Volume 4, Article 04, 2026
DOI:10.25950/6c2e5798

Large language models as priors for low-data chemical and materials discovery

Commentary by

Kevin Maik Jablonka
Laboratory of Organic and Macromolecular Chemistry (IOMC), Friedrich Schiller University Jena, Humboldtstrasse 10, 07743 Jena, Germany
Jena Center for Soft Matter (JCSM), Friedrich Schiller University Jena, Philosophenweg 7, 07743 Jena, Germany
Center for Energy and Environmental Chemistry Jena (CEEC Jena), Friedrich Schiller University Jena, Philosophenweg 7a, 07743 Jena, Germany
Helmholtz Institute for Polymers in Energy Applications Jena (HIPOLE Jena), Lessingstraße 12–14, 07743 Jena, Germany
Correspondence: mail@kjablonka.com

Adrian Mirza
Helmholtz-Zentrum Berlin für Materialien und Energie GmbH, Hahn-Meitner-Platz 1, 14109, Berlin, Germany
Helmholtz Institute for Polymers in Energy Applications Jena (HIPOLE Jena), Lessingstraße 12–14, 07743 Jena, Germany

on

1. **Chemical reasoning in LLMs unlocks strategy-aware synthesis planning and reaction mechanism elucidation**
   A. M. Bran, T. A. Neukomm, D. Armstrong, Z. Jončev, and P. Schwaller, Matter, 9(5):102812 (2026)
2. **Bayesian Optimization of Catalysis with In-Context Learning**
   M. C. Ramos, S. S. Michtavy, A. D. White, and M. D. Porosoff, ACS Cent. Sci., 12(5):599 (2026)

Received June 27, 2026
Published July 14, 2026

Statement of Significance

The chemical sciences often operate in the low-data regime, yet practitioners typically hold prior beliefs about what should be feasible or expected. Such priors have seldom been incorporated into modeling approaches in the chemical sciences. Large language models (LLMs) can offer a route to do so: recent findings indicate that they carry priors shaped by their architecture and training data that can make them more data-efficient. On top of that, they accept inputs in natural language and thus enable the use of data that could not have been used for modeling before. We discuss two recent works that leverage this prior—using LLMs to score fuzzy criteria in synthesis planning, and as in-context surrogate models for Bayesian optimization—and argue that fully delivering on this promise will require better ways to evaluate fuzzy objectives and to reproduce results.

The discovery of a material or molecule class is, by definition, a rare event. In machine-learning terms, this means that we often have little data for those cases. And yet we rarely start empty-handed: faced with a new

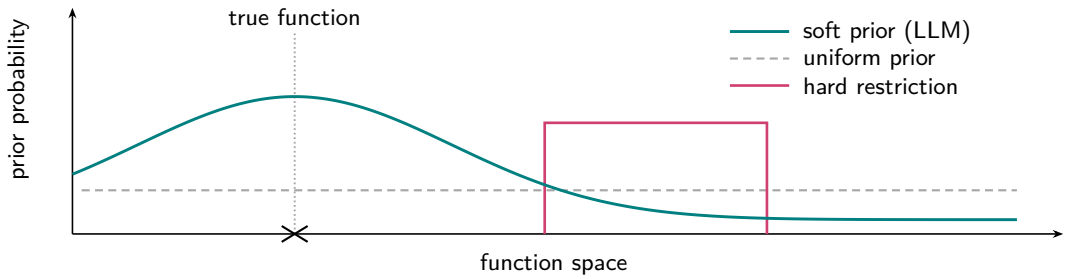

Figure 1: Inductive biases as priors over candidate functions; the dotted line marks the true function. A *uniform* prior (grey) spreads belief evenly and so carries little mass on any particular function; with few data it cannot concentrate on the truth. A *hard restriction* (purple red) places uniform mass on a compact set of functions and none outside it; if the truth falls outside that set, as here, it is excluded no matter how much data is collected. A *soft* inductive bias (teal)—what an LLM can provide—keeps support across the whole space while concentrating most of its mass near the truth, combining the coverage of the uniform prior with the focus of the restriction. This is what makes it useful when data are scarce. Illustration inspired by Wilson [1] and Goldblum et al. [2].

problem, a chemist already holds strong beliefs about what should be feasible or expected. We have largely failed to put those beliefs into our models, starting them instead from scratch, as if we knew nothing.

In general, a good model is expressive enough to represent many kinds of functions while also having a preference—an inductive bias—for the "right" ones [1] (Fig. 1).

