Chemical language models for de novo drug design: challenges and opportunities

Francesca Grisoni[1,2]*

[1]Eindhoven University of Technology, Institute for Complex Molecular Systems and Dept. Biomedical Engineering, Eindhoven, Netherlands.
[2]Centre for Living Technologies, Alliance TU/e, WUR, UU, UMC Utrecht, Netherlands.
*Email to: f.grisoni@tue.nl

Abstract

Generative deep learning is accelerating de novo drug design, by allowing the construction of molecules with desired properties on demand. Chemical language models – which generate new molecules in the form of strings – have been particularly successful in this endeavour. Thanks to advances in natural language processing methods and interdisciplinary collaborations, chemical language models are expected to become increasingly relevant in drug discovery. This minireview provides an overview of the current state-of-the-art of chemical language models for de novo design, and analyses current limitations, challenges, and advantages. Finally, a perspective on future opportunities is provided.

Introduction

Chemical biology is populated with linguistics analogies [1]: the genetic code is transcribed and translated, and cells communicate with each other by sending and receiving signals. Molecules can be considered as the elements of a 'chemical language' [1]. Like human language, chemical language possesses a *syntax*: finite elements (atoms, like words) can relate (bind) to one another only in specific ways to form 'chemically valid' molecules (Fig. 1a). Molecules also have *semantical properties*: based on which elements are present and how they are connected, different high-level properties (*e.g.*, physicochemical, biological) will emerge (Fig. 1b).

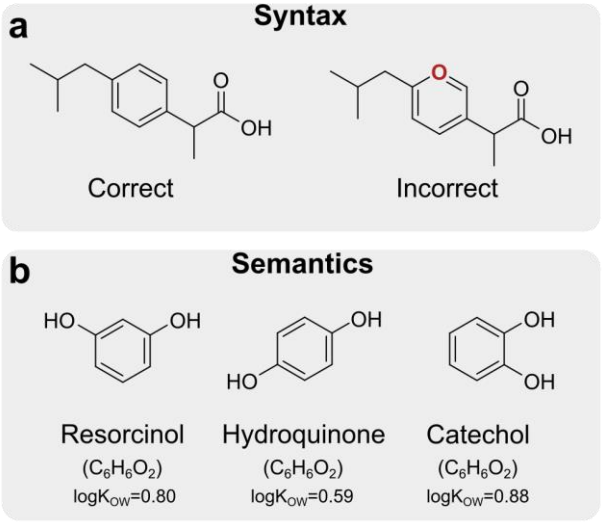


Figure 1. The 'chemical language'. (a) *Syntax*. Only certain combinations of atoms and bonds will lead to 'chemically valid' molecules. (b) *Semantics*, molecular properties emerging depending on what atoms are present and how they are connected to each other. Depicted, three molecules with the same chemical formula, but different semantical properties: resorcinol, an antiseptic and disinfectant, hydroquinone, a skin lightening agent, and catechol, a toxic molecule.

Learning the chemical language is crucial for de novo drug design [2], which addresses the difficult question of how to generate molecules from scratch that are chemically valid (*syntax*) and possess desired pharmacological properties (*semantics*). De novo design is confronted with an extremely vast 'chemical universe', estimated to contain up to 10^{60} drug-like molecular entities one could synthesize [3], which makes extensive enumeration practically impossible. De novo design has highly benefitted from the recent artificial intelligence (AI) *renaissance,* in the form of deep learning [4]. 'Generative' deep learning allows for generating 'raw' representations of molecules (*e.g.*, molecular graphs), and circumvents the need for molecule assembly and construction rules of conventional design algorithms [5].

Among the many flavours of deep learning for drug discovery, chemical language models (CLMs) [6], [7] have spearheaded AI-driven de novo design (Fig. 2). CLMs borrow and adapt algorithms developed for natural language processing to learning the chemical language. This is allowed by the usage of string notations, such as Simplified Molecular Input Line Entry Systems (SMILES [8]) strings (Fig. 2).

In the last few years, CLMs have been successful in designing experimentally-validated bioactive molecules [7], [9]–[11], and are providing increasing evidence of their capacities to explore the uncharted biochemical matter. This mini-review will focus on CLMs for de novo molecule design, although many other exciting applications have been reported [12]. After discussing the state-of-the-art of CLM-driven de novo design, this paper describes current gaps and future opportunities in the field of drug discovery.

