https://doi.org/10.1038/s44286-025-00241-7

Why your next breakthrough needs fewer experiments

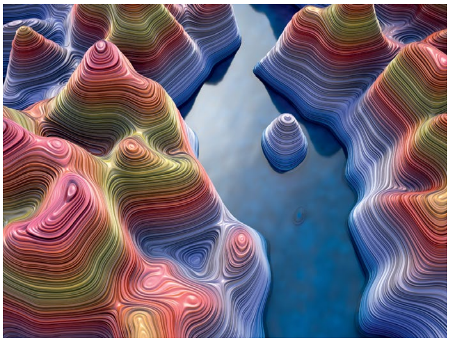
***Joel Paulson* discusses how selecting the most informative data can accelerate breakthroughs in chemical engineering.**

Chemical engineers are no strangers to scale: million-gallon reactors, megawatt electrolyzers and petabyte-scale industrial datasets. But in the age of artificial intelligence, size is not a strategy but just a helpful starting point. When experiments are slow, expensive or resource-intensive, it is not about how much data you have, but how wisely you choose it.

This idea is not new. A recent example from the machine learning community is so-called 'scaling laws', which describe how model performance grows with more parameters, data tokens and compute budget. These laws drove an arms race of ever-larger systems in natural language processing. But then came Chinchilla: a 2022 study from DeepMind that flipped the script. It showed that balanced growth – not just brute-force scale – was the key. A smaller model trained on more data outperformed a much larger model, all while using the same compute budget[1]. This showed that how resources are allocated can matter a lot more than raw scale.

So, what does this have to do with chemical engineering? At its core, the Chinchilla study is not a story about having more data – it is a story about using limited resources effectively. DeepMind held the compute budget fixed and reallocated it: smaller models, more strategically chosen data. That kind of trade-off, allocating limited resources where they deliver the most value, is second nature to chemical engineers. Our bottlenecks are not graphics processing unit (GPU) hours – they are reactor runs, instrument time and sleepless graduate students. Even though we cannot afford to brute force our way through parameter space, we can be smart about which measurements we take next.

That is exactly what techniques such as optimal experimental design[2] and its modern (useful and flexible) cousin, Bayesian optimization[3], are built to do. Rather than blindly collecting data, they help you pick the next measurement based on what you already know. Each step becomes a decision: observe, update your belief, then ask the most informative next question. It is like navigating a mountain range – not by mapping every single inch, but by following the contours most likely to lead you uphill.



And it works. In metal–organic framework discovery, a recent study searched a space of 47,740 candidates with approximately 1% of candidates (around 480 evaluations) to find all 7 Pareto-optimal materials for carbon dioxide capture and separation[4]. That is, months of simulation time reduced to days. In another example, a closed-loop campaign targeting photostable donor–acceptor molecules found high-performing structures after synthesizing roughly 30 compounds – less than 1.5% of a 2,200-member design space[5]. The algorithm learned chemistry as it optimized, something that is nearly impossible if you try to brute force your way through every option. And in catalyst design, synthesizing just 33 aluminum complexes was enough to uncover the top performers for stereoselective polymerization, from a design space of a possible 576 candidates[6]. This leads to less trial-and-error and more strategic precision.

Across domains, a pattern emerges. Many researchers casually cite a '30% rule': roughly 30% of the right data can deliver 90–95% of full-data performance. It is not just hearsay. A meta-analysis in medical imaging found that a deep learning model for classifying COVID-19 from computed tomography (CT) scans reached 95% of its final accuracy using only 30% of labeled data[7]. The exact number varies – it depends on noise, dimensionality and prior knowledge – but the message is clear: the information curve flattens quickly[8].

Chemical engineering problems, from materials design to process control, are particularly ripe for this kind of data efficiency. We may not routinely have giant datasets, but we do have physics-based constraints, mechanistic models and control over how we design experiments. These are exactly the ingredients that make smart data strategies work. Whether you are discovering new electrolytes, designing modular reactors or optimizing separation trains, embedding sequential design into the workflow can drastically reduce cost, time and environmental impact.

So, let's stop chasing data just because we can. The next breakthrough will not come from simply pouring more experiments into the hopper – it will come from asking sharper questions and listening closely to each answer. In a world of rising energy prices, supply-chain disruptions and mounting sustainability challenges, that is a numbers game every chemical engineer should want to play.

Joel A. Paulson

Chemical and Biomolecular Engineering,
The Ohio State University, Columbus, OH, USA.
e-mail: paulson.82@osu.edu

Published online: 22 July 2025

References

1. Hoffmann, J. et al. In *Advances in Neural Information Processing Systems 35* (eds Koyejo, S. et al.) 30016–30030 (NeurIPS, 2022); https://go.nature.com/4jnSAU
2. Fisher, R. A. *The Design of Experiments* 9th edn (Macmillan, 1971).
3. Paulson, J. A. & Tsay, C. *Curr. Opin. Green Sustain. Chem.* **51**, 100983 (2025).
4. Comlek, Y., Pham, T. D., Snurr, R. Q. & Chen, W. *npj Comput. Mater.* **9**, 170 (2023).
5. Angello, N. H. et al. *Nature* **633**, 351–358 (2024).
6. Wang, X. Q. et al. *Nat. Commun.* **14**, 3647 (2023).
7. Wu, X., Chen, C., Zhong, M., Wang, J. & Shi, J. *Med. Image Anal.* **68**, 101913 (2021).
8. Wu, Y., Walsh, A. & Ganose, A. M. *Digital Discov.* **3**, 1086–1100 (2024).

Competing interests

The author declares no competing interests.

CREDIT: XIA YUAN/ MOMENT/GETTY