Deep learning for low-data drug discovery: hurdles and opportunities

Derek van Tilborg[1,2,+], Helena Brinkmann[1,+], Emanuele Criscuolo[1,+], Luke Rossen[1,+], Rıza Özçelik[1,2,+], Francesca Grisoni[1,2,*]

[1]Institute for Complex Molecular Systems (ICMS), Department of Biomedical Engineering, Eindhoven University of technology, PO Box 513, 5600 MB Eindhoven, The Netherlands.

[2]Centre for Living Technologies, Alliance TU/e, WUR, UU, UMC Utrecht, Princetonlaan 6, 3584 CB, Utrecht, The Netherlands.

[+]These authors contributed equally to the work.

*Corresponding author: f.grisoni@tue.nl

Abstract

Deep learning is becoming increasingly relevant in drug discovery, from *de novo* design to protein structure prediction and synthesis planning. However, it is often challenged by the small data regimes typical of certain drug discovery tasks. In such scenarios, deep learning approaches – which are notoriously 'data-hungry' – might fail to live up to their promise. Developing novel approaches to leverage the power of deep learning in low-data scenarios is sparking great attention, and future developments are expected to propel the field further. This minireview provides an overview of recent low-data-learning approaches in drug discovery, analyzing their hurdles and advantages. Finally, we venture to provide a forecast of future research directions in low-data learning for drug discovery.

Introduction

In recent years, artificial intelligence in the form of deep learning has permeated the molecular sciences. Deep learning – based on artificial neural networks with multiple processing layers [1] – has demonstrated remarkable potential in numerous applications, such as protein structure prediction [2] and organic reaction planning [3]. Deep learning has three main advantages: (a) it can learn complex and highly non-linear patterns from data [1], (b) it can be trained on a wide range of molecular representations (*e.g.*, SMILES strings [4] and graphs, Fig. 1b), and (c) it can be adapted to various types of training regimes, facilitating the development of tailored models for diverse applications. These aspects open novel modeling avenues compared to using human-engineered features only [5]. Yet, its transformative potential has primarily been harnessed in data-rich settings where extensive datasets are readily available (*e.g.*, [2], [3]). This relates to the fact that deep learning approaches optimize millions (or even billions) of neural-network parameters, which is made more robust by training on large datasets. Drug discovery, on the contrary, is often a low-data endeavor.

Due to costs and time limitations, typical drug discovery datasets are comprised of only several hundreds of molecules – a number drastically smaller compared to other deep learning applications. Moreover, drug discovery datasets are often characterized by limited structural diversity, and insufficient 'negative data' (*e.g.*, inactive molecules) [6], which restricts the information accessible for learning. Finally, the need to represent molecules as 'computer-readable' formats inevitably leads to information loss (*e.g.*, about molecular systems dynamics), which might hamper the performance, for instance, with highly non-linear structure-activity landscapes [7], [8], [9], [10].

Despite these limitations, neural networks shine for their 'adaptable' nature; *e.g.*, to different types of inputs and modeling tasks [5], and to manipulation of internal model representations [11] – which unlocks novel avenues compared to traditional methods. Hence, deep learning for drug discovery bears incredible potential to extract relevant information from complex molecular systems and perform tasks that are not accessible via traditional computational approaches. Its application in low-data regimes faces unique challenges that demand innovative approaches.

Stemming from these observations, this minireview delves into existing deep learning approaches for drug

