Episode Book

EPISODE #5957

The Scientific Plugins Hiding in Plain Sight

PUBLISHED	14 October 2026 at 13:47 UTC
DURATION	21:43
LISTEN	myweirdprompts.com/episode/scientific-plugins-chatgpt-directory-09166/
HOSTS	Corn, Herman, Hilbert
SCRIPT MODEL	deepseek-flash
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.4
SOURCES	23 cited by the research (listed at the end)
BOOK GENERATED	19 October 2026 at 17:41 UTC

SYNOPSIS

ChatGPT's plugin directory now lists 5,723 plugins with 238,000 callable tools, yet only 52% have useful descriptions. The leaderboard rewards commercial SaaS with clean schemas — Ashby, Supabase, Miro, Stripe — while scientific tools like Wolfram and Consensus don't crack the top rankings. Meanwhile, the genuinely capability-expanding drug discovery work — LIDDIA, MADDIE, DrugPilot — lives in arXiv papers that haven't migrated into the directory at all. This episode explores the split between the visible commercial layer and the deeper research layer, and asks whether general-purpose assistants are quietly becoming the access layer for serious scientific infrastructure. We dig into measurable impact: LEADS showing 20.8% time savings in study selection versus a systematic review finding GenAI missed 68-96% of studies during searching. The evidence cuts both ways, and the honest numbers are more interesting than the marketing.

TRANSCRIPT



Corn

Daniel has been building with MCP since it was a scrappy little protocol nobody had heard of, and he's been watching the ChatGPT plugin library the way some people watch a tide gauge. Not because he wants to use every plugin. Because he thinks the directory is the clearest public signal of where agentic AI is actually going.



Herman

And what caught his eye specifically?



Corn

The scientific corner of it. He expected finance and analytics, everybody expects finance and analytics. What surprised him was how substantial the ecosystem around scientific computing, drug discovery, and protein structure modelling has become. Highly specialised disciplines, and yet there's a whole shelf of tools for them.



Herman

Right, and his actual question underneath that is the interesting one. Not which plugins exist. Which of them have stopped being demos.



Corn

He wants the distinction between integrations that make existing software a bit easier to use, and integrations that expand what a researcher can do. Measurable time savings, new discoveries, previously impractical experiments becoming feasible. And then the big one. Whether general-purpose assistants are quietly becoming the access layer for serious scientific infrastructure.



Herman

That last part is the one I keep chewing on, because there's a version of it that's obviously true and a version that's mostly marketing, and they look identical from the outside.



Corn

Then let's start with what the directory actually is, because the number problem shows up immediately.



Herman

It does, and it's worse than most people realise. So plugins now bundle capabilities into reusable workflows across both ChatGPT and Codex, and the two products share one universal plugin directory. Under the hood an app needs an MCP server that defines its tools, and then optionally a web component that renders in an iframe. ChatGPT implements the open MCP Apps UI standard, so a UI you build once runs on any host that supports it.



Corn

Which is the smart part of the design. Build the interface once, it travels.



Herman

It travels, and that's why the counts exploded. Agent Discoverability is tracking five thousand seven hundred and twenty-three ChatGPT plugins as of today. Up four hundred and sixty-nine in seven days.



Corn

In a week.



Herman

Two hundred and thirty-eight thousand callable tools across them, nine thousand and eighty-six visible brands. And here's the part that should make anyone cautious, only fifty-two percent of the documented tools have descriptions that are actually useful.



Corn

So half the directory is a name and a shrug.



Herman

Half the directory is a name and a shrug. But the counting problem is worse than that. The MCP App Store indexed three thousand seven hundred and fifty-six ChatGPT apps back in September. The community awesome list counted one thousand six hundred and twenty-four apps across forty-four categories in July.



Corn

Three trackers. Three numbers. Overlapping dates.



Herman

And none of them is wrong, exactly. They're counting different things. One counts apps, one counts plugins, one counts MCP servers that may or may not be surfaced as plugins. So when somebody says the ecosystem has five thousand entries, ask them what an entry is.



Corn

A directory that can't agree on its own size is a directory that's still being built.



Herman

I'd put it more kindly, but yes. And the interesting question isn't the total anyway. It's the shape.



Corn

So if we can't agree on how many plugins there are, what can we say about the ones that matter for science?