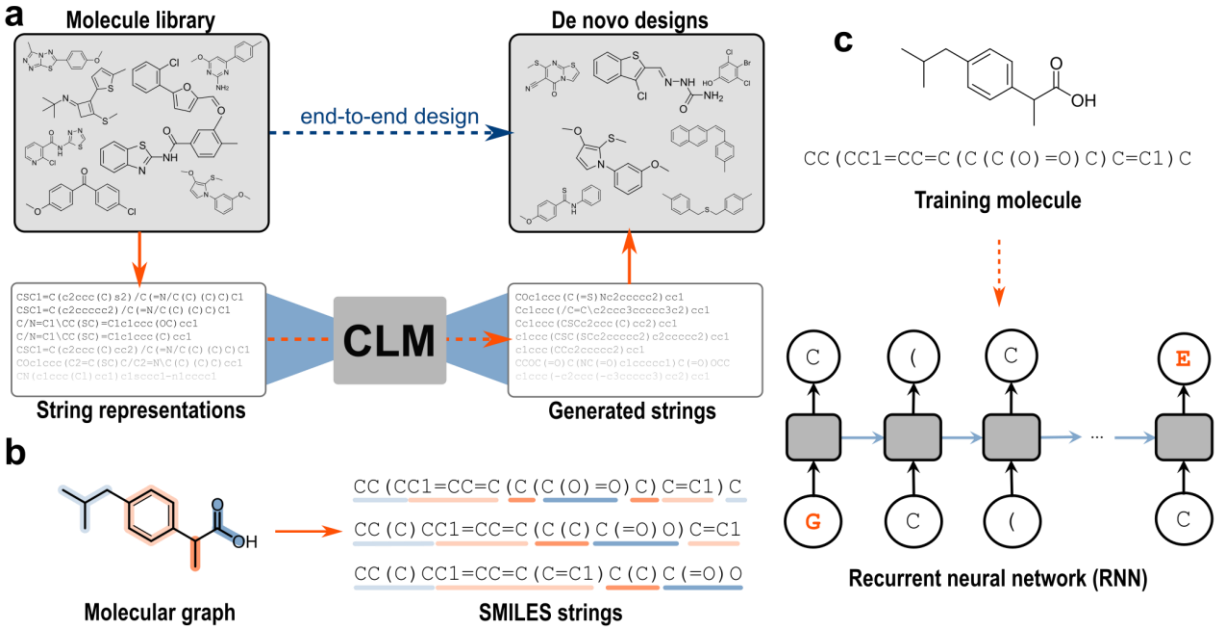


Figure 2. Overview of chemical language models (CLMs) for de novo drug design. (a) Starting from a string representation of molecules (*e.g.*, Simplified Molecular Input Line Entry System [SMILES] string), CLMs can be used to generate novel molecules possessing desired properties on-demand, in an 'end-to-end' fashion. (b) Overview of the SMILES algorithm, where atoms are indicated with their atomic letters, bonds and branching with symbols and ring opening/closure with numbers. Colours indicate the correspondence between substructures and elements of the string. Multiple SMILES for the same molecule can be obtained. (c) Example CLM in the form of recurrent neural networks (RNNs), which are trained to predict (and generate) the next character in a molecular string. 'G' and 'E' indicate the start and the end of the sequence, respectively. Once trained, the CLM can generate new strings starting from the 'go' ('G') character.

State of the art

Molecular string representations: advantages and drawbacks.

Molecular string representations were originally developed for database storage and molecule identification [13], but they have found a *renaissance* thanks to deep learning algorithms for sequence processing [12]. The most popular molecular string representations for de novo design are the following (Table 1):

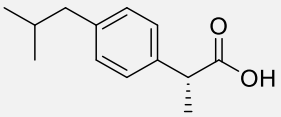
- *Simplified Molecular Input Line Entry Systems (SMILES)* [8]. SMILES strings are obtained by converting H-depleted molecular graphs into a string where atoms are indicated with their atomic letters, bonds, and branching with symbols and ring opening/closure with numbers. SMILES are non-univocal, as they can be obtained from any non-H atom by traversing the molecular graph in any chosen direction (Fig. 2b). To obtain a univocal SMILES string, canonicalization algorithms are necessary [14]–[16]. Several studies have shown the beneficial effect of using multiple SMILES for the same molecule [17]–[19] to artificially inflate the number of samples used for CLM training ('data augmentation').
- *DeepSMILES* [20] were proposed as an improvement of SMILES, to address unbalanced parentheses and ring closure pairs which cause invalid syntax. DeepSMILES have been applied to predict drug-target binding affinity [21], but their difficult syntax limits molecule generation compared to SMILES strings [19].
- *Self-referencing embedded strings (SELFIES)* [22] are built upon 'semantically constrained graphs' so that each symbol in the string can be used to convert it into a unique graph. Unlike SMILES, every SELFIES string corresponds to a valid chemical graph.

Each representation can be thought of as a different 'chemical language', characterized by its own syntactic rules to be preserved to generate chemically valid molecular entities. SELFIES bypasses the need to learn the chemical syntax (as these strings always correspond to valid molecules), which has been described as an advantage of this language [22], [23]. A recent study suggested that learning the syntax of SMILES strings allows invalid molecules to be filtered out, and to retain de novo designs matching the target chemical space better than SELFIES [24]. This agrees with findings from the natural language processing domain [25], highlighting the benefits of syntax learning to achieve better semantical properties.

All in all, the superior performance of either SMILES, SELFIES, or DeepSMILES seems to depend on the chosen application, with small differences in general [24], [26]–[29]. InChI notations (describing chemical substances via layers of information separated by "/", Table 1) were also used in combination with CLMs, but performed substantially worse than SMILES, due to a more complex syntax that includes counting and arithmetic [30].

Molecular strings constitute an ideal representation for molecule generation, due to the ease of producing text compared to more complex entities like graphs [29]. However, linear notations also possess certain drawbacks, since atoms that are close in the molecular graph might be located distant from each other in the corresponding string (*e.g.,* due to the presence of rings and branches). This might be the reason why bidirectional learning strategies [31] and infusion of 'linguistic knowledge' [32], [33] have been shown to improve CLMs.

