rows of $\mathcal{M}_k$ with elements of $\mathcal{A}$, the map $u_1 \mapsto u_1 M_r$ is multiplication by r in $\mathcal{A}$, so the sample reads $(\tilde{u},\ \tilde{u}\, r + \tilde{g})$ with $\tilde{u}$ uniform — ring-LWE over the étale algebra $\mathcal{A}$, equivalently rank-k module-LWE whose coefficient matrix is confined to (a uniform element of) the torus. At the deployed twist $\alpha = x$ the algebra is, in addition, a familiar ring in its own right: substituting $x = y^2$ in the defining polynomial of the base ring gives $\mathcal{A} = R_q[y]/(y^2 - x) \cong \mathbb{Z}_q[y]/(y^{2n}+1) = \mathbb{Z}_q[y]/\Phi_{4n}(y)$ in the power-of-two family and $\mathbb{Z}_q[y]/(y^{2n} - y^n + 1) = \mathbb{Z}_q[y]/\Phi_{6n}(y)$ in the second — the cyclotomic ring of degree nk in each case (the identification is a $\mathbb{Z}_q$-algebra isomorphism by the dimension count, both sides being free of rank $2n$ over $\mathbb{Z}_q$). Concretely, over the category-3 ring Φ_{1152} the algebra $\mathcal{A}$ is the Φ_{2304} ring. The row-truncated sample is therefore ring-LWE in a standard cyclotomic ring of doubled degree, with the error distribution untouched by the identification. The only joint structure is the secret r shared across the two samples, the ordinary multiple-sample setting of LWE. There is moreover no subring shortcut in either sample: the trace descent $\mathrm{Tr}_{\mathcal{A}/R_q}$ yields module-LWE of rank k in dimension nk and no reduction to dimension n, because the relation is additive rather than multiplicative and no norm map applies. The *untruncated* matrix sample, by contrast, would expose k correlated short secrets $T_j\mathrm{coef}(r)$ under a common B, with T_j the public matrix of multiplication by y^{j-1}; the naive attempt to apply B^{-1} and isolate $T_j\mathrm{coef}(r)$ fails because $B^{-1}F$ is not short, but the correlation is genuine structure and would deserve dedicated scrutiny. The deployed scheme never exposes it.

4.8 IND-CPA security

Theorem 1 (IND-CPA). *For every PPT adversary $\mathcal{B}$ against the IND-CPA security of* PKE *there are PPT adversaries $\mathcal{B}_1, \mathcal{B}_2$ with*

$$\mathbf{Adv}^{\mathsf{ind\text{-}cpa}}_{\mathsf{PKE}}(\mathcal{B}) \ \leq\ \mathbf{Adv}^{\mathsf{dec}}_{\mathsf{pp}}(\mathcal{B}_1) \ +\ \mathbf{Adv}^{\mathsf{cLWE}}_{\mathsf{pp},2}(\mathcal{B}_2),$$

and the running times of $\mathcal{B}_1, \mathcal{B}_2$ are those of $\mathcal{B}$ up to additive polynomial overhead.

Proof. We proceed by games, on the deployed scheme of Section 4.3: the public key is $(\rho,\, t[1,:])$ and the challenge ciphertext is $\big(\mathrm{Compress}_d(c_1), \mathrm{Compress}_d(c_2[1,:])\big)$; since compression is a public deterministic function, it suffices to argue on $(c_1, c_2[1,:])$.

Game G_0. The real IND-CPA game.

Game G_1. Replace $t[1,:]$ by a uniform $u_1 \leftarrow\!\!\$\ U(R_q^k)$, keeping A as generated. By Remark 5 the published row is literally a truncated NTC instance, so a distinguisher truncates a full $\mathsf{NTC\text{-}Dec}$ challenge to its first row and $|\Pr[G_0] - \Pr[G_1]| \le \mathbf{Adv}^{\mathsf{dec}}_{\mathsf{pp}}(\mathcal{B}_1)$.

Game G_2. Replace $\big(c_1 - \Delta\mathrm{coef}(m_0),\, c_2[1,:]\big)$ by uniform elements of $R_q^k \times R_q^k$. In G_1 both A and u_1 are uniform and independent of everything else, and

$$\big(c_1 - \Delta\mathrm{coef}(m_0),\ c_2[1,:]\big) \;=\; \big(A\,\mathrm{coef}(r) + f_1,\ u_1 M_r + g_1\big),$$

which is exactly the truncated $\mathsf{cLWE}_{\mathcal{A}}$ instance with $\ell = 2$ and common secret r: the first sample truncated to its first column under $B_1 = A$, the second truncated to its first row under a uniform B_2 with $\mathrm{row}_1(B_2) = u_1$. Hence $|\Pr[G_1] - \Pr[G_2]| \le \mathbf{Adv}^{\mathsf{cLWE}}_{\mathsf{pp},2}(\mathcal{B}_2)$.

In G_2 the value c_1 is a uniform element of R_q^k shifted by $\Delta\mathrm{coef}(m_0)$, hence uniform and independent of m_0, and $c_2[1,:]$ is uniform and independent of m_0. The challenge bit is therefore information-theoretically hidden and $\Pr[G_2] = 1/2$. □

Remark 9. The reduction is tight and invokes neither Heuristic 6 nor the non-split condition: those are needed for the *search* problem. Non-splitness remains necessary upstream, since without it the CRT attack recovers the key outright and $\mathsf{NTC\text{-}Dec}$ is false.

Remark 10 (Samples consumed by the reduction). The reduction above is stated directly for the deployed row-truncated scheme, and both hops consume only what it publishes: the key hop truncates a full $\mathsf{NTC\text{-}Dec}$ challenge, and the ciphertext hop consumes the column- and row-truncated cLWE samples, which Remark 8 identifies as module-LWE and as ring-LWE over $\mathcal{A}$ with a shared secret. The untruncated reference design of Remark 11 would instead consume the full matrix sample, whose column correlation is exactly the structure flagged in Section 5.7; no statement in this paper relies on it. The IND-CCA lift below is unchanged, with the message space R_2 providing n bits of entropy, ample for a 256-bit KEM key at every category. What truncation does change is the *attack* lattice, which acquires dimension $2nk$ and volume q^{nk} — exactly the $m^* = 1$ regime that Section 5.1 identifies as optimal for the attacker on the untruncated instance. No security is lost.

A worst-case route, deferred. Widening the key distribution to the smoothing parameter of the key lattice would make the public key statistically uniform and remove the NTC-Dec assumption altogether, leaving IND-CPA on module-LWE alone. The construction, the regularity conjecture it rests on, the modulus class for which that conjecture is not needed, and its parameters — a factor 3.4 on the public key and 4.0 on the ciphertext — are developed in Appendix F.

4.9 From PKE to a CCA-secure KEM

We apply the Fujisaki–Okamoto transform with implicit rejection, $\mathrm{FO}^{\not\perp}$. Let G, H be hash functions and KDF a key derivation function, all modelled as random oracles, and let z be a secret seed stored in sk.

Algorithm 4 $\mathsf{Encaps}(\mathsf{pk})$

Input: pk
Output: ciphertext c, session key K
1: $m \leftarrow\!\!\$\ U(\mathcal{A}_p)$
2: $(\bar{r}, \bar{K}) := G\big(m, H(\mathsf{pk})\big)$
3: $c := \mathsf{Enc}(\mathsf{pk}, m; \bar{r})$
4: **return** $\big(c,\ K := \mathsf{KDF}(\bar{K}, c)\big)$

Algorithm 5 $\mathsf{Decaps}(\mathsf{sk}, c)$

Input: sk (including the implicit-rejection seed z), ciphertext c
Output: session key
1: $m' := \mathsf{Dec}(\mathsf{sk}, c)$; $\quad (\bar{r}', \bar{K}') := G\big(m', H(\mathsf{pk})\big)$
2: **if** $\mathsf{Enc}(\mathsf{pk}, m'; \bar{r}') = c$ **then**
3: **return** $\mathsf{KDF}(\bar{K}', c)$
4: **else**
5: **return** $\mathsf{KDF}(z, c)$
6: **end if**

Theorem 2 (IND-CCA, ROM). *Let δ be the failure probability of Proposition 2 and let $\mathcal{B}$ be an adversary against the IND-CCA security of* KEM *making at most q_G queries to G and q_H to H. Then there is a PPT adversary $\mathcal{B}'$ with*

$$\mathbf{Adv}^{\mathsf{ind\text{-}cca}}_{\mathsf{KEM}}(\mathcal{B}) \;\le\; 2\,\mathbf{Adv}^{\mathsf{ind\text{-}cpa}}_{\mathsf{PKE}}(\mathcal{B}') \;+\; q_G\,\delta \;+\; \frac{c_1 q_G + c_2 q_H}{p^{nk}},$$

for small absolute constants c_1, c_2 depending on the FO variant.

This is the standard bound; we state it in this shape rather than with fixed constants because these differ across formulations of $\mathrm{FO}^{\not\perp}$, and because a QROM

statement requires a separate argument. Two consequences for parameter selection. First, the term $q_G\,\delta$ forces a failure probability well below the security level: for $\lambda = 128$ and $q_G \le 2^{64}$ one targets $\delta < 2^{-140}$, i.e. the tail of $\|D\|_\infty$ in (3) must be cut at roughly six to seven standard deviations. Second, the message space $\mathcal{A}_p$ has p^{nk} elements, which is ample entropy for any reasonable (n, k, p).

4.10 Sizes

With row truncation the public key is nk ring coefficients plus a seed and the ciphertext is $2nk$ coefficients before compression, against nk^2 and $nk(k+1)$ for the untruncated design. Concrete byte counts, together with the parameter sets that produce them, are collected in Section 6.1.

Remark 11 (The untruncated scheme, and the reasons against proposing it). Replacing the three levers of Section 4.1 by their trivial settings — publish all k rows of t and of c_2, encode p-adically rather than in the high bits, and use a single noise width η throughout — gives the design one obtains by reading NTC naively. Its public key is nk^2 coefficients and its ciphertext $nk(k+1)$, dominated by the nk^2 term of c_2; at the security categories of Section 6.1 this amounts to 1.91–3.78 kB of pk and 2.81–5.62 kB of ct, that is, a factor 2.3 to 2.7 on pk and 1.5 to 2 on ct against the scheme above, at *equal* security. The comparison is worth making once because it locates the cost precisely: it is the nk^2 term of c_2, i.e. the price of the matrix reading, and Section 5.1 shows that the attacker never uses it.

5 Cryptanalysis and concrete methodology

A new assumption without a worst-case reduction is worth exactly what its cryptanalysis is worth. This section therefore does three things: it maps the structural attack surface of NTC and records which reductions are available to an adversary; it fixes a single estimator, calibrated against published figures for Kyber and Dilithium, in terms of which every security number in this paper is expressed; and it states, with the arithmetic, what the estimator does *not* model. It also discharges the two design choices that the scheme of Section 4 took on credit — the column truncation and the decoupled noise widths. A reader willing to take Table 10 on faith may proceed directly to the parameter sets of Section 6.

5.1 The structural attack surface

We collect the reductions that an adversary can perform for free, since each one constrains the parameter space.

$k = 1$ is NTRU. At $k = 1$ we have $\mathcal{A} = R_q$, the regular representation is the identity, and the relation collapses to $t - A = es^{-1}$. By Lemma 5 the public shift A is neutral, so NTC-Dec becomes decisional NTRU. Every parameter set must therefore satisfy, at a minimum, the constraints known for NTRU; in particular the overstretched analysis of Section 5.6 is not optional.

The commutative reading is strictly weaker, and must be excluded. Conditioning A to lie in the torus $M(\mathcal{A})$ collapses the instance to $b := t - A = es^{-1} \in \mathcal{A}$, at which point the norm map $N_{\mathcal{A}/R_q}(a) = \det M_a$ turns $bs = e$ into $N(b)N(s) = N(e)$: an NTRU instance over R_q in dimension n rather than nk. The overstretched threshold would then be governed by n, which is fatal. This is why Definition 2 insists that A be uniform over *all* of $\mathcal{M}_k$, and why any implementation shortcut that biases A towards the torus is forbidden. With A uniform the relation is a difference rather than a quotient and no norm map applies.