Herman

We can say they're not where you'd expect to find them. Start with the categories that are dense and completely unsurprising. Analytics and Data, Finance and Trading. FactSet, LSEG, MSCI, CB Insights, Aiera, Datarails. Hex, Deepnote, MotherDuck, ThoughtSpot, Omni, PostHog, Mixpanel, Amplitude.



Corn

Those are obvious because the pitch is legible in one sentence. Connect the assistant to the data, ask questions, get answers.



Herman

And nobody needs domain expertise to see why that works. If you're already paying for FactSet, having the assistant reach into it is a straight line.



Corn

The finance shelf is the easy sell. Everybody understands wanting faster answers about money.



Herman

Now here's the twist that I think Daniel half-suspected. The directory's own ranking logic does not reward scientific tools. Agent Discoverability scores plugins, and the top of the ChatGPT leaderboard is Ashby at ninety-six, Supabase at ninety-three, Miro at ninety, Shopify at eighty-eight, Stripe at eighty-five.



Corn

So the winners are commercial SaaS with tidy tool descriptions.



Herman

That's exactly what the score is measuring. Discoverability rewards clean schema, clear descriptions, obvious naming. Wolfram and Consensus don't crack the leaderboards.



Corn

Wolfram, which does symbolic computation. Consensus, which synthesises the literature.



Herman

Neither of them scores like a project management tool. Which tells you the ranking is a map of commercial integration, not scientific depth. If you browsed that leaderboard looking for the frontier of research tooling, you'd walk away thinking there isn't one.



Corn

That's the trap. The surface of the directory is optimised for a different question than the one Daniel's asking.



Herman

And it's worth saying out loud what the visible scientific entries actually are. Wolfram appears as add computation and knowledge. Symbolic computation, curated knowledge, formal math inside chat, from the team behind Wolfram Alpha and Mathematica. Consensus retrieves papers from an indexed academic corpus, generates a synthesis, and then shows you a Consensus Meter, which is a little readout of how much the retrieved papers agree with each other.



Corn

That meter is the interesting design choice. It's not just giving you a paper list, it's giving you a sense of the state of the disagreement.



Herman

Because the disagreement is the finding, half the time. But here's the thing, and this is what I want to get to. Neither of those two is a drug discovery tool. Wolfram is general computation, Consensus is literature search. The capability-expanding work in drug discovery is almost entirely absent from the plugin directory.



Corn

So where is it?



Herman

In the arXiv agent literature. Which is a completely different ecosystem, with different incentives, and it hasn't migrated into the directory yet. LIDDIA, revised in October, generates molecules meeting key pharmaceutical criteria on over seventy percent of thirty clinically relevant targets. It identified a novel candidate on a receptor called AR, NR3C4.



Corn

Seventy percent on thirty targets. That's not a demo number.



Herman

It's a real number, and so is MADDIE, which is a multi-agent engine that was tested across fifty protein targets. DrugPilot reports task completion of ninety-eight percent on simple tasks, ninety-three and a half on multi-tool, and sixty-four on multi-turn scenarios.



Corn

That last number is the honest one.



Herman

That last number is the honest one, and I appreciate that they published it. Everything degrades when you chain it. There's a survey from October of last year, forty-five pages, that claims to be the first comprehensive work presenting real-world implementations and quantifiable impacts of agentic systems in operational drug discovery. It reports workflows that once took months now taking hours.



Corn

Now taking hours, with what in the loop?



Herman

That's the question the survey doesn't fully answer, and we'll come back to it. But the mechanism part is real. AlphaFold 3, released back in May of twenty twenty-four by DeepMind and Isomorphic Labs, predicts joint structures of proteins, nucleic acids, small molecules, ions, modified residues. That's the foundation everybody builds on.



Corn

And the building on top is where it gets interesting.



Herman

It is. There's Bayesian refinement of AlphaFold models for in-silico drug design, published in July. There's DrugCLIP, which does genome-wide virtual screening using contrastive learning on AlphaFold-predicted structures.



Corn

So AlphaFold gave you the structure, and now the field is trying to turn structures into decisions.



Herman

That's the whole second act. And it's contested. There's a preprint from July that evaluates AlphaFold 3 across diverse application areas and documents specificity gaps and training-data dependence. The framing is basically, how well do the new capabilities actually translate into application tasks.



Corn

And the answer isn't a clean yes.