Table 1. Examples of molecular strings for ibuprofen.

Type	Representation
Structure (2D molecular graph)	
IUPAC Name	2-[4-(2-methylpropyl)phenyl]propanoic acid
SMILES	CC(C)Cc1ccc(cc1)[C@@H](C)C(=O)O
DeepSMILES	CCC)Ccccccc6))[C@@H]C)C=O)O
SELFIES	[C][C][Branch1][C][C][C][C][=C][C][=C][Branch1][Branch1][C][=C][Ring1][=Branch1][C@@H1][Branch1][C][C][C][=Branch1][C][=O][O]
InChI	1S/C13H18O2/c1-9(2)8-11-4-6-12(7-5-11)10(3)13(14)15/h4-7,9-10H,8H2,1-3H3,(H,14,15)
InChIKey	HEFNNWSXXWATRW-UHFFFAOYSA-N

De novo design with chemical language models.

Many 'flavours' of deep learning have been used for chemical language modelling [12], [34], and recurrent neural networks (RNNs) with memory cells [35], [36] have found widespread usage [29], [37], [38]. RNNs are often trained to generate one character at a time, based on the preceding portions of the molecular string (Fig. 2c). In this way, they can become generative tools to produce molecular strings de novo. Other popular CLM architectures are (a) variational autoencoders (VAE) [30], constituted by an encoder that converts molecular strings to latent vectors and a decoder that converts latent vectors back to molecular strings, and (b) generative adversarial networks (GANs) [39], constituted by a generator network that produces novel molecular strings and a discriminator network aiming to distinguish between the generated molecules and existing molecules. Alternative deep learning approaches have been proposed for de novo design (*e.g.*, molecular graph generation [40, p.], [41] or fragment-based assembly [42]), but they have not been shown to outperform CLMs [29], [37], [38].

CLMs can be grouped into three categories (Fig. 3):

- *Distribution-learning* [37], whereby new molecules are generated to populate the same chemical space of the training set. Distribution-learning algorithms are usually evaluated for their abilities to match the properties of the training set, *e.g.*, via Kullback-Leibler [43] (KL) divergence (*e.g.*, to compare the distribution of computed physicochemical properties), or the Fréchet ChemNet Distance [44] (FCD), which measures the similarity in terms of chemical and biological properties.
- *Goal-directed generation* [37], in which molecules are generated aiming to optimize one or more goals. In this paper, the frequent connotation of the term is considered, *i.e.*, models that leverage scoring functions to quantify the molecule conformity to the end goal. In particular, scoring functions are used to iteratively improve the generated molecules. This can be achieved via reinforcement learning, in which the actions taken by the model are steered towards promising solutions via a reward. Commonly used scoring functions for CLMs are the similarity to known active molecules, predicted bioactivity, and computed physicochemical properties (*e.g.*, [45], [46]).
- *Conditional generation*. Conditional molecule generation can be considered somewhat intermediate between goal-directed (via a scoring function) and distribution-learning algorithms. It tackles the task of generating new molecules satisfying designated properties, by learning a joint semantic space between experimentally determined properties and corresponding molecular structures. The desired set of properties can be used as an input 'prompt' for molecule generation. By forming latent representations capturing both desired properties (*e.g.*, a desired three-dimensional shape [47], gene-expression signature [48], and protein target [49], [50]) and corresponding molecular structure in an end-to-end fashion, these algorithms allow a goal-directed generation that bypasses the need of scoring-function engineering.

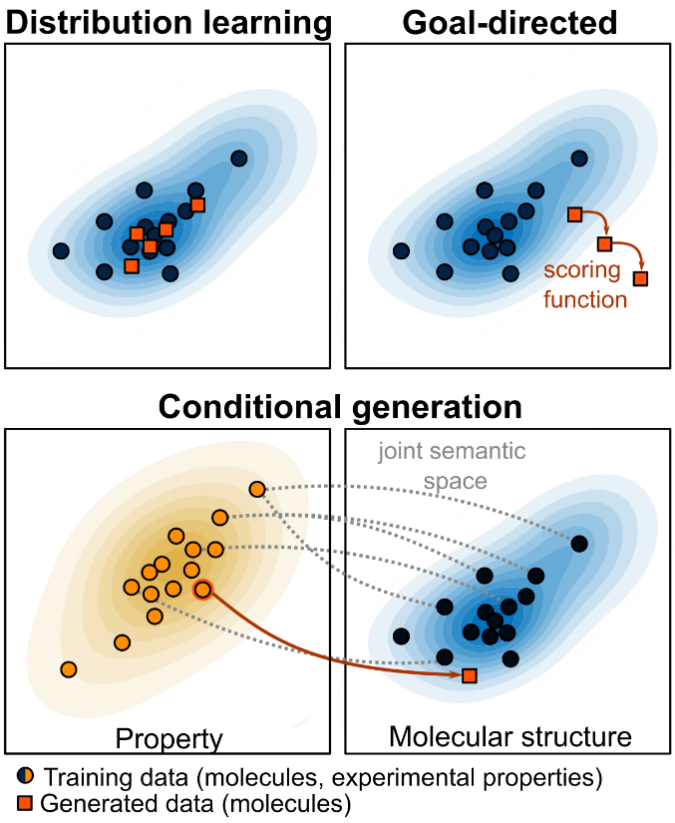


Figure 3. Categories of CLMs for molecule generation. *Distribution learning* algorithms aim to generate molecules that match the chemical distribution of the training set. *Goal-directed* algorithms generate the best possible molecule(s) to satisfy a predefined goal, specified via a scoring function. *Conditional generation* algorithms design molecules that are conditioned on desired properties, trained via a joined semantic space.