The adjugate linearisation is not competitive. An adversary may multiply the relation by $\mathrm{adj}(M_s)$ to obtain $N(s)\, t = M_s A \,\mathrm{adj}(M_s) + e\,\mathrm{adj}(M_s)$, which is linear in the entries of $\mathrm{adj}(M_s)$. This trades the nonlinearity for a squared secret — the entries of $\mathrm{adj}(M_s)$ are degree-$(k-1)$ forms in those of s, of size η^{k-1} — inside an inflated dimension. At $k = 2$ the resulting instance has a secret of variance $\Theta(\eta^2)$ in the same dimension, hence a strictly worse root-Hermite requirement than the direct primal attack, and we discard it. For $k \geq 3$ the inflation is worse still.

CRT and slot attacks are blocked by non-splitness, and only by it. If α were split, $\mathcal{A}$ would contain a factor isomorphic to R_q and shortness could be tested one CRT slot at a time, recovering the key. Non- splitness (Definition 1, Lemma 1) makes $\mathcal{A}$ a product of the large fields $\mathbb{F}_{q^k}$, so shortness is a global condition in coefficient space. This is the *only* place in this work where non-splitness is used: it is an upstream hardness condition, not an ingredient of any reduction.

Column subselection: the attack uses one column. The primal lattice of Definition 3 may be truncated to $m \leq k$ retained columns, giving dimension $nk(1+m)$ and volume q^{nkm}. At every parameter set in this paper the optimum is $m^* = 1$. This is the numerical face of a structural fact: the $k^2 - k$ additional error coordinates carry no fresh entropy about s, so they inflate pk and ct without contributing to hardness. The scheme of Section 4 simply stops publishing them, and its truncated lattice coincides with the $m^* = 1$ regime, so no security is lost by the truncation — a claim that this observation, rather than an additional assumption, justifies.

Rigidity, and the meaning of "solving". By Corollary 1 and Heuristic 6, the short solutions form the planted module $\mathcal{P} = \{(us, ue) : u \in R_q\}$ of rank n inside dimension $nk(1+k)$, and under the Gaussian heuristic the first minimum outside $\mathcal{P}$ sits at $\approx \sqrt{d/2\pi e}\, q^{k/(k+1)}$. All parameter sets are chosen so that $\sqrt{d}\max(\beta_s, \beta_e)$ is well below this value, so that recovering *any* short solution is recovering the key up to a short scalar.

5.2 The primal estimator, and its calibration

Security against lattice reduction is measured by the primal uSVP attack under the standard 2016 estimate [1]: blocksize β suffices when

$$\sqrt{\beta/d}\,\|v\| \;\le\; \delta_\beta^{2\beta-d-1}\,\mathrm{vol}^{1/d}, \qquad \delta_\beta = \Big(\tfrac{(\pi\beta)^{1/\beta}\,\beta}{2\pi e}\Big)^{1/(2(\beta-1))}, \tag{6}$$

with core-SVP costs 0.292β classically [3] and 0.265β quantumly [22]. Two lattices are estimated and the *minimum* of the two reported:

- the NTC lattice of Definition 3, optimised over the number $m \le k$ of retained columns: $d = nk(1+m)$, $\mathrm{vol} = q^{nkm}$, target $\|v\| \approx \sigma\sqrt{nk(1+m)}$;
- the $\mathsf{cLWE}_{\mathcal{A}}$ lattice of Definition 8, optimised over the number $m_s \le nk+nk^2$ of samples: $d = nk + m_s + 1$, $\mathrm{vol} = q^{m_s}$, Kannan embedding.

The estimate is only as good as its calibration, so we run it on the module-LWE instances of Kyber, whose core-SVP figures are published, using each category's actual secret width ($\eta_1 = 3$ for Kyber-512, $\eta_1 = 2$ above):

Table 3. Calibration of the primal estimator on Kyber's published core-SVP figures.

	Kyber-512	Kyber-768	Kyber-1024
blocksize β	406	624	874
estimated (this work)	118.6	182.2	255.2
published	118	181	254
discrepancy	+0.6	+1.2	+1.2

Agreement is within 1.2 bits, and the estimator is uniformly mildly optimistic; this is the accuracy at which every figure in Section 6 should be read. We stress that the discrepancy is sensitive to the assumed secret width: taking $\eta = 2$ for Kyber-512 instead of its actual $\eta_1 = 3$ moves the estimate to 111.5 and would degrade the apparent calibration to 6.5 bits.

And its behaviour under reduction. Calibration checks that we compute the same estimate others do, not that the estimate is right. Running single-shot BKZ-β key recovery on scaled-down instances of the KEM's own key lattice shows the rescaled form of the estimate tracking reduction to about three blocksizes, while the unrescaled form — the one that reproduces the figures of Section 6.1 — runs 9 to 20 blocksizes high, i.e. 3 to 6 bits in the unsafe direction at toy dimensions. The measurements, the width question they turn on, and the resulting reading of the KEM figures are in Appendix A; the conclusion — the quoted figures are not conservative — is carried into Section 5.4 and Section 6.1.

5.3 Attacks on cousin assumptions, and their applicability

NTC belongs to the NTRU-with-errors family, and two members of that family have been broken. Neither break transfers, but we set out why instead of assuming it, since both attacks target exactly the regime — a planted structure under an overstretched modulus — that Section 5.6 is about.

MiNTRU*, and the dense sublattice.* The somewhat-homomorphic scheme of [16] rests on MiNTRU: given

$$A \;=\; S^{-1}\big(G - E\big) \bmod q,$$

with G the *gadget* matrix and E short, recover S. Lee and Wallet [24] recover S at the smallest parameters of [16] in about 15 hours of fplll/BKZ 2.0 [12] on one laptop core, against a claimed 100 bits, by applying the overstretched analysis of Kirchner and Fouque [21] — which those authors had not carried out — to the lattice the instance defines.

The mechanism is therefore the one this section already models, and the two features that make it bite are both absent here. First, G is a fixed, public, *short* matrix, so the planted sublattice is spanned by the rows of $(S,\, G - E)$ and inherits the gadget's density; in NTC the corresponding object is $M_s A + e$ with A uniform over the whole of $\mathcal{M}_k$, and the planted module is the rank-n $\mathcal{P}$ of Corollary 1, of density $1/(2k)$ on the published lattice — which is precisely the quantity Section 5.6 measures rather than assumes. Second, [16] needs an overstretched q to obtain homomorphic depth at all, so the instance sits on the wrong side of the fatigue point by construction; the sets of Section 6.1 are selected to sit 13 to 19 bits below theirs. The lesson we take is not that NTC is safe from this attack but that the attack is the one its parameter selection is organised around.

iNTRU *over the integers.* Barthel, Müller and Roşie [2] break the integer case of iNTRU — the same relation with $R = \mathbb{Z}$ and the gadget vector $g_i = 2^i$ — by elementary lattice reduction, and sketch a generalisation to polynomial rings and to *low-dimensional* MiNTRU with *small* noise. The integer case has dimension $O(\log q)$; NTC has $nk \geq 512$, carries no gadget, and its inhomogeneous term is uniform rather than structured. The sketched generalisation is explicitly to the low-dimensional small-noise corner, which is not the one instantiated here.

Galois descent, and the necessity of uniform A. The richer structure of $\mathcal{A}$ opens one avenue the two attacks above do not: $\mathcal{A} = R_q[y]/(y^k - \alpha)$ carries the automorphism group of the extension, generated at $k = 2$ by $\sigma : (a_0, a_1) \mapsto (a_0, -a_1)$, and $\sigma(s)$ is as short as s, so an adversary might hope to enlarge the planted module by adjoining Galois images of the witness, making the dense sublattice denser.

The hope fails, and the obstruction is again the uniformity of A. Writing $D := \mathrm{diag}(1, -1)$, one has $M_{\sigma(a)} = D M_a D^{-1}$ for every $a \in \mathcal{A}$, so conjugating

the defining relation $t\,M_s = \mathrm{row}_1(M_s A) + e_1$ by D gives

$$\sigma(t)\,M_{\sigma(s)} \;=\; \mathrm{row}_1\big(M_{\sigma(s)}\,DAD^{-1}\big) \;+\; \tilde{e}, \qquad \|\tilde{e}\|_\infty = \|e_1\|_\infty.$$

The Galois image of the witness is therefore short, but it is a witness for the instance $(DAD^{-1}, \sigma(t))$, not for (A, t). Since A is uniform on $\mathcal{M}_k$ we have $DAD^{-1} \neq A$ except on a negligible fraction of instances, so the conjugate contributes no vector to the lattice the adversary actually holds, and $\mathcal{P}$ is not enlarged. Had A been drawn from the torus, or from any D-conjugation-invariant set, the conclusion would reverse; this is a second reason, independent of the norm map of Section 5.1, why Definition 2 insists on uniformity over all of $\mathcal{M}_k$.

We claim no exhaustiveness. What we claim is that the two published breaks in this family proceed by dense-sublattice discovery under an overstretched modulus, that this is the attack Section 5.6 models and now measures, and that the obvious algebraic route the torus opens up — Galois descent — closes for a reason that is checkable in one line.

5.4 Unequal secret and error widths

Both schemes use secrets and errors of different widths — deliberately so, since Section 4.1 shows that widening χ_r costs only linearly in the noise budget while χ_f costs quadratically. The $\|v\|$ appearing in (6) is then not a single $\sigma\sqrt{d}$, and the standard remedy is to rescale the secret block of the lattice by $\omega := \sigma_e/\sigma_s$ before applying the estimate, so that

$$d = n_{\mathrm{sec}} + m + 1, \qquad \mathrm{vol} = q^m\,\omega^{\,n_{\mathrm{sec}}}, \qquad \|v\| \approx \sigma_e\sqrt{d}.$$

This matters concretely, and the direction it points in is the awkward one. In the KEM the lattice secret is $s = 1 + p\,s'$, not s': with $p = 2$ and $\chi_{s'} = \mathcal{B}_3$ one has $\sigma_s^2 = p^2 \cdot \frac{3}{2} = 6$. The error, however, is *not* scaled by p: the compact scheme publishes $e_1 \leftarrow\!\$\ \chi_f^k$ and not $p\,e_1$ (Remark 5), so with $\chi_f = \mathcal{B}_2$ one has $\sigma_e^2 = \eta_f/2 = 1$ and the two widths differ by a factor $\sigma_s/\sigma_e = 2.45$. This is the largest such ratio anywhere in the paper, and it is what makes the choice of estimate matter here and not for Kyber, whose secret and error have the same width so that $\omega = 1$ and the two forms coincide.

Table 4. Effect of the secret-block rescaling on the KEM primal estimates, with the Kyber peers under the same estimator.

	NTC-KEM-512	NTC-KEM-768	NTC-KEM-1024
unrescaled, $\|v\|^2 = nk(\sigma_s^2 + \sigma_e^2)$ (as in Section 6.1)	118.6	185.1	256.1
rescaled, $\omega = \sigma_e/\sigma_s$	112.4	176.7	251.4
difference	−6.1	−8.5	−4.7
Kyber peer, same estimator	118.3	182.5	256.1

The rescaling therefore *lowers* the estimate rather than raising it, and the figures of Section 6.1 are the unrescaled ones. Which of the two to believe is no longer a matter of taste: Section 5.2 measures it, and the rescaled form is the one that tracks reduction, to about three blocksizes across three toy cells, while the unrescaled form runs 9 to 20 blocksizes high. Read that way the KEM sits some 5 bits below its Kyber peer rather than at parity with it, and the gap is asymmetric precisely because $\omega = 1$ for Kyber.

We do not revise the parameter sets on the strength of three cells at dimension ≤ 192, and we note the standing objection that the rescaling is exact only for a spherical secret. What we do withdraw is the claim that the quoted figures are conservative: on the evidence available they are optimistic, by the amounts above. Narrowing σ_s/σ_e — widening χ_f rather than $\chi_{s'}$, which would widen the ratio — is the lever that would recover the difference, at a cost in the noise budget that Section 4.4 would have to re-price.

5.5 The hybrid attack, and the necessity of calibrating the model

Howgrave-Graham's hybrid attack [20] guesses g coordinates of a narrow secret by meet-in-the-middle and solves the residual lattice problem in dimension $d-g$. It bites hardest exactly where the levers of Section 4.1 push, so it must be modelled rather than assumed away. Two models are in circulation and they disagree by tens of bits.

The strict model, ruled out by the Kyber calibration. Suppose the attacker runs a fresh uSVP reduction on the residual lattice for each guess. The cost is $\max\big(0.292\,\beta(d-g),\ g\,H(\chi_s)/2\big)$ optimised over g, with H the Shannon entropy per coefficient. Applied to Kyber-512 this predicts a hybrid gain of 22 bits over the primal attack, which the published analyses of that scheme exclude. We therefore do not use this model to select parameters, and report it only as a stress column.

