



Generative Chemistry Models Still Need Medicinal Chemists to Ask the Right Questions

Sipho Lewandowski^{1*}, Lakshmi Gagnon², Chloe Avetisyan³

¹Department of Pharmacy, Sharavati Institute of Pharmacy, Nigeria

²Department of Pharmacy, Orinoco University Centre for Pharmacy, Greece

³Department of Pharmacy, Balkan University Centre for Pharmacy, Malaysia

Article Info

Volume 1, Issue 1, 2026

doi: [10.62715/ijpsdd.2026.v1.i1.pp23-28](https://doi.org/10.62715/ijpsdd.2026.v1.i1.pp23-28)

Received: 15 February 2026

Accepted: 23 April 2026

Published: 22 June 2026

Abstract

Generative models now produce valid, novel and often synthesisable molecular structures in large numbers, and several have contributed to compounds that reached experimental testing. The attention this has attracted tends to settle on the generator, as though the hard part of molecular design were the proposing of structures. This perspective argues the opposite. The capacity to enumerate candidates has become abundant, while the capacity to specify what should be enumerated has not, and that specification is a medicinal chemistry task that no current model performs. A generative run is defined by its scoring function, its constraints and its starting point, and each encodes judgements about which target to pursue, which selectivity matters, what metabolic liability is tolerable in the intended indication, and which measured property the team is willing to trade away. Models optimise what they are given, so a poorly posed objective produces compounds that score well and mean nothing. The article sets out the argument, takes seriously the counter-case that models will absorb these judgements as data accumulate, and draws implications for how design projects are organised and how medicinal chemists are trained. The claim is not that automation has been oversold, but that its value depends on a skill the field has not systematically taught.

Keywords: *Generative models, De novo design, Medicinal chemistry, Scoring functions, Design objectives, Drug discovery*

© 2026 Sipho Lewandowski et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

In the space of a few years, generating molecular structures stopped being difficult. Recurrent networks trained on known compounds produce valid structures and can be focused on a chemical series (Segler *et al.*, 2018a). Variational autoencoders map structures into a continuous space in which optimisation can be carried out (Gómez-Bombarelli *et al.*, 2018). Reinforcement learning steers a generator towards a desired profile (Olivecrona *et al.*, 2017). The outputs are valid, largely novel, and increasingly synthesisable, and a widely discussed study reported a kinase inhibitor designed with such methods and tested in animals within a few weeks (Zhavoronkov *et al.*, 2019).

* Corresponding author: Sipho Lewandowski, Department of Pharmacy, Sharavati Institute of Pharmacy, Nigeria. E-mail: sipho.lewandowski@sharavati.edu

© 2026 Sipho Lewandowski et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The enthusiasm that followed centred on the generator, as if proposing structures had been the obstacle. It was not. Chemists have never had difficulty drawing molecules; the difficulty has always been in deciding which molecules are worth making, and that decision is made before a generator is started, in the objective it is given. This perspective argues that the design question, not the generative engine, now limits what these methods deliver, and that posing it well is a medicinal chemistry skill which the current organisation of projects neither rewards nor teaches.

2. What the Models Actually Optimise

Every generative run is defined by three things: a scoring function, a set of constraints and a starting point. The model searches for structures that score well under the function, within the constraints, from where it was told to begin. It has no access to anything outside that specification. If the score rewards predicted affinity for a single protein, the model will find structures that the affinity predictor likes, including structures that exploit the predictor's weaknesses. If a synthesisability term is added, using a heuristic such as the fragment-based accessibility score of [Ertl and Schuffenhauer \(2009\)](#) or a learned estimate of complexity ([Coley *et al.*, 2018](#)), the output becomes more makeable but no more relevant.

This is not a criticism of the models; it is their nature as optimisers. The point has been made from within the field. [Walters and Murcko \(2020\)](#) observed that the published demonstrations rarely show that a generative method found something a competent chemistry team would not have, and proposed comparisons that would settle the matter. Benchmarks such as those of [Brown *et al.* \(2019\)](#) and [Polykovskiy *et al.* \(2020\)](#) measure validity, novelty, diversity and the ability to hit a stated objective. They cannot measure whether the objective was worth hitting, because that depends on the project and not on the model.