Each of these approaches comes with distinct advantages and limitations. Distribution-learning approaches allow end-to-end learning and generation, thereby not requiring scoring functions to steer the molecular design. Although models are evaluated by comparing the properties of the generated molecules with those of the training set molecules, no indication is provided of the quality of individual designs. This requires human-engineered post-hoc ranking and/or filtering procedures to narrow down the list to promising molecules, thereby partially reducing the advantages of these end-to-end pipelines.

On the other hand, goal-directed studies provide a direct indication of the quality of both the population and single molecules, via the scoring function. However, several studies have pointed out the challenges of goal-directed generation [51]–[53], due to (a) the difficulty in condensing complex chemical properties (*e.g.*, bioactivity, drug-likeness, and synthesizability) into single scoring functions, (b) model shortcuts, in which the generator exploits features unique to the scoring function it was optimized for, and (c) limited structural diversity due to biases induced by scoring functions. Here, the data and models used to develop scoring functions becomes essential to avoid failures [51]. Finally, the predicted synthesizability of the designs is in certain cases lower than for distribution-learning algorithms [54].

Conditional generation approaches have found relatively little application in drug discovery compared to distribution learning and goal-directed methods. By being devoid of 'externally computed' scoring functions, these approaches might (a) overcome the shortcomings of their discriminative counterparts (*e.g.*, bioactivity prediction models), (b) capture complex structure-objective relationships by forming latent associations between the desired properties and the corresponding structure(s), and (c) preserve the end-to-end learning character of distribution learning algorithms. Despite promising, these methods have found no experimental application to date. Finally, it is yet to be demonstrated how well they can explore regions of the chemical space that are not well represented in the training set (*e.g.*, out-of-distribution generation) to motivate their usage compared to simpler distribution learning or goal-directed generation algorithms.

Exploring uncharted regions in the chemical space with language models.

With the 'chemical universe' estimated to contain way more drug-like molecules than there are stars in the Milky Way ([3], [55]), we are in need of methods to efficiently chart 'dark' chemical matter. CLMs bear great promise to navigate the chemical space and explore sparsely populated regions ([6], [18], [24], [29]), thanks to their ability to produce in theory millions of molecules in one go, without requiring human-engineered rules.

Deep neural networks are notoriously 'data hungry', and drug discovery datasets are notoriously small (*e.g.*, in the order of ~10^1 to ~10^4 known molecules possessing the relevant biological activity). Transfer learning has found a widespread application to leverage small datasets for chemical space exploration. Transfer learning is a two-step procedure, aiming to transfer knowledge acquired by solving one task to another, related, task. In the first step ('pretraining'), CLMs are usually trained on a large set containing 10^5 to 10^6 molecules (*e.g.*, via next-character prediction, Fig. 2b). In the second step, the generic CLM is 'fine-tuned' using a smaller set of molecules possessing desired properties (*e.g.*, bioactivity on a certain pharmacological target). CLMs have shown promise to navigate the chemical space in low training data regimes [18], [24], *e.g.*, to design natural-product-inspired bioactive molecules [11] and learn multiple properties simultaneously [10], [29].

The efficiency of CLMs to navigate the chemical space depends on multiple factors simultaneously [18], [24], [29]. The minimum number of molecules required to train a robust model is linked to the complexity of the target molecules [24], [56]: the higher the complexity and the heterogeneity, the more data will be required. The structural diversity of training molecules will also reflect in the 'broadness' of the chemical space explored, *e.g.*, in terms of the structural diversity and molecular scaffolds of the designs [18]. Fine-tuning with 10-10^2 molecules has been shown to lead to experimentally-determined bioactive designs [9], [11], but might require more careful post-hoc filtering/ranking procedures to consider high-quality designs only [56]. SMILES augmentation increases the CLM performance [17]–[19], with a diminishing return when increasing the augmentation folds (*e.g.*, after 10- to 20-fold augmentation [18], [24]). SMILES augmentation is particularly beneficial with small training sets (less than 10,000 molecules [19]), while its effect plateaus for large datasets of structurally complex molecules (*e.g.*, more than 500,000 molecules), potentially with the risk of 'over-enumeration' and quality decrease [24]. The number of necessary epochs for fine-tuning also affects the 'semantical' quality of the designs, and depends on the dataset size and diversity [18], [56], while hyperparameter tuning seems to have little effect on the performance of CLMs overall [24].