The Babai/GSA model. The correct picture, standard since Wunderer [30], is that BKZ runs *once* on the residual basis, and that per guess the attacker performs a nearest-plane call which succeeds when the smallest Gram–Schmidt norm exceeds the noise standard deviation by a constant factor τ:

$$b^*_{\min} \;=\; \delta_\beta^{1-d'}\,\mathrm{vol}^{1/d'} \;\geq\; \tau\,\sigma, \qquad d' := d-g, \tag{7}$$

which determines the required δ_β, hence β, hence the cost $\max\big(0.292\beta,\ gH/2\big)$, again optimised over g and over the number of samples. The single free constant τ must be fitted. Calibrating on Kyber, and reporting the quantity $\mathrm{cost}_{\mathrm{hyb}} - \mathrm{cost}_{\mathrm{primal}}$ (negative means the hybrid is cheaper, i.e. the attacker gains):

Table 5. Calibration of the hybrid constant τ on Kyber: hybrid minus primal cost, in bits (negative means the hybrid is cheaper).

τ	1.0	1.5	2.0	3.0
Kyber-512	−19.6	−6.6	+2.8	+17.8
Kyber-768	−27.9	−11.6	+3.2	+22.2

At $\tau = 1.0$ the model would hand the attacker some 20 to 28 bits on Kyber, which the published analyses exclude; at $\tau = 3.0$ it concedes nothing at all and is uninformative. The crossing — the value at which the hybrid ceases to threaten Kyber — lies between 1.5 and 2.0, and we adopt $\tau = 2.0$, consistent both with the published consensus and with the theoretical picture in which nearest-plane decoding succeeds once $b^*_{\min}$ exceeds σ by roughly a factor of two. We note explicitly that $\tau = 2.0$ sits *at* the crossing and not comfortably beyond it, which is why we tabulate the sensitivity of the resulting figures to this one constant instead of describing it.

Consequences. The width question at the KEM instances and the replacement of the geometric series assumption by the Chen–Nguyen profile simulator — checked against the measured Gram–Schmidt profiles of Section 5.6 — are developed in Appendix B; the outcome is what follows.

Under the profile simulator the hybrid binds nowhere, and the primal figures of Section 6.1 are the reference at every parameter set in this paper; the GSA closed form is retained as a stress model only, and it is in that capacity that the deployed-width KEM figures above should be read.

5.6 Dense sublattice discovery and the fatigue threshold

The one attack specific to planted constructions is dense sublattice discovery (DSD), which past a modulus threshold — the *fatigue point* — finds the plant faster than (6) predicts [21,11]. This is the binding *upper* constraint on q; it closes in on the correctness requirement from the other side, and the resulting window is what Section E discusses. Since the whole point of the torus structure is to widen that window, this subsection carries more of the paper's weight than any other, and we therefore measure the threshold rather than extrapolate it.

The relevant density is that of the published *lattice.* A point of bookkeeping first, because it is easy to get wrong. The full NTC lattice has dimension $nk(1+k)$ and carries a plant of rank n (Corollary 1), a density of $1/\big(k(k+1)\big)$. But the scheme does not publish that lattice: it exposes a single row (Section 4.1), so what an attacker actually holds is

$$\Lambda \;=\; \big\{(s,e)\;:\;L_1(s)+e \equiv 0 \bmod q\big\}, \qquad \dim\Lambda = 2nk, \quad \det\Lambda = q^{nk}, \quad \operatorname{rank}\mathcal{P} = n, \tag{8}$$

whose density is $1/(2k)$, that is $1/4$ at $k = 2$, against NTRU's $1/2$. Both formulas agree at $k = 1$. Structurally, therefore, the deployed lattice is an NTRU lattice of half-dimension nk whose plant is k times sparser — and it is this factor, not the factor $k(k+1)/2$ of the full lattice, that the simulation below must be run against.

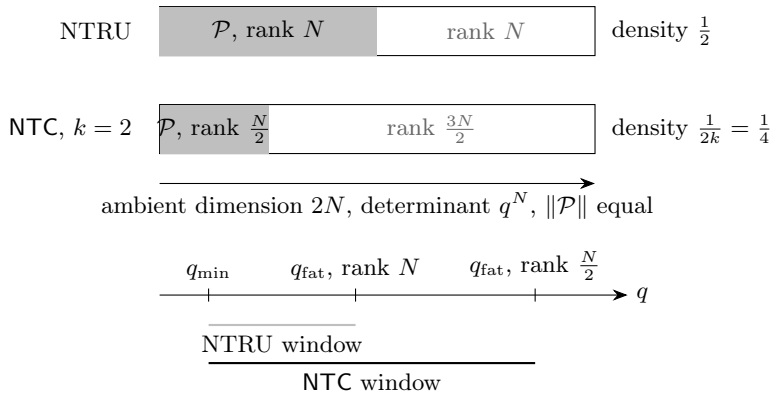

Fig. 1. The structural claim of this section. Top: on the published lattice (8) of dimension $2N$, at equal determinant and equal planted-vector norm, the NTC plant occupies rank $N/2$ against NTRU's rank N. Bottom: the sparser plant pushes the fatigue point — the modulus past which dense-sublattice discovery beats key recovery — upward, widening the usable window between the correctness lower bound $q_{\min}$ and the fatigue upper bound. The predicted gain is 4.5 to 6.3 bits of modulus (Table 7); the measured brackets confirm the ordering at every modulus reduced (Table 9).

Method. We adapt the DSD-PT estimator of Ducas and van Woerden [11] to a q-ary lattice carrying a dense sublattice of *arbitrary* rank. Their model is stated for $\dim = 2n$ with a plant of rank n, so that the ambient half-dimension, the plant rank and their difference all coincide; the generalisation requires disentangling the three. Following their Claim 4.2, the recursion on the volumes of the intersections $\mathcal{P} \cap \Lambda_{[0:s)}$ projects a dual vector living in dimension $s+1$ onto the l-dimensional span of the intersection, so the quantity to be generalised is the *co-rank* $d - r$ and not the half-dimension. Concretely, for ambient dimension d and plant rank r the recursion runs over $s = d-1, \ldots, d-r+1$ with intersection rank $s-(d-r)$, and the projection loss term becomes $\frac{1}{2}\big(\psi(\frac{s+1-(d-r)}{2}) - \psi(\frac{s+1}{2})\big)$. The average log-volume of the plant is that of a rank-one R-module in R^K with $K = 2k$ ring coordinates; for the rings of Section 2.1, all of whose embeddings are complex,

$$\mathbb{E}\big[\ln \mathrm{vol}(\mathcal{P})\big] \;=\; \tfrac{n}{2}\Big(\ln\big(n\sigma^2\big) + \psi(K)\Big), \tag{9}$$

which specialises at $K = 2$ to the circulant formula of [11] up to the single real embedding carried by the circulant ring; unequal secret and error widths are handled by numerical integration of the corresponding weighted χ^2.

Validation. Two checks, both necessary before any of the numbers below may be read. First, instantiated on the NTRU geometry ($d_q = d_1 = n$, $r = n$, $K = 2$) the generalised estimator reproduces the original one *exactly*: over a grid of moduli and blocksizes the largest discrepancy on the DSD probability is 0 to machine precision. Second, the fatigue point it predicts for ternary circulant NTRU is

Table 6. Fatigue points predicted by the generalised estimator for ternary circulant NTRU, against the published fit $q_{\text{fat}} \approx 0.004\, n^{2.484}$.

n	79	101	127	151
q_{fat}	233	427	744	1086
$q_{\text{fat}}/n^{2.484}$	0.0045	0.0045	0.0044	0.0042
β at the crossing	20	29	48	64

against the published $q_{\text{fat}} \approx 0.004\, n^{2.484}$ and a crossing blocksize in the range 30–60: both the constant and the blocksize regime are reproduced.

The density experiment. Fixing the lattice dimension *and* the secret vector norm and varying only the plant rank isolates the effect the torus is meant to produce. At $k = 2$ the NTC plant vectors have squared norm $4 \cdot (N/2) \cdot \sigma^2 = 2N\sigma^2$, identical to NTRU's, so the comparison is clean:

Table 7. Predicted fatigue points at fixed lattice dimension and fixed planted-vector norm, varying only the plant rank.

$N = nk$	160	192	224
q_{fat}, plant of rank N (NTRU)	1216	1919	2819
q_{fat}, plant of rank $N/2$ (NTC)	27751	85049	215509
gain (bits of modulus)	4.51	5.47	6.26

The sparser plant does raise the fatigue point, and the gain *grows* with the dimension rather than settling at a constant factor. This is the quantitative form of the structural claim made in Section 1.1.

The parameter sets. Applied to the sets of Section 6.1, at the deployed modulus the DSD event never triggers — the attack proceeds by secret-key recovery, at a blocksize agreeing with the primal estimate of Section 5.2 to within 5% — and the fatigue point sits far above:

Table 8. Fatigue margins of the KEM parameter sets.

	NTC-KEM-512	NTC-KEM-768	NTC-KEM-1024
q	7681	10369	13313
β at q, Pr[DSD]	389, 0	615, 0	855, 0
q_{fat}	$6.9 \cdot 10^7$	$7.4 \cdot 10^8$	$5.3 \cdot 10^9$
margin	13.1 bits	16.1 bits	18.6 bits

The KEM margins supersede the 4.2/5.2/6.1 bits that a rescaling of the ternary fit yields: the heuristic estimate was conservative by roughly 4 to 8 bits of q_{fat} from category 2 to category 5, for the reason now made precise — it charges NTC the plant density of NTRU.

Experimental confirmation. The figures above are now backed by lattice reduction, not by the predictor alone. We generate instances of the published lattice (8) at ring degree n and admissible modulus $q \equiv 2n + 1 \bmod 4n$, together with NTRU instances at the *same* ambient dimension $2N$, the *same* determinant q^N, the *same* modulus and the *same* planted-vector norm — writing $N = nk$, every coarse invariant coincides and the two families differ in the rank of the plant (N against $N/2$) and in nothing else. Each instance is reduced by progressive BKZ (LLL, then blocksizes $3, 4, \ldots$ at eight tours each, the schedule the estimator assumes), and after every tour the basis is tested for vectors of the planted module. The test is exact rather than a norm heuristic: v lies in the rational span of $\mathcal{P}$ if and only if $v_i w_j - v_j w_i = 0$ in $\mathbb{Q}[x]/(x^n + 1)$ for every pair of ring coordinates, w being the planted witness. Following [11] the first event is scored SKR when the vector found is no longer than w and DSD when it is longer; a tour producing both at once is scored SKR, a rule applied to both families alike and which can only *under*-count DSD. The reduction is fplll's enumeration-based BKZ [12]; every reported vector is re-verified to lie in the lattice in exact integer arithmetic, so a numerical accident can lose an event but cannot invent one.

Table 9. Measured DSD-against-SKR outcomes and fatigue brackets, fplll BKZ reduction at fixed dimension, determinant and planted-vector norm.

$N = nk$ (ambient dimension $2N$)	96	128
NTRU, plant of rank N: DSD first	109/109	9/9
NTC, plant of rank $N/2$: DSD first	4/109	0/6
q_{fat}, rank N	≤ 673	≤ 2689
q_{fat}, rank $N/2$	≥ 19489	≥ 28289
measured gain (bits of modulus)	≥ 4.86	≥ 3.40
predicted gain	4.32	4.24

At every modulus reduced, the rank-N plant is discovered as a dense sublattice and the rank-$N/2$ plant is not. Neither crossing falls inside the grid — NTRU is already overstretched at the cheapest modulus within reach and NTC is still not overstretched at the dearest — so the gain is bounded rather than located, from the side that does not flatter NTC. Where the comparison can be made cell by cell it goes the same way: at $N = 128$ the estimator assigns the six NTC moduli DSD probabilities of $0.00, 0.01, 0.19, 0.48, 0.81, 0.96$ as q rises, and all six reductions returned SKR, an outcome of probability 0.003 under the model. The estimator therefore calls fatigue *earlier* than reduction delivers it, here as on the NTRU controls below. The ordering the torus was designed to produce is observed on reduced bases, at the plant density the deployed lattice actually carries, and the predictor errs on the safe side of it.

Harness calibration, and limitations. Run on circulant NTRU — the geometry for which the source model publishes experimental fatigue points — the same harness brackets the published crossings to within 20% at $n = 79$ and 2% at $n = 101$, early in both cases, which is the conservative direction; the measured Gram–Schmidt slopes at the event match the estimator's Z-shape to three digits. The calibration data and the three limitations of the simulation — grid coarseness, degenerate crossings at $\beta < 20$, and unmodelled lucky-lift events — are recorded in Appendix C.

