PERELMAN

Automated Quantitative Meta-Analysis of Scientific Literature
via Vision-Language Models

Daniil Merkulov[1,2] Danil Sherki[2] [co-authors TBD]

[1]Skoltech [2]AI for Science Lab, Sber d.merkulov@skoltech.ru

The Data Extraction Bottleneck

Scientific literature grows at >2 M papers/year. For materials science, drug discovery, and physics, critical quantitative results are **not in tables or abstracts — they live in figures**.

Motivating Example

A systematic review of **NMC811 cathode materials** (Savina & Abakumov) required processing **548 papers** with **950+ experiments** — months of manual work for a PhD student.

Key insight: ∼40–60% of quantitative results appear *only in figures*, not in text or tables. Text-based retrieval methods (Elicit, SciSpace) miss this data entirely.

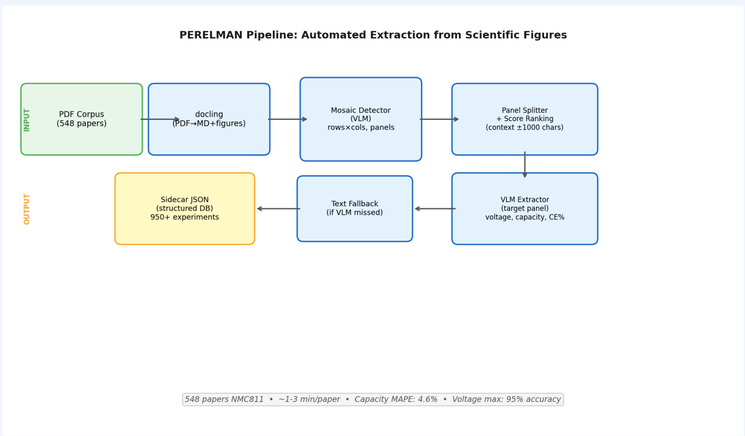


Fig. 1: PERELMAN pipeline — from PDF corpus to structured database

Competitive Landscape

System	Task	Expert?	Output
PERELMAN	**Meta-analysis corpus**	**No**	**Struct. DB**
AutoResearchClaw	Write a paper	Topic only	Draft paper
Elicit/SciSpace	Q&A over abstracts	No	Text answers
Manual review	Meta-analysis	Full process	Tables

Unique niche: PERELMAN extracts *numbers from graphs* — a capability that no text-based RAG can replicate.

Applications

- **Materials Science:** capacity/voltage DB from battery papers → regression models linking synthesis params to performance
- **Drug Discovery:** IC50/EC50 from pharmacology graphs
- **Physics:** critical temperatures, band gaps, phase diagrams
- **General pattern:** any domain with characteristic graph types → one domain-specific prompt → process entire corpus

Cost estimate

$0.01–0.05 per paper (VLM calls via OpenRouter).
548 papers ≈ $5–25 total — negligible vs. months of manual annotation.

Pipeline Overview

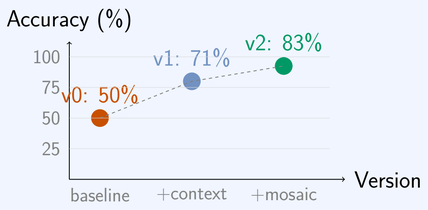
PERELMAN is a multi-stage VLM pipeline for automated figure-based data extraction:

PDF Corpus
(548 papers)
↓
docling — PDF →
↓
Mosaic Detector (VLM)
↓
Panel Splitter +
↓
VLM Extractor
(target panel only)
voltage, range
↓
Text Fallback
(LLM on paper text)
↓
Sidecar JSON —
Structured Database

Key innovations:

- **docling:** OCR + structure preservation (better than PyMuPDF)
- **Mosaic Detector:** custom VLM prompt for multi-panel figures (a/b/c/d)
- **Context-aware prompting:** ±1000 char window around figure references (caption + in-text mentions) — improves VLM extraction precision
- **Zero-shot domain adaptation:** NMC811 synonyms + baseline-series logic

Accuracy Progression



Each ablation stage adds a key module: v0 = direct VLM on figure; v1 = + context-aware prompting (+21pp); v2 = + mosaic detection and baseline-series logic (+12pp).

Results on NMC811 Corpus

Evaluation: 24 papers, ground truth from Savina & Abakumov.

