

Using Oghma to construct ML-ready databases for the pharma industry

Problem statement

Accurate predictive models of physical and chemical properties play an increasingly important role across the R&D of many technology sectors. Models enable the rapid screening of potential experimental strategies by limiting the range of possible conditions to explore. The result is a drastic saving of the experimentation time and costs, and a significant increase in the discovery throughput. In particular, solubility predictions are critical for the chemical and pharma industry, since the understanding of the solvation behaviour of a novel active pharmaceutical ingredient informs the entire development process, from drug discovery to manufacturing.

The construction of machine-learning (ML) solubility models depends tightly on the availability of highly curated experimental data. These are scattered across the published literature in an unstructured format (text, pictures, tables, etc.). Although manual data extraction yields accurate datasets, it is manually intensive and not scalable to the volume of data produced today.

The challenge

APC has been working on developing ML solubility models to drive their R&D. The models are based on a manually collected dataset from APC, which includes approximately 12,000 entries extracted from 162 data-dense publicly available documents, also manually resourced. The estimated time for the data extraction alone is about 370 hours. The grand challenge was *to produce a significantly expanded database mined from a large collection of documents*. This poses three main challenges:

- 1) Find the primary sources,
- 2) Perform automated extraction at an accuracy comparable with manual work,
- 3) Scale the process up to a large corpus of documents, while keeping accuracy

The solution

Oghma Select agentic system enabled the identification of 20,000 primary documents bearing a high chance to contain relevant solubility information. These were automatically downloaded and processed through *Oghma Extract*, which then compiled a new dataset of about 30,000 entries. Quality control was performed on the original APC primary sources (162 documents – control dataset), for which the extracted data were verified independently by APC: 96% of the entries were within a log difference of 0.3 with another 1.6% between 0.3 and 0.6. A ML model, independently constructed by APC on the control dataset, returned an R^2 coefficient within 10% of that of the original model based on manually curated data.

Breakthrough results

Metric	Without Oghma	With Oghma	Improvement
Dataset size (entries)	12,078	25,600	Two-fold enlargement of the dataset
Documents identified and processed	162	20,000	120x documents identified and analysed
Number of unique molecules (solutes)	101	253	2.5x increase in dataset diversity
Extraction time per 100 entries	3 hours	2.8 minutes	60x entry acquisition speed
Total extraction time (control database)	372.8 hours	12 hours	162 documents processed in 10 weeks against 20,000 documents processed in half a day.

Competitive Advantage

Oghma agentic system enables the extraction of solubility data at scale. This goes beyond solving a simple capacity problem. In fact, with the exhaustion of data-dense documents, additional entries have to be extracted from a wider corpus containing much more fragmented information. The new database extends to 253 molecules (from the original 101), many of which are only rarely investigated, so that a much larger corpus of primary sources has to be considered (in our case 20,000). Sourcing and processing that volume of documents is beyond manually extraction capacity: at the APC manual extraction rate, 20,000 documents can be processed in 5.2 years.

“The collaboration with Oghma highlights the increased value that can be obtained through the use of agentic-enabled search strategies. The use of these strategies is providing great support to our scientists in the development of new and more sophisticated models”

Brian Glennon, CTO and Co-founder, APC

Note on APC

Funded in 2011, APC Ltd. (www.approcess.com) is a privately owned Irish company with 250 staff based in Cherrywood in Dublin. APC partners with the global pharmaceutical and biopharmaceutical sector to develop better, more reliable processes for the manufacture of their medicines. Its sister company, VLE Therapeutics Ltd, focuses on the manufacture of advanced therapeutics.