5.7 Unmodelled attacks and effects

What follows is what the analysis above does not cover.

Algebraic attacks specific to the torus. We have excluded the norm map, the adjugate linearisation and slotwise CRT attacks (Section 5.1), but the non-split torus is a rich object and we make no claim of exhaustiveness. The trace descent $\mathrm{Tr}_{\mathcal{A}/R_q}$, in particular, yields module-LWE of rank k in dimension nk with no reduction to dimension n (Remark 8), but that is an observation, not a proof that no descent helps.

The column correlation in $\mathsf{cLWE}_{\mathcal{A}}$. The full matrix sample of Definition 8 exposes k correlated short secrets $T_j\mathrm{coef}(r)$ under a common B, where T_j is public. Applying B^{-1} fails because $B^{-1}F$ is not short, but the correlation is genuine structure and would deserve dedicated scrutiny — for the reference design. The deployed schemes never expose it: the reduction of Theorem 1 consumes only the column- and row-truncated samples, which are module-LWE and ring-LWE over $\mathcal{A}$ with a shared secret (Remark 8).

Decryption-failure attacks. We enforce $\delta \leq 2^{-140}$, and δ is now computed from the exact distribution of D_1 rather than from a Gaussian matched to its variance (Section 4.5); the correction is 5 to 12 bits and every set still clears the constraint. What remains unmodelled is the *use* an adversary could make of failures: $q_G\delta$ bounds the probability that one occurs, not the information a chosen-ciphertext attacker extracts from a correlated sequence of them, and the directional-failure literature for Kyber-like schemes has no counterpart here.

Multi-instance and multi-target settings. All figures are single-target. Definition 7 records that exposing several t_j under a common s must be re-estimated separately; the KEM uses $\ell = 1$, the most conservative form, and keys must not be shared across primitives (Remark 13).

Implementation attacks. Side channels, fault injection against the rejection loop, and the timing behaviour of the two distinct number-theoretic transforms are out of scope.

Attack / effect	Modelled how	Status
Primal uSVP	2016 estimate, min over two lattices	1.2 bits on Kyber; **loose by** 3–6 **bits under reduction at** $2nk \leq 192$
Column subselection	optimised over $m \leq k$	$m^* = 1$ at every set
Hybrid MitM	Babai/GSA, $\tau = 2.0$	calibrated on Kyber; GSA over-states $b^*_{\min}$ by 1.31× against measurement and ceases to bind under a simulator; stress model only, see §5.5
Dense sublattice disc.	adapted DvW simulator	BKZ run, $2N \leq 256$; **ordering only**
Norm / adjugate / CRT	excluded structurally	see Section 5.1
$\mathsf{cLWE}_{\mathcal{A}}$ columns	—	untruncated variant only; not consumed by the deployed scheme
Decryption failure	exact tail, saddlepoint	Gaussian was optimistic by 5–12 bits; all sets still clear 2^{-140}

Table 10. Summary of the cryptanalytic model. Entries in bold are the ones a parameter set cannot be considered validated without. The dense-sublattice row is the one that carries the most weight: the attack is simulated, and the simulation is checked against actual lattice reduction on NTC instances (Section 5.6). What that experiment establishes is the *ordering* — the sparser plant survives moduli at which NTRU's is already discovered — at dimensions and blocksizes far below the deployed ones; the *size* of the margins quoted for the parameter sets still rests on the predictor alone.

6 Parameter sets for the KEM

We instantiate the scheme of Section 4.3 at NIST security categories 1, 3 and 5. Parameters are chosen by maximising the primal figure of Section 5, evaluated on both the NTC and the $\mathsf{cLWE}_{\mathcal{A}}$ lattice, subject to three constraints: the exact decryption failure probability of Section 4.5 must satisfy $\delta \leq 2^{-140}$; q must be prime with $q \equiv M+1 \pmod{2M}$, so that $\alpha = x$ is non-split (Lemma 3); and the set must keep a margin of at least two bits below the dense-sublattice fatigue estimate of Section 5.6. The hybrid attack does not enter the objective: under the profile simulator of Section 5.5 it binds at no parameter set, and the GSA closed form is tracked as a stress model only.

6.1 The sets

The search converges to $\eta_{s'} = 3$ ($H \approx 2.33$ bits/coord, matching Kyber-512's secret entropy), $\eta_r = 12$ (wide, since χ_r costs only linearly in the noise budget), and a per-category η_f — 3, 2 and 4 at categories 1, 3 and 5 — set jointly by the exact failure constraint of Section 4.5 and the security columns, as the tuning below (Table 12) details for category 1.

Table 11. KEM parameter sets.

	NTC-KEM-512	**NTC-KEM-768**	**NTC-KEM-1024**
NIST category	1	3	5
ring	$x^{256}+1$	Φ_{1152}	$x^{512}+1$
n, k	256, 2	384, 2	512, 2
q	7681	10369	13313
$\log_2 q$	12.91	13.34	13.70
$(\eta_f, \eta_r, \eta_{s'})$	$(3, 12, 3)$	$(2, 12, 3)$	$(4, 12, 3)$
compression d	12	12	12
failure δ, exact	2^{-149}	2^{-158}	2^{-157}
NTC primal, unrescaled	119.7	185.1	256.1
NTC primal, rescaled	115.6	176.7	251.4
cLWE primal, unrescaled	120.0	186.3	258.4
cLWE primal, rescaled	116.2	178.1	254.0
security (min)	**115.6**	**176.7**	**251.4**
fatigue margin	13.1 bits	16.1 bits	18.6 bits
public key	864 B	1376 B	1824 B
ciphertext	1536 B	2304 B	3072 B
secret key	224 B	320 B	416 B

Three reading conventions, each fixed by a finding of Section 5. No hybrid rows appear: the Babai/GSA model over-states $b^*_{\min}$ by a factor 1.31 against measured Gram–Schmidt profiles, and under a profile simulator the attack does not bind at any set here (Section 5.5), so the primal attack is the reference throughout. The primal figures are given in both forms, since Section 5.2 finds the rescaled one to be the one reduction follows, and the minimum is taken over the rescaled column. And the category-1 set is the outcome of a tuning against the exact failure probability of Section 4.5, detailed next; the rescaled figures at categories 3 and 5 are reported, but no such adjustment is needed there.

How the category-1 set was tuned. The natural starting point is the *baseline* candidate $(\eta_f, \eta_r, \eta_{s'}) = (2, 12, 3)$ at $d = 11$: the optimum a selection against the Gaussian δ of Section 4.4 and the unrescaled estimate would return. Under the rescaled form of (6) — the one Section 5.2 finds reduction to follow — that candidate stands at 112.4 rather than 118.6, some 6 bits below its Kyber-512

peer. The gap is entirely the asymmetry $\sigma_s/\sigma_e = 2.45$, which is 1 for Kyber, and narrowing it means widening χ_f. That is the expensive lever: σ_f^2 multiplies five of the eight terms of (5), and raising η_f from 2 to 4 at the baseline q and d drives the exact δ of Section 4.5 from 2^{-144} to 2^{-93}. It has to be paid for elsewhere, and the adopted set pays with one bit of compression, which quarters the rounding variance and buys back far more than η_f costs:

Table 12. Tuning of the category-1 set against the exact failure probability (all rows at $q = 7681$, $\eta_{s'} = 3$; security columns rescaled).

	η_f	η_r	d	$\log_2 \delta$ (Gaussian)	$\log_2 \delta$ (exact)	NTC	cLWE	**min**	ct
baseline	2	12	11	−156.3	−144.2	112.4	113.0	112.4	1408 B
adopted	3	12	12	−164.4	−148.7	115.6	116.2	**115.6**	1536 B
rejected	4	8	12	−140.9	*−128.6*	118.3	115.6	—	1536 B

(All at $q = 7681$, $\eta_{s'} = 3$; security columns rescaled.) The adopted set buys 3.2 bits for 128 bytes of ciphertext, and changes two parameters: η_f from 2 to 3, and one bit of compression. The third row is instructive and is why the compression bit is spent where it is. Selecting against the *Gaussian* δ of Section 4.4 makes $(4, 8, 3)$ look like the better set — it reaches 118.3 on the NTC column, exactly Kyber-512's figure — but the exact tail of Section 4.5 puts its δ at $2^{-128.6}$, so it violates the correctness constraint by 11 bits and is not admissible. Widening χ_f to $\eta_f = 4$ is simply out of reach at this modulus.

Parity is not reached. Under the exact δ, a search over $q \in [6 \cdot 10^3, 3 \cdot 10^4]$, $\eta_f \le 6$, $\eta_r \le 16$, $\eta_{s'} \le 4$ and $d \le 12$ tops out at 115.9, at $q = 10753$ with $(\eta_f, \eta_r, \eta_{s'}) = (6, 14, 3)$: three bits below Kyber-512 is where this ring and this correctness constraint leave category 1, and the adopted set gives up only 0.3 of those three in exchange for keeping q, η_r and $\eta_{s'}$ where they were. Closing the rest would take a change of ring, not a change of widths.

The basis of the category-1 label. Three bits below Kyber-512 under a shared estimator is a statement about an exponent; the label "category 1" is a statement about cost, with the floor defined as key search on AES-128. The core-SVP exponent deliberately discards both the polynomial gate factors and the memory cost of sieving, which every gate-level accounting prices in the tens of bits, and it is under such an accounting — not under a reading of bare exponents as gate counts — that Kyber-512 itself clears the floor. We therefore state the dependence rather than leave it implicit: NTC-KEM-512 is category 1 under any costing of sieving in which Kyber-512 is category 1 with three bits to spare, and the claim we actually defend is the relative one, parity minus three bits under one calibrated estimator. A reader who requires the full Kyber-512 margin at

category 1 should either await the change of ring above or deploy NTC-KEM-768; we do not present the label as unconditional.

One further property of the adopted set is worth more than the 3.2 bits. On the baseline candidate the two forms of the estimate disagreed by 6.2 bits (118.6 against 112.4), so which one to believe decided the category; on the adopted set the minimum moves by 3.5 bits (119.7 against 115.6) and, more to the point, the two lattices now sit within 0.6 bits of each other rather than being separated by whichever column the modelling choice favours. The tuning narrows the anisotropy that made the question sharp.

Design observations, deferred. Why $k = 2$ is forced by the ciphertext accounting, why Φ_{1152} wins at category 3 and only there, why the encryption randomness settles wide at $\eta_r = 12$, and the two-transforms trade-off are collected in Appendix E.

7 Implementation outline

Every ring in this paper admits a full number-theoretic transform, and the operation counts are within small factors of their Kyber counterparts: the price of the torus is a factor ≤ 4 on ring multiplications, partially offset by Karatsuba in $\mathcal{A}$, and a reduced lazy-reduction headroom at the larger moduli. Appendix G develops the transforms, the arithmetic in $\mathcal{A}$, per-primitive operation counts, vectorisation and constant-time considerations, and lists the constants that only an implementation can settle (Appendix G.7); Section 8 settles the ones our implementation reaches, by measurement.

8 Performance benchmarking

Two C implementation trees accompany this paper:[4] a portable reference tree, whose ring product is the schoolbook $O(n^2)$ convolution and whose inversion is a scalar constant-time divstep, and an AVX2 tree implementing the transforms of Appendix G. The AVX2 tree replaces three things at once: the schoolbook product by the number-theoretic transform, the scalar inversion by a lane-local one in the transform domain, and the Keccak core by a four-way SIMD permutation. Both trees, and an independent Python reference, emit byte-identical known-answer vectors, so the two columns of Table 13 are the same function computed two ways.

All figures are median reference cycles per operation, read from `rdtscp`: a warm-up, then repeated runs under a fixed wall-clock budget of 0.4 s per operation with at least 15 iterations, on a 12th-generation Intel Core i7-1265U (invariant TSC at 2.688 GHz, so these are cycles at the base frequency, not at the current core clock), compiled with gcc 11.4.0; measured 2026-08-14. The peers of Table 14 are timed by the same harness in the same binary, through

[4] `https://github.com/Naibeye/ntc_suite`

liboqs [13] on its own x86_64/AVX2 backends where it ships them, so no cross-tool calibration is involved.