Wilson and co-workers identified one such inductive bias in transformers, even untrained ones: they prefer simple functions [2]. When one randomly samples sequences of characters from these models, the sequences are simpler than randomly assembled ones. On top of this architectural effect, LLMs carry a prior shaped by all the data they have seen in training. Two recent works show how to leverage this prior in the chemical sciences [3, 4], in two complementary ways (Fig. 2).

Strategic reasoning for reaction planning

Synthesis planning is one area of strategic reasoning that relies on expert knowledge, with broad implications such as in drug discovery, where the goal is to find the most efficient route to a target compound. Current systems use models that predict single-step transformations. One can then couple those single-step predictions to a search algorithm, such as Monte Carlo tree search [5], to find feasible synthesis pathways. A good synthesis, however, also involves strategy. Expert chemists know when to use a protecting group, or not to carry a complex stereocenter to the final transformation. This is very hard to implement with a search algorithm, and properties such as "feasibility" tend to be too fuzzy to hard-code.

Bran and co-workers address this by using an LLM for the fuzzy scoring steps. In their Synthegy system, they guide synthesis planning with natural language [6]. With a prompt such as "highly feasible synthesis routes with high overall yields consider potential side reactions and byproducts, while ensuring that no unnecessary reactions are performed", they could filter out infeasible reactions from their synthesis routes. Crucially, the LLM never generates routes or predicts single steps—a classical search algorithm does that—and is used only as an evaluator for these fuzzy criteria. To evaluate the results, they recruited a panel of 36 experts and found that, in a double-blind A/B study, the experts' assessments agree with Synthegy in more than 71% of cases. The same principle extends beyond route planning: the authors also use the LLM to score elementary electron-pushing steps, guiding the search toward plausible reaction mechanisms.

These results indicate that LLMs can evaluate and rerank candidates against fuzzy criteria expressed in natural language, drawing on the priors built up during training.

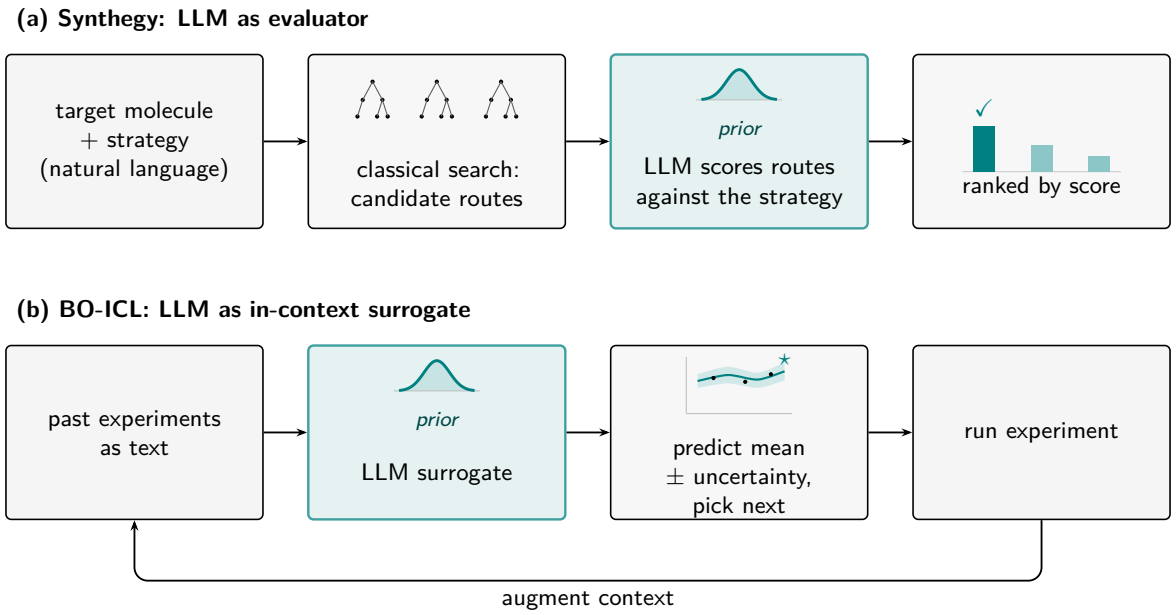

Figure 2: Both works place the LLM at the same point—as the carrier of the prior—but they use it differently. (a) In Synthegy, a classical search generates the candidate synthesis routes and the LLM only evaluates them against a strategy stated in natural language; the LLM never generates chemical structures itself. (b) In BO-ICL, the LLM is the surrogate model of Bayesian optimization: past experiments are supplied as natural-language context, the LLM predicts each outcome and its uncertainty ($\star$ marks the next pick), and the result is appended to the context.

