

NLRP3_V4_Whitespace_and_Trajectory

Table of Contents

NLRP3 — V4 Structural Occupancy and Trajectory.....	1
I. What was measured	2
I.0 The largest correction in this run, stated before anything rests on it	2
I.1 The corpus is a floor, and here is by how much	3
I.2 What is out of scope by election, and therefore not measured	3
II. The binding constraint is claim scope, not corpus size	4
II.1 The excluded population is not a random sample	4
III. The occupancy grid	4
III.1 Most of the emptiness is not opportunity.....	5
III.1a Where the readable whitespace actually is	5
III.2 What the axes actually reached	6
III.3 The fence cells	6
III.4 Ownership is a floor, and the floor is low	7
IV. Direction, indication, modality.....	7
V. Trajectory	8
V.1 The foundational tail is a named absence	8
VI. The overlap probes	8
VI.1 The sequence probe ran, and then disqualified itself	8
VI.2 The chemistry probe did not return	9
VI.3 What that leaves	9
VII. The asset strip	10
VIII. The regulatory floor is zero, and the detector was proven first.....	10
IX. What this package does not establish.....	11
X. Verification.....	11

NLRP3 — V4 Structural Occupancy and Trajectory

Fence Line Package. A named configuration of the V4 methodology, not a variant of it. Claim-strength scoring (KC) is **omitted by election** under §3.0, so no sentence in this

document asserts how strong any claim is. What it asserts is *where the claims sit* and *what the instruments could and could not see*.

This is candidate generation and structural occupancy. It is never freedom to operate. No cell verdict, genus hit or sequence hit in this document is a clearance opinion.

I. What was measured

Discovery returned 1083 candidate rows. What survived ingest and reconciliation is a corpus of 780 patent families and 2126 publications, assembled from a discovery slate the operator approved before any query ran: 9 core target terms and 11 compound code-names, with 11 string collisions rejected at the gate. The operator added 0 terms of their own, so the recall of everything downstream is bounded by that approved slate and by nothing wider.

Of those families, 522 carry NLRP3-directed subject matter. The remainder are named rather than discarded: 127 admitted on a broadening token and nothing else, 93 with no signal, 34 naming NLRP3 only inside a multi-gene panel with no selectivity evidence, 2 where a competing inflammasome sensor (NLRP1, NLRC4, AIM2, NLRP6) dominates the claim scope, 1 mixed, and 1 homonym. Those buckets and the on-target set sum to 780, which is the corpus total — the parts cannot silently disagree with the whole.

I.0 The largest correction in this run, stated before anything rests on it

The first bucket above did not exist until late in the analysis, and it took 127 families out of the cohort — 0.1957 of the 649 this package had been treating as on-target.

The vocabulary was built on a split rule that the playbook is explicit about: a term may be **broad enough to admit** a family into the corpus and must be **strict enough to decide** whether that family is actually about the target, and those are not the same pattern. A third consumer — the disposition screen — had taken the admitting pattern, so tokens inside it were casting votes they were never approved to cast: pyroptosis, which is the gasdermin-executed terminal programme the operator had **elected out of scope**, and bare inflammasome, where the gate approved the phrase *NLRP3 inflammasome* and not the noun — NLRP1, NLRC4, AIM2 and pyrin all form inflammasomes.

Every one of the 127 entered through a pyroptosis or inflammasome discovery slice, and 0 entered through a slice naming the sensor or any of its assets. That is a measured zero, not an unreachable branch: the same detector run across the surviving on-target bucket returns 402. None of the displaced families carries a sensor string or an asset code-name anywhere in its readable text. They are in the corpus and out of the cohort, which is why they are counted above rather than deleted.

The finding is not that a fifth of the cohort was wrong. It is that **an admission rule was being read as an evidence rule**, and those must never be the same object. Every count in this document is computed after the separation was enforced.

I.1 The corpus is a floor, and here is by how much

911 ingest **requests** were submitted — a request is not a family, and the distinction is not pedantry. 505 requests reached full analysis and resolve to 495 distinct families; 288 arrived bibliographic-only, resolving to 285 families. The reason recorded for that deferral — that their members carry no English claims — was measured on a **sample**, not on all of them, and is carried here as the sampled attribution it is rather than as a property of all 288. Together that is 780 families, which is the corpus.

