Doing mathematics with AI agents: A case study

Rational torsion on simple genus two Jacobians: a human–AI collaboration

Andrew V. Sutherland (MIT)

Joint work with Jennifer S. Balakrishnan, Filip Najman, and Ari Shnidman
Slides created by Claude Fable 5.1, reviewed by GPT Astra 6

arXiv:2608.28543 · notebook: github.com/AndrewVSutherland/Genus2TorsionNotebook

Mordell–Weil

Let A be an abelian variety over a number field K.

Theorem (Mordell 1922 for elliptic curves over $\mathbb{Q}$; Weil 1928 in general)

$A(K)$ is a finitely generated abelian group:

$$A(K) \simeq \mathbb{Z}^r \oplus A(K)_{\mathrm{tors}}, \qquad A(K)_{\mathrm{tors}} \text{ finite.}$$

Two very different invariants: no algorithm is known for the rank r; the torsion subgroup is computable for any given A (reduce mod good primes, then height bounds).

Fix $K = \mathbb{Q}$ and let A vary over all abelian varieties of dimension g.
Which finite groups occur as $A(\mathbb{Q})_{\mathrm{tors}}$?

Finiteness for each A says nothing about a uniform bound.

Dimension one: Ogg's conjecture and Mazur's theorem

For $E/\mathbb{Q}$, Beppo Levi (1908) and Ogg (1970s) conjectured the list; Mazur proved it.

Theorem (Mazur 1977)

$E(\mathbb{Q})_{\mathrm{tors}}$ is one of the fifteen groups

$$\mathbb{Z}/n \ \ (1 \le n \le 10 \text{ or } n = 12), \qquad \mathbb{Z}/2 \times \mathbb{Z}/2n \ \ (1 \le n \le 4).$$

Each occurs for infinitely many $\overline{\mathbb{Q}}$-isomorphism classes.

Why it works: a point of order N on E is a rational point on the modular curve $X_1(N)$.

- $X_1(N)$ has genus 0 exactly for the N on the list (plus the $X_1(2, 2n)$ analogues).
- Mazur: for the other N, $X_1(N)(\mathbb{Q})$ consists of cusps only (Eisenstein ideal).
- Merel (1996): uniform bound over number fields of degree d.

Dimension two: three kinds of abelian surface

Over $\overline{\mathbb{Q}}$, every principally polarized abelian surface is a product $E_1 \times E_2$ or the Jacobian $J = \mathrm{Jac}(X)$ of a genus-two curve $X\colon y^2 = f(x)$, $\deg f \in \{5, 6\}$.

Over $\mathbb{Q}$ there are further cases: Weil restrictions of elliptic curves over quadratic fields, and geometric Jacobians whose curve has a Mestre obstruction to descent, even when the abelian surface itself descends. This talk is about Jacobians of curves defined over $\mathbb{Q}$.

	description	largest known before this paper
products $E_1 \times E_2$ over $\mathbb{Q}$	$E_1(\mathbb{Q})_{\mathrm{tors}} \times E_2(\mathbb{Q})_{\mathrm{tors}}$	read off from Mazur: $[2, 2, 8, 8]$
geometrically **split** J	$J_{\overline{\mathbb{Q}}} \sim E_1 \times E_2$	Howe–Leprévost–Poonen 2000: $[2, 2, 4, 8]$
geometrically **simple** J	no elliptic factor over $\overline{\mathbb{Q}}$	Elkies 2024: $[2, 2, 2, 10]$

$G \hookrightarrow J$ is a rational point on a Siegel modular *threefold* $\mathcal{A}_2(G)$; only finitely many G have a component not of general type (Borisov 1999), and rational points on such threefolds are beyond current methods.

