Teaching the Rules to Learn:

How ARTUR Finds What Rules Missed — and Why That Changes the Economics of Data Curation

Chee Ping Ng, Founder
AvatarsBio, The Netherlands

SUMMARY

Every scientific dataset has a quality problem. Not a dramatic one — usually it is quieter than that. A measurement recorded in the wrong units. A label that means one thing in one lab and something different in another. An assay description written in plain language that no computer can parse without help.

This paper describes a practical experiment: we took a dataset of nearly 3,000 records from a major public chemistry database and asked two systems to clean it — one using carefully written computer rules, the other using a general-purpose AI. The question was not which one was better. The question was: what can they learn from each other?

The result is ARTUR — a hybrid workflow for scientific data curation, named in the tradition of Per Artursson, whose foundational work established the cell-based permeability assay at the heart of this dataset. Rules perform the scalable and auditable work, AI acts as a diagnostic layer, and human review focuses only on the records that genuinely require attention. In this study, the workflow reduced the number of records requiring focused human review by approximately 95%, while costing less than two US dollars to run.

2,747	95%	137	$1.77
records processed	agreement between methods	meaningful disagreements	total AI API cost

95%	~90%	↑	16
reduction in human review records	projected API saving vs full LLM†	review quality — less fatigue	reference compounds

† Projection based on running LLM on flagged subset only.

THE PROBLEM: CLEANING SCIENTIFIC DATA IS HARDER THAN IT LOOKS

When scientists share data publicly — through databases like ChEMBL or PubChem — they describe their experiments in ordinary language. One researcher might write that a molecule was tested for its ability to cross a cell membrane 'in the apical-to-basolateral direction.' Another might simply note the result with no directional information at all. Both entries end up in the same database, side by side.

Before that data can be used to train a predictive model, compare results across labs, or support a drug development decision, it needs to be standardised. That standardisation — deciding what each record means, converting units, recognising what type of experiment was run — is called curation. And it is slow, expert-intensive work.

The standard approach is to write rules: if the description contains this word, classify it this way. If the unit looks like this, convert it using that factor. Rules are fast and consistent, but they are brittle. They break the moment someone phrases something in a way the rule-writer did not anticipate.

THE EXPERIMENT

We built a curation pipeline focused on a specific type of measurement: how well a drug compound passes through a layer of intestinal cells — a test that predicts whether a drug taken by mouth will actually reach the bloodstream. This is one of the most commonly run experiments in early drug development, and public databases contain thousands of records for it.

Our pipeline — ARTUR — tackled three specific cleaning tasks:

- Is this record actually from the right type of experiment?
- What units was the measurement recorded in, and can we convert it to a standard?
- Was the experiment run in one direction through the cell layer, the other direction, or is it measuring a ratio of both?

We ran both a rule-based system and an AI (a large language model, given a carefully written scientific prompt) through all 2,747 records and compared what each one decided.

WHAT THE DATA LOOKS LIKE AFTER CURATION

The figure below shows the curated results — the scientific output of the pipeline. Reference compounds land where biology predicts they should: digoxin and atenolol flag as likely candidates for active cellular transport, verapamil registers as highly permeable, and talinolol — which has the most records in the dataset — shows a directional transport pattern consistent with published literature.