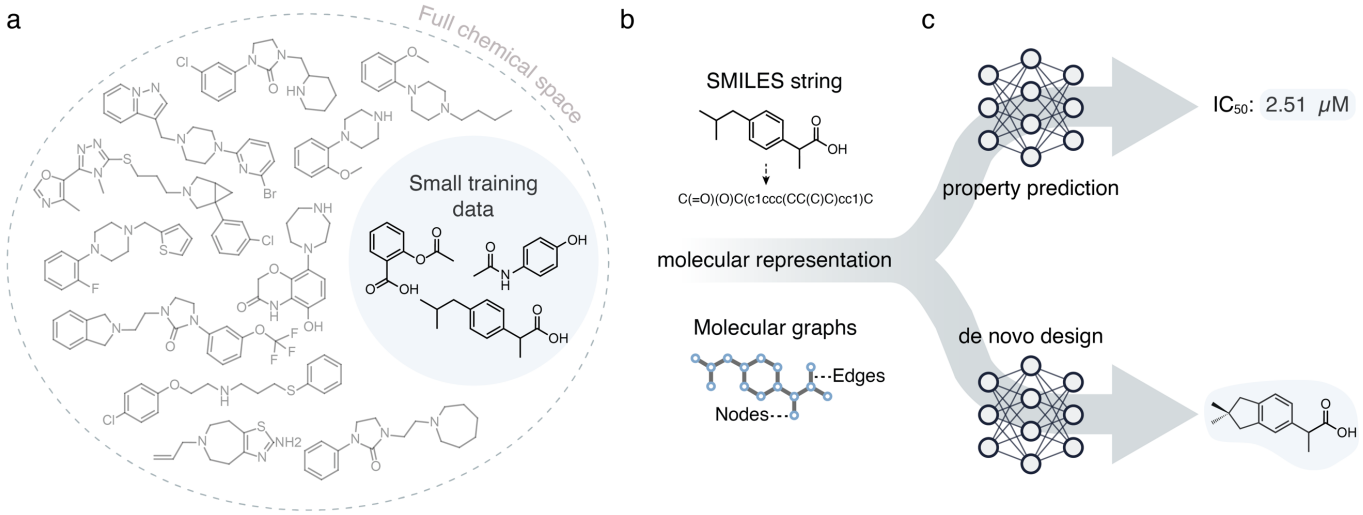

Figure 1. Simplified overview of deep learning for drug discovery. a. Molecular data relevant to the prediction task at hand is often limited in size, quality, and diversity, especially compared to the vastness of the full explorable 'chemical space'. **b**. Molecular representations are used to encode information on the molecular structure in a 'machine-learnable' format. Well-established examples are SMILES strings (where two-dimensional molecular information is converted into a string format) and molecular graphs (where nodes and edges represent atoms and bonds, respectively). **c.** Two key tasks in drug discovery are molecular property prediction (whereby a property like bioactivity is predicted from the molecular structure) and *de novo* design (whereby structures with desirable properties are generated from scratch).

discovery with limited data availability, with a focus on bioactivity prediction (*i.e.*, how to predict if and how a ligand will interact with one or more macromolecular targets) and *de novo* design (*i.e.,* how to design novel bioactive molecules from scratch, Fig. 1). While extensive reviews on how machine learning is employed for drug discovery exist [5], [12], this minireview provides a structured overview of deep learning for low-data drug discovery, with special emphasis on recent approaches, their advantages and limitations, and the opportunities ahead.

Strategies for deep learning in low-data scenarios

Small data scenarios in drug discovery can affect the performance of deep learning and call for tailored methods. In what follows, we discuss some approaches that have been successful in alleviating limitations of the low-data regimes typical of drug discovery. The approaches have been grouped into the following categories:

1. *Data augmentation*. The number of samples in the training data is artificially inflated by leveraging known properties of the chemical system investigated.
2. *Muti-stage training strategies*. Instead of performing a given task in 'one-go', these strategies rely on several training phases to steer the model towards the desired performance.
3. *Context-enrichment*. Extra information (context) is given to the model by providing different, additional inputs or by incorporating auxiliary prediction tasks.

For each of these approaches, the main advantages and disadvantages are discussed, and summarized in Table 1.

Data augmentation

The term 'data augmentation' refers to approaches that artificially inflate the data available for training. This is usually done by generating multiple (and different) instances of the same molecule to be used as input for a deep learning model (Fig. 2). Data augmentation can be applied to the entire training set, or selectively to mitigate the presence of imbalanced classes, *e.g.*, lack of negative data.