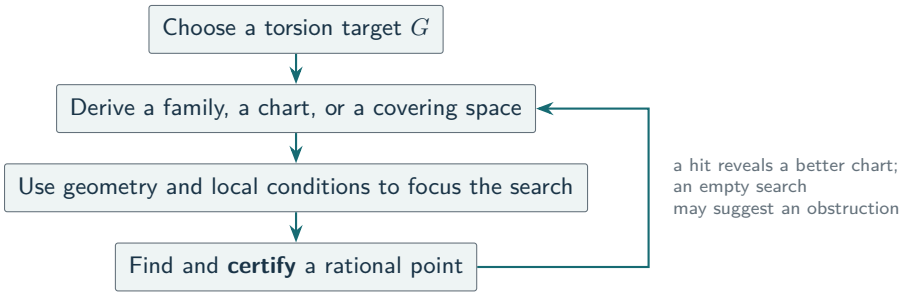
Conjecturally $\#J(\mathbb{Q})_{\mathrm{tors}}$ is uniformly bounded, but not even a conjectural list exists.

Before this paper

- Genus two is concrete: explicit models, divisor arithmetic, order-n classes on the curve via $f = h^2 - c\,(x - r)^n$, and explicit square conditions for division by 2.
- Databases: LMFDB genus-two curves and extended Booker–Sutherland database; 65 torsion groups realized by geometrically simple Jacobians.
- Constructions: Flynn, Leprévost, Howe, Platonov–Petrunin, Nicholls, Elkies.
- Largest prime order: 31 (found in 2026 in a 2022 GitHub repository by a survey Epoch AI ran with GPT-5.6 Sol).
- Total: 67 groups known for geometrically simple J; record order 80.

Goals: realize new groups, decide which occur infinitely often, understand the loci where large torsion appears.

A search program with mathematical feedback



In practice a funnel: parametrize $\to$ cheap $\#J(\mathbb{F}_p)$ divisibility prefilter $\to$ exact `TorsionSubgroup` $\to$ simplicity certificate.

An empty search is useful when its scope and its explanation are recorded.

The headline: order 96

Theorem

Let $T = \{a, b, c, u, v\}$ be positive integers with distinct squares such that $a^2 + b^2 + c^2 = u^2 + v^2$ and $a^4 + b^4 + c^4 = u^4 + v^4$. Then $X_P \colon y^2 = x \prod_{t \in T}(x - t^2)$ has $[2, 2, 2, 12] \hookrightarrow \mathrm{Jac}(X_P)(\mathbb{Q})$.

Three low-height points give non-isomorphic curves with $J(\mathbb{Q})_{\mathrm{tors}} \simeq [2, 2, 2, 12]$ exactly and J geometrically simple; the first $[2, 2, 2, 12]$ on *any* genus-two Jacobian over $\mathbb{Q}$:

$$y^2 = x\,(x - 120^2)(x - 143^2)(x - 266^2)(x - 218^2)(x - 241^2).$$

The whole proof is one identity:

$$(x - a^2)(x - b^2)(x - c^2) \;=\; x\,(x - u^2)(x - v^2) - a^2 b^2 c^2.$$

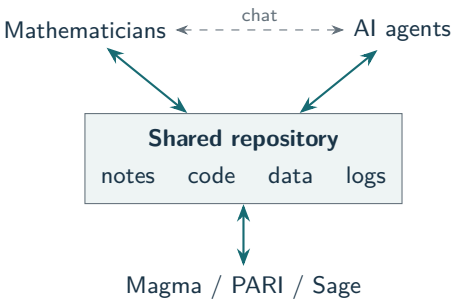
$S \subset \mathbb{P}^4$ is of **general type** but contains a genus-one curve through the point $P_1 = [120 : 143 : 266 : 218 : 241]$ of the curve above (Choudhry 2016; $E \colon y^2 = x^3 - 21x - 20$, rank 1): infinitely many Jacobians containing $[2, 2, 2, 12]$.

Twelve months of coevolution

- **July 2025**, ICERM workshop "Algebraic Points on Curves": project starts. o3, Claude Opus 4, Gemini 2.5 Pro tried out of curiosity; *none contributed*.
- **Spring 2026**: a shared GitHub repository becomes the lab notebook. Notes, code, logs for every route; humans and agents both write to it.
- **Summer 2026**: GPT 5.6 Sol, Claude Opus 4.8/5, Fable 5 (and codex as PR reviewer) make substantive contributions. Agents run overnight campaigns.
- **Final phase**: the authors believe the project is finished and let the agents "off the leash" into the repository with little oversight, since any discovery can be certified easily.
- **August 2026**: human audit of every computational claim; a public verification suite (eight Magma scripts, under one minute); paper written entirely by the authors.