118 requests failed outright — the upstream index returned no rows for them. Every one of those 118 failed requests carries **no family identifier at all**, so the number of distinct *families* lost is **unknown**. It is at most 118 and it is not zero. Every population in this document is a floor by that unknown amount, and stating it as a family count would be inventing a precision the record does not have.

Discovery sharding under-reached its own truncated parent in 5 of 9 sharded cells, so no per-jurisdiction comprehensiveness claim is made anywhere in this package. 1 of those cells **saturated its own shard** — the NLRP3 × CN cell — which means that cell is UNMEASURED rather than counted, and **no absence falling inside it may be reported as a measured zero**. That constraint is applied where it bites, in §III.2 below.

There is also no independent classification route to check recall against. The CPC symbol carrying more NLRP3 publications than any other holds 66 of them against 11508 publications on the symbol — 0.57 per cent. No CPC symbol carries "inflammasome" in its title at all. The classification system has no concept for this target's biology, so the term slate is the only instrument, and no independent instrument checks its recall.

I.2 What is out of scope by election, and therefore not measured

The mechanism axis stops at the sensor and its proximal inputs. By operator election (gate 2, option B), 4 terminal nodes — caspase-1, GSDMD, IL-1 β and IL-18 — are **out of scope**. They were not searched, not binned and not placed on the grid.

This is load-bearing, and the ontology requires every reader-facing surface to say it: those nodes are **not measured-empty**. Nothing in this package may be read as evidence that they are absent, thin, open, or whitespace. The terminal-cytokine arm of this biology — where the approved products and their regulatory floors sit — is outside the fence this package draws, not inside it and empty.

Concretely: 3 approved products act on that elected-out arm — canakinumab, anakinra and rilonacept, all IL-1-axis blockers, including biologics carrying BPCIA exclusivity floors, and all of them used in the CAPS spectrum this package reports families on. **Every absence statement below is an absence within the elected scope**, never an absence from the biology.

II. The binding constraint is claim scope, not corpus size

Playbook §5.3 requires families to be binned by **claim scope**, and states that a title alone is insufficient. On this corpus that rule is expensive.

Of the 522 on-target families, 456 carry claim text, 2 carry only model-derived claim summaries, 2 carry only a title, and 192 carry an abstract and no claim scope at all. At publication grain — and this statistic is over **all** 2126 binned publications, including those of off-target, no-signal and panel-only families, not over the on-target set this paragraph otherwise describes — 1207 were binned with no claim evidence whatsoever, a share of 0.568.

By operator ruling the therapeutic grid counts **only claim-backed families**. That leaves 341 families in the grid, 156 excluded for want of English claim scope, and 25 held out because their direction could not be resolved. The diagnostic cohort holds 0 families — a **measured zero** against a reachable ontology value, `imaging_or_diagnostic`, which no family claimed. Those cohorts sum to 522, the on-target total. Excluded families keep their row, their cohort label and their basis: **excluded is counted, never deleted**.

II.1 The excluded population is not a random sample

This matters because the 156 are structurally different from the families that remain, and the contrast only shows as a proportion. 117 of the 156 excluded families are single-member — a share of 0.75 — against 141 of 511 claim-backed families, a share of 0.2759. Stated as bare counts they look alike; stated as shares the excluded population is the more single-member by a wide margin. They are predominantly Chinese-origin applications arriving bibliographic-only by the ingest contract, and they cluster on the recent frontier rather than in the historical tail.

The analytic ceiling this imposes, stated once and carried everywhere: this grid is an occupancy map of the **English-claim-bearing corpus**. Every OPEN cell below is open *in that corpus*. Chinese-origin early-stage filings are systematically under-represented in it, and no cell may be read as an absence of filings worldwide.

III. The occupancy grid

The grid is built on 2 of the ontology's 4 axes — mechanism node and modality — with direction and indication used to cohort rather than to place. The ontology was committed and hashed **before the corpus existed**, with 11 mechanism values and 17 modality values.