Herman

The answer is, sometimes, and less often than the reviews suggest. That's the pattern I keep seeing. The capability is real, the performance on an actual task is variable.



Corn

So here's what I'm hearing. The scientific ecosystem is real and it's growing. But it's split in two. A thin visible layer of general-purpose tools in the directory, Wolfram and Consensus, and then a much deeper layer of specialised agent research happening in papers that haven't become plugins at all.



Herman

That's the split, and I'd go further. The plugin library is not yet the map of scientific AI. It's the map of commercial AI with a few scientific postcards pinned to the corner.



Corn

A postcard is cheaper than a lab.



Herman

Right, and that's why the counts look healthy while the thing Daniel actually cares about is still mostly in research code. Which takes us to the impact question, because that's where the evidence cuts both ways.



Corn

Which impact question?



Herman

Whether any of this is measurably reducing research time, or producing discoveries, or making impossible experiments possible. And the honest answer is that the closest measurable evidence isn't plugin evidence at all. It's model-level and agent-benchmark-level.



Corn

Give me the strongest version of the positive case.



Herman

LEADS. A medical literature foundation model, published in Nature Communications. Trained on six hundred and thirty-three thousand samples drawn from twenty-one thousand systematic reviews.



Corn

That's a lot of reviews.



Herman

It's the whole visible literature of evidence synthesis. And then they did a user study with sixteen clinicians and researchers from fourteen institutions, which is small but it's real humans doing real work. Experts using LEADS hit zero point eight one recall in study selection, versus zero point seven eight working alone. Twenty point eight percent time savings. In data extraction they went from zero point eight zero to zero point eight five accuracy, with twenty-six point nine percent time savings.



Corn

So roughly a quarter of the grunt work, gone.



Herman

Roughly a quarter of the grunt work, gone, with humans still in the loop doing the judging. That's the shape of the win.



Corn

Now the countervailing case. Because I know there is one.



Herman

There is one, and it's ugly. A systematic review in Research Synthesis Methods looked at GenAI for evidence synthesis and found it missed sixty-eight to ninety-six percent of studies during searching. Median, ninety-one percent.



Corn

It missed nine out of ten studies.



Herman

Nine out of ten. And the conclusion is blunt. The current evidence does not support GenAI use in evidence synthesis without human involvement or oversight. Then a separate review on data extraction found accuracy ranging from forty-seven percent to ninety-nine point nine, with numerical data much worse, forty-seven to eighty-eight percent, and categorical data much better, seventy-four to ninety-six percent.



Corn

So it can read the paper but not the table.



Herman

It can read the paper and it will confidently misread the table. Numbers are where it falls apart, and numbers are the entire point in half of these workflows.



Corn

And that gets us back to the distinction Daniel drew. Integrations that make existing software easier to use versus integrations that expand what you can do. Wolfram in chat is the first kind.



Herman

Wolfram is a real accessibility win. Natural language into symbolic computation means a biologist who never learned Mathematica syntax can get a symbolic answer. That's useful and it is not a new capability. The math was always there.



Corn

And LIDDIA, MADDIE, those are the second kind. They generate molecules that didn't exist.



Herman

They propose candidates, which is a capability, not an interface. But they're not in the plugin directory. They're in research code and papers, and the gap between a paper and a tool a lab can rely on is enormous.



Corn

What's in that gap?



Herman

Three things, if you believe the deployment paper from March. It cites over ten thousand active MCP servers and ninety-seven million monthly SDK downloads as of early this year, which is enormous. And then it names three missing production primitives. Identity propagation. Adaptive tool budgeting. Structured error semantics.



Corn

Say what those mean in practice.



Herman

Identity propagation means when the agent calls a tool on your behalf, the tool knows it's you, with your permissions. Not the agent, not a generic service account, you. Today that's mush. Adaptive tool budgeting means the agent knows how much of each tool it can afford to call before it stops being worth it. Structured error semantics means when a tool fails, it fails in a way the model can actually recover from, instead of returning a sentence that says something went wrong.



Corn

So the demo works and the production thing doesn't, because the demo never had to handle a failure at three in the morning.



Herman

And there's a security tax on top of it that nobody wants to talk about. A security analysis from January found MCP's architecture amplifies attack success rates by twenty-three to forty-one percent compared to equivalent non-MCP integrations.



Corn

Compared to the same integration without MCP.