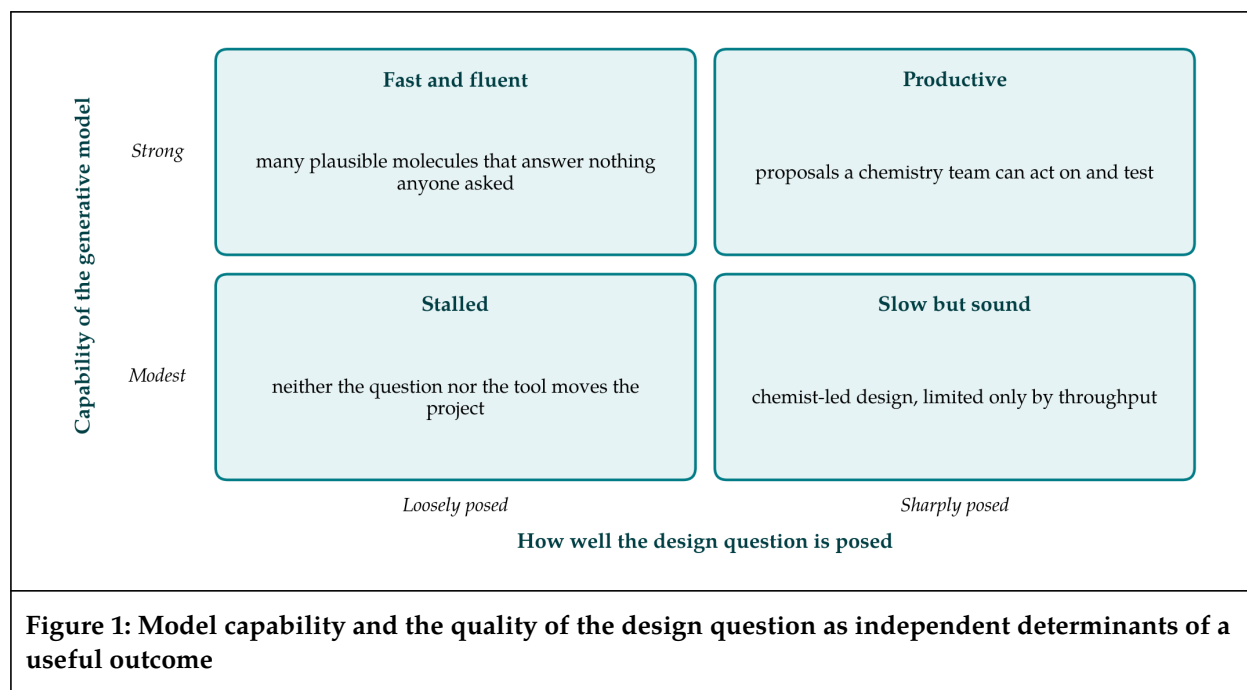
Table 1 illustrates the pattern across several ways a design question is commonly posed, the quantity the model then optimises, the failure that results, and what a chemist contributes when the question is set more carefully.

Table 1: How the posing of a design question shapes what a generative run returns			
Question as posed	What the model optimises	Characteristic failure	Chemist's contribution
Make potent binders of this protein	A single predicted affinity	Exploits the predictor; ignores selectivity	Names the off-targets that matter
Improve this property profile	A weighted sum of predictions	Optimises the weights, not the project	Sets what may be traded for what
Find novel scaffolds	Distance from the training set	Novel and unmakeable, or novel and known-bad	Distinguishes useful from merely new
Keep it synthesisable	A complexity heuristic Ertl and Schuffenhauer, 2009	Easy chemistry, irrelevant compounds	Judges route risk against value
Match known actives	Similarity to the training series	Reproduces the series and its liabilities	Decides when to leave the series
Maximise ligand efficiency	A ratio metric Hopkins <i>et al.</i>, 2014	Small fragments with no prospects	Fixes the stage at which it applies

3. The Argument: Specification Is the Scarce Skill

Posing a design question well requires knowledge that is tacit, project-specific and largely unwritten. It means knowing which selectivity ratio against a related enzyme is a genuine safety requirement and which is a habit of the therapeutic area. It means knowing that a basic amine will be tolerated in one indication and not in another, that a particular metabolic soft spot can be left alone because the intended dose is low, that a measured potency in one assay should be discounted because the substrate

concentration is far from physiological. These are the judgements that make a design objective correspond to the problem the project actually has. Figure 1 places the two dimensions, the capability of the tool and the sharpness of the question, against each other, and the region of interest is the one where both are high.



The attrition literature supports the emphasis. Candidates fail predominantly for lack of efficacy and for safety findings rather than for want of potency (Waring *et al.*, 2015), and the long decline in research productivity has been attributed in part to screening and design systems that answer the wrong question efficiently (Scannell *et al.*, 2012). A generator optimising predicted affinity accelerates exactly the part of the process that was never the bottleneck. Bender and Cortés-Ciriano (2021) make a related point when they distinguish what artificial intelligence has demonstrably achieved in drug discovery from what has been claimed on its behalf, and locate most of the gap in the framing of problems and the quality of data rather than in the algorithms.