The grid holds 187 cells. 45 carry at least one family; 142 are empty. Verdicts: 20 COVERED, 2 PARTIAL, 23 THIN, 142 OPEN, summing to 187 — the accounting closes.

III.1 Most of the emptiness is not opportunity

Of 142 OPEN cells, 113 are **not readable as whitespace**. They rest on an axis or corpus artifact rather than on a measured absence of filings, and §6.2 forbids reading them as opportunity. 29 survive that screen.

The reasons overlap — a cell may carry several, so they do not partition the population and their counts may not be summed. Over those 113 screened cells the highest-frequency reason is ARTIFACT_MECHANISM_UNRESOLVED_AT_DOMAIN_GRAIN at 56 cells, with ARTIFACT_COL_UNPOPULATED_AXIS_VALUE immediately behind it at 55, across 5 distinct reasons. **That margin is 1 cell**, so the domain grain is the leading reason and not a dominant reason, and the runner-up — an unpopulated modality column — is printed here rather than left implicit.

Taking the domain grain on its own terms: only 93 of 341 therapeutic families resolve to a specific NLRP3 structural node — a domain resolution rate of 0.2727. 230 sit on sensor_unpartitioned, which means the claim reaches the sensor and names no domain, and 18 resolve to no node at all: 248 families, the clear majority of the cohort, are **unresolved at domain grain**.

Below a domain resolution rate of 0.5, an empty cell at a *specific sensor domain* reports what claims recite, not what the field has filed. 68 cells across the whole grid carry ARTIFACT_MECHANISM_UNRESOLVED_AT_DOMAIN_GRAIN, of which 56 fall inside the screened population above — the remainder sit on **occupied** cells, which were never candidates for whitespace in the first place.

The artifact gates on the 4 partitions of the NLRP3 protein itself — the NACHT/NBD domain, the pyrin domain, the leucine-rich repeats and the oligomerisation interface. Only 27 of 341 therapeutic families resolve to any of them, a sensor-domain resolution rate of 0.0792, lower again than the all-node rate above. After the artifact is applied, the number of readable whitespace cells at a sensor-protein domain is 0.

That is the headline correction of this analysis, and its size is this. Before the artifact was applied, 64 cells satisfied every other readability test; 35 of those sat at one of the sensor-protein domains, and 9 at the NACHT/NBD domain that the MCC950 chemotype binds — including the NACHT/NBD small-molecule cell itself. Applying the artifact removed 35 cells and left 29. So the pre-artifact grid did show readable whitespace at the very domain the chemotype binds, and this grid does not. The emptiness there is our inability to read domain from claims, not the field's failure to file. A correction whose size the reader cannot see is not a disclosed correction, which is why both states are printed here, not merely the post-screen grid.

III.1a Where the readable whitespace actually is

The artifact reaches sensor-domain cells and no others, and that distinction is load-bearing rather than technical: reading *NEK7* off a claim is a different and easier act than reading *NACHT* rather than *LRR*.

But the consequence is an identity, not a discovery, and it is reported as one. The readability predicate already refuses sensor-domain cells, catch-all cells and unresolved-axis cells. The survivors are therefore non-sensor resolved nodes *by construction*: 29 of the 29 readable cells sit at such a node, and their being equal is a theorem about the predicate, not a measurement of the field. The artifact records that identity explicitly as True so that no reader is sold a tautology as a finding.

What *is* measured is how far the survivors spread: they fall across 4 distinct mechanism nodes — the NEK7 licensing kinase, the ionotropic upstream input, the ASC adaptor and the NLRP3 transcript. **Whatever readable whitespace this instrument can see is beside the sensor rather than on it, because it cannot see on the sensor at all.** That is a statement about the instrument's reach, and it is what is available.

III.2 What the axes actually reached

8 mechanism values are occupied by a resolved, specific node, out of 10 occupied in total — the difference being this analysis's own catch-all and unresolved bins. 11 modality values are occupied resolved, out of 12. **Quote the resolved counts.** Quoting the totals would count our failure to resolve as the field's spread.