Herman

Same functionality, different plumbing, more successful attacks. And their proposed extension, MCPSec, cuts attack success from fifty-two point eight percent down to twelve point four.



Corn

So the fix works and it isn't standard yet.



Herman

The fix works in the paper. Whether it's in the thing you installed is a different question. And this matters more for scientific tools, because a literature search that gets poisoned is annoying. An instrument controller or a clinical dataset that gets poisoned is a different category of problem.



Corn

Which is interesting, because Daniel framed the whole thing as an access layer. General-purpose assistants becoming the front door to sophisticated infrastructure.



Herman

And that's happening. It's just that the front door currently opens onto a lobby with a lot of commercial SaaS in it and a small research annex off to the side. The access layer is real. The sophistication behind it is thinner than the directory's size suggests.



Corn

So augmentation works and autonomy doesn't. That's the one-line version.



Herman

With an asterisk, which is that we don't have plugin-level evidence for either. LEADS is a model. LIDDIA is an agent benchmark. Nobody has published the study that says a researcher using Consensus inside ChatGPT saved eleven hours this month. The evidence is one layer down.



Corn

Does that bother you?



Herman

It bothers me professionally. The mechanism is the same, so I'd expect the direction to hold. But I've been burned before by expecting a result to survive one layer of indirection, and I'll say so rather than pretend the study exists.



Corn

And the failure modes here are weirdly specific.



Hilbert

The word you keep using is wrong.



Herman

Which word?



Hilbert

Plugin. You keep saying plugin like it's a thing. My niece works in a lab, and she'd say you're describing a door.



Corn

Go on.



Hilbert

She rang me two weeks ago. Her lab tried one of those AI research assistants for a literature review. It cited a paper. Title, authors, journal, DOI. The DOI resolved to nothing. The paper does not exist. It was never published. It was never written.



Herman

That's a hallucinated citation, which we've all seen.



Hilbert

It's worse than a hallucinated citation. The title was specific. It was on topic, it had the right words in it, and the postdoc spent a full day trying to find it through interlibrary loan before he gave up.



Corn

A full day on a paper that doesn't exist.



Hilbert

Two days across the lab, if you count the second person who looked. And that's the part you can't put in a benchmark. It isn't that the tool lied. It's that the lie was good enough to be chased.



Herman

That is the weak spot I'd worry about more than the obvious one. An obviously wrong answer costs you nothing. A plausible wrong answer costs you a day.



Hilbert

My niece called me specifically to ask if I'd heard of the journal.



Corn

And had you?



Hilbert

I told her I'd look into it. I was in publishing for a while, so I know how to check these things.



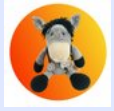
Herman

Were you in publishing?



Hilbert

Circulation. I was a circulation consultant for a regional magazine. It folded. Not because of the circulation. But I know what a real journal looks like. I have opinions about impact factors.



Herman

Of course you do.



Hilbert

She checks every citation by hand now. All of them. She says it's slower. I told her that's probably fine.



Corn

Is it fine?



Hilbert

It's what she's doing, and nothing has gotten past her since.



Corn

The tools are getting better at sounding right.



Herman

That's the thing. Sounding right is exactly the property that makes the failure expensive, because a researcher's whole workflow is built on trusting the citation.



Corn

And the fix is a human checking each one by hand. Which is the least futuristic possible solution to a problem created by the most futuristic possible tool.



Hilbert

I don't think there's anything wrong with that. You check the thing you're going to build on. That's the whole job.



Herman

Nobody puts that on the box though. The box says synthesise the literature. It doesn't say synthesise the literature and then verify every reference manually before you trust a sentence of it.



Corn

Which is the honest picture. Augmentation, not autonomy, and the human in the loop is doing the least glamorous part.



Herman

That's what Daniel asked directly, whether there are measurable reductions in research time or previously impractical experiments becoming feasible. The answer is that the reductions are real when a human is judging and fictional when nobody is.



Corn

Let's do the correction, because I think there's one belief that needs killing.



Herman

There is, and it's the one most people hold without noticing they hold it. The belief is that the size of the plugin library measures scientific capability. Five thousand seven hundred and twenty-three plugins, two hundred and thirty-eight thousand tools, so science must be deep in there somewhere.



Corn

It doesn't. The leaders on discoverability are Ashby, Supabase, Miro, Shopify, Stripe. Wolfram and Consensus aren't up there. The directory's surface tells you about commercial integration. The scientific depth is somewhere else, mostly in research code.