Table 13. Median cycles per operation across the two implementation trees.

set	op	reference	AVX2	speed-up
NTC-KEM-512	KeyGen	1 896 780	81 926	23.2×
	Encaps	696 298	94 084	7.4×
	Decaps	922 398	110 322	8.4×
NTC-KEM-768	KeyGen	3 784 506	116 026	32.6×
	Encaps	1 149 712	143 444	8.0×
	Decaps	1 607 884	155 416	10.3×
NTC-KEM-1024	KeyGen	7 750 060	121 090	64.0×
	Encaps	2 090 652	157 856	13.2×
	Decaps	2 964 712	187 692	15.8×

KeyGen gains the most at every set, because the routine that dominated it — the constant-time inversion of s in $\mathcal{A}$ — is the one the transform domain makes cheap: what was a divstep over the whole ring becomes an independent inversion in each lane. The speed-ups also grow with the category, as they must: the reference product is quadratic in the ring dimension and the transform quasilinear, so the ratio between the columns widens as n does.

Table 14. The KEM against the NIST-track peers, per category. Sizes in bytes, bandwidth = pk + ct; median cycles per operation, NTC-KEM rows from the AVX2 tree, peers through liboqs.

algorithm	**pk**	**ct**	**BW**	**KeyGen**	**Encaps**	**Decaps**
category 1						
ML-KEM-512	800	768	1 568	14 866	16 902	19 722
NTC-KEM-512	864	1 536	2 400	81 926	94 084	110 322
BIKE-L1	1 541	1 573	3 114	326 594	72 300	1 248 282
FrodoKEM-640	9 616	9 752	19 368	1 712 472	1 893 166	1 850 194
HQC-1	2 241	4 433	6 674	2 672 416	5 362 516	8 085 698
McEliece-348864	261 120	96	261 216	79 500 120	24 908	260 164
category 3						
ML-KEM-768	1 184	1 088	2 272	24 144	25 756	30 754
NTC-KEM-768	1 376	2 304	3 680	116 026	143 444	155 416
BIKE-L3	3 083	3 115	6 198	916 744	164 958	3 801 658
FrodoKEM-976	15 632	15 792	31 424	3 616 182	3 874 266	3 811 984
HQC-3	4 514	8 978	13 492	8 074 570	16 253 790	25 025 574
category 5						
ML-KEM-1024	1 568	1 568	3 136	33 670	35 594	42 566
NTC-KEM-1024	1 824	3 072	4 896	121 090	157 856	187 692
BIKE-L5	5 122	5 154	10 276	2 316 680	345 732	9 431 424
FrodoKEM-1344	21 520	21 696	43 216	6 527 558	7 002 278	6 753 240
HQC-5	7 237	14 421	21 658	19 528 498	38 886 848	61 238 658

Two readings of Table 14. The cycle columns place the AVX2 tree within a small factor of ML-KEM — 5.5×, 4.8× and 3.6× its KeyGen at categories 1, 3 and 5, with Encaps and Decaps in the same bracket — and ahead of every non-module-LWE peer on every operation at every category, usually by one to two orders of magnitude. The gap to ML-KEM narrows as the category grows, which is the expected shape: the fixed price of the torus (the factor ≤ 4 on ring multiplications of Appendix G.3, and the Keccak share of each operation) is diluted as the ring work grows. The bandwidth column, unlike the cycle columns, is implementation-independent: NTC-KEM sits between ML-KEM and BIKE at every category, at roughly 1.5× ML-KEM, which is the position the row truncation of the public key and ciphertext was designed to buy.

The measurements also settle the third, and per Appendix G.7 the most deployment-relevant, of the questions the operation counts leave open: whether the radix-3 stage of the Φ_{3N} transform makes category 3 disproportionately slow. It does not. NTC-KEM-768 runs at 0.96×, 0.91× and 0.83× the cycles of NTC-KEM-1024 on the three operations — more expensive per coefficient, since its ring is 3/4 the size, but comfortably between categories 1 and 5 in absolute terms — so the single-transform alternative of Appendix E is not warranted on performance grounds. The first two questions — how much of the factor ≤ 4 instruction-level parallelism absorbs, and the price of the reduced lazy-reduction

headroom — concern individual routines that this end-to-end harness does not isolate; the aggregate ratios to ML-KEM above bound their combined effect, and a per-routine profile is left to future work.

Conclusion and open problems

The question this work set out to answer was whether anything useful lives between NTRU and module-LWE: between a planted construction whose economy is paid for by an overstretched regime, and a pseudorandom one whose worst-case guarantee is paid for by size. NTC is one answer. It keeps NTRU's plant and changes what the plant sits in, confining the secret to a non-split maximal torus of $\mathrm{GL}_k(R_q)$ and taking the instance to be a noisy conjugation rather than a quotient. The two structural consequences are the ones that motivated the design and that we have been able to establish: the planted module occupies a fraction $1/(2k)$ of the published lattice's dimension instead of NTRU's $1/2$, and the norm-map descent that would collapse the problem to dimension n is unavailable once A is uniform over the full matrix algebra. Both point in the same direction, towards a higher fatigue threshold and hence a wider usable modulus window than NTRU enjoys.

Both claims are now backed by a dedicated simulation rather than by rescaling a fit from a different density regime: Section 5.6 adapts the dense-sublattice estimator of [11] to arbitrary plant rank, validates the adaptation against its authors' own NTRU figures, and finds a gain of 4.5 to 6.3 bits of modulus at equal lattice dimension — growing with dimension — which leaves the parameter sets 13 to 19 bits below their fatigue point. The prediction that the sparser plant is the harder one to discover is also, now, an observation: running BKZ on small instances of the published lattice and on NTRU controls of identical dimension, determinant, modulus and planted-vector norm, the rank-N plant is found as a dense sublattice at every modulus reduced and the rank-$N/2$ plant at essentially none. The predictor is, however, still used far beyond the dimensions in which either its authors or we have checked it, and the experiment settles the ordering rather than the size of the gain; the margins should be read with that caveat, made precise in Section 5.6.

The instantiations show that the structure is not merely of theoretical interest. Under a single estimator, calibrated to 1.2 bits against Kyber's published core-SVP figures, the KEM reaches public keys within 1.08 to 1.16 times Kyber's and ciphertexts 2.0 to 2.1 times Kyber's. The remaining gap is concentrated in the KEM ciphertext and is traceable to the nk^2 term that the matrix reading imposes on c_2.

The status of the assumption

NTC has no worst-case reduction, and neither does any member of the NTRU-with-errors family to which it belongs. Its decision form must be assumed separately from its search form, because a single key yields a single instance, and

the torus orbit (codimension $k^2 - k$) is too small to rerandomise it. The rigidity theorem that makes the search problem well posed rests on the Gaussian heuristic, and the fatigue margins that bound the modulus from above come from a predictor validated against published NTRU figures but used beyond the dimensions and blocksizes in which it was checked. Lattice reduction has now been run on NTC instances, and it confirms the structural claim the margins rest on — a sparser plant is discovered later — but only at $2N \leq 256$ and at crossing blocksizes an order of magnitude below the deployed ones, so it corroborates the direction of the margins and not their size. What we have supplied is a first round of cryptanalysis, structured so that a reader can see precisely which attacks are modelled and which are not (Table 10); what we have not supplied is the years of scrutiny that would make the assumption load-bearing.

Open problems

Five items, in decreasing order of the weight we attach to them. First, the fatigue experiment at deployed scale: the reduction of Section 5.6 settles the ordering at $2N \leq 256$, and the single most valuable follow-up is the same measurement at dimensions where the crossing blocksizes approach the deployed ones, which would turn the margins from predicted into observed. Second, Conjecture 1: a regularity lemma for the completely split rings the transform uses would upgrade the variant of Appendix F from conjectural to proved, and the modulus class $q \equiv 3, 5 \pmod 8$ of Remark 16 suggests the obstruction is an artefact of the proof rather than of the object. Third, a reduction between the search and decision forms of the assumption, or a rerandomisation argument that would let one imply the other; the torus orbit is too small as it stands, and we do not know whether that is essential. Fourth, whether $k \geq 3$ buys anything once the ciphertext accounting that forces $k = 2$ here is absent — the planted density falls to $1/(2k)$ on the published lattice, and a primitive free of the ciphertext accounting is the natural place to test whether the fatigue gain survives the coarseness of the ring grid. Fifth, an implementation: Appendix G.7 lists the three constants that only measurement can settle.

References

1. Erdem Alkim, Léo Ducas, Thomas Pöppelmann, and Peter Schwabe. Post-quantum key exchange – a new hope. In *USENIX Security Symposium*, 2016.
2. Jim Barthel, Volker Müller, and Răzvan Roşie. On the (M)iNTRU assumption in the integer case. In *Provable and Practical Security (ProvSec)*, 2021.
3. Anja Becker, Léo Ducas, Nicolas Gama, and Thijs Laarhoven. New directions in nearest neighbor searching with applications to lattice sieving. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2016.
4. Daniel J. Bernstein, Chitchanok Chuengsatiansup, Tanja Lange, and Christine van Vredendaal. NTRU prime: Reducing attack surface at low cost. In *Selected Areas in Cryptography (SAC)*, 2017.

5. Joppe W. Bos, Léo Ducas, Eike Kiltz, Tancrède Lepoint, Vadim Lyubashevsky, John M. Schanck, Peter Schwabe, Gregor Seiler, and Damien Stehlé. CRYSTALS-Kyber: A CCA-secure module-lattice-based KEM. In *IEEE European Symposium on Security and Privacy (EuroS&P)*, 2018.
6. Jung Hee Cheon, Duhyeong Kim, Taechan Kim, and Yongha Son. A new trapdoor over module-NTRU lattice and its application to ID-based encryption. IACR Cryptology ePrint Archive, Report 2019/1468, 2019.
7. Chitchanok Chuengsatiansup, Thomas Prest, Damien Stehlé, Alexandre Wallet, and Keita Xagawa. ModFalcon: Compact signatures based on module-NTRU lattices. In *ACM Asia Conference on Computer and Communications Security (ASIACCS)*, 2020.
8. Sidoine Djimnaibeye, Aminata Ngom, and Djiby Sow. Post-quantum cryptography for healthcare: A number theory based two-factor mutual authentication and key exchange protocol over lattices for TMIS. *Advances & Applications in Discrete Mathematics*, 41(1), 2024.
9. Sidoine Djimnaibeye et al. NAIBI: Binding reconciliation KEMs and ephemeral key agreement over non-split commutative algebras. IACR Cryptology ePrint Archive, 2026.
10. Léo Ducas, Eike Kiltz, Tancrède Lepoint, Vadim Lyubashevsky, Peter Schwabe, Gregor Seiler, and Damien Stehlé. CRYSTALS-Dilithium: A lattice-based digital signature scheme. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2018.
11. Léo Ducas and Wessel P. J. van Woerden. NTRU fatigue: How stretched is overstretched? In *ASIACRYPT*, 2021.
12. The FPLLL development team. fplll, a lattice reduction library, Version 5.4. Available at `https://github.com/fplll/fplll`, 2023.
13. Douglas Stebila and Michele Mosca. Post-quantum key exchange for the Internet and the Open Quantum Safe project. In *Selected Areas in Cryptography (SAC)*, 2016. liboqs available at `https://github.com/open-quantum-safe/liboqs`.
14. Eiichiro Fujisaki and Tatsuaki Okamoto. Secure integration of asymmetric and symmetric encryption schemes. In *CRYPTO*, 1999.
15. Joel Gärtner. NTWE: A natural combination of NTRU and LWE. In *Post-Quantum Cryptography (PQCrypto)*, 2023.
16. Nicholas Genise, Craig Gentry, Shai Halevi, Baiyu Li, and Daniele Micciancio. Homomorphic encryption for finite automata. In *ASIACRYPT*, 2019.
17. Jeffrey Hoffstein, Jill Pipher, and Joseph H. Silverman. NTRU: A ring-based public key cryptosystem. In *Algorithmic Number Theory (ANTS-III)*, 1998.
18. Dennis Hofheinz, Kathrin Hövelmanns, and Eike Kiltz. A modular analysis of the Fujisaki-Okamoto transformation. In *Theory of Cryptography Conference (TCC)*, 2017.
19. Abiodoun Clément Hounkpevi, Sidoine Djimnaibeye, and Michel Seck. EagleSign: A new post-quantum ElGamal-like signature over lattices. Submission to the NIST post-quantum cryptography standardization process, 2023.
20. Nick Howgrave-Graham. A hybrid lattice-reduction and meet-in-the-middle attack against NTRU. In *CRYPTO*, 2007.
21. Paul Kirchner and Pierre-Alain Fouque. Revisiting lattice attacks on overstretched NTRU parameters. In *EUROCRYPT*, 2017.
22. Thijs Laarhoven, Michele Mosca, and Joop van de Pol. Finding shortest lattice vectors faster using quantum search. *Designs, Codes and Cryptography*, 2015.
23. Adeline Langlois and Damien Stehlé. Worst-case to average-case reductions for module lattices. *Designs, Codes and Cryptography*, 2015.