Unpopulated axis values are split by kind, because only the former kind says anything about the field. 1 mechanism value and 4 modality values are recorded as **measured zeros**: reachable, and no family claimed them. That label must be read against the saturated CN discovery cell named in §I.1 — 3 of those modality values returned at most a single row inside it, so their emptiness is **not separable from that truncation**, and this package treats them as floors rather than as findings about the field. 1 modality value is **unreachable by construction** — admission to it routes a family into the diagnostic cohort, so the therapeutic grid can never contain it, and its emptiness is guaranteed rather than observed. The figure draws 10 mechanism lanes and 12 modality lanes and names the unpopulated values in its footer instead of drawing empty rows for them.

The grid of record and the drawn grid are different objects, and every count in this document is on the grid of record. The grid of record is 187 cells; the exhibit draws 120. The 67 difference are cells built on an unpopulated axis value: counted everywhere in this document, and not drawn, because an all-dot lane beside the real unresolved bucket invites exactly the whitespace misread this analysis exists to prevent. A reader counting cells in the figure will not reach 187, and should not.

III.3 The fence cells

A cell is a **drawn fence** only if both of its axes resolved. 24 occupied cells qualify, holding 63 of 341 therapeutic families between them. The largest is `sensor_pyd|solid_form` at 10 families; the runner-up is `sensor_transcript|process_intermediate` at 8.

Set that against the largest occupied cell in the grid overall, which is `sensor_unpartitioned|solid_form` at 106 families, taken over 45 occupied cells, with `sensor_unpartitioned|unresolved` behind it at 60. The biggest concentration in this

corpus sits on the catch-all node. **It is not a mechanism finding and it must not be read as one.**

III.4 Ownership is a floor, and the floor is low

328 of 341 therapeutic families carry no resolved assignee group. 3 groups resolve at all, covering 13 families, and the enumeration closes: Roche at 6, Novartis at 5, Genentech at 2.

45 grid cells carry an explicit lower-bound flag on their assignee roster. That figure carries **no information**, and is shown with that warning attached: with the owner axis this sparse the flag fires on every occupied cell, so the count is the occupied-cell count under another name. The artifact records the identity as True.

No concentration, consolidation or competitive-position claim is made anywhere in this package. The instrument cannot support such a claim.

IV. Direction, indication, modality

Direction. 278 of 341 therapeutic families claim suppression. 43 claim level modulation — changing abundance rather than per-molecule activity. 13 claim no direction, being structures, salts, forms or processes, and 0 are unresolved on the axis — the claim-backed cohort having already had its direction-unresolved families held out. 7 claim **potentiation**: deliberate inflammasome activation, pursued in oncology and adjuvancy, and a real slice rather than a theoretical possibility.

Indication, on 2 tiers that are never summed. The genetically validated tier is small and rare-disease shaped: Wikidata confirms 4 gene-disease associations for NLRP3, every one a monogenic CAPS-spectrum disorder plus a non-syndromic deafness locus. 22 therapeutic families sit on that tier — 22 on the CAPS spectrum and 0 on hereditary hearing loss, the latter a measured zero against a reachable value.

The pharmacologically pursued tier carries 245 families and looks nothing like it. rheumatologic_inflammatory leads at 108 families, metabolic_hepatic follows at 35, over 10 ranked values **within that tier** — the genetically validated values are ranked separately and never folded in, which is what keeping the tiers apart means in practice. A further 63 families are unresolved and 11 recite no indication at all — the latter absent by the nature of a composition claim, not absent from the world.

Modality. The corpus is a chemistry corpus. 190 of 341 therapeutic families are a chemical article, its solid form, or its manufacture: 132 solid form, 36 process or intermediate, 22 small molecule claimed per se. 101 are unresolved. Everything else is thin: 14 antibody or scaffold, 11 natural product, 9 peptide, 5 RNAi, 4 cell therapy, 3 active immunisation, 2 gene editing, 2 antisense.

V. Trajectory

These bands are computed over the 652 **on-target** families, not over the 780-family corpus, and are labelled accordingly.

The on-target set spans filing years 2005 to 2026. 7 families filed in 2005–2009, 19 in 2010–2014, 154 in 2015–2019, and 472 in 2020–2026 — 0.7239 of the on-target set. In the therapeutic cohort, 233 of 341 dated families fall in that last band.