Herman

The second belief underneath it, that these tools are either working or not working, is also wrong. The evidence splits cleanly. With human oversight you get measurable time savings, a fifth to a quarter of the work in some tasks. Without it you get a median ninety-one percent miss rate. It's not one verdict. It's two, depending on whether a person is in the loop.



Corn

The honest place to land is that the access layer is real, the science is real, and the gap between demo and tool is still the story. Which raises the open question, do the scientific tools migrate into the plugin layer, or do they stay in research code and specialised platforms?



Herman

It depends on whether identity propagation, adaptive tool budgeting and structured error semantics get solved, and whether the security tax gets paid down. Those three primitives are what separate a demo from something a lab can build on. Until they exist, the serious tools stay where the serious engineers are, which is not the plugin directory.



Corn

The shape of the transition might be the finding. The researchers getting value are the ones staying in the loop. Maybe that isn't a limitation waiting to be overcome. Maybe that's what this actually looks like for a while.



Herman

A tool that needs a human checking its citations is still a tool that saves you twenty-five percent of your week. That's not nothing. It's just not the story being sold.



Corn

Thanks to Hilbert Flumingtop, who produces this show and checks the things he's going to build on.



Herman

If you want more of this, try episode fourteen, AGI's Crossroads; episode twelve, The AI Breakthrough; and episode ten, How ASR Went From Frustration To ... Whisper Magic. This has been My Weird Prompts.



Corn

If you want to send us a prompt of your own, the line is send us your own prompt on Telegram at t dot me slash MWP listener bot.



Herman

We'll be back soon.



Corn

See you tomorrow.

SOURCES

What the research for this episode read before the script was written. Primary sources first.

1. **OpenAI, 2025-10-06 (updated 2025-11-13) [primary]**
https://openai.com/index/introducing-apps-in-chatgpt/

2. **ChatGPT Learn, current [primary]**
https://learn.chatgpt.com/docs/plugins

3. **OpenAI Developers, current [primary]**
https://developers.openai.com/apps-sdk/quickstart

4. **OpenAI Developers, current [primary]**
https://developers.openai.com/apps-sdk/mcp-apps-in-chatgpt

5. **Agent Discoverability MCP Directory, 2026-10-14**
https://agentdiscoverability.com/track

6. **MCP App Store, updated 2026-09-19**
https://mcpapp.net/chatgpt-apps

7. **community list, last updated 2026-07-02**
https://github.com/rdmgator12/awesome-chatgpt-apps

8. **community directory, current**
https://chatgptplugins.directory/

9. **LIDDIA, revised 2025-10-13**
https://arxiv.org/abs/2502.13959v3

10. **MADDIE, 2026-10-12**
https://arxiv.org/abs/2610.16975v1

11. **DrugPilot, revised 2025-07-28**
https://arxiv.org/abs/2505.13940v2

12. **AI Agents in Drug Discovery survey, 2025-10-31**
https://arxiv.org/abs/2510.27130v1

13. **Bridging Protocol and Production, 2026-03-12**
https://arxiv.org/abs/2603.13417v1

14. **MCP security analysis, 2026-01-24**
https://arxiv.org/abs/2601.17549v1

15. **AlphaFold 3, Nature, 2024**
https://www.nature.com/articles/s41586-024-07487-w

16. **AF3 review, 2026**
https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2026.1739303/full

17. **AF3 capability gaps, 2026-07-13**
https://www.biorxiv.org/content/10.64898/2026.07.13.738147v1.full.pdf

18. **Bayesian refinement of AlphaFold models, 2026-07-13**
https://pubs.acs.org/doi/10.1021/acs.jctc.6c00868

19. **LEADS foundation model, Nature Communications, 2025**
https://pubmed.ncbi.nlm.nih.gov/40993125/

20. **GenAI evidence synthesis review, Research Synthesis Methods, 2025**
https://pubmed.ncbi.nlm.nih.gov/41626912/

21. **LLM data extraction review, J. Biomedical Informatics, 2026**
https://pubmed.ncbi.nlm.nih.gov/42501879/

22. **Consensus App review, 2025**
https://pmc.ncbi.nlm.nih.gov/articles/PMC12318603/

23. **CASRAI Consensus guide, current**
https://casrai.org/guides/consensus-ai-search-engine-guide