Curating Huntington's Disease Perturbation Data with Claude Code as an Agentic Extraction Workflow

Viktoria Andreeva (1), Irina Murasheva (1), Jeff Aaronson (2), Rick Blevins (3), Oleg Stroganov (1)
(1) Rancho BioSciences LLC, San Diego, CA, USA; (2) CHDI Management/CHDI Foundation, Princeton, NJ, USA; (3) Bridlewood Consulting LLC, Flower Mound, TX, USA

Background

HDinHD (Huntington's Disease in High Definition; HDinHD.org) is an open, online data portal designed to support and accelerate HD research. It provides synthesized access to diverse HD experimental datasets, community-developed analysis tools, and curated, standardized data processed through reproducible pipelines.

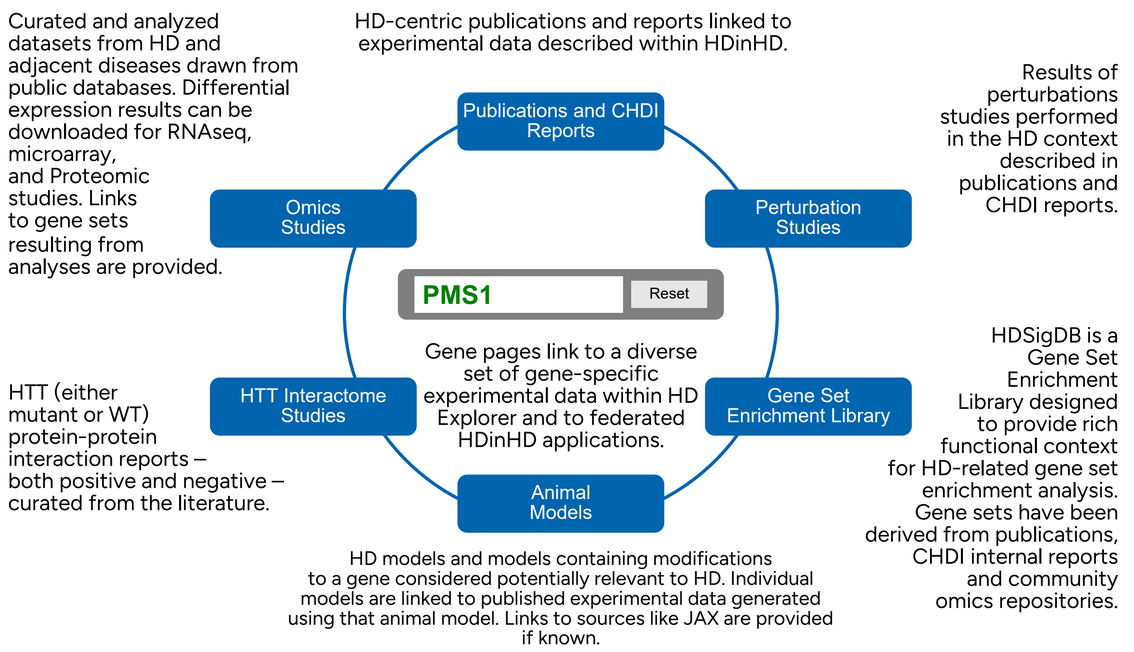
HD Explorer, a core component of HDinHD, integrates HD experimental data curated from the literature, internal CHDI reports and community omics repositories. Researchers can browse and search across omics studies, animal models, gene sets, publications and CHDI reports, the Huntingtin Interactome (HINT), and **Perturbation Studies (PerturbDB)** [1].

PerturbDB holds over **14,000 therapeutic intervention reports** of genetic and pharmacological perturbation experiments from over **1,600 publications and reports**. The HDinHD team actively monitors the literature to keep the portal current, but manual curation of PerturbDB is labor-intensive given the inherently complex, multidimensional nature of perturbation experiments and their results.

Prior work, and the question here

In prior work we validated a Claude Code skill-based agentic pipeline for curating protein–protein interactions (PPI), achieving high union recall (**93.7%**) and cost-effective throughput (**$5–8 per article**) versus manual curation. Here we extend this approach to PerturbDB's more complex schema to test whether the strategy generalizes beyond PPI curation.