The most common data augmentation strategy has been applied to Simplified Input Molecular Line System (SMILES [4]) strings. SMILES strings are one of the most common ways to represent a molecular structure for deep learning. They encode two-dimensional aspects of the molecular structure (*i.e.*, atom connectivity and type, and bond type) and, optionally, stereochemistry in the form of a string, by traversing the molecular graph and annotating the encountered atoms and bonds with predefined characters (Fig. 2a). Any heavy (non-hydrogen) atom can be used as a starting point, and, therefore, one molecule can have multiple valid and different SMILES strings.

Table 1. Summary of the discussed approaches, with definitions, selected references, benefits, and limitations.

Strategy	Application	Definition	Benefits	Limitations
Data augmentation	Molecular string (*e.g.*, SMILES) augmentation [13]	Usage of multiple molecular strings for the same molecule, as input to the model.	Improvement in the quality and diversity of *de novo* designs [13], [14], and in QSAR model performance [15].	Few text-based augmentation approaches [16] have been explored to date. Decreasing gains when increasing the augmentation level [14], [17].
	Conformer augmentation [18]	Usage of multiple 3D conformations of the same molecule, as model input.	Potential to overcome the limitations of using a single, minimum-energy conformation [19].	Might perform worse than or on par with simpler methods [20], [21], and depend on the conformer generation method [20].
Multi-stage training	Transfer learning	Leveraging knowledge gained from one training task to improve the model performance on a different, but related, task with fewer data available.	Allows leveraging large training sets before focusing on smaller and task-specific datasets, often with increased performance.	The efficacy depends on the chosen pretraining set, transfer learning strategy, and molecular representation [22]. Possible undesirable 'pretraining bias' in the fine-tuning task, *e.g.*, for *de novo* design [23].
	Reinforcement learning	Steers the actions taken by a model towards promising solutions via a reward function.	A well-designed reward function can steer designs towards desired molecular properties, while requiring no training data.	Appropriate reward functions might be difficult to devise for complex problems, and models could exploit their weaknesses [24], [25].
	Active learning	Iterative train-predict-test cycles aimed to improve the model using (fewer) experimental datapoints.	Can overcome the lack of diversity and work in small-data regimes [26].	Might be time-inefficient and requires sustained collaboration between experimental and computational groups. Performance depends on many factors to be tuned on a case-by-case basis.
Context-enriched training	Multi-modal learning [27]	Leveraging multiple input types (*e.g.*, different molecular representations, or textual descriptions) to enhance the model performance on a given task.	Can be used to combine 'partial' molecular representations and their individual strengths.	It is not straightforward to choose how to combine modalities. Modality competition might arise [28].
	Multi-task learning [29]	Training a model to predict multiple outputs (*e.g.*, molecular properties of a given molecular input) simultaneously.	Can leverage information sharing between tasks and improve the performance with little labelled data [30].	It is complicated to tune the impact of each task. Requires the tasks to be related and does not necessarily outperform single-task models [30].