Bayesian optimization

Such chemical priors—and the ability to incorporate fuzzy "hints"—also matter for optimization. The most commonly used flavor is Bayesian optimization, in which a surrogate model is trained iteratively to predict the expected outcome of each measurement alongside an uncertainty. Using the prediction and uncertainty estimate, an acquisition function then proposes the experiment that gives the best bet at reaching the optimum in the next iteration. Despite numerous studies, reviews, and packages on Bayesian optimization, reliably convincing results remain surprisingly difficult to achieve in practice.

One reason is that we typically begin observations completely ignorant of previous optimizations or general scientific understanding. Practitioners often use Gaussian processes initialized with a few datapoints for the task at hand, but seldom draw on learnings from similar experiments or even simple chemical heuristics to guide the optimization.

Ramos and co-workers provide a different paradigm—again relying on the LLM as prior [7]. The authors implement the LLM as the surrogate model using in-context learning. In in-context learning, no model weights are updated, and the "learning" happens directly "in the context" by providing training examples in the prompt. A useful side effect is that inputs can be provided in natural language, which could be used to incorporate fuzzy intuitions to further tune the model's prior. The authors demonstrate Bayesian optimization with in-context learning on four datasets. They also consider experimental procedures in natural text—a relevant application, as many modeling efforts currently rely on simplified tabularized representations of experimental procedures—and at least match the performance of Gaussian process regression. They further perform prospective validation by optimizing catalysts for the reverse water-gas shift reaction, where their LLM-based optimization approached the thermodynamic yield limit within a handful of iterations, well ahead of a random search.

Conclusions and outlook

Both works rest on the same idea: by carrying prior knowledge and accepting inputs in plain language, LLMs let us do more with less data—and let us simply tell a model what we already know. This may be one of the field's central promises, and curiously one of its least explored.

Two obstacles might explain why. The first is evaluation: there is still no scalable, systematic way to score performance on the soft objectives these methods target, and progress is hard to judge without it. The works by Bran and Ramos make first steps in this direction — but the fundamental fact that things that are hard to specify are hard to evaluate remains. For instance, there is no ground truth for what counts as "good strategy". Bran and co-workers handle this carefully, with their expert panel—and even so the experts disagree with one another about as much as they disagree with the system, the more so on the hardest routes. That is less a shortcoming of their setup than a property of the objective: one cannot cleanly score a system against a target the experts themselves do not agree on. The temptation is then to fall back on whatever is easiest to measure—a version of McNamara's fallacy, where we optimize the proxy we can quantify and quietly drop the objective we actually care about [8].

The second is reproducibility. This is especially visible in the Bayesian-optimization work: the models in the preprint were unavailable a short time later, the peer-reviewed version uses a different set, and some of those are already gone as we write this comment. And it is not only a matter of access: in the synthesis work, the capability itself is tied to scale—strategy-aware reasoning appears only in the newest and largest models, and smaller ones do no better than chance. Either way, the prior we lean on is a moving target, and—with proprietary models—one we often do not control. And it might be even worse than a moving target: the performance of these models seems to depend, in ways we do not yet fully understand, on factors such as how a problem is represented or phrased, and they can behave in strange ways—arriving at correct answers through what philosophers call epistemic luck [9–14]. This matters here because, when an LLM rediscovers a catalyst, it is hard to distinguish true chemical reasoning from memorization. Such failures can be probed—Ramos and co-workers scramble their labels and find that their method then does no better than random search, evidence that it uses a real signal rather than recall—but these checks are still the exception. Even the uncertainty calibration that Bayesian optimization relies on depends on the base model, and a post-trained model can be less calibrated than the base one [15].

Beyond this, there is a more fundamental tension. Optimization identifies local extrema within a given candidate space; discovery seeks to extend that space beyond what was anticipated. The use of a prior is most suited when data are scarce. However, in discovery we often want to leave the familiar behind—and an LLM's prior pulls back toward what is already common in its training data. Ramos and co-workers see this directly: the pretraining that speeds up their optimization also biases it toward literature-familiar catalysts. This brings us to the deeper question of whether any model or system can ask new questions or create new knowledge, or only recombine what it has already seen [16, 17].

The opportunity is nonetheless real: LLMs let us fold the tacit [18], hard-to-formalize knowledge that defines so much of chemistry—experimental procedures, rules of thumb, intuitions about what should work—directly into our models. Whether we deliver on it depends less on better models than on whether we, as a community, learn to evaluate these methods rigorously and to preserve the ground we gain.

Acknowledgments

We thank Martiño Ríos-García for feedback on the manuscript. We acknowledge support from the Helmholtz Association within the framework of the Helmholtz Foundation Model Initiative (project SOL-AI) and support from the Carl Zeiss Foundation.