These rules of thumb provide a good indication of what to expect from CLMs for chemical space exploration based on data availability and structural complexity. In general, evaluating CLMs and the quality of their designs is complex and more challenging than with predictive models [24], [37], [38], and often involves seemingly contradictory objectives (*e.g.*, maximizing the similarity to known bioactive molecules while achieving structural novelty [56]). Thus, a careful assessment of the designs is recommended on a case-by-case basis, by leveraging domain expertise and auxiliary computational tools (*e.g.*, pharmacophore models, molecular dynamics). The time- and cost requirements of chemical synthesis constitute a bottleneck to evaluating the quality of CLM designs on large scales. In the future, 'self-driving' labs will constitute a solution to swiftly explore the chemical space guided by CLMs.

Chemical language modelling: gaps and opportunities

In recent years, deep learning has taken drug discovery by storm, offering new opportunities to design new molecules de novo. With small molecules still being the "brick and mortar" [57] of the global pharmaceutical industry, chemical language models are here to stay. CLMs are providing increasing

evidence of their capacities to explore the uncharted biochemical matter, also thanks to the ease of generation of molecular strings and the flexibility of application to a multitude of tasks. Advances in language processing algorithms and the incorporation of medicinal chemistry expertise are expected to further propel the capabilities of CLMs in drug discovery.

CLMs are often evaluated for their capability to optimize 'toy' properties, *e.g.*, the calculated octanol-water partitioning coefficient, molecular weight, or the quantitative estimate of drug-likeness (QED [58]). These objectives capture the ability to generate molecules fulfilling predefined criteria, but fail to capture the complexity of real-world drug discovery and might lead to trivial solutions [52], [59]. Existing benchmarks for de novo design (*e.g.*, GuacaMol [37] and MOSES [38]) are a solution to ensure comparability between approaches developed independently, although not fully addressing the quality of the generated compounds [37]. Given the complexity of evaluating the goodness of de novo designs computationally, experimental validation constitutes the ultimate 'proof of the pudding'. Only a few prospective applications of CLMs have been published this far ([9]–[11], [60]), due to the complementary expertise required, and the time and cost investment. Interdisciplinary collaborations between deep learning practitioners, cheminformaticians, and medicinal chemists will be the key to bringing CLMs into real-world deployment. Automated synthesis platforms might constitute a solution to accelerate de novo design driven by CLMs [10], despite potentially limiting the chemical space accessible for synthesis.

Conditional generation algorithms are expected to increase in relevance in the years to come. These methods might overcome limitations of existing scoring functions (*e.g.*, which struggle in the presence of activity cliffs or non-additivity [61], [62]) and allow generating molecules matching certain criteria by design. Among them, structure-based design bears particular promise, by generating molecules matching electrostatic and shape features of certain binding pockets, and potentially start addressing de novo design for unexplored macromolecular targets. Structure-based de novo design has found an underwhelming prospective application, potentially due to limitations and bias in existing protein-ligand affinity datasets [63]. Achieving a fine-grained control on multiple properties of de novo designs bears great potential for uncharted applications, such as polypharmacology or selectivity.

'Few-shot' learning approaches combined with large-scale pretrained chemical language models [64] are expected to further boost prospective applications of CLMs. Moreover, improving the ability of CLMs to propose synthesizable molecules is expected to increase their practical relevance for drug discovery [54]. Extending chemical languages to more complex molecular entities also bears great promise to advance the potential of generative deep learning in chemistry, *e.g.*, for proteins and peptides containing non-natural amino-acids, crystals, and supramolecular chemistry. Future extensions of SELFIES to address challenges of current molecular string representations have been thoroughly discussed recently [23] and they might inspire variants of SMILES and DeepSMILES, too.

Deep learning models like CLMs and beyond are expected to have an increasingly relevant role in drug discovery. Besides improving time- and cost-efficiency, deep learning will accelerate our capacity to explore uncharted regions in the chemical space, as well as to formulate and verify exciting new scientific hypotheses for drug discovery. In the future, joined forces among AI experts, chemists, and biologists will allow designing innovative algorithms imbued with scientific knowledge and gathering of new scientific insights into human biology driven by AI.

Acknowledgments

The Institute for Complex Molecular Systems (ICMS, TU/e) and the Centre for Living Technologies (Alliance TU/e, WUR, UU, UMC Utrecht) are acknowledged for support. I thank Michael Moret and Rıza Özçelik for valuable discussions on chemical language models.

References

[1] P. Bralley, 'An Introduction to Molecular Linguistics', *BioScience*, vol. 46, no. 2, pp. 146–153, 1996, doi: 10.2307/1312817.

[2] D. C. Elton, Z. Boukouvalas, M. D. Fuge, and P. W. Chung, 'Deep learning for molecular design—a review of the state of the art', *Mol. Syst. Des. Eng.*, vol. 4, no. 4, pp. 828–849, 2019, doi: 10.1039/C9ME00039A.

[3] C. M. Dobson, 'Chemical space and biology', *Nature*, vol. 432, no. 7019, pp. 824–828, Dec. 2004, doi: 10.1038/nature03192.

[4] Y. LeCun, Y. Bengio, and G. Hinton, 'Deep learning', *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] G. Schneider and U. Fechner, 'Computer-based de novo design of drug-like molecules', *Nat. Rev. Drug Discov.*, vol. 4, no. 8, pp. 649–663, Aug. 2005, doi: 10.1038/nrd1799.