Figure 1. HD Explorer Portal



Methods

A curated PerturbDB record is a multi-level data object, not a single row. One perturbation **Experiment** contains one or more treatment and control **Arms**, and each arm carries several measured **Outcomes**. Capturing one publication therefore means building a set of interdependent records that must stay mutually consistent — substantially more complex than the single-row PPI records of our earlier work.

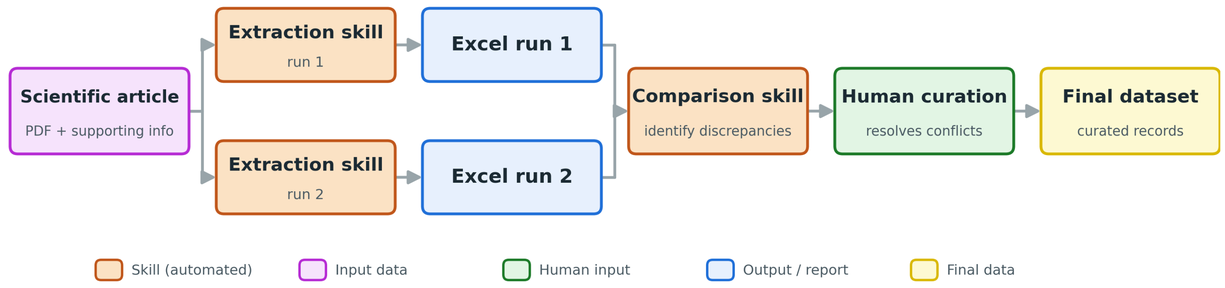
We developed a Claude Code skill-driven pipeline, adapted from our PPI workflow, with an **Extraction skill** that processes PDFs and populates all three levels in a phased manner, and a **Comparison skill** that evaluates dual runs and classifies discrepancies as major or minor. Skills are reusable, human-readable prompt instructions that Claude Code follows during curation: refined iteratively from subject matter expert (SME) feedback (Figure 2), then applied in production as two independent runs with automated comparison (Figure 3).

Validation follows a three-phase design. In Phase 1 (Reproducibility Training), 20 previously curated articles are processed in batches of five with dual runs per batch, and the skills refined until dual-run agreement reaches acceptable thresholds. In Phase 2 (External Alignment), a distinct set of 20 articles is processed with dual runs and compared against existing curated PerturbDB data, generating agreement and error-rate metrics by level. In Phase 3 (Pilot Production), five new, not-yet-curated publications are processed with dual-run extraction and SME review, capturing time, cost, and error rates.

Figure 2. Skill development workflow
Iterative skill refinement with SME feedback



Figure 3. Production curation workflow
Dual-run extraction with automated comparison



Extraction skill architecture

A phased pipeline over each PDF: survey of all reported experiments → commitment to extraction targets before extraction begins → record-level extraction at all three levels → programmatic validation. Locking targets up front prevents post-hoc exclusion bias; validation catches broken links between levels before review.

Results

Twenty previously curated HD articles were processed with the production workflow, each extracted twice by independent runs. The two runs were compared with one another and against the existing curated PerturbDB records, and every resulting difference was adjudicated against the article PDF to establish which side was correct.

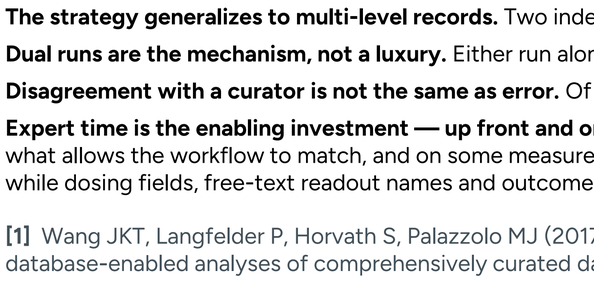


Figure 4. Run-to-run reproducibility
97.5% of 3,170 compared fields agree, and performance is consistent across all levels of the data model.