Output of the agents: about 6,850 files, 300 MB: 1,082 Magma scripts, 148 Python, 329 notes, 3,390 run logs. Preserved publicly, warts and all.

The shared notebook carried the context



Each agent session starts empty. What survived from one session to the next:

- **Route triple**: `notes/<route>.md`, `code/<route>.m`, `results/<route>*.log`.
- **Dated notes, never retro-edited**; corrections are new notes plus an errata index.
- **Skills**: the methodology distilled into loadable documents.
- **Two-certificate rule**: no result without exact torsion *and* geometric simplicity.

Dated failures, corrections and reusable methods are the part that carried over. Several results came from an agent reviving a route that another agent, or a human, had abandoned.

One night, July 17–18, 2026



Thirteen human messages in thirty hours, most of them one line. Three carried the steering:

- "make a ranked list of the ten torsion subgroups . . . most likely to be realizable, develop a strategy for realizing each . . . fan out subagents"
- "I asked GPT 5.6 codex to review the repo and give its own top-10 list . . . let me know what you think"
- "the machine is now 50% free . . . proceed with your plan"

What the agent actually did

1. Read the draft, all notes, code and data; fanned out subagents; ranked ten targets by existence evidence (split realizations), absence of local obstructions, unfinished machinery in the repo, prize value.
2. Reviewed GPT 5.6's rival ranking against the notes with six verification subagents: one target already realized, one plan aimed at a lane closed by an impossibility theorem, confidence inflated by reading "no local obstruction" as evidence. Conceded a miss of its own: $[5,5]$ is the smallest open group.
3. Launched ten production lanes. The $[2,2,2,12]$ lane proved a square-class identity in the $[2,2,2,6]$ chart that explained its 59 near misses and 0 hits, switched to a neighbouring one-parameter family of charts, and found a parameter point with all four square conditions met.
4. Re-verified from scratch: exact torsion $[2,2,2,12]$; Frobenius certificates at $p = 37, 73, 113, 149$; QM excluded by the Laga–Schembri–Shnidman–Voight bound.

The model it found: $y^2 + (x^2+1)y = 756900x^6 + 737595570x^5 + \cdots + 830742747091037849$.

From a point to a surface: the human half

- **July 18**: a second curve the same evening (renewed parameter search, found independently by two lanes); both are isolated points in an ugly chart.
- **July 20**: Ari Shnidman rewrites the moduli problem as a surface over $\mathbb{P}^2_{A,B,C}$ cut out by two square conditions. An agent verifies the formulas, finds that *both branch sextics factor* and share a conic, sieves to height 20000, finds a third curve, and notices every orbit is $\{3 \text{ values}\} \times \{2 \text{ values}\}$.
- **July 20**, note `22212_symmetric_model.md`: the surface is $a^2+b^2+c^2 = u^2+v^2,\ a^4+b^4+c^4 = u^4+v^4$ and the curve is $y^2 = x\prod(x-t^2)$. Ten-line proof of the torsion.
- **July 23**: the displayed witness becomes $y^2 = x(x-120^2)(x-143^2)(x-266^2)(x-218^2)(x-241^2)$.
- **Later**: Choudhry's multigrade paper supplies the elliptic curve on S.

Neither half came out of the other's method: the surface did not come from the search machinery, and the search was not guided by the surface.

Checks and balances

- **Cross-model review.** Fable reviewed GPT 5.6's plan; codex reviewed every pull request (a negative-residue bug in a C sieve, a hard-coded scratch path in a "reproducibility" script, a verifier with no assertions). Agreement between models sharing one notebook is weak evidence; a disagreement that ends in a checkable correction is the product.
- **Errata index**, publicly preserved. Retracted working claims: "two new $[2, 22]$ curves" (isomorphic to known ones); cyclic $[40]$ as a record (Elkies 2001); an "infinite family" from one seed on a degenerate branch; a "genus 7" verdict that was only a lower bound; conductors from `genus2red` silently missing their 2-part.
- **Human audit** (Aug 22): every computational claim mapped to a script; one displayed curve differed from the certified one and was fixed.