4.6%
Capacity MAPE
(n=20 matched pairs)

83%
Coverage
(20/24 papers)

95%
Voltage max
accuracy (±0.1V, n=21)

86%
Voltage min
accuracy (±0.1V, n=21)

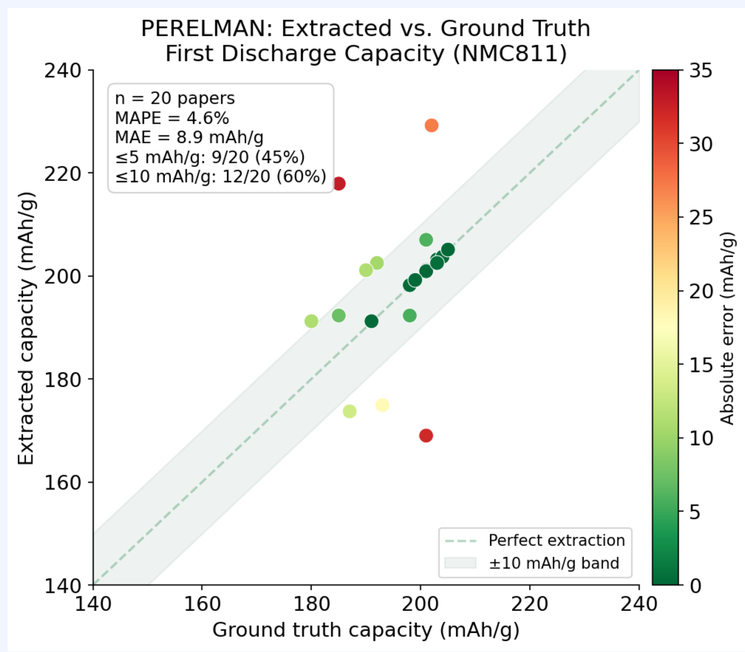


Fig. 2: Extracted vs. ground-truth capacity (n=20, MAPE=4.6%)

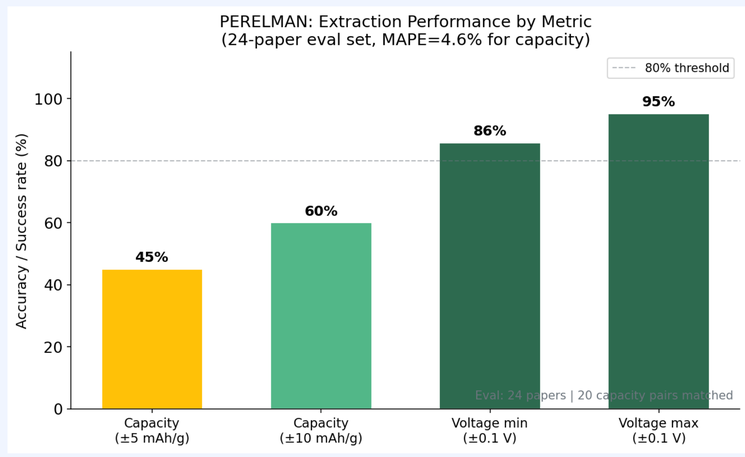


Fig. 3: Accuracy by metric — voltage extraction is highly precise

Error Analysis

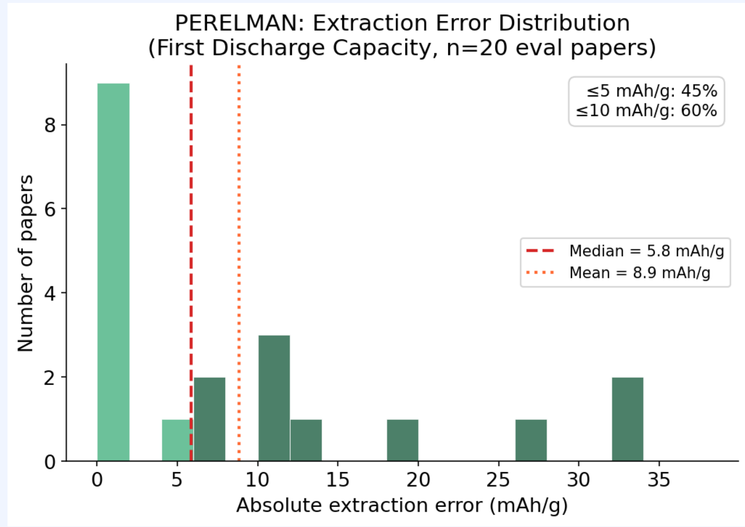


Fig. 4: Error distribution — median 5.8 mAh/g, 9/20 papers <2 mAh/g error

Failure modes (17% miss):

- Non-standard panel layouts (3-column, overlapping)
- Extremely low-resolution figures
- Papers reporting only normalized capacity (C/C_0)

Conclusion

PERELMAN reduces quantitative literature curation from **months to hours**, enabling data-driven modeling at the scale of entire research domains.

- **4.6% MAPE** — publication-quality precision
- **83% coverage** — high throughput on real corpus
- **Domain-agnostic architecture** — adapt with one prompt
- **Open pipeline** — docling + OpenRouter VLMs

Future work: generalization to drug discovery / physics corpora; active learning for prompt refinement; integration with downstream regression models.

Demo dataset: NMC811 database (548 papers, 950+ experiments) available on request.

Key References

- Savina & Abakumov (2024). NMC811 systematic review. *[corpus source]*
- Auer et al. (2024). docling — Document Understanding & Conversion. arXiv:2408.09869
- Submission #316, AI4X-AC 2026, Singapore