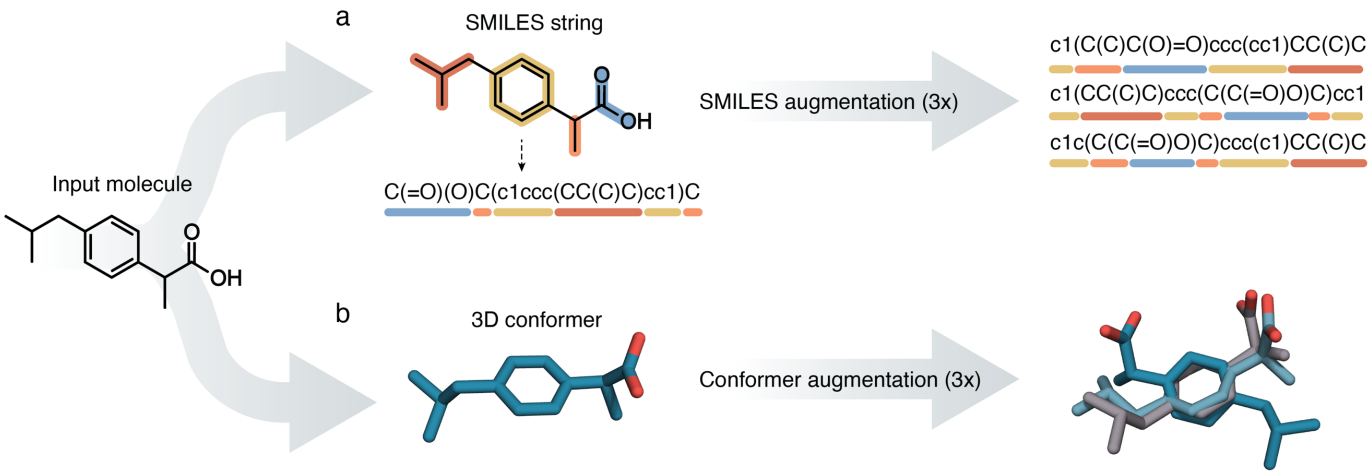

Figure 2. Selected data augmentation strategies for molecules. a. SMILES augmentation. The number of molecules available for training can be artificially inflated by using multiple different SMILES strings representing the same molecule. SMILES strings encode atom types and their connectivity into a character string in a non-univocal manner. **b.** Conformer-based augmentation. Three-dimensional (3D) molecular conformers can be augmented by using multiple conformations (*e.g.*, with different favorable energies) for training, instead of just the minimum energy conformer.

This characteristic of SMILES strings can be used for data augmentation, by representing the same molecule in the training set with *n* different SMILES strings, usually generated at random. SMILES augmentation [13] has been shown beneficial to improve the performance of quantitative structure-activity relationship (QSAR) models [15], as well as to improve the quality of *de novo* design models [13], with the magnitude depending on the structural complexity of the training molecules [17]. The performance improvement of SMILES augmentation plateaus with increasing the number of SMILES per molecule [14], [17] leading to progressively smaller performance gains with increasing computational cost.

Other molecular representations often used for deep learning, like molecular graphs or molecular descriptors are less suited to data augmentation. Molecular graphs, for instance, encode molecular topology (atoms and bonds) and properties (atom properties and/or coordinates, and bond properties) in a permutation-invariant manner. In other words, every molecule (given a set of atom and bond properties to capture) maps to a unique molecular graph, rendering a 'SMILES-like' augmentation impossible. Similarly, molecular descriptors map one-to-one with the molecular representation they are computed from, rendering permutation-based augmentation unfeasible. A useful, albeit less explored, approach to circumvent this limitation is by considering 3D conformations and performing 'conformer-based augmentation'. Here, the same molecule is described by several, different 3D conformations, which enables data augmentation with molecular graphs and/or molecular descriptors that capture 3D information [5]. How to aggregate information on different conformations is, however, still far from trivial, and several approaches can be explored [31]. To date, the benefits of using multiple conformations are not fully evident [20], [21], and the performance might be affected by the chosen conformer generator [20]. Finally, 3D-based approaches do not necessarily outperform more well-established ones that consider only 2D molecular information (*e.g.*, [10]).

Other molecule augmentation strategies exist, *e.g.*, by calculating molecular descriptors on molecular fragments obtained by pre-defined decomposition rules [32], or by adding noise (mask, swap, deletion, and fusion) to existing SMILES strings [33]. These strategies have, however, found limited application to date. Advances from the deep learning domain might help further boosting the performance in low data scenarios, *e.g.*, for SMILES augmentation [16].

Multi-staged training

While 'conventional' deep learning approaches rely on extensively labeled datasets to learn a given task in 'one-go', other training paradigms have been developed to address the challenges posed by limited data scenarios. These strategies will herein be referred to as 'multi-stage', since they iteratively adapt to the information contained in multiple datasets or tasks (usually utilized in a stepwise manner for training) to improve model performance. The three most common multi-staged strategies are (Fig. 3):

- Transfer Learning, which leverages knowledge gained from one task to improve the model performance on a different, but related, task (*e.g.*, [34], [35]) (Fig. 3a).