Profile of the errors: code bugs, unsupported novelty claims, overgeneralization from one example, misread tool output. None survived certification.

Research payoff, and sharper questions

- Six new torsion groups for geometrically simple genus-two Jacobians over $\mathbb{Q}$ (67 $\to$ 73); two new for split ones.
- New record order 96; infinitely many Jacobians containing $[2,2,2,12]$, from genus-one curves on a surface of general type.
- Geometry of several moduli spaces: $\mathcal{A}(2,2,4,4)$ Calabi–Yau with dense points; $M(2,2,2,8)$ a K3 with dense points; the $[2,2,20]$ curve unique in its family by Chabauty.
- Certified tables, a one-minute verification suite, and a frozen notebook to continue from, dead ends included ($\mathbb{Z}/35$, $[5,5]$, $[2,24]$, $[4,16]$).

For many of the largest groups, every known point of $\mathcal{A}_2(G)$ lies on a **special locus**: extra endomorphisms, or a marked torsion class *on the curve* (S° and Elkies' family are of this kind). Are these loci more accessible to search, or genuinely richer? Fifteen open questions in the paper, from "realize $[5,5]$" to "prove some $\mathcal{A}_2(G)(\mathbb{Q})$ is empty".

Takeaways that generalize

1. **Context is capital.** Each agent session starts empty. The shared structured notebook turned weeks of failures into an overnight success. Invest in the record.

2. **Let agents search; let humans extract.** Broad, cheap, unglamorous exploration is where the agents excelled here; recognizing that the point lay on a beautiful surface was the human contribution. Keep the two in a feedback loop.

3. **Make verification cheap and non-negotiable.** "Off the leash" was safe only because every output of value is a certifiable witness. Decide the certificate before the search.

4. **Use more than one model, adversarially.** Mostly for calibration, sometimes for ideas. Require disagreements to end in a checkable correction.

5. **Oversee where the errors are.** Novelty against the literature, scope of claims, interpretation of tool output. Keep an errata index; never retro-edit.

Six new groups for geometrically simple Jacobians

group	order	how
$[2,2,2,12]$	96	the surface S; infinitely many Jacobians contain the group
$[2,2,20]$	80	variant of Elkies' 5-torsion construction; Chabauty at $p=3$: *unique* in its family
$[2,2,4,4]$	64	$\mathcal{A}(2,2,4,4)$ is a Calabi–Yau threefold, open in $E \times_Y E'$ for a 2-isogeny of elliptic surfaces; rational points Zariski dense
$[2,2,2,8]$	64	an order-8 class $[P-\infty]$ on the curve $\Leftrightarrow$ the K3 surface $(\sum_{i<j} x_i x_j)^2 = 4abcd$, the universal elliptic curve over $X_1(8)$; Zariski dense
$[2,4,8]$	64	$y^2 = x(3x-5)(5x+333)(16x^2+23x+576)$: the quadratic factor gives 2-rank 3; found by the search strategy of §4
$[6,6]$	36	contact-6 chart $f = h_6^2 - (x-1)^6$ plus independent 3-torsion

Also new: generic ($\mathrm{End} = \mathbb{Z}$) examples of $[2,2,14]$; and $[11]$, $[2,2,4,4]$ for geometrically split Jacobians.

Tables: 73 groups (simple), 77 groups (split); every row certified by a Magma script.

Three collaborators, looking back

"The thing that most changed my opinion about the capabilities of AI and how useful it is, was when I started using Codex (and later also Claude Code), which allowed the AI to code and run massive computations. ... I let multiple agents code and run programs on the server and have them each write down notes of what they did and what they proved."