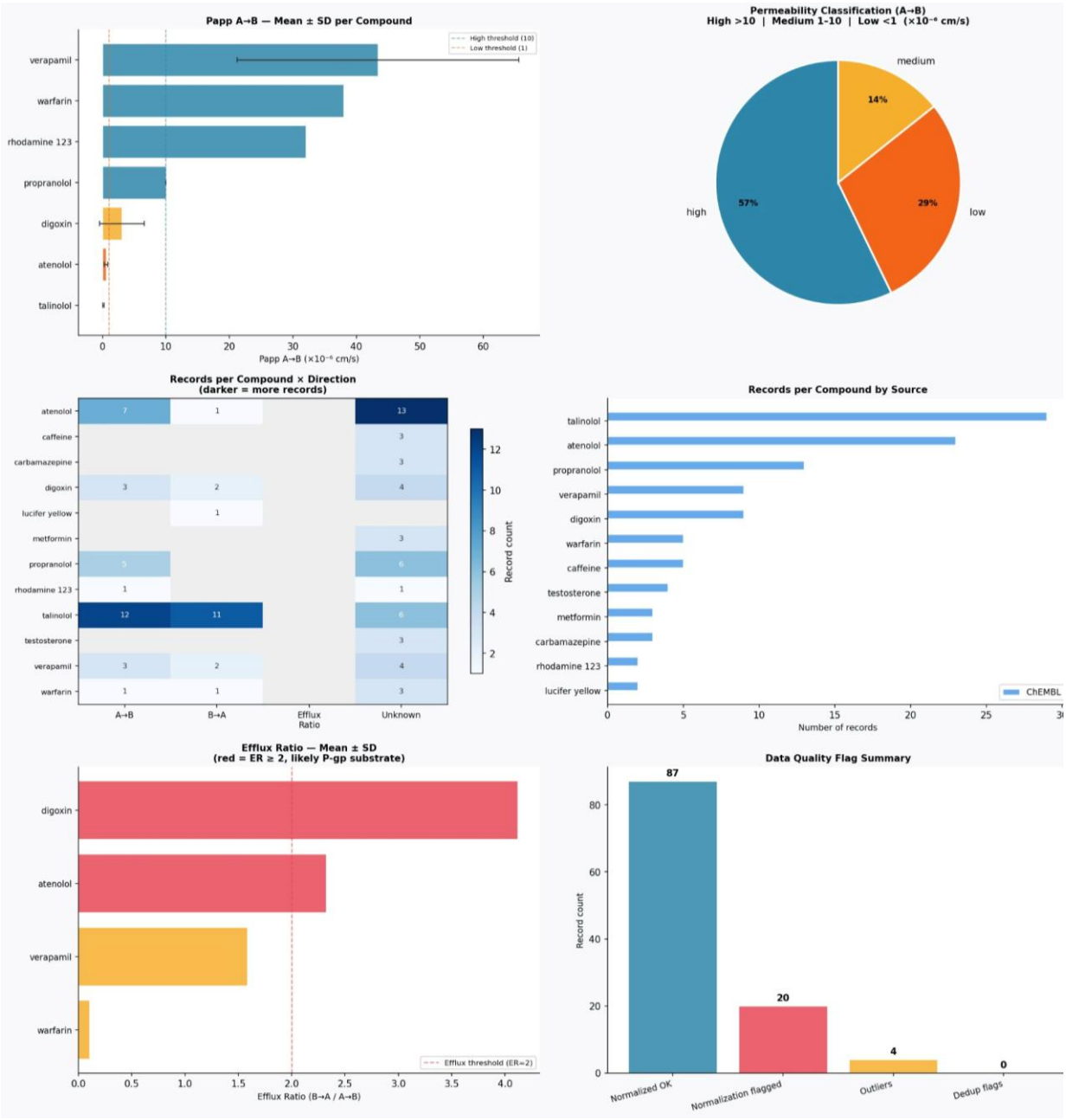

Figure 1. Curated Caco-2 permeability results across 16 reference compounds. The pipeline produces biologically coherent outputs — a prerequisite for any downstream modelling use.

WHAT ARTUR NOTICED THAT THE RULES MISSED

The two systems agreed on 95% of records. But the 5% where they disagreed turned out to be the most informative part of the experiment. In each of the three curation steps, the disagreements were not random — they clustered in ways that pointed to specific gaps in the rule-based system.

Curation Step	What ARTUR noticed	Why it matters
Experiment type	Found 13 additional valid records the rules missed — their descriptions used phrasing the keyword list did not anticipate	If those records are genuine, the rules are leaving data on the table

Units	Was less confident than the rules on unusual formatting — typographic variants the rules handle precisely	Rules outperform here; AI is useful only as a fallback for formats not yet in the lookup table
Experiment direction	Identified 26 additional records as measuring a transport ratio — the rules only found 4	Most scientifically important gap: those 26 records may be incorrectly labelled in the rule-based output

> ***"The disagreements are not errors — they are a map of where the rules need to improve."***

A DIFFERENT WAY TO THINK ABOUT AI IN DATA CURATION

Most conversations about AI in science ask the same question: can AI replace the expert? That is usually the wrong question. In the context of data curation, a more productive question is: can AI tell us where our existing process is failing?

