



Chemical Education

Division of Chemical Education

A Division of the Swiss Chemical Society

DIY a Molecule: Generative Models in Chemistry

Magdalena Lederbauer

*Correspondence: M. Lederbauer, E-mail: magled@mit.edu
Department of Chemical Engineering and Schwarzman College of Computing, Massachusetts Institute of Technology, 77 Massachusetts Ave, 02139 Cambridge MA, USA.

Abstract: This column introduces generative artificial intelligence and its application to molecular design. We contrast generative models with predictive models that most chemists have already encountered, build up the intuition behind conditional generation and explore how discrete diffusion models treat molecular design as a ‘fill-in-the-blank’ problem. Using a recently developed generative fragment-based drug discovery model, we provide a companion web application where chemists can interactively generate, evaluate and visualize novel molecules.

Keywords: Digital chemistry · Fragment-based drug design · Generative artificial intelligence · Machine learning

Chemistry, at its core, is a science of creating. Through most of its history, this act of creation took place in the laboratory: the discovery of new molecules relied on physical intuition, incremental structural modifications and high-throughput screening.^[1] Recently, however, the computer has become a creative venue in its own right. As noted in a previous column in this series, structural representations in chemistry have long been among its sharpest intellectual tools since a drawn formula anticipates what is possible, not just what is known.^[2] Furthermore, statistical algorithms trained on millions of chemical structures can now use those very representations to computationally propose plausible new molecules and materials.^[3] This is generative artificial intelligence. This column provides a conceptual introduction to how these models work and explores how discrete diffusion treats molecular design as a ‘fill-in-the-blank’ puzzle, illustrated by a companion web application which readers can explore.

What Makes a Model ‘Generative’?

To appreciate what generative models do, it helps to contrast them with the kind of machine learning most chemists already know. A *discriminative* model takes a molecular structure as input and predicts a label or value, for instance: Will this compound inhibit a kinase? What is its lipophilicity $\log P$?^[3] These tasks are useful, but they all answer the same question: ‘What are the properties of this molecule?’

A generative model asks something different. Rather than mapping structures to labels, it learns the probability distribution $p(X)$ of structures themselves and draws new samples from it. Trained on a large library of drug-like compounds, for instance, a model produces molecules that resemble the training data distribution, sharing statistical patterns in ring sizes, valences, and bonds. In other words, the model learns what molecules in that library look like and generates (arbitrarily many) more of them.^[3]

The most useful generative models in chemistry are *conditional*. In practice, we rarely want just any molecule but ones with specific properties, for example high binding affinity, synthetic accessibility and drug-likeness. This is where we make use of Bayes’ rule. Writing Y for a desired property and X for a molecule, the conditional distribution we want to sample from is:

$$p(X|Y) \propto p(Y|X) \cdot p(X) \quad (1)$$

$p(X)$ is the so-called *prior*, which is what the model has learned about the distribution of molecules in general. $p(Y|X)$ is the *likelihood*, in practice a scoring function evaluating how well a candidate satisfies the desired property. Their product is the *posterior*: the narrower region of chemical space containing molecules that are both structurally plausible and match the target (see Fig. 1).^[4]

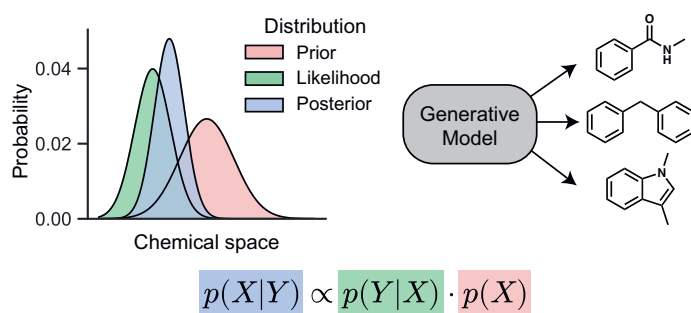


Fig. 1. How generative models target chemical space. Left: the prior $p(X)$ (pink) represents all molecules the model has encountered; the likelihood $p(Y|X)$ (green) scores candidates against a desired property; their product, the posterior $p(X|Y)$ (blue), is the region of chemical space where both overlap, i.e. molecules that are structurally plausible and satisfy the target property. Right: in practice, a generative model samples from this posterior to propose novel candidate structures.

This conditioning signal Y can take many forms: a predicted property score, a substructure constraint or even a direct experimental measurement. The latter is particularly compelling: Mass spectrometry, for instance, is routinely used to identify unknown compounds, where a molecule is fragmented and chemists interpret the resulting spectrum to determine the structure. A recently developed model called DiffMS^[5] frames this as a generative problem: Given a fragmentation pattern as the conditioning signal, it is trained to generate candidate structures that are consistent with that mass spectrum.

Discrete Diffusion Fills in the Blanks

The generative models described above need a concrete mechanism to produce new structures. One such approach is *masked*

Would you like to publish a Chemical Education topic here?

Please contact: Prof. Catherine Housecroft, University of Basel, E-mail: Catherine.Housecroft@unibas.ch

diffusion. The general idea, which is defining a process that gradually destroys structure in data and training a neural network to reverse it, was established in the context of image generation.^[6] For images, ‘destruction’ typically means adding Gaussian noise to pixels, where the network learns to denoise, and new images are generated starting from pure noise and running the process in reverse. For discrete sequences such as molecular strings (for example SMILES, which were introduced in a related recent column^[7]), the analogous destruction process is masking: progressively replacing symbols with a [MASK] placeholder until the entire molecule is hidden, then learning to recover it (Fig. 2a).^[8]

What makes this particularly well-suited to molecular design is that the model ‘sees’ the entire sequence at once, including masked positions. This stands in contrast to autoregressive models like GPT, which generate sequences strictly left-to-right, each token conditioned only on what came before.^[9] In masked diffusion, any part of the molecule can be fixed as context while the rest is generated, which maps naturally onto how fragment-based drug discovery works, as we will see next.

Molecules as Fragments

To see these principles in practice, we turn to GenMol,^[11] a discrete diffusion model trained on roughly a billion molecular struc-

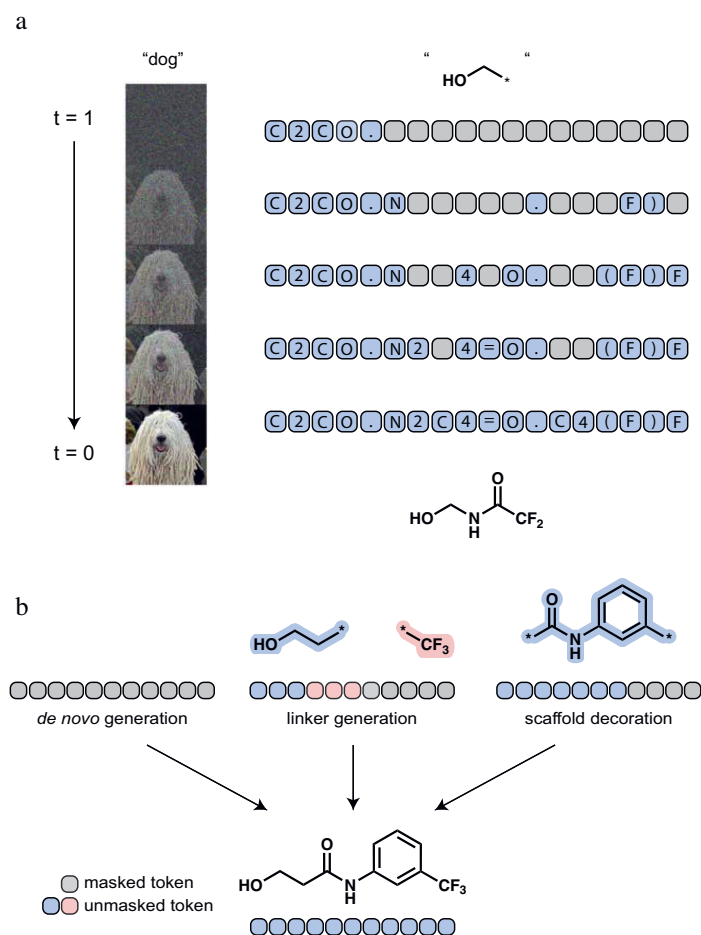


Fig. 2. Discrete diffusion for generating molecular structures. (a) During training, a clean molecular sequence is progressively masked (bottom to top); the model learns to reverse this process. At generation time, the model starts from a fully masked sequence and iteratively fills in tokens (top to bottom) until a complete structure is created. The image of a Komondor dog is included as an analogy for image-based diffusion models, where pixels are noised and denoised rather than tokens masked and unmasked.^[10] (b) The same model handles multiple drug discovery tasks by changing which fragments are provided as context (in colour) and which are left for the model to fill in (grey). Adapted from Lee *et al.* Ref. [11].

tures. Rather than encoding molecules atom-by-atom as in standard SMILES, GenMol uses the so-called SAFE representation,^[12] which decomposes a molecule into its constituent fragments and treats them as the basic units of generation (see Fig. 2a).

With this representation, generating a molecule becomes a fill-in-the-blank problem: some fragments are fixed as context and the model generates the rest. In *de novo generation*, the whole molecule is masked. In *scaffold decoration*, the core ring system is fixed. In *linker design*, two terminal fragments are held constant while the model generates the bridge (see Fig. 2b). The same logic applies to optimization: in fragment remasking, one fragment of an existing molecule is replaced with mask tokens and regenerated conditioned on the rest, allowing the model to iteratively improve an existing structure one fragment at a time.^[11]

Try it Yourself

We developed a companion web application to test the model’s capabilities directly (<https://mlederbauer.com/posts/2026-03-31-genai-fbdd/>). The user can select a task, draw, or type an input fragment in the embedded molecular editor, mark attachment points with [*] and click ‘Generate’. The model returns a gallery of candidate molecules annotated with computed physicochemical properties, such as a drug-likeness score (QED), molecular weight, logP and the number of hydrogen bond donors and acceptors. Generated molecules are projected onto a two-dimensional map of chemical space, using PubChem as a reference library, so the user can immediately see where new structures land relative to known compounds.^[13]

Two generation parameters of GenMol are worth exploring. *Temperature* works analogously to its thermodynamic counterpart of the Boltzmann distribution: low values confine sampling to high-probability, drug-like structures, while high values allow the model to explore more unusual regions of chemical space at the cost of occasional implausibility. Randomness controls which fragments get unmasked first during generation, introducing a second source of variation on top of temperature. Try out the application and test the model’s limits yourself!

Outlook

Generative molecular artificial intelligence has matured considerably over the past decade and several companies and research groups are now actively deploying it in drug discovery and materials design pipelines. In this column, we introduced how generative models learn distributions over chemical space, how masked diffusion generates molecules by filling in masked fragments and how both ideas come together in models such as GenMol. One important caveat is that a high predicted property score is not a synthesis plan, since models can propose structures that are difficult or impossible to make,^[14] and many structures presented as novel turn out to be close variants of known compounds.^[15] These tools are best used with ‘chemical judgment’, scepticism even. The best way to build it is to use these tools, test their limits and pay attention to where they fail. Have fun creating!

Acknowledgement

I want to thank Stefan P. Schmid for his support and discussions while preparing this manuscript. The GenMol model weights used in the web application are made available and licensed by the NVIDIA Corporation under the NVIDIA Open Model License.

Received: March 31, 2026

- [1] J. Hughes, S. Rees, S. Kalindjian, K. Philpott, *British J. Pharmacology* **2011**, *162*, 1239, <https://doi.org/10.1111/j.1476-5381.2010.01127.x>.
- [2] A. Togni, *CHIMIA* **2023**, *77*, 353, <https://doi.org/10.2533/chimia.2023.353>.
- [3] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360, <https://doi.org/10.1126/science.aat2663>.

- [4] P. Dhariwal, A. Nichol, *NeurIPS*, **2021**, <https://doi.org/10.48550/arXiv.2105.05233>. For an accessible derivation, see S. Dieleman, 2022, <https://benanne.github.io/2022/05/26/guidance.html>, accessed March 31, 2026.
- [5] M. Bohde, M. Manjrekar, R. Wang, S. Ji, C. W. Coley, *ICML*, **2025**, <https://doi.org/10.48550/arXiv.2502.09571>.
- [6] J. Ho, A. N. Jain, P. Abbeel, *NeurIPS*, **2020**, <https://doi.org/10.48550/arXiv.2006.11239>.
- [7] M. Lederbauer, *CHIMIA*, **2025**, *79*, 174, <https://doi.org/10.2533/chimia.2025.174>.
- [8] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, R. van den Berg, *NeurIPS*, **2021**, <https://doi.org/10.48550/arXiv.2107.03006>.
- [9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, I. Polosukhin, *NeurIPS*, **2017**, <https://doi.org/10.48550/arXiv.1706.03762>.
- [10] Image credits: <https://www.flickr.com/photos/petsadviser-pix/16601213761>, accessed March 31, 2026.
- [11] S. Lee, K. Kreis, S. Prasad Veccham, M. Liu, D. Reidenbach, Y. Peng, S. Paliwal, W. Nie, A. Vahdat, *ICML* **2025**, <https://doi.org/10.48550/arXiv.2501.06158>.
- [12] E. Noutahi, C. Gabellini, M. Craig, J. S. C. Lim, P. Tossou, *Dig. Disc.* **2024**, *3*, 796, <https://doi.org/10.1039/D4DD00019F>.
- [13] M. Awale, R. Van Deursen, J.-L. Reymond, *J. Chem. Inf. Model.* **2013**, *53*, 509, <https://doi.org/10.1021/ci300513m>.
- [14] A. Nigam, R. Pollice, A. Aspuru-Guzik, *Dig. Disc.* **2022**, *1*, 390, <https://doi.org/10.1039/D2DD00003B>.
- [15] Y. Ivanenkov, B. Zagrebnyy, A. Malyshev, S. Evteev, V. Terentiev, P. Kamya, D. Bezrukov, A. Aliper, F. Ren, A. Zhavoronkov, *ACS Med. Chem. Lett.* **2024**, *14*, 901, <https://doi.org/10.1021/acsmchemlett.3c00041>.