— Filip Najman

"Looking back, it's almost hard to believe that in the summer and fall of 2025, LLMs would readily hallucinate Magma commands. ... By the end of this project, we saw how powerful AI agents could be for mathematical research, in particular when guided by precisely posed questions and ample context."

— Jennifer Balakrishnan

"By the end, I admit to feeling a disturbing loss of control over the math when seeing what the LLMs can do. ... But I learned that to make maximal long term use of the tools, it is best to go slowly and systematically, always trying to understand why any improvement or finding is really working so as to make the best 'next move.' "

— Ari Shnidman

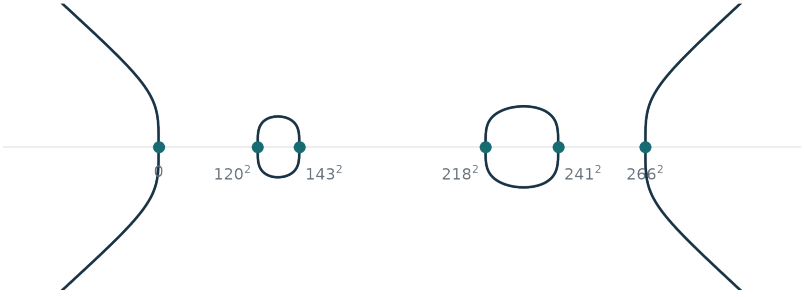
Since the paper: a living census

genus-2-torsion.github.io (Filip Najman, October 2026), seeded with the paper's tables.

- Every curve re-verified from scratch in Magma (exact torsion; a certificate of geometric simplicity or an explicit splitting), in three classes: geometrically simple, split over $\mathbb{Q}$, simple over $\mathbb{Q}$ but split over $\overline{\mathbb{Q}}$. LMFDB links; submissions by GitHub issue, verified automatically.
- **Three new geometrically simple groups**, all found by Najman: $[5,5]$ (order 25, the smallest open group of the paper's Question 1), $[3,12]$, $[4,12]$. Now **76** groups; smallest open group $\mathbb{Z}/35$; record order still 96.
- **Infinitely many, exactly:** certified for 60 of the 76 groups (an explicit family on a rational or rank-one base, a strict Frobenius prime, a residue class on which $\gcd \#J(\mathbb{F}_p)$ pins the group). For $[3,12]$ the base is a rank-one elliptic curve of conductor 990 in the order-12 chart.
- Infinitude still open (isolated examples, several with real multiplication): $[27]$, $[29]$, $[31]$, $[33]$, $[34]$, $[36]$, $[39]$, $[40]$, $[2,18]$, $[2,22]$, $[2,26]$, $[2,28]$, $[4,12]$, $[5,5]$, $[6,6]$, $[2,2,20]$.
- Split side: 78 groups, record order 128, and eleven groups proved *impossible* for geometrically split Jacobians (Mazur plus the Galois action on the endomorphism algebra).

$$y^2 = x\,(x-120^2)(x-143^2)(x-266^2)(x-218^2)(x-241^2)$$

$$J(\mathbb{Q})_{\mathrm{tors}} \simeq \mathbb{Z}/2 \times \mathbb{Z}/2 \times \mathbb{Z}/2 \times \mathbb{Z}/12$$



real locus; differences of the six Weierstrass points generate $J[2](\mathbb{Q}) \simeq (\mathbb{Z}/2)^4$; vertical axis compressed

---

paper: arXiv:2608.28543 · certificates: github.com/AndrewVSutherland/Genus2Torsion
notebook, including the discovery transcript:
github.com/AndrewVSutherland/Genus2TorsionNotebook