[6] M. H. Segler, T. Kogej, C. Tyrchan, and M. P. Waller, 'Generating focused molecule libraries for drug discovery with recurrent neural networks', *ACS Cent. Sci.*, vol. 4, no. 1, pp. 120–131, 2018.

[7] W. Yuan *et al.*, 'Chemical space mimicry for drug discovery', *J. Chem. Inf. Model.*, vol. 57, no. 4, pp. 875–882, 2017.

[8] D. Weininger, 'SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules', *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, 1988.

[9] D. Merk, L. Friedrich, F. Grisoni, and G. Schneider, 'De Novo Design of Bioactive Small Molecules by Artificial Intelligence', *Mol. Inform.*, vol. 37, no. 1–2, p. 1700153, Jan. 2018, doi: 10.1002/minf.201700153.

[10] F. Grisoni *et al.*, 'Combining generative artificial intelligence and on-chip synthesis for de novo drug design', *Sci. Adv.*, vol. 7, no. 24, p. eabg3338, 2021.

[11] M. Moret, M. Helmstädter, F. Grisoni, G. Schneider, and D. Merk, 'Beam Search for Automated Design and Scoring of Novel ROR Ligands with Machine Intelligence', *Angew. Chem. Int. Ed.*, vol. 60, no. 35, pp. 19477–19482, Aug. 2021, doi: 10.1002/anie.202104405.

[12] H. Öztürk, A. Özgür, P. Schwaller, T. Laino, and E. Ozkirimli, 'Exploring chemical space using natural language processing methodologies for drug discovery', *Drug Discov. Today*, vol. 25, no. 4, pp. 689–705, Apr. 2020, doi: 10.1016/j.drudis.2020.01.020.

[13] W. J. Wiswesser, 'Historic development of chemical notations', *J. Chem. Inf. Comput. Sci.*, vol. 25, no. 3, pp. 258–263, 1985.

[14] D. Weininger, A. Weininger, and J. L. Weininger, 'SMILES. 2. Algorithm for generation of unique SMILES notation', *J. Chem. Inf. Comput. Sci.*, vol. 29, no. 2, pp. 97–101, 1989.

[15] D. G. Krotko, 'Atomic ring invariant and Modified CANON extended connectivity algorithm for symmetry perception in molecular graphs and rigorous canonicalization of SMILES', *J. Cheminformatics*, vol. 12, no. 1, p. 48, Aug. 2020, doi: 10.1186/s13321-020-00453-4.

[16] N. M. O'Boyle, 'Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI', *J. Cheminformatics*, vol. 4, no. 1, p. 22, Sep. 2012, doi: 10.1186/1758-2946-4-22.

[17] E. J. Bjerrum, 'SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules', *ArXiv170307076 Cs*, May 2017, Accessed: Nov. 23, 2021. [Online]. Available: http://arxiv.org/abs/1703.07076

[18] M. Moret, L. Friedrich, F. Grisoni, D. Merk, and G. Schneider, 'Generative molecular design in low data regimes', *Nat. Mach. Intell.*, vol. 2, no. 3, pp. 171–180, 2020.

[19] J. Arús-Pous *et al.*, 'Randomized SMILES strings improve the quality of molecular generative models', *J. Cheminformatics*, vol. 11, no. 1, p. 71, Nov. 2019, doi: 10.1186/s13321-019-0393-0.

[20] N. O'Boyle and A. Dalke, 'DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures', *ChemRxiv*, Sep. 2018, doi: 10.26434/chemrxiv.7097960.v1.

[21] H. Öztürk, E. Ozkirimli, and A. Özgür, 'WideDTA: prediction of drug-target binding affinity'. arXiv, Feb. 04, 2019. doi: 10.48550/arXiv.1902.04166.

[22] M. Krenn, F. Häse, A. Nigam, P. Friederich, and A. Aspuru-Guzik, 'Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation', *Mach. Learn. Sci. Technol.*, vol. 1, no. 4, p. 045024, Nov. 2020, doi: 10.1088/2632-2153/aba947.

[23] M. Krenn *et al.*, 'SELFIES and the future of molecular string representations'. arXiv, Mar. 31, 2022. doi: 10.48550/arXiv.2204.00056.