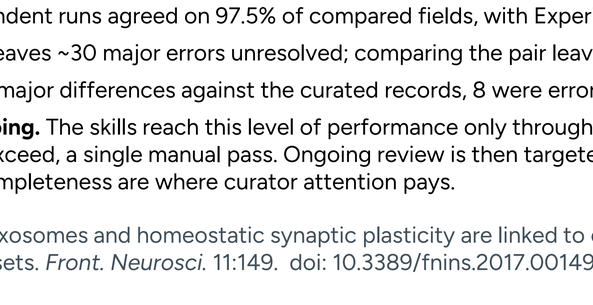


Figure 5. Errors by strategy
Of the same 58 major issues, one run alone leaves ~30 unresolved; two runs leave 5.

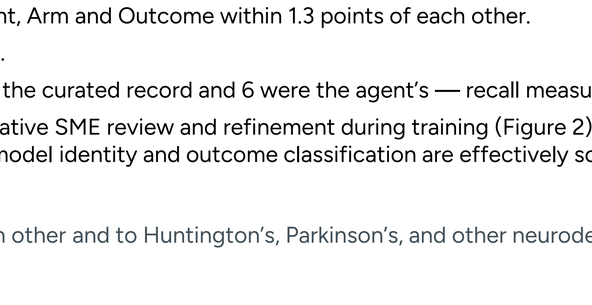


Figure 6. Recall vs. curated records (run 1)
Experiments and Arms are found almost completely; Outcomes are the gap.

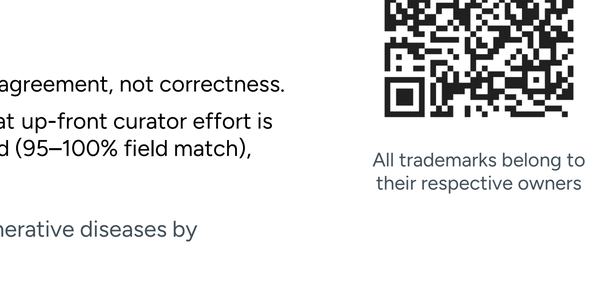


Figure 7. Source of major differences
Where the records differ, the curated record was wrong more often than the agent.

Consistent with itself

97.5% field agreement over 3,170 compared fields; 7 of 20 articles had zero real conflicts. Half the genuine conflicts sit in phenotype direction, outcome value and contrast type — 1.2% of all cells.

Two reads beat one

Run 1 was more accurate on 9 articles, run 2 on 6, 3 tied — **neither run is the better one**. 53 of 58 adjudicated major issues were recoverable because one of the two runs held the correct value.

Most gaps are ours to close

Of 65 major issues, **49** were a guideline already in place not being followed (25) or a result plainly in the paper being missed (24) — both fixable in-house. Only 9 needed an SME decision; 7 could not be classified from the available notes.

Conclusions

Extending this agentic, skill-based workflow from single-row PPI records to PerturbDB's multi-level record structure tests whether the strategy generalizes to more complex biomedical data structures. As with PPI, LLM extraction serves as a cost-effective first-pass tool that shifts curator effort toward targeted verification, as expert review remains essential for interpretive judgment. Results across all three phases will support a hybrid, scalable strategy for maintaining and expanding resources like PerturbDB within the HDinHD portal.

The strategy generalizes to multi-level records. Two independent runs agreed on 97.5% of compared fields, with Experiment, Arm and Outcome within 1.3 points of each other.

Dual runs are the mechanism, not a luxury. Either run alone leaves ~30 major errors unresolved; comparing the pair leaves 5.

Disagreement with a curator is not the same as error. Of 18 major differences against the curated records, 8 were errors in the curated record and 6 were the agent's — recall measures agreement, not correctness.

Expert time is the enabling investment — up front and ongoing. The skills reach this level of performance only through iterative SME review and refinement during training (Figure 2); that up-front curator effort is what allows the workflow to match, and on some measures exceed, a single manual pass. Ongoing review is then targeted: model identity and outcome classification are effectively solved (95–100% field match), while dosing fields, free-text readout names and outcome completeness are where curator attention pays.

All trademarks belong to their respective owners

[1] Wang JKT, Langfelder P, Horvath S, Palazzolo MJ (2017) Exosomes and homeostatic synaptic plasticity are linked to each other and to Huntington's, Parkinson's, and other neurodegenerative diseases by database-enabled analyses of comprehensively curated datasets. *Front. Neurosci.* 11:149. doi: 10.3389/fnins.2017.00149