Another project: torsion of elliptic curves over number fields

Merel (1996): $\Phi(d) = \{E(K)_{\mathrm{tors}} : [K : \mathbb{Q}] = d\}$ is finite for every d. Determining it:

d	who, when	$\#\Phi(d)$	sporadic (finitely often)
1	Mazur 1977	15	none
2	Kenku–Momose, Kamienny (four papers, 1986–1992)	26	none
3	Derickx, Etropolski, van Hoeij, Morrow, Zureick-Brown 2021	26	$\mathbb{Z}/21$
4	Derickx–Najman 2025	38	none
5	Najman, September 2026	35	$\mathbb{Z}/28$, $\mathbb{Z}/30$, $\mathbb{Z}/2\times\mathbb{Z}/18$
6	Najman, arXiv:2611.03933, November 2026	48	$\mathbb{Z}/25$, $\mathbb{Z}/37$, $\mathbb{Z}/4\times\mathbb{Z}/12$, $\mathbb{Z}/2\times\mathbb{Z}/24$
7–10	Najman, forthcoming		

Degree 6: the 44 groups occurring infinitely often were known (Derickx–Sutherland); $\mathbb{Z}/25$ and $\mathbb{Z}/37$ each arise from a unique curve over a unique sextic field; $\mathbb{Z}/2 \times \mathbb{Z}/24$ is new. New ingredient: a formal immersion criterion at $p = 2$ via the Cartier operator. Degrees 6–9 will form one paper of about 120 pages, degree 10 a separate one.

From the AI disclosure of arXiv:2611.03933

"After completing [the quartic and quintic classifications] and giving talks about this work, I was asked several times whether we could do such a classification for number fields of degree $d > 5$. My answer was that I thought it was not possible as it would require explicit computations that did not seem feasible to me. To check whether this was correct, in mid-October 2026 I asked AI whether this was possible, and whether it could do it. To my surprise, it was able to complete the classifications . . . not only in degree 6, but later also in degrees 7, 8, 9 and finally 10. After some initial hesitation about releasing primarily AI-generated mathematics, I decided that the results were simply too interesting and the mathematics too beautiful to just stay on my computer."

— Filip Najman

"OpenAI's GPT-6.1 Sol produced almost all of the mathematical content of the first version of the proof and carried out most of the computations. Anthropic's Claude Opus 5.5 reviewed that version, independently recomputed the relevant computations and the explicit examples, wrote the draft of the present manuscript and organized the code repository. I carefully reviewed the paper line by line and rewrote most parts of it. I understand it completely, and take full responsibility for the content of the paper."

— Filip Najman

Filip, by email

"I got these results and was reading the proofs of what AI did, and I was really in awe. There is this great argument to improve and make computationally much more efficient formal immersions at 2, but also so many small smart simple arguments which when you read them are obvious and simplify and clarify so many of the arguments I've been doing for a decade. And I thought should this really not be shared, why, for what purpose? I thought it was completely crazy.

I think it's beautiful math, I enjoyed reading and thinking about it, and it should be shared. I think a lot of the conversation is focused on who gets credit for what, but I personally don't care that much about that, so I thought I won't really think about that, and will just share whatever I obtain."

— Filip Najman, November 2026

Another project: the elliptic curve rank leaderboard

elliptic-rank.icarm.cloud (ICARM, NSF): "Can we find small elliptic curves of high rank?" A curve is submitted with witness points whose independence is certified by 2-descent; 6,552 curves from 55 submitters since May 2026. Motivated by the Elkies–Watikins paper *Elliptic curves of large rank and small conductor*, initially populated with known examples of rank up to 29 [Elkies–Klagsbrun 2025].

On August 19 I received a text from Levent Alpöge:

```
y^2 = x^3 - 26149266969750243961233899047871157627419207130526434470813999930 7 x
    + 53706092008565228442861216100591608869938724607894717195123911907190573781715962031272981146771206
```

"with my friend Ava Howell (and my other friend claude of course!) please please please keep ultraconfidential . . . i did think you might like it so texting :p"

— Levent

"You could quietly submit this to elliptic-rank.icarm.cloud. Honestly, this is cooler than JC as far as I am concerned."

— my reply

Submitted the next day as entry 273: rank 30, a new record;[1] rank 31 followed on August 23.

[1] Elkies and Klagsbrun found a rank-30 curve in September 2025 but did not announce it until August 28, 2026 when they posted to the leaderboard.

Equations and nothing else?