References

[1] A. G. Wilson, "Deep Learning is Not So Mysterious or Different", *arXiv preprint arXiv: 2503.02113*, 2025,

[2] M. Goldblum, M. Finzi, K. Rowan, and A. G. Wilson, "The No Free Lunch Theorem, Kolmogorov Complexity, and the Role of Inductive Biases in Machine Learning", *arXiv preprint arXiv:2304.05366*, 2023, https://arxiv.org/abs/2304.05366.

[3] N. Alampara, A. Aneesh, M. Ríos-García, A. Mirza, M. Schilling-Wilhelmi, A. A. Aghajani, M. Sun, G. Prastalo, and K. M. Jablonka, "General-Purpose Models for the Chemical Sciences: LLMs and Beyond", *Chemical Reviews*, 126(4):2484–2549, 2026, http://dx.doi.org/10.1021/acs.chemrev.5c00583.

[4] M. C. Ramos, C. J. Collison, and A. D. White, "A review of large language models and autonomous agents in chemistry", *Chemical Science*, 16(6):2514–2572, 2025, http://dx.doi.org/10.1039/D4SC03921A.

[5] M. H. S. Segler, M. Preuss, and M. P. Waller, "Planning chemical syntheses with deep neural networks and symbolic AI", *Nature*, 555(7698):604–610, 2018, https://doi.org/10.1038/nature25978.

[6] A. M. Bran, T. A. Neukomm, D. Armstrong, Z. Jončev, and P. Schwaller, "Chemical reasoning in LLMs unlocks strategy-aware synthesis planning and reaction mechanism elucidation", *Matter*, 9(5):102812, 2026, https://doi.org/10.1016/j.matt.2026.102812.

[7] M. C. Ramos, S. S. Michtavy, A. D. White, and M. D. Porosoff, "Bayesian Optimization of Catalysis with In-Context Learning", *ACS Cent. Sci.*, 12(5):599–615, 2026, https://doi.org/10.1021/acscentsci.5c02418.

[8] N. Alampara, M. Schilling-Wilhelmi, and K. M. Jablonka, "Lessons from the trenches on evaluating machine learning systems in materials science", *Computational Materials Science*, 259:114041, 2025, https://www.sciencedirect.com/science/article/pii/S0927025625003842.

[9] M. Ríos-García, N. Alampara, and K. M. Jablonka, "Semantic Content Determines Algorithmic Performance", *arXiv preprint arXiv: 2601.21618*, 2026,

[10] K. M. Jablonka, "Clever Materials: When Models Identify Good Materials for the Wrong Reasons", *arXiv preprint arXiv: 2602.17730*, 2026,

[11] M. Ríos-García, N. Alampara, C. Gupta, I. Mandal, S. Mannan, A. A. Aghajani, N. M. A. Krishnan, and K. M. Jablonka, "AI scientists produce results without reasoning scientifically", *arXiv preprint arXiv: 2604.18805*, 2026,

[12] Z. Qiu, W. Liu, H. Feng, Z. Liu, T. Z. Xiao, K. M. Collins, J. B. Tenenbaum, A. Weller, M. J. Black, and B. Schölkopf, "Can Large Language Models Understand Symbolic Graphics Programs?", *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, OpenReview.net, 2025, https://openreview.net/forum?id=Yk87CwhBDx.

[13] M. Nezhurina, L. Cipolina-Kun, M. Cherti, and J. Jitsev, "Alice in Wonderland: Simple Tasks Showing Complete Reasoning Breakdown in State-Of-the-Art Large Language Models", *arXiv preprint arXiv: 2406.02061*, 2024,

[14] J. Ichikawa and M. Steup, "The Analysis of Knowledge", *The Stanford Encyclopedia of Philosophy*, ed. by E. N. Zalta and U. Nodelman, Summer 2026, Metaphysics Research Lab, Stanford University, 2026.

[15] OpenAI, *GPT-4 Technical Report*, *arXiv preprint arXiv:2303.08774*, https://arxiv.org/abs/2303.08774.

[16] S. Thais, R. Trotta, N. Suri, E. Sullivan, V. Poonamallee, T. N. Narong, R. Croft, and N. Hartman, *AI for Science Needs Scientific Alignment*, https://philsci-archive.pitt.edu/28744/.

[17] F. Chollet, "On the measure of intelligence", *arXiv preprint arXiv:1911.01547*, 2019,

[18] M. Polanyi, *The tacit dimension*, eng, Reproduction en fac-similé, Chicago: University of Chicago press, 2009.