24. Changmin Lee and Alexandre Wallet. Lattice analysis on MiNTRU problem. IACR Cryptology ePrint Archive, Report 2020/230, 2020.
25. Vadim Lyubashevsky, Chris Peikert, and Oded Regev. On ideal lattices and learning with errors over rings. In *EUROCRYPT*, 2010.
26. Vadim Lyubashevsky and Gregor Seiler. NTTRU: Truly fast NTRU using NTT. *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2019.
27. National Institute of Standards and Technology. Module-lattice-based digital signature standard, 2024. FIPS 204.
28. National Institute of Standards and Technology. Module-lattice-based key-encapsulation mechanism standard, 2024. FIPS 203.
29. Damien Stehlé and Ron Steinfeld. Making NTRU as secure as worst-case problems over ideal lattices. In *EUROCRYPT*, 2011.
30. Thomas Wunderer. Revisiting the hybrid attack: Improved analysis and refined security estimates, 2016. IACR Cryptology ePrint Archive, Report 2016/733.

A The primal estimator under reduction, and the width question

To test the estimate itself, and not merely our agreement with others in computing it, we run it against key recovery on scaled-down instances of the KEM's own key lattice — the secret $s = 1 + ps'$ with $s' \leftarrow\!\!\$\ \mathcal{B}_3$ and the error $e_1 \leftarrow\!\!\$\ \mathcal{B}_2$ unscaled, so that the secret block is genuinely wider than the error block. For each cell we locate the smallest blocksize at which a *single* BKZ-β on an LLL-reduced basis returns the planted (s, e); the progressive schedule used in Section 5.6 would not do here, since reaching β progressively also runs every smaller blocksize and the resulting figure is not what (6) is a statement about. In every run the vector recovered was a rotation of the planted witness, of exactly its length, so what is measured is key recovery and not a sublattice artefact.

Table 15. Measured single-shot BKZ-β key-recovery blocksizes on scaled-down instances of the KEM key lattice, against the two forms of the estimate (6).

$(n,\, q)$, ambient dimension $2nk$	$(32, 193)$, 128	$(40, 401)$, 160	$(48, 1249)$, 192
β measured (single BKZ-β)	50.7	57.5	48.0
(6) with $\|v\|^2 = nk(\sigma_s^2 + \sigma_e^2)$	71	69	57
(6) with the secret block rescaled by ω	58	58	50

Two things follow. The rescaled form of Section 5.4 tracks reduction closely — within about three blocksizes across the three cells — while the unrescaled form, which is the one that reproduces the figures of Section 6.1 to 0.3 bits, sits 9 to 20 blocksizes high. At these dimensions, therefore, the estimate as we apply it *over*-states the blocksize an attacker needs, by 3 to 6 bits of core-SVP cost. That is the unsafe direction, and we record it as such; against it stand the facts that the toy dimensions are 128 to 192 against the deployed 1024 to 2048, that (6) is

an asymptotic whose root-Hermite factor is not meaningful below $\beta \approx 50$, and that a sufficient condition is expected to be loose. What the measurement does not support is treating the unrescaled figures as conservative.

Remark 12 (The choice of σ_e). The instances above take $\sigma_e^2 = \eta_f/2 = 1$, the width of the error the compact scheme actually publishes: Remark 5 records that the p-scaling of the errors is dropped under the MSB encoding, and this deployed width is also the one Section 5.4 uses. The p-scaled width $\sigma_e^2 = p^2 \cdot 1 = 4$ belongs to the reference design of (2), which no deployed parameter set instantiates. The choice is not cosmetic, and it fixes the sign of the conclusion: at $\sigma_e = 2$ the rescaling of that section would *raise* the estimate, licensing a reading of its figures as a conceded margin, whereas at the deployed $\sigma_e = 1$ it *lowers* them — which is why that section withdraws the "conservative" reading instead of claiming a margin. Every KEM figure quoted in this paper is computed at the deployed width; the reference width appears only where the hybrid model is stress-tested (Section 5.5).

B Hybrid model: widths and the profile simulator

Sensitivity to the constant τ

For the KEM the closed form must be read with the width question of Remark 12 settled. At the deployed error width $\sigma_e^2 = \eta_f/2$ the Babai/GSA hybrid at $\tau = 2.0$ evaluates to 100.4/165.3/248.2 at the three categories — below the primal everywhere, i.e. binding everywhere — while at the reference-design width $\sigma_e^2 = p^2\eta_f/2$ it evaluates to 121.3/196.0/296.7 and binds nowhere. (The same code reproduces the Kyber calibration above to about a bit, so the spread is a property of the instance, not of the model's implementation.) Taken at face value the deployed-width figures would pull category 1 from 115.6 down to 100.4. They should not be taken at face value, and the next paragraph is why: the closed form rests on the GSA, and against measured Gram–Schmidt profiles the GSA credits the nearest-plane step with a decoding radius that reduction does not deliver.

Replacing the geometric series assumption

All of the above inherits the GSA. That assumption can now be checked rather than adopted, because Section 5.6 records the Gram–Schmidt profile of every reduced basis at the moment of its event. Against the smallest Gram–Schmidt norm those bases actually exhibit, the GSA closed form $\delta_\beta^{1-d'}\mathrm{vol}^{1/d'}$ is too *large* by a factor 1.31 on average over 34 non-degenerate cells — 1.44 at $\beta \in [25, 34]$, 1.26 at $\beta \in [45, 60]$ — while the Chen–Nguyen BKZ 2.0 simulator, seeded with the q-ary Z-shape, reproduces it to 0.98 and 1.09 on the same two bands. The GSA error is in the direction that flatters the attacker: it credits the nearest-plane step with a decoding radius that reduction does not deliver, because it

models a pure geometric series where a q-ary basis in fact retains a plateau near 1.

Recomputing the hybrid with the simulator in place of the closed form, the attack ceases to bind anywhere: at $\tau = 2.0$ its cost rises above the primal one at all three KEM sets, and the optimum degenerates into pure meet-in-the-middle guessing. We do not quote the resulting figures, since once the nearest-plane step never succeeds at a usable blocksize the number is set by where the guessing optimum lands rather than by lattice reduction. The conclusion we do draw is that the sensitivity tabulated above, and the σ_e ambiguity with it, are artefacts of the GSA: under a profile model that matches measurement, the primal attack of Section 5.2 is the binding constraint at every parameter set in this paper. The caveat of Section 5.6 applies again — the simulator is checked at $\beta \leq 60$ in dimension ≤ 256 and used at $\beta \approx 400$ in dimension ≥ 1024.

C Fatigue measurement: harness calibration and limitations

Calibration of the harness itself

Run on circulant NTRU — the geometry for which [11] publish experimental fatigue points, and against which the estimator was validated above — the same code returns

Table 16. Calibration of the reduction harness on circulant NTRU, against the published experimental fatigue points of [11].

n	79	101
q_{fat} measured, grid bracket	239–331	307–431
q_{fat} measured, logistic fit	282	427
q_{fat}, estimator of this section	235	438
q_{fat}, published fit $0.004\,n^{2.484}$	207	381
measured/estimator	1.20×	0.98×
β measured / predicted, at the non-degenerate end	40.2/41.0	32.0/37.0

The estimator is 20% early at $n = 79$ and exact to 2% at $n = 101$, so what discrepancy there is shrinks with the dimension — and it is early, which is the conservative direction. The geometry underneath the blocksize agrees as well: at the two cells whose crossing blocksize is not degenerate, the slope of the middle half of the Gram–Schmidt log-profile recorded at the event matches the estimator's Z-shape to three digits (0.02662 against 0.02671, and 0.02777 against 0.02786), which tests the geometric series assumption directly rather than through its consequences.

Limits of the simulation

Three limitations, stated because the numbers above are otherwise easy to overread. *(i)* The experiment settles the *ordering* and not the *magnitude*. Almost every cell it could afford crosses at $\beta < 20$, the regime item *(iii)* below flags as degenerate: there LLL alone can return the whole planted module together with the key, so which of the two events is recorded turns on which plant vector happens to come out shortest rather than on any ordering in time. Re-locating the same crossing on an indicator immune to that tie-break — whether the basis ends up holding more than the key line at all — gives a gain of 1.68 bits at $N = 96$ instead of ≥ 4.86. Under either reading the rank-$N/2$ plant is discovered at a strictly larger modulus than the rank-N one; the size of the gap is what the degeneracy leaves open, and separating the two readings needs dimensions where the crossings are not degenerate. *(ii)* The model is validated by its authors up to $n \approx 200$ and $\beta \approx 60$, and predicted to $n = 499$; the parameter sets sit at $nk = 512$–2048 with crossing blocksizes ranging from 130 to about 390 across the KEM sets. It is the same model used further than it has been checked, and our own experiment reaches only $2N \leq 256$ with blocksizes in the tens, since it runs on enumeration rather than sieving; it narrows the gap without closing it. *(iii)* Pr[DSD] is not monotone in q: in small dimension the estimator saturates at $\beta \to 2$ and the ratio falls back, so the fatigue point must be located by a first-crossing search rather than by bisection, and crossings occurring at $\beta < 20$ must be discarded as degenerate. Both effects are visible in the reduction data and not only in the model — at $n = 79$ the measured DSD share runs $0.00, 0.88, 1.00, 0.50$ as q increases through the crossing, falling back at the largest modulus exactly where the crossing blocksize has collapsed to 2, and at that blocksize the measured Gram–Schmidt slope has the wrong sign against the Z-shape. Finally, as in the source model, only DSD-PT events are modelled and the "lucky lift" events are not; that they do not lower the successful blocksize was checked for NTRU, not for NTC.

D The exact failure probability

Section 4.4 bounds δ by a Gaussian tail matched to the variance (5). That bound is not safe as stated, for two reasons that compound, and since δ enters Theorem 2 through $q_G\delta$ we compute it properly here.

The terms entering the convolution. Read off (4) coordinate by coordinate, D_1 is a sum of *products* of independent, finitely supported variables:

$$\begin{aligned} D_1 \;=\; & \textstyle\sum_{k\mathsf{E}_R} f_i r_i \;+\; g_a \;+\; p\sum_{k\mathsf{E}_R} g_i s'_i \;+\; f_b \;+\; p\sum_{\mathsf{E}_\alpha} f_i s'_i \\ & +\; u_a \;+\; p\sum_{k\mathsf{E}_R} u_i s'_i \;+\; u_b \;+\; p\sum_{\mathsf{E}_\alpha} u_i s'_i \;+\; \tfrac{1}{2}\textstyle\sum_{\mathsf{E}_\alpha} \varepsilon_i s'_i, \end{aligned}$$

with $f, g \leftarrow\!\$\ \chi_f$, $r \leftarrow\!\$\ \chi_r$, $s' \leftarrow\!\$\ \chi_{s'}$, $\varepsilon \leftarrow\!\$\ U(\{\pm 1\})$ and u the decompression error. A product of two centred binomials has a markedly heavier tail than

a Gaussian of the same variance, and no number of summands removes that at the six-to-seven standard deviations $\delta \le 2^{-140}$ requires. Each factor being finitely supported, however, every summand has an exactly computable moment generating function, hence so does D_1; we obtain the tail from the saddlepoint (Lugannani–Rice) expansion applied to the exact cumulant generating function, and check the method on a sum whose tail is known in closed form, where it agrees to 0.5 bits out to twelve standard deviations.

And the rounding term. Separately, (5) prices the decompression error at $V = (q/2^d)^2/12$, the variance of a *continuous* uniform. The true error $x - \mathrm{Decompress}_d(\mathrm{Compress}_d(x))$ is discrete and, at these (q, d), has a somewhat larger variance; at $q \le 2^{14}$ it can simply be enumerated. This is the smaller of the two effects but not a negligible one, since u appears with multiplicity $\mathsf{E}_\alpha + k\mathsf{E}_R$.

Table 17. Decryption failure probability: the Gaussian approximation against the exact convolution, on the baseline category-1 candidate $(\eta_f, \eta_r, \eta_{s'}) = (2, 12, 3)$, $d = 11$, and the categories 3 and 5 sets.