That is exactly what ARTUR demonstrated. The AI did not outperform the rules overall. In one step — unit interpretation — the rules were clearly more reliable. But the AI read scientific language more flexibly, and where it diverged from the rules, it pointed to real gaps: descriptions phrased in ways the rules had not anticipated, and experimental types the rules were systematically miscategorising.

The result is a hybrid approach where rules do the heavy lifting at scale — fast, consistent, auditable — and AI acts as a diagnostic layer that flags where those rules should be updated. A scientist then reviews only the flagged disagreements, rather than manually checking thousands of records.

THE SWEET SPOT: WHERE COST, COVERAGE, AND QUALITY INTERSECT

Nvidia's VP of applied deep learning made headlines recently stating that for his team, AI compute now costs more than human labour. Uber's CTO has reportedly exhausted the company's entire 2026 AI budget by April. These are symptoms of the same underlying problem: AI deployed without a design principle for efficiency.

ARTUR offers a different model. The efficiency is not accidental — it is the result of a deliberate design decision: rules first, AI only where rules are weakest, human review only on flagged disagreements. That architecture produces a compounding three-layer saving:

Saving	Estimate	Note
Human review time	~95% reduction	137 flagged records reviewed instead of 2,747. At ~2.5 minutes per record, that is approximately 108 hours saved.
Review quality	Higher	Reviewers stay in the high-attention zone throughout. Manual review of 2,747 records degrades after 20–50 records as fatigue sets in — errors in record 2,500 are far more likely than in record 50.
API cost — diagnostic use	$1.77 total	Full LLM pass across all 2,747 records. Less than two dollars to identify gaps that would otherwise take days.

API cost — optimised†	~90% lower	Running LLM only on records the rules flag as uncertain reduces API cost further. Projection based on ~10–20% subset processing.

† Projected saving — not yet empirically validated. Review time estimate assumes ~2.5 min per record under attentive conditions.

Full manual review is thorough but slow and quality-limited by fatigue. Full LLM is faster but introduces systematic errors on formatting tasks that compound at scale. ARTUR hits the sweet spot — concentrating both AI and human effort only where they add the most value.

BROADER RELEVANCE

This approach is not specific to drug permeability data. The same pattern applies wherever scientists are trying to standardise free-text experimental records at scale: toxicology databases, clinical trial registries, genomics repositories, materials science datasets. The underlying challenge is always the same — human language is richer than any keyword list.

Importantly, ARTUR required no specialised training. It was given a clear scientific prompt written by a domain expert — no labelled datasets, no machine learning pipeline, no dedicated infrastructure. This makes the approach accessible to individual researchers and small teams who need practical curation tools without institutional resources.

Recent work from researchers at EMBL-EBI, published in the Journal of Cheminformatics in early 2026, reached a similar conclusion using a purpose-built trained model: combining expert knowledge with automated methods outperforms either alone. ARTUR suggests the same result is achievable with a much lighter toolset — and a design principle that keeps costs firmly under control.

KEY TAKEAWAYS

- ARTUR — a hybrid rule-based and AI workflow — achieved 95% agreement across 2,747 scientific records.
- The 5% disagreements were scientifically meaningful: a map of where the rules needed refinement, not random errors.
- AI was most valuable in interpreting ambiguous scientific language, while rules remained stronger for precise formatting tasks.
- Reviewing only 137 flagged disagreements reduced focused human review by ~95%, concentrating attention where scientific judgment mattered most.
- Total AI API cost for 2,747 records was USD 1.77. Running the LLM only on flagged subsets could reduce this further (~90% projected).
- The value of AI in this workflow is not that it is cheap enough to replace the rules — it is that it is cheap enough to audit them.
- No training data or ML infrastructure was required — only domain expertise translated into a well-written prompt. ARTUR — a hybrid rule-based and AI curation workflow — achieved 95% agreement across 2,747 scientific records.