

LPA_V4_Whitespace_and_Trajectory

Table of Contents

LPA (Lipoprotein(a)) — V4 Whitespace and Trajectory Analysis.....	1
Read-me guardrails	1
I. Scope ontology.....	2
The therapeutic / diagnostic line	2
II. The whitespace map	3
The largest cell is the least flattering one	3
Occupancy on the resolved axes.....	3
OPEN and THIN cells — and the ones that are artifacts.....	3
III. The structure that does resolve	4
The transcript fence is crowded, recent, and not the sponsors' alone	4
The assembly node is measured, and it is separate	4
IV. Trajectory vector	4
Direction of travel	5
V. Limits	5
VI. Knowledge graph.....	5
VII. What the next run should do	6

LPA (Lipoprotein(a)) — V4 Whitespace and Trajectory Analysis

Target: LPA / lipoprotein(a) · **Corpus:** 1260 publications across 376 families · **Registry snapshot:** 2026-09-16

Read-me guardrails

Read these before any number below.

- **This is candidate generation and structural occupancy, never freedom to operate.** Nothing here is a clearance opinion. A cell described as OPEN or THIN is a statement about what this corpus, on these axes, could resolve — not a statement that anyone may build there.
- **Genus recall is a lower bound.** Where a genus is reported it understates coverage; it never overstates it.

- **Sequence identity is not a lower bound and errs in the opposite direction.** It must never share a clause or a sentence with a genus recall figure, because the two carry opposite error senses and a reader who averages them will be wrong in an unsafe direction.
 - **Comparison is not reliance.** Co-location on these axes does not assert claim overlap between any two estates.
 - **The two overlap fences in this run are UNMEASURED.** Sequence fence: UNMEASURED. Chemistry fence: UNMEASURED. 0 of 2 returned a usable result, so **no overlap number is licensed in either direction** — neither "these estates overlap" nor "these estates are clear".
 - **Owner names are a floor.** Families whose assignee did not resolve are counted but unnamed, so every roster in this document is a lower bound on the parties present.
 - **KC (claim-strength) analysis is deliberately OMITTED** under playbook §3.0; the decision is recorded in data/kc_decision.json as OMIT. No claim-strength or relative-quality assertion appears anywhere in this document, and none should be inferred from cell occupancy.
-

I. Scope ontology

2 axes, defined **before** the corpus was assembled and never permitted to cross-infer: **mechanism node** (where in the Lp(a) pathway an agent acts) and **modality / chemotype** (what kind of molecule acts there). The ontology declares 4 axes in total — direction and indication are used to separate cohorts rather than to place a family on the map — and the 2 grid axes carry 8 mechanism values and 15 modality values. The pre-corpus commit timestamp precedes the corpus build, which is what makes the axes a hypothesis the data can falsify rather than a summary of what happened to be found.

Mechanism and modality are **coupled in biology** — an siRNA acts on a transcript, an apheresis column acts on a particle. That coupling is reported below as an observed correlation. It is deliberately **not** built into the grid, because encoding it would make the grid unable to show a modality appearing at an unexpected node.

2 values sit **outside** the pre-specified ontology by design. apo_a_unspecified is a terminal catch-all meaning "mentions apo(a)/Lp(a) but resolved to no specific node"; unresolved_* values mean the screen could not decide. Both are carried as first-class rows rather than dropped, because a dropped row is an invisible limitation.

The therapeutic / diagnostic line

Lp(a) is **measured** far more often than it is **lowered**, and an immunoassay antibody is indistinguishable from a therapeutic antibody on a modality axis. A family therefore enters the therapeutic cohort only on **positive matched evidence in claim scope** — claims plus LLM claim summaries, never title or abstract, because a diagnostic kit routinely recites "useful in the treatment of atherosclerosis" as background. The matched span is published per family.

Of 141 on-target families: **88 therapeutic**, 29 diagnostic, 19 with measurement direction and no claim scope to test, and 5 whose direction did not resolve. These sum to 141. Diagnostic and undetermined families are **counted and set aside**, never deleted. 3 families recite measurement *and* therapeutic use and are retained as therapeutic — that is what a companion-diagnostic clause inside a treatment patent looks like.

II. The whitespace map

The grid is mechanism node × modality over the therapeutic cohort only: 160 cells, of which **26 are occupied**.

The largest cell is the least flattering one

43 of 88 therapeutic families sit on the apo_a_unspecified catch-all row, and a further 2 on unresolved_mechanism_node. That is a statement about this method, not about the field: 427 of 1260 publications were read **without any claim text**, because claims were never extracted or recite a chemical structure naming no target. The catch-all row must be read as *unresolved*, not as protein-targeting.

43 families across 17 cells carry a resolved, non-catch-all mechanism node. Every conclusion below rests on that population, and its size relative to 88 is the single most important limitation in this document.

Occupancy on the resolved axes

The field spreads across **5 resolved mechanism nodes** and **10 resolved modality classes**. The occupied totals are higher (7 and 11) because they include this analysis's own catch-all and unresolved bins; quoting those as field diversity would count our failure to resolve as somebody else's spread.

mechanism node	therapeutic families
apo_a_unspecified (catch-all — not a finding)	43
LPA_transcript	23
apo_a_protein	8
apo_a_apoB_assembly	6
LPA_gene_locus	3
Lp_a_particle_extracorporeal	3
unresolved_mechanism_node	2

OPEN and THIN cells — and the ones that are artifacts

134 cells are empty. **Only 35 of them are readable as whitespace**. The remaining 99 are axis or corpus artifacts under playbook §6.2 — combinations that are biologically incoherent, or that this corpus could not have populated — and reporting them as opportunity would be the single easiest way to mislead a reader of this document. 18

occupied cells carry a THIN verdict, meaning they rest on so few families that a single reclassification would move them.