"The Elkies–Klagsbrun construction of a rank 29 elliptic curve over the rationals came with a geometric explanation, but the new rank 30 and rank 31 elliptic curves came with equations and nothing else. We didn't learn anything really."

— Peter Scholze, Heidelberg Laureate Forum 2026

"The precise claim is about elliptic K3 surfaces over $\mathbb{Q}$ with Picard rank 19 and trivial Galois action on the full NS lattice. I have identified the complete list conditional on modular curve genus $g \leq 6$, including one with $g = 5$. If a K3 surface has a $\mathbb{Q}(T)$ fibration of rank 17, it must have $g \geq 7$, or be one of $\{X_{948}, X_{1008}, X_{1020}$ (infinite family), X_{1032}, X_{1092} (infinite family), $X_{1152}\}$. ... I am preparing a nicely organized website with a searchable database of all these surfaces, including explicit maps from the modular curve. ... I am also writing up the accompanying human-readable proof of completeness, and formalizing the proof in Lean."

— Samuel Li, ICARM Zulip, October 2026

Rank 32

Leaderboard entry 4239, 15 October 2026: "Specialization of a rank 17 fibration of the known K3 surface of discriminant 1152. Discovered by Noam Elkies, Zev Klagsbrun, and Samuel Li." At $T = -2868768/362209$ in the elliptic fibration of X_{1152}

$$y^2 + (Lx + B)y = x(x + D)(x + E), \qquad B = pE + qD + pq(L - p - q),$$

$$L = 466548T^2 - 828066T + 32416, \qquad p = -478380T^2 - 1076788T + 103008,$$

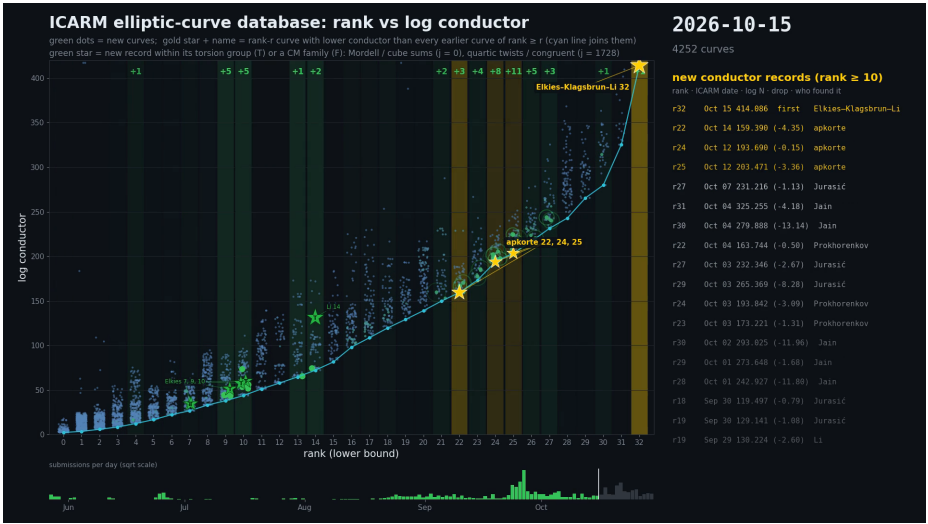
$$q = 82612T^2 - 180205T + 81622,$$

$$D = -26666192028T^4 - 91335649060T^3 + 11566653808T^2 + 239872553168T - 53374213248,$$

$$E = -47423161625T^4 + 1185250025 74T^3 - 104259187849T^2 + 76235905664T - 24372268764.$$

An explicit-formula bound (S. Hackfort, all primes below $6 \cdot 10^{10}$, certified with Arb) gives analytic rank ≤ 32 under GRH, so with 32 independent points the rank is exactly 32 under GRH and BSD.

This morning, 4 November 2026: a second rank-32 curve, entry 6534, a fibre of the same surface, found by Blair Butler, Lewis Combes and Shir Levav-Porat, "search performed by Claude Code Opus 5.5"; a Selmer-group bound (class groups computed under GRH) gives rank exactly 32 under GRH.



animation by Michael Rubinstein