	NTC-KEM-512	NTC-KEM-768	NTC-KEM-1024
σ_{D_1} from (5)	128.9	169.0	219.3
σ_{D_1} with the exact rounding term	132.7	171.9	220.5
$\log_2 \delta$, Gaussian (5)	−156.3	−165.4	−161.5
$\log_2 \delta$, Gaussian, exact variance	−147.3	−159.8	−159.7
$\log_2 \delta$, exact distribution	**−144.2**	**−158.0**	**−156.9**

(The category-1 column is the *baseline* candidate — the optimum a selection against the Gaussian δ returns, cf. Table 12 — so that the three effects can be read on the set they disqualify; the adopted set is at −148.7.) The variance approximation costs 2 to 9 bits and the non-Gaussian tail a further 2 to 3, so the Gaussian δ is optimistic by 5 to 12 bits. Every set still clears $\delta \le 2^{-140}$, which is the conclusion that matters, but at category 1 the Gaussian reading shows a margin of 16 bits where the exact one leaves 4; that is what makes the constraint bind in the tuning of Section 6.1, and it is why a set selected against the Gaussian δ there turns out to be inadmissible.

E Design observations for the KEM

Constraints forced by the algebra

Two features of NTC force the shape of the scheme of Section 4.3.

The trapdoor rescales by s, hence a p-adic encoding. The only secret operation available to the receiver is multiplication by M_s. Whatever is extracted is therefore scaled by s: if the sender encoded a message as $\Delta \cdot m$ with $\Delta = \lceil q/2 \rfloor$, the

receiver would recover $\Delta \cdot sm$, and $\|sm\|_\infty$ is far too large for $\Delta\, sm$ to avoid wrap-around modulo q. The Kyber-style encoding is thus structurally unavailable, and we follow the NTRU route instead: all parasitic terms are made $\equiv 0 \bmod p$ and the message is read modulo a small prime p.

The encryption randomness must lie in $\mathcal{A}$. Decryption closes only if the sender's randomness commutes with M_s. By Lemma 6 the commutant of a generic matrix meets $M(\mathcal{A})$ in the scalars only, so the randomness must be taken in $\mathcal{A}$ itself, not as an arbitrary short matrix of $\mathcal{M}_k$.

$k = 2$ is forced. The ciphertext costs $nk(k+1)$ coefficients, so at fixed ciphertext budget one maximises nk by minimising k. Concretely $(n, k) = (256, 3)$ and $(512, 2)$ have identical ciphertexts but 171 against 235 bits of security, and the same domination recurs at every size.

Φ_{1152} at category 3, and only there. The second ring family is *worse* for noise ($\mathsf{E}_R = 1.5N$). It wins at category 3 for an unrelated reason: the residue class $q \equiv 1153 \pmod{2304}$ contains a prime as early as $q = 3457$, whereas $q \equiv 513 \pmod{1024}$ reaches its first prime only at 7681. A 12-bit modulus instead of a 15-bit one is worth far more than the 50% variance penalty. The advantage does not generalise: within the same family, $N = 192$ yields only 80.9 bits, and $N = 768$ reaches 391 bits for a 9 kB ciphertext, which is dominated by the power-of-two set at 257 bits and 5.62 kB. Category 3 is the one place where the grid is favourable.

Wide encryption randomness. The optimum settles at $\eta_r = 12$, far wider than η_f or $\eta_{s'}$. This is not an accident of the search: by Section 4.1 the randomness χ_r enters the noise budget only linearly, whereas χ_f enters quadratically, so entropy is cheapest to buy on the r side. The practical price is $\lceil \log_2(2\eta_r + 1) \rceil$ bits of randomness per coefficient at encryption.

Two transforms. Categories 1 and 5 use $x^n + 1$ while category 3 uses Φ_{1152}: an implementation covering all three carries two distinct number-theoretic transforms. If that is unacceptable, category 3 can be served by the category-5 set at a cost of roughly 0.8 kB per ciphertext, in the manner of deployments that ship only two tiers.

Open parameter tension

The modulus is squeezed from both sides. Correctness (3) demands $q \gtrsim p\, \eta^2 \sqrt{nk}$ times a tail margin. The hardness analysis, on the other hand, imposes an *upper* bound on q through the dense sublattice discovery (fatigue) threshold. The two constraints close in on one another and the resulting window must be verified numerically before any parameter set is fixed.

Two circumstances work in favour of a nonempty window, both specific to the matrix reading. By Corollary 1 the planted sublattice has rank n inside a lattice

of dimension $nk(1+k)$, a fraction $1/\big(k(1+k)\big)$ instead of the $1/2$ of NTRU, which pushes the fatigue point up substantially; and the norm attack, which pinned the overstretched ceiling to dimension n in the commutative reading, does not apply.

Remark 13 (Do not share keys across primitives). The pair (A, t) must not be reused between the KEM and a signature scheme built on the same relation, without an explicit joint assumption, even though both keys would be truncated NTC instances: signatures reveal short linear relations on s and hint-correlated functions of the target error that the IND-CCA analysis above does not model.

F Statistical key uniformity, and a conjectural worst-case route

Theorem 1 rests on $\mathsf{NTC\text{-}Dec}$, an assumption introduced in this work and carrying no worst-case support. An alternative is available, and it is the one Stehlé and Steinfeld [29] took for NTRU: widen the key distribution until the public key is *statistically* uniform. The assumption then does not appear in the analysis at all, and IND-CPA rests on $\mathsf{cLWE}_{\mathcal{A}}$ alone — which, in the row-truncated form that the scheme of Section 4.3 actually publishes, is exactly module-LWE by Remark 8 and therefore inherits the worst-case reduction to module-SIVP of Langlois and Stehlé [23]. This subsection states the criterion, prices it, and is explicit about the one step that is not yet rigorous.

The key lattice. Fix A and a candidate public key $h \in R_q^k$, and set

$$\Lambda_{A,h} \;:=\; \big\{(s,e) \in \mathbb{Z}^{nk} \times \mathbb{Z}^{nk} \;:\; e \equiv h\, M_s - \mathrm{row}_1(M_s)\, A \pmod q\big\}, \qquad (10)$$

a lattice of dimension $d = 2nk$ and determinant q^{nk}: for each s the error e is determined modulo q. Comparing with KeyGen, the public key $t[1,:]$ equals h exactly when the sampled pair (s, e_1) lies in $\Lambda_{A,h}$, so

$$\Pr\big[\, t[1,:] = h \,\big] \;\propto\; \rho_\sigma\big(\Lambda_{A,h}\big),$$

and $t[1,:]$ is near-uniform on R_q^k as soon as this Gaussian mass is essentially independent of h. That is the classical smoothing criterion. Note that $\Lambda_{A,h}$ is the same lattice, up to the choice of h, whose planted sublattice the analysis of Section 5.6 concerns; the present subsection is what happens when the plant is widened out of existence.

Lemma 9 (Regularity). *Let $\varepsilon \in (0, \frac{1}{2})$ and suppose $\sigma \geq \eta_\varepsilon(\Lambda_{A,h})$ for every $h \in R_q^k$. Then the public key produced by KeyGen with (s, e_1) drawn from a discrete Gaussian of width σ is within statistical distance 2ε of uniform on R_q^k.*

Proof. For $\sigma \geq \eta_\varepsilon(\Lambda)$ the Gaussian mass of every coset satisfies $\rho_\sigma(\Lambda + c) \in \big[\frac{1-\varepsilon}{1+\varepsilon}, 1\big] \cdot \rho_\sigma(\Lambda)$. The lattices $\Lambda_{A,h}$, as h ranges over R_q^k, partition $\mathbb{Z}^{nk} \times \mathbb{Z}^{nk}$ into q^{nk} classes of equal determinant, so the ratio of the largest to the smallest mass is at most $\frac{1+\varepsilon}{1-\varepsilon}$, and the induced distribution on h is within 2ε of uniform in statistical distance. □

Lemma 10 (Smoothing bound). *Write $d = 2nk$. Under the Gaussian heuristic for the dual minimum, $\lambda_1(\Lambda^*) \approx \sqrt{d/(2\pi e)}\, q^{-1/2}$, and the standard bound $\eta_\varepsilon(\Lambda) \le \sqrt{\ln(2d(1+1/\varepsilon))/\pi}\,/\,\lambda_1(\Lambda^*)$ gives*

$$\eta_\varepsilon(\Lambda_{A,h}) \;\le\; c(nk)\,\sqrt{q}, \qquad c(nk) \;=\; \sqrt{\frac{e\,\ln\bigl(4nk(1+1/\varepsilon)\bigr)}{nk}}. \tag{11}$$

At $\varepsilon = 2^{-128}$ this is $c = 0.715,\ 0.508,\ 0.415,\ 0.360$ for $nk = 512,\ 1024,\ 1536,\ 2048$.

Remark 14 (The convention $s \equiv 1 \bmod p$ costs a factor p). Row truncation requires $s = 1 + p\,s'$ (Remark 7), so the secret is confined to a coset of $p\mathcal{A}$ rather than ranging over $\mathcal{A}$. Smoothing the coset lattice requires width $p\,\eta_\varepsilon$, i.e. $\sigma \ge p\,c(nk)\sqrt{q}$, exactly as the analogous constraint $f = pf' + 1$ costs Stehlé and Steinfeld a factor p. All figures below include it.

Theorem 3 (IND-CPA from module-LWE alone, conditional). Assume Conjecture 1. *Instantiate the scheme of Section 4.3 with (s, e_1) drawn from a discrete Gaussian of width $\sigma \ge p\,c(nk)\sqrt{q}$ and with the encryption distributions χ_r, χ_f unchanged. Then for every PPT adversary $\mathcal{A}$ there is a PPT $\mathcal{B}$ with*

$$\mathbf{Adv}^{\mathsf{ind\text{-}cpa}}_{\mathsf{PKE}}(\mathcal{A}) \;\le\; 2\,\mathbf{Adv}^{\mathsf{mlwe}}_{nk,q,\chi_r}(\mathcal{B}) \;+\; 2^{-127},$$

and no NTC*-type assumption is used.*

Proof. Game G_1 replaces $\mathsf{pk} = (\rho, t[1,:])$ by (ρ, u) with $u \leftarrow\!\!\$\ U(R_q^k)$. By Lemmas 9 and 10 together with Remark 14, the two are within statistical distance $2\varepsilon = 2^{-127}$, so $|\varepsilon_1 - \varepsilon_0| \le 2^{-127}$; this step is information-theoretic and replaces the invocation of NTC-Dec in the proof of Theorem 1. In G_1 the public key carries no information about (s, e_1), and the remainder of the argument is the ciphertext-indistinguishability step of Theorem 1 verbatim, which appeals only to $\mathsf{cLWE}_{\mathcal{A}}$. By Remark 8 the truncated sample published by the scheme is exactly a module-LWE sample of rank k in dimension nk, whence the stated bound. Composing with Langlois–Stehlé [23] gives IND-CPA from the worst-case hardness of module-SIVP. The appeal to Lemma 9 is where Conjecture 1 enters, and it is the only place it does. □

The price. Widening the key inflates the decryption noise, which forces q up, which in turn forces σ up through (11). The system closes. Retaining the noise expression (5) with $\sigma_s = \sigma_{e_1} = \sigma$ and $\sigma_r = \sigma_f = \sigma_{\mathrm{sml}}$, and dropping compression (which the tight noise budget no longer permits), the dominant terms give $\sigma_{D_1} \approx \sigma\,\sigma_{\mathrm{sml}}\sqrt{6n}$, so that $\delta \le 2^{-140}$ requires $\sigma\,\sigma_{\mathrm{sml}}\sqrt{6n} \le q/58.4$. Combining with $\sigma \ge p\,c(nk)\sqrt{q}$ yields $q \ge \bigl(58.4\,p\,c(nk)\,\sigma_{\mathrm{sml}}\sqrt{6n}\bigr)^2$.