There is a further consideration specific to optimisers. Any sufficiently strong search will exploit the weaknesses of its objective, finding structures that satisfy the letter of the score while violating its intent. Protecting against that requires someone who can look at a proposed molecule and see what is wrong with it for reasons the score does not contain, which is a description of an experienced medicinal chemist. The skill is diagnostic as much as creative, and it is exercised after the run as well as before it.

4. The Counter-Case

The strongest argument against this position is historical. Chess and protein structure prediction were both held to require irreducible human judgement until models absorbed enough of it from data, and structure prediction in particular now operates at a level that has changed how structure-based design begins (Jumper *et al.*, 2021). Retrosynthetic planning, long treated as the preserve of trained intuition, has been performed competitively by learned systems (Segler *et al.*, 2018b). If route design can be learned, the argument runs, so can objective design, once enough project histories are recorded in a usable form.

The counter-case has force, and the response is not that design judgement is mysterious. It is that the data needed are of a kind the industry does not generate. Chess offers unlimited self-play and structure prediction a large public archive of solved structures. The record of which design objectives were correct is held in failed internal programmes, is rarely published, is seldom documented in a form that survives the reorganisation of a department, and accumulates at the rate of a few terminated projects per company per year. Schneider *et al.* (2020) set out what would have to be built for autonomous design to work, and the list is dominated by data infrastructure and closed experimental loops rather than by better

generators.

A second line of counter-argument holds that the question matters less when synthesis and testing are cheap, because a poorly specified run can simply be tested and corrected. Automated synthesis and screening have made some of this feasible, and the identification of an antibacterial compound by screening a large virtual library with a trained classifier shows what throughput can do when the property of interest is directly measurable (Stokes *et al.*, 2020). But most properties that decide a programme, chronic tolerability above all, are not measurable at that speed or cost, and for those the specification cannot be replaced by iteration.

5. Implications for Practice and Training

If specification is the scarce input, three things follow for how projects are run. The scoring function deserves the scrutiny that a synthetic route receives: written down, justified term by term, reviewed by the team, and revised explicitly when the project's understanding changes, rather than inherited from a previous programme. The chemists who hold the relevant judgement should be involved in constructing it rather than presented with the output of a run, which also addresses the common complaint that generated compounds look unappealing to the people asked to make them. And negative results should be recorded in a form that captures the objective in force at the time, since that is the only way the data needed for the counter-case will ever exist.

For training, the implication is a shift in emphasis. Medicinal chemistry teaching has long centred on the properties of molecules, with the influential property-based analyses of Leeson and Springthorpe (2007) shaping what students learn to look for. That remains necessary, but it is no longer sufficient. Students should also be taught to construct and criticise objective functions: to say which property predictions are reliable enough to optimise against, how to weight competing terms, how to encode a constraint that is really a hard requirement, and how to detect that an optimiser has gamed a score. These are teachable skills, they belong in a computational laboratory alongside the mechanics of running a model, and at present they are mostly acquired by accident.

6. Conclusion

Generative chemistry models have made the proposing of molecules abundant and, in doing so, have made the posing of design questions the limiting step. The models optimise what they are told to optimise, and the content of that instruction, which target, which selectivity, which liability is tolerable, which property may be traded, comes from medicinal chemistry judgement that is neither encoded in the training data nor measured by current benchmarks. That may change if the record of past programmes is ever captured systematically, and the field should work towards it. Until then, the useful question is not whether these models will replace medicinal chemists, but whether projects are set up so that the chemists' judgement reaches the objective function at all.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- Bender, A. and Cortés-Ciriano, I. (2021). [Artificial intelligence in drug discovery: What is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet.](#) *Drug Discovery Today*, 26(2): 511-524. <https://doi.org/10.1016/j.drudis.2020.12.009>
- Brown, N., Fiscato, M., Segler, M.H.S. and Vaucher, A.C. (2019). [GuacaMol: Benchmarking models for de novo molecular design.](#) *Journal of Chemical Information and Modeling*, 59(3): 1096-1108.

<https://doi.org/10.1021/acs.jcim.8b00839>

Coley, C.W., Rogers, L., Green, W.H. and Jensen, K.F. (2018). *SCScore: Synthetic complexity learned from a reaction corpus*. *Journal of Chemical Information and Modeling*, 58(2): 252-261.

<https://doi.org/10.1021/acs.jcim.7b00622>

Ertl, P. and Schuffenhauer, A. (2009). *Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions*. *Journal of Cheminformatics*, 1: 8.