Every cell carries its qualifiers and an explicit `readable_as_whitespace` flag, so no cell's emptiness can be quoted without its reason.

III. The structure that does resolve

The transcript fence is crowded, recent, and not the sponsors' alone

LPA_transcript holds 23 therapeutic families: **16 siRNA against 3 antisense**. The dates are the substance. The ASO families run from 2001 to 2025, while the siRNA families begin in 2016 and run to 2025 — a fence built almost entirely in the last decade around one mechanistic idea.

It is **not** the property of the clinical sponsors. The siRNA cell names 7 distinct assignee groups, and a further 7 families in that cell have **no resolvable assignee at all**, so the roster is a floor and the true number of parties standing inside this fence is higher than any name list here.

The assembly node is measured, and it is separate

apo_a_apoB_assembly holds 6 families, 4 of them small molecule. This node was absent from an earlier build of this analysis and was closed by targeted re-ingest rather than disclosed as an absence — which matters, because it is the node that makes the trajectory argument possible. An agent blocking the assembly of apolipoprotein(a) with apolipoprotein B-100 interferes with **formation of the particle** rather than **production of its distinctive protein**. It is a categorically different fence, and a result about one does not transfer to it.

IV. Trajectory vector

5 named clinical assets are in scope. They occupy **3 cells across two mechanism nodes**: 4 at LPA_transcript and 1 at apo_a_apoB_assembly. 3 of the 5 are siRNA and therefore share a single cell.

That shared cell is the whole trajectory finding. A result speaking to the **consequence of lowering Lp(a)** reaches all 5. A result speaking to **a particular way of lowering it** reaches only the cell it lands in. The first Phase III topline for this target was reported on 2026-09-04; the remaining programmes read out years later, and at the registry snapshot the most advanced of the non-sponsor siRNA programmes stands at Phase II — explicitly not parity with the others.

Ownership does not follow the sponsor. For **2 of the 5 assets**, the composition estate sits with the **platform originator rather than the trial sponsor**. Attribution here is

lineage-level — assignee, mechanism and modality agreement only — because the corpus never names these compounds. It must not be read as claim coverage of any asset.

Direction of travel

Filing activity at the transcript node runs to 2025, and the corpus has a hard horizon: 28 publications could not be retrieved at all. That absence is **not random** — it falls disproportionately on oral small molecules and Chinese applicants, which is to say on precisely the newest and most divergent entrants. The trajectory stated here is therefore a floor on how fragmented the field has become, not a measurement of it.

V. Limits

Stated plainly, because several of these are large.

1. **Both overlap fences are UNMEASURED** (0 of 2 usable). The chemistry fence cannot be built at all: 0 Markush scaffolds were extracted corpus-wide, so no genus compiles. Any statement about estate-to-estate overlap is outside this document's reach.
 2. **427 of 1260 publications (0.339 of the corpus) carry no claim text**, and 9 rest on an LLM claim summary with no claim text beside it. Claim scope for those is unread.
 3. **38 on-target families (27.0% of them) rest on a title or abstract mention with nothing in claim scope**. They are counted as present in the field, not as evidence of what is claimed.
 4. **28 publications are permanently unreachable** and 222 discovery rows named a US document that could not be resolved to a publication number.
 5. **Claim scope is screened by deterministic regex over claim text plus LLM claim summaries**, which is a deviation from the playbook's mandated LLM claim-scope read *and* from its hard gate "never derive a claim-scope fact by regexing prose". The mitigating measurement: 9 publications are claim-summary-only while 824 carry real claim text, so the named undercount mechanism reaches a minority of rows. It remains a gate deviation, not a judgement call.
 6. **Discovery searched title and abstract only**. Every count in this document is a floor.
 7. **29 diagnostic and 19 undetermined families are excluded** from every grid figure. A reader comparing this map to a raw family count will not reconcile without them.
-

VI. Knowledge graph

data/lpa_knowledge_graph.json carries 160 nodes and 384 edges over targets, mechanism nodes, modalities, cells, families, assignee groups and clinical assets.

The graph is validated on **properties, not totals**: 0 self-edges, 0 duplicate triples, 0 dangling endpoints, and node and edge counts derived from the serialised collections so a published headline cannot drift from the artifact. This check exists because a previous V4 run passed two adversarial reviews on an edge-count headline where a third of the edges were a family pointing at itself.

8 axis values are unoccupied — ontology values that **no** family in the therapeutic cohort occupies, 3 on the mechanism axis and 5 on the modality axis. They are declared rather than pruned, because a dropped row is an invisible limitation.

An unoccupied axis value is an artifact, not whitespace. 2 of the 8 are unreachable by construction: admission to them routes a family out of the therapeutic cohort altogether, so their emptiness is guaranteed by the cohort rule before any data is seen and says nothing whatever about the field. Only the remaining 6 are measured-zero values. Reading either kind as an opportunity is the easiest available way to mislead a reader of this document, which is why they are not drawn on the map — they are named in the figure footer instead, split by kind.

VII. What the next run should do

- **Measure the chemistry fence.** It needs Markush scaffold extraction to succeed on at least the composition families of the 5 clinical assets; at 0 scaffolds corpus-wide there is nothing to compile.
- **Recover the horizon.** 28 unreachable publications bias against oral small molecules and Chinese entrants — the two groups most likely to change the shape of the map.
- **Attack the catch-all.** 43 families are unresolved for want of claim text; targeted claim extraction there would do more for this analysis than any new discovery shard.
- **Resolve the assignee floor.** 7 families in the single most important cell have no resolvable owner.