Proposition 3 (A parameter set for the conjectural route). *Take $R = \mathbb{Z}[x]/\Phi_{2304}$, $n = 768$, $k = 2$, $p = 2$, $\sigma_{\mathrm{sml}} = 1$, and*

$$q = 10858753,$$

which is prime and satisfies $q \equiv M+1 \pmod{2M}$ *for* $M = 2304$*, so that* $\alpha = x$ *is non-split (Lemma 3). Then* $\sigma = 2737$ *meets* (11) *with the factor of Remark 14; the decryption failure probability satisfies* $(q/4)/\sigma_{D_1} = 14.61$*, hence* $\delta \le 2^{-140}$*; and the module-LWE instance underlying Theorem 3 costs* 193.6 *bits under the estimator of Section 5.2. Public key and ciphertext are* 4640 *and* 9216 *bytes.*

Table 18. The conjectural worst-case variant against the computational sets of Section 6.1.

	this variant (Prop. 3)	Section 6.1 category 1	Section 6.1 category 3
nk	1536	512	768
q	10858753	7681	10369
σ (key)	2737	2.45	2.45
σ_{sml}	1	1	1
security (shared est.)	193.6 (module-LWE)	115.6	176.7
key assumption	*none* (statistical)	NTC-Dec	
public key	4640 B	864 B	1376 B
ciphertext	9216 B	1536 B	2304 B

Read against the category-3 set, which reaches 176.7 bits under the same rescaled estimate, the worst-case variant costs a factor 3.4 on the public key and 4.0 on the ciphertext at some 17 bits more security; read against category 1 the factors are 5.4 and 6.0, but that comparison is between different security levels and we do not rely on it. Two further points. The dimension $nk = 1536$ is forced rather than chosen: at $nk = 1024$ the same constraints yield only 114.5 bits, so the ring grid wastes the interval in between, and an intermediate ring would recover part of the cost. And the penalty is markedly milder than the corresponding one for NTRU, where the Stehlé–Steinfeld parameters are far from competitive with the practical scheme. The reason is structural: the NTC key lattice already has dimension $2nk$ rather than $2n$, so the smoothing width $c(nk)\sqrt{q}$ carries a factor $1/\sqrt{nk}$ in place of $1/\sqrt{n}$.

Remark 15 (The plant disappears, and with it the overstretched regime). At the parameters of Proposition 3 the secret vector has norm $\|(s, e_1)\| \approx \sigma\sqrt{2nk} = 151693$, whereas the Gaussian heuristic puts the first minimum of the ambient lattice at $\sqrt{2nk/(2\pi e)}\sqrt{q} = 44194$. The "secret" is thus 3.4 times *longer* than a generic shortest vector: there is no unusually short vector to find, no dense sublattice to discover, and consequently no fatigue point. The entire analysis of Section 5.6, and the tension of Section E between correctness and the overstretched threshold, are vacuous for this variant. What replaces them is the ordinary hardness of module-LWE and module-SIS on a random q-ary lattice.

Remark 16 (Remaining non-rigorous steps). Lemma 10 uses the Gaussian heuristic for $\lambda_1(\Lambda^*)$, and Lemma 9 requires the resulting bound to hold *for every* h,

not merely on average. Supplying a proven lower bound on the dual minimum, uniformly over h, is the entire technical content of the Stehlé–Steinfeld argument, and it is there that they require $x^n + 1$ to factor into few irreducible polynomials modulo q: with two factors of degree $n/2$ the exceptional set is bounded by $q^{-n/2}$, whereas with n linear factors the natural union bound degrades to n/q.

The gap is not marginal, and a number makes that plain. At the parameters of Proposition 3 one has $n = 768$ and $q = 10858753$, so $n/q = 2^{-13.8}$, against the 2^{-127} that Theorem 3 asserts: the available argument supports a statistical distance 113 bits weaker than the one claimed. The gap is therefore recorded below as a conjecture, not as a caveat.

Conjecture 1 (Regularity for completely split R_q). For R_q completely split and $\sigma \geq p\, c(nk)\sqrt{q}$ as in Lemma 10, the bound $\sigma \geq \eta_\varepsilon(\Lambda_{A,h})$ holds for *every* $h \in R_q^k$ with $\varepsilon = 2^{-128}$, over the choice of A, except with negligible probability.

Everything in this subsection that is stated unconditionally — the criterion (11), the pricing, Proposition 3, Remark 15 — is independent of the conjecture; Theorem 3 is not.

A second route, which needs no conjecture. The obstruction is complete splitting, and complete splitting is imposed by $M \mid q - 1$, which the instantiation of Section 6.1 adopts for the number-theoretic transform. But that constraint, and the non-splitness condition $q \equiv M + 1 \pmod{2M}$ beside it, are both artefacts of the *other* branch of this paper. The non-splitness of α exists to block the CRT attack on NTC (Section 5.1), and this variant does not assume NTC; the full transform is a speed choice. Neither is a security requirement here.

Dropping both, one may choose q so that R_q has exactly the shape Stehlé and Steinfeld require. In the power-of-two family with $n = 2^a$, a direct computation gives that

$$q \equiv 3 \text{ or } 5 \pmod 8 \quad \Longrightarrow \quad x^n + 1 \equiv f_1 f_2 \pmod q, \quad \deg f_1 = \deg f_2 = n/2,$$

both irreducible — exactly two factors, so the exceptional set is $q^{-n/2}$ and Lemma 9 applies unconditionally. The price is the transform: with two factors of degree $n/2$ only one level of the Cooley–Tukey recursion is available, and multiplication falls back on Karatsuba over $\mathbb{F}_{q^{n/2}}$. For a variant whose entire purpose is to trade efficiency for a worst-case guarantee, and which already pays 3.4× on the public key, that is the cheaper of the two prices. We have not carried out the parameter search for such a set, and note only that the route exists and requires no new mathematics.

G Implementation considerations

We have not implemented the scheme, so this section contains no measurements. What it does contain is the structural cost analysis that an implementer needs before starting: which transforms are required, what arithmetic in $\mathcal{A}$ costs relative to a plain ring multiplication, where the moduli chosen in Section 6 help

or hurt, and which operations carry secret-dependent data. Appendix G.7 states what only an implementation can settle.

G.1 Two transforms

Every ring in this paper admits a full number-theoretic transform, since the non-splitness condition $q \equiv M + 1 \pmod{2M}$ of Lemma 3 refines the full-splitting condition $M \mid q - 1$. The transform lengths and their factorisations are:

Table 19. Transform lengths and butterfly structures of the deployed rings.

ring	n	$M = \mathrm{ord}(x)$	butterflies
KEM, category 1 $x^{256} + 1$	256	512	radix-2
KEM, category 3 Φ_{1152}	$384 = 2^7 \cdot 3$	1152	radix-2 and radix-3
KEM, category 5 $x^{512} + 1$	512	1024	radix-2

The power-of-two rings need nothing beyond a standard negacyclic transform. The Φ_{3N} family, popularised in this context by NTTRU [26], requires one radix-3 stage in addition; it is well-trodden ground, but it is a second code path. An implementation covering all three categories therefore carries two transforms, and Section E records the alternative: serve category 3 with the category-5 parameter set, at roughly 0.8 kB of extra ciphertext, and ship a single power-of-two transform.

G.2 Arithmetic in $\mathcal{A}$, and the price of the twist

In the transform domain $\mathcal{A} \cong \prod_{i=1}^{n} \mathbb{F}_{q^k}$ by Lemma 1, and at $k = 2$ each slot is $\mathbb{F}_{q^2} = \mathbb{F}_q[y]/(y^2 - \zeta_i)$ with ζ_i the evaluation of $\alpha = x$ at slot i. A product in $\mathcal{A}$ is therefore n independent products in $\mathbb{F}_{q^2}$, each of which costs three $\mathbb{F}_q$-multiplications by Karatsuba plus one multiplication by the constant ζ_i:

$$(a_0{+}a_1y)(b_0{+}b_1y) = \big(t_0{+}\zeta_i t_1,\ t_2{-}t_0{-}t_1\big), \qquad t_0 = a_0b_0, \quad t_1 = a_1b_1, \quad t_2 = (a_0{+}a_1)(b_0{+}b_1).$$

A multiplication in $\mathcal{A}$ thus costs $4n$ base-field multiplications against n for a plain ring product: a factor of four, with the ζ_i foldable into precomputed tables.

The twist deserves a remark, because it is the one design choice where the arithmetic and the noise analysis agree. Multiplication by $\alpha = x$ is, in the coefficient domain, a negacyclic shift in the power-of-two rings — a signed index permutation, hence free — and in Φ_{3N} a shift with the wrap rule $x^{N+r} = x^{N/2+r} - x^r$, hence one extra addition per wrapped coefficient. In the transform domain it is a pointwise multiplication by a known constant. So $\alpha = x$ is essentially free to compute, and by Section 2.1 it costs a factor $\mathsf{E}_\alpha/\mathsf{E}_R = 2$ in noise growth against 10 or 26 for the smallest admissible constant twists. A constant α would be marginally cheaper arithmetically and far more expensive in the noise budget; there is no trade-off to make.

G.3 Operation counts

We count length-n transforms and pointwise products of length n, the two units that dominate. Throughout, the matrix A and the row u_1 are expanded from the seed *directly in the transform domain*, as in Kyber and Dilithium, so they cost no transforms and need never be stored: the transient working set is $k^2 n$ coefficients.

Table 20. Transform and pointwise-product counts per primitive.

	operation	transforms	pointwise products
KEM	KeyGen	4	8 (+ one inversion in $\mathcal{A}$)
	Encaps	6	8
	Decaps	5	4 (+ a full Enc for re-encryption)

One entry needs comment. KEM key generation needs M_s^{-1}, that is, an inversion in $\mathcal{A}$. Done naively this is n inversions in $\mathbb{F}_{q^2}$; done through the norm, $a^{-1} = \bar{a}/N(a)$ with $N(a) = a_0^2 - \zeta_i a_1^2 \in \mathbb{F}_q$, it becomes n inversions in $\mathbb{F}_q$, and Montgomery's batch trick reduces those to a *single* field inversion plus $O(n)$ multiplications. Key generation should therefore not be inversion-dominated.

G.4 Moduli, lazy reduction, and vectorisation

The KEM moduli need care. All three fit comfortably in signed 16-bit lanes — the largest half-modulus is 6656 for $q = 13313$ — so Kyber's 16-bit vectorisation strategy applies. What shrinks is the headroom for lazy reduction, that is, the number of Montgomery products that may be accumulated before a reduction is forced:

Table 21. Lazy-reduction headroom against Kyber.

	NTC-KEM-512	NTC-KEM-768	NTC-KEM-1024	Kyber
q	7681	10369	13313	3329
$\lceil \log_2 q \rceil$	13	14	14	12
accumulations before overflow	4	3	2	9

At category 5 only two products may be accumulated, against nine for Kyber. This does not change the asymptotics and does not threaten correctness, but it removes a standard optimisation from the innermost loop and is the most likely source of a constant-factor gap against Kyber in a vectorised implementation. An implementer with a free hand might prefer to spend a few bits of the security margin — which Section 5.4 shows is conservative by 5 to 8 bits anyway — on a smaller modulus.

G.5 Sampling

Three distributions must be sampled, and one choice made earlier pays off here.

The masking randomness $y \leftarrow\!\!\$\ U(S_{\gamma_1-1})^{nk}$ uses $\gamma_1 \in \{2^{16}, 2^{18}, 2^{17}\}$, a power of two by construction. Uniform sampling on a power-of-two interval is exact bit extraction from a SHAKE stream, with no rejection and no bias correction — unlike a general interval, which would need either rejection or a modular reduction with a bias argument.

The secrets use $\eta = 2$, five values per coefficient, which does require rejection; the standard technique of drawing four bits and rejecting the values ≥ 15 before reducing modulo 5 applies verbatim.

G.6 Constant-time considerations

Two places carry secret-dependent control flow and must be handled.

Invertibility resampling. KeyGen resamples s until it is invertible in $\mathcal{A}$. By Lemma 2 the failure probability is at most $n\,q^{-k}$, which is below 2^{-17} at every KEM set.

Implicit rejection in Decaps. The $\mathrm{FO}^{\not\perp}$ transform re-encrypts and compares; both the comparison and the selection between $\mathsf{KDF}(\bar{K}', c)$ and $\mathsf{KDF}(z, c)$ must be constant-time, as for any FO-based KEM. This is standard, but it is also where deployed implementations have historically gone wrong, and the decryption-failure rate $\delta \leq 2^{-140}$ of Section 4.4 is what makes the analysis of the rejection branch meaningful.

G.7 Measurements left to an implementation

The analysis above fixes the operation counts but not the constants, and three questions can only be answered by measurement. First, whether the factor of four for $\mathcal{A}$-multiplication is realised in practice or partly absorbed by better instruction-level parallelism, since the three Karatsuba products are independent. Second, how much the reduced lazy-reduction headroom of Appendix G.4 actually costs in a vectorised transform. Third, whether the radix-3 stage of the Φ_{3N} transform makes category 3 enough slower than categories 1 and 5 to justify the single-transform alternative of Section E. We regard the third as the question most likely to change a deployment decision. Section 8 reports measurements from the two C trees accompanying this paper and settles the third question in the negative; the first two await a per-routine profile.