<https://doi.org/10.1186/1758-2946-1-8>

Gómez-Bombarelli, R., Wei, J.N., Duvenaud, D., Hernández-Lobato, J.M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T.D., Adams, R.P. and Aspuru-Guzik, A. (2018). *Automatic chemical design using a data-driven continuous representation of molecules*. *ACS Central Science*, 4(2): 268-276. <https://doi.org/10.1021/acscentsci.7b00572>

Hopkins, A.L., Keserü, G.M., Leeson, P.D., Rees, D.C. and Reynolds, C.H. (2014). *The role of ligand efficiency metrics in drug discovery*. *Nature Reviews Drug Discovery*, 13(2): 105-121.

<https://doi.org/10.1038/nrd4163>

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P. and Hassabis, D. (2021). *Highly accurate protein structure prediction with AlphaFold*. *Nature*, 596(7873): 583-589. <https://doi.org/10.1038/s41586-021-03819-2>

Leeson, P.D. and Springthorpe, B. (2007). *The influence of drug-like concepts on decision-making in medicinal chemistry*. *Nature Reviews Drug Discovery*, 6(11): 881-890. <https://doi.org/10.1038/nrd2445>

Olivecrona, M., Blaschke, T., Engkvist, O. and Chen, H. (2017). *Molecular de-novo design through deep reinforcement learning*. *Journal of Cheminformatics*, 9: 48. <https://doi.org/10.1186/s13321-017-0235-x>

Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., Kurbanov, R., Artamonov, A., Aladinskiy, V., Veselov, M., Kadurin, A., Johansson, S., Chen, H., Nikolenko, S., Aspuru-Guzik, A. and Zhavoronkov, A. (2020). *Molecular Sets (MOSES): A benchmarking platform for molecular generation models*. *Frontiers in Pharmacology*, 11: 565644.

Scannell, J.W., Blanckley, A., Boldon, H. and Warrington, B. (2012). *Diagnosing the decline in pharmaceutical R&D efficiency*. *Nature Reviews Drug Discovery*, 11(3): 191-200.

<https://doi.org/10.1038/nrd3681>

Schneider, P., Walters, W.P., Plowright, A.T., Sieroka, N., Listgarten, J., Goodnow, R.A., Fisher, J., Jansen, J.M., Duca, J.S., Rush, T.S., Zentgraf, M., Hill, J.E., Krutoholow, E., Kohler, M., Blaney, J., Funatsu, K., Luebke, C. and Schneider, G. (2020). *Rethinking drug design in the artificial intelligence era*. *Nature Reviews Drug Discovery*, 19(5): 353-364. <https://doi.org/10.1038/s41573-019-0050-3>

Segler, M.H.S., Kogej, T., Tyrchan, C. and Waller, M.P. (2018). *Generating focused molecule libraries for drug discovery with recurrent neural networks*. *ACS Central Science*, 4(1): 120-131.

<https://doi.org/10.1021/acscentsci.7b00512>

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

Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackermann, Z., Tran, V.M., Chiappino-Pepe, A., Badran, A.H., Andrews, I.W., Chory, E.J., Church, G.M., Brown, E.D., Jaakkola, T.S., Barzilay, R. and Collins, J.J. (2020). *A deep learning approach to antibiotic discovery*. *Cell*, 180(4): 688-702.

<https://doi.org/10.1016/j.cell.2020.01.021>

Walters, W.P. and Murcko, M. (2020). *Assessing the impact of generative AI on medicinal chemistry*. *Nature Biotechnology*, 38(2): 143-145. <https://doi.org/10.1038/s41587-020-0418-2>

Waring, M.J., Arrowsmith, J., Leach, A.R., Leeson, P.D., Mandrell, S., Owen, R.M., Pairaudeau, G., Pennie,

W.D., Pickett, S.D., Wang, J., Wallace, O. and Weir, A. (2015). [An analysis of the attrition of drug candidates from four major pharmaceutical companies.](#) *Nature Reviews Drug Discovery*, 14(7): 475-486. <https://doi.org/10.1038/nrd4609>

Zhavoronkov, A., Ivanenkov, Y.A., Aliper, A., Veselov, M.S., Aladinskiy, V.A., Aladinskaya, A.V., Terentiev, V.A., Polykovskiy, D.A., Kuznetsov, M.D., Asadulaev, A., Volkov, Y., Zholus, A., Shayakhmetov, R.R., Zhebrak, A., Minaeva, L.I., Zagribelnyy, B.A., Lee, L.H., Soll, R., Madge, D., Xing, L., Guo, T. and Aspuru-Guzik, A. (2019). [Deep learning enables rapid identification of potent DDR1 kinase inhibitors.](#) *Nature Biotechnology*, 37(9): 1038-1040. <https://doi.org/10.1038/s41587-019-0224-x>