Liveness in that cohort: 138 families have at least one live granted member, 293 have at least one pending member, and 7 are dead in every member.

V.1 The foundational tail is a named absence

This package originally recorded the discovery foundation as wholly absent. **That was too strong, and the re-test this analysis had itself mandated corrected it.**

4 foundational anchor families were admitted, and they are the founding groups themselves: Martinon and Petrilli of the Lausanne laboratory that defined the inflammasome, filing from 2005; Núñez and Kanneganti at Michigan; Alnemri at Jefferson; Harton at South Florida. The people are in this corpus.

What is genuinely absent is narrower and should be stated narrowly: **the CIAS1/cryopyrin identification of 2001 and the inflammasome-concept filings of 2002 themselves** — priority documents earlier than this corpus reaches at all. The gate-approved historical vocabulary was queried across the 4 jurisdictions this pipeline searches — CN/EP/US/WO — which are not every jurisdiction, and the earliest material any of it returned post-dates those publications.

The absence is named here, in its corrected and narrower form, so that the trajectory above is read as *the filing record the term slate reaches* and never as the history of the field.

VI. The overlap probes

Occupancy says who has filed *where*. It does not say whether filings in different cells claim overlapping subject matter. Independent probes address that, and **they are never combined into a single number**: genus recall is a lower bound, sequence identity is not and errs in the opposite direction, so a blended figure would take its error direction from whichever leg happened to dominate. The blended value in the data is therefore null, by rule.

VI.1 The sequence probe ran, and then disqualified itself

The sequence fence is MEASURED: 206 alignment edges, 205 candidate, 1 own, 0 indeterminate, with 5 coverage gaps. Its alignment floor holds on this target's own edges — minimum length 53, median 114, 0 edges below the production floor.

Then the §6.1a relevance audit was applied, and the layer failed it. 91 distinct publications were hit; 0 of the resolved titles are on-target. They are PCSK9 dosing patents, CLDN18.2 CAR-T constructs, tau antibody shuttles, an antithyroidin transporter antibody. The query side carries immunoglobulin constant-region and hinge motifs. **This layer is measuring antibody scaffold adjacency, not NLRP3 scope.**

Its verdict is recorded as MEASURED_NON_EVIDENTIARY, with evidentiary = False. That is a **third state**, distinct from both of the others: the probe ran and returned, so it is not unmeasured; and its returns do not bear on this target, so it is not evidence. Collapsing it into either neighbour would misreport it.

The consequence is enforced rather than merely stated: **no cell verdict in this grid takes a sequence edge as an input.**

The probe was also narrow by rule before it was narrow by outcome. Sponsors are selected per resolved assignee group, and 3 groups resolved, so 328 families in the assignee-unresolved slice were never eligible for it.

VI.2 The chemistry probe did not return

The chemistry fence is UNMEASURED, result state PENDING_AT_DEADLINE after 2718 seconds. 50 therapeutic publications carry extracted compounds; 12 were ranked as candidates and 8 were actually submitted, that being the probe's own hard cap. It reports 0 genus-overlap edges.

Note the narrowing that a returned probe would have carried anyway: 8 publications of 50 compound-bearing ones. Even a successful run would have spoken for a sample, and would have had to say so.

That count is not a measured zero. It is the number of edges a probe that never returned was able to report, and the two readings are not interchangeable. The genus overlap between the chemical estates in this corpus was **never measured**, and nothing in this document says it was measured and found empty.

VI.3 What that leaves

The sequence probe ran and was disqualified on relevance. The chemistry probe did not return. **Neither leg is evidentiary**, and every one of the 187 grid cells carries OVERLAP_UNMEASURED_OCCUPANCY_ONLY to say so.

This grid is therefore an **occupancy map only**. It reports where claims sit. It does not report whether they overlap, and no reader may infer non-overlap from an adjacent empty cell.

VII. The asset strip

11 compound code-names were approved at the gate. They are 8 molecules: MCC950, CRID3 and CP-456773 are one molecule, and dapansutrile and OLT1177 are one molecule. Treating codes as assets would have inflated this strip by 0.375.

6 of 8 have corpus name-evidence; 2 are **measured absent** from this corpus, which is a recall statement about the corpus and not a claim that no such filings exist. 6 of 8 asset rows carry an asserted identity.

MCC950 leads at 27 naming families, ahead of dapansutrile at 18, over 8 assets.

Naming is not owning, and the naming parties are not all of a kind. Measured across the 6 assets that have any, only 1 has a majority-academic naming roster. The leader MCC950 sits at an academic share of 0.5833; the runner-up dapansutrile at 0.1667, its roster mostly commercial. A blanket characterisation of these parties as academic would be false, and is not made.

What holds for all of them is the rule: a party appearing on a family that names a compound may be using it as a tool or citing it as a comparator. **No ownership is asserted from a naming relationship anywhere in this package.**

VIII. The regulatory floor is zero, and the detector was proven first

0 regulatory exclusivity floors exist for this target, state MEASURED_ZERO.

That zero was earned rather than assumed. The Orange Book leg returned 0 rows for NLRP3-directed products — and the **same query shape** returned 5 rows on its positive control, proving the detector fires before its zero was reported. The Open Targets leg returned 12 suggestions of which 0 are tier-verified; every one came back unverified and model-derived, so no sponsor and no phase anywhere in this package rests on it. 0 of 8 assets carry a registry-verified development phase.

No product directed at the elected scope is approved — the sensor and its proximal inputs. Inside that scope there is no regulatory floor to clear, which means the fence there is **purely patent-structural** — and this package has just reported that the structural layer is occupancy-only, because neither overlap probe is evidentiary.

This is not a statement that the biology has no approved therapy. As §I.2 records, 3 approved products — canakinumab, anakinra and rilonacept — act on the terminal arm that is **out of scope by election**, and they were never in range of the detector above. The zero is a fact about the elected scope and about the Orange Book leg that measured it, not about the field.

IX. What this package does not establish

Stated plainly, because absence of a finding must be distinguishable from absence of a check.

- **No claim-strength assertion.** KC is omitted by election (§3.0).
- **No overlap finding, in either direction.** One probe was disqualified on relevance, the other never returned.
- **No freedom-to-operate opinion.** Not for any cell, any family, any asset.
- **No worldwide absence claim.** The grid is the English-claim-bearing corpus; 156 on-target families are outside it and they are not randomly distributed.
- **No ownership or concentration claim.** 328 of 341 therapeutic families have no resolved owner.
- **No census.** 118 families failed ingest and are absent from every count here.
- **No mechanism claim from the largest cell.** 248 of 341 families are unresolved at domain grain.

X. Verification

The Stage-A vocabulary layer was mutation-tested across 9 hazard classes: 24 mutations caught, 0 survived. Defects in this package's own fixtures were found by that harness rather than by review.

Catch rates are reported **per layer** and never blended (§7c.4), across 7 layers, each against its own denominator:

layer	caught	mutations	what a mutation here corrupts
terms	24	24	the Stage-A vocabulary and the admit/decide separation
disposition	6	6	which families are on target
cohort	4	4	the claim-scope line that sets the denominator
grid	5	5	cell verdicts, qualifiers and the whitespace flag
fences	8	8	probe states, edge counts and the measured/unmeasured line
report_prose	8	8	the delivered sentences themselves
ledger	7	7	closure records and their negative controls

No total is given and no mean is taken: a blended figure over layers this different is a dilution, not a summary, so the artifact records blended as None — by rule, not by omission. 0 mutations went unnoticed by any gate. Read that number rather than the rates — a rate bounds these gates against the mutations someone thought to write, and says nothing about the ones nobody did. The report_prose layer exists because a guard rewritten to replace a tautology has to be tested against the real defect, or a correct rewording and a live defect look alike to it (§7c.9).

The findings ledger holds 34 rows with 0 open must-fix. 6 operator decisions are recorded in `data/operator_decisions.json`, including the ruling that set the claim-scope denominator and the agent-raised stall this run recorded against itself.

Figure: `figures/nlrp3_structural_boundary_heatmap.png` — panel A the occupancy grid, panel B the asset strip. Its render gate asserts that the families plotted equal the cohort counted.