[24] M. A. Skinnider, R. G. Stacey, D. S. Wishart, and L. J. Foster, 'Chemical language models enable navigation in sparsely populated chemical space', *Nat. Mach. Intell.*, vol. 3, no. 9, pp. 759–770, Sep. 2021, doi: 10.1038/s42256-021-00368-1.
[25] J. Russin, J. Jo, R. C. O'Reilly, and Y. Bengio, 'Compositional generalization in a deep seq2seq model by separating syntax and semantics'. arXiv, May 23, 2019. doi: 10.48550/arXiv.1904.09708.
[26] '[2010.09885] ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction'. https://arxiv.org/abs/2010.09885 (accessed Jul. 27, 2022).
[27] V. B. Siramshetty, D.-T. Nguyen, N. J. Martinez, N. T. Southall, A. Simeonov, and A. V. Zakharov, 'Critical Assessment of Artificial Intelligence Methods for Prediction of hERG Channel Inhibition in the "Big Data" Era', *J. Chem. Inf. Model.*, vol. 60, no. 12, pp. 6007–6019, Dec. 2020, doi: 10.1021/acs.jcim.0c00884.
[28] K. Rajan, A. Zielesny, and C. Steinbeck, 'DECIMER: towards deep learning for chemical image recognition', *J. Cheminformatics*, vol. 12, no. 1, p. 65, Oct. 2020, doi: 10.1186/s13321-020-00469-w.
[29] D. Flam-Shepherd, K. Zhu, and A. Aspuru-Guzik, 'Language models can learn complex molecular distributions', *Nat. Commun.*, vol. 13, no. 1, p. 3293, Jun. 2022, doi: 10.1038/s41467-022-30839-x.
[30] R. Gómez-Bombarelli *et al.*, 'Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules', *ACS Cent. Sci.*, vol. 4, no. 2, pp. 268–276, Feb. 2018, doi: 10.1021/acscentsci.7b00572.
[31] F. Grisoni, M. Moret, R. Lingwood, and G. Schneider, 'Bidirectional Molecule Generation with Recurrent Neural Networks', *J. Chem. Inf. Model.*, vol. 60, no. 3, pp. 1175–1183, Mar. 2020, doi: 10.1021/acs.jcim.9b00943.
[32] I. Lee and H. Nam, 'Infusing Linguistic Knowledge of SMILES into Chemical Language Models'. arXiv, Apr. 19, 2022. doi: 10.48550/arXiv.2205.00084.
[33] M. J. Kusner, B. Paige, and J. M. Hernández-Lobato, 'Grammar Variational Autoencoder'. arXiv, Mar. 06, 2017. doi: 10.48550/arXiv.1703.01925.
[34] X. Liu, A. P. IJzerman, and G. J. P. van Westen, 'Computational Approaches for De Novo Drug Design: Past, Present, and Future', in *Artificial Neural Networks*, H. Cartwright, Ed. New York, NY: Springer US, 2021, pp. 139–165. doi: 10.1007/978-1-0716-0826-5_6.
[35] S. Hochreiter and J. Schmidhuber, 'Long Short-Term Memory', *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
[36] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, 'On the Properties of Neural Machine Translation: Encoder-Decoder Approaches'. arXiv, Oct. 07, 2014. doi: 10.48550/arXiv.1409.1259.
[37] N. Brown, M. Fiscato, M. H. S. Segler, and A. C. Vaucher, 'GuacaMol: Benchmarking Models for de Novo Molecular Design', *J. Chem. Inf. Model.*, vol. 59, no. 3, pp. 1096–1108, Mar. 2019, doi: 10.1021/acs.jcim.8b00839.
[38] D. Polykovskiy *et al.*, 'Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models', *Front. Pharmacol.*, vol. 11, p. 1931, 2020, doi: 10.3389/fphar.2020.565644.
[39] G. L. Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P. L. C. Farias, and A. Aspuru-Guzik, 'Objective-Reinforced Generative Adversarial Networks (ORGAN) for Sequence Generation Models'. arXiv, Feb. 06, 2018. doi: 10.48550/arXiv.1705.10843.
[40] Y. Li, L. Zhang, and Z. Liu, 'Multi-objective de novo drug design with conditional graph generative model', *J. Cheminformatics*, vol. 10, no. 1, p. 33, Jul. 2018, doi: 10.1186/s13321-018-0287-6.
[41] Z. Zhou, S. Kearnes, L. Li, R. N. Zare, and P. Riley, 'Optimization of Molecules via Deep Reinforcement Learning', *Sci. Rep.*, vol. 9, no. 1, p. 10752, Jul. 2019, doi: 10.1038/s41598-019-47148-x.
[42] W. Jin, R. Barzilay, and T. Jaakkola, 'Junction Tree Variational Autoencoder for Molecular Graph Generation', in *International Conference on Machine Learning*, Jul. 2018, pp. 2323–2332. Accessed: Aug. 26, 2021. [Online]. Available: https://proceedings.mlr.press/v80/jin18a.html
[43] S. Kullback and R. A. Leibler, 'On Information and Sufficiency', *Ann. Math. Stat.*, vol. 22, no. 1, pp. 79–86, 1951.
[44] K. Preuer, P. Renz, T. Unterthiner, S. Hochreiter, and G. Klambauer, 'Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery', *J. Chem. Inf. Model.*, vol. 58, no. 9, pp. 1736–1741, Sep. 2018, doi: 10.1021/acs.jcim.8b00234.
[45] M. Olivecrona, T. Blaschke, O. Engkvist, and H. Chen, 'Molecular de-novo design through deep reinforcement learning', *J. Cheminformatics*, vol. 9, no. 1, p. 48, Sep. 2017, doi: 10.1186/s13321-017-0235-x.
[46] T. Blaschke *et al.*, 'REINVENT 2.0: An AI Tool for De Novo Drug Design', *J. Chem. Inf. Model.*, vol. 60, no. 12, pp. 5918–5922, Dec. 2020, doi: 10.1021/acs.jcim.0c00915.

[47] M. Skalic, J. Jiménez, D. Sabbadin, and G. De Fabritiis, 'Shape-Based Generative Modeling for de Novo Drug Design', *J. Chem. Inf. Model.*, vol. 59, no. 3, pp. 1205–1214, Mar. 2019, doi: 10.1021/acs.jcim.8b00706.
[48] O. Méndez-Lucio, B. Baillif, D.-A. Clevert, D. Rouquié, and J. Wichard, 'De novo generation of hit-like molecules from gene expression signatures using artificial intelligence', *Nat. Commun.*, vol. 11, no. 1, pp. 1–10, 2020.
[49] D. Grechishnikova, 'Transformer neural network for protein-specific de novo drug generation as a machine translation problem', *Sci. Rep.*, vol. 11, no. 1, p. 321, Jan. 2021, doi: 10.1038/s41598-020-79682-4.
[50] M. Skalic, D. Sabbadin, B. Sattarov, S. Sciabola, and G. De Fabritiis, 'From Target to Drug: Generative Modeling for the Multimodal Structure-Based Ligand Design', *Mol. Pharm.*, vol. 16, no. 10, pp. 4282–4291, Oct. 2019, doi: 10.1021/acs.molpharmaceut.9b00634.
[51] M. Langevin, R. Vuilleumier, and M. Bianciotto, 'Explaining and avoiding failure modes in goal-directed generation of small molecules', *J. Cheminformatics*, vol. 14, no. 1, p. 20, Apr. 2022, doi: 10.1186/s13321-022-00601-y.
[52] P. Renz, D. Van Rompaey, J. K. Wegner, S. Hochreiter, and G. Klambauer, 'On failure modes in molecule generation and optimization', *Artif. Intell.*, vol. 32–33, pp. 55–63, Dec. 2019, doi: 10.1016/j.ddtec.2020.09.003.
[53] 'Testing the Limits of SMILES-based De Novo Molecular Generation with Curriculum and Deep Reinforcement Learning | bioRxiv'. https://www.biorxiv.org/content/10.1101/2022.07.15.500218v1.abstract (accessed Oct. 12, 2022).
[54] W. Gao and C. W. Coley, 'The Synthesizability of Molecules Proposed by Generative Models', *J. Chem. Inf. Model.*, vol. 60, no. 12, pp. 5714–5723, Dec. 2020, doi: 10.1021/acs.jcim.0c00174.
[55] 'Counting the Stars in the Milky Way', *HuffPost*, Mar. 17, 2014. https://www.huffpost.com/entry/number-of-stars-in-the-milky-way_b_4976030 (accessed Oct. 12, 2022).
[56] S. Amabilino, P. Pogány, S. D. Pickett, and D. V. S. Green, 'Guidelines for Recurrent Neural Network Transfer Learning-Based Molecular Generation of Focused Libraries', *J. Chem. Inf. Model.*, vol. 60, no. 12, pp. 5699–5713, Dec. 2020, doi: 10.1021/acs.jcim.0c00343.
[57] 'Can Molecular Modeling Overcome The Limitations Of Drug Discovery AI?' https://www.drugdiscoveryonline.com/doc/can-molecular-modeling-overcome-the-limitations-of-drug-discovery-ai-0001 (accessed Jul. 27, 2022).
[58] G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan, and A. L. Hopkins, 'Quantifying the chemical beauty of drugs', *Nat. Chem.*, vol. 4, no. 2, pp. 90–98, Feb. 2012, doi: 10.1038/nchem.1243.
[59] J. Meyers, B. Fabian, and N. Brown, 'De novo molecular design and generative models', *Drug Discov. Today*, vol. 26, no. 11, pp. 2707–2715, Nov. 2021, doi: 10.1016/j.drudis.2021.05.019.
[60] X. Li, Y. Xu, H. Yao, and K. Lin, 'Chemical space exploration based on recurrent neural networks: applications in discovering kinase inhibitors', *J. Cheminformatics*, vol. 12, no. 1, p. 42, Jun. 2020, doi: 10.1186/s13321-020-00446-3.
[61] D. van Tilborg, A. Alenicheva, and F. Grisoni, 'Exposing the limitations of molecular machine learning with activity cliffs.', *J. Chem. Inf. Model.*, vol. 62, no. 23, pp. 5938–5951, Dec. 2022, doi: 10.1021/acs.jcim.2c01073.
[62] K. Kwapien, E. Nittinger, J. He, C. Margreitter, A. Voronov, and C. Tyrchan, 'Implications of Additivity and Nonadditivity for Machine Learning and Deep Learning Models in Drug Design', *ACS Omega*, vol. 7, no. 30, pp. 26573–26581, Aug. 2022, doi: 10.1021/acsomega.2c02738.
[63] M. Volkov *et al.*, 'On the Frustration to Predict Binding Affinities from Protein–Ligand Structures with Deep Neural Networks', *J. Med. Chem.*, vol. 65, no. 11, pp. 7946–7958, Jun. 2022, doi: 10.1021/acs.jmedchem.2c00487.
[64] H. Abdel-Aty and I. R. Gould, 'Large-Scale Distributed Training of Transformers for Chemical Fingerprinting', *J. Chem. Inf. Model.*, Oct. 2022, doi: 10.1021/acs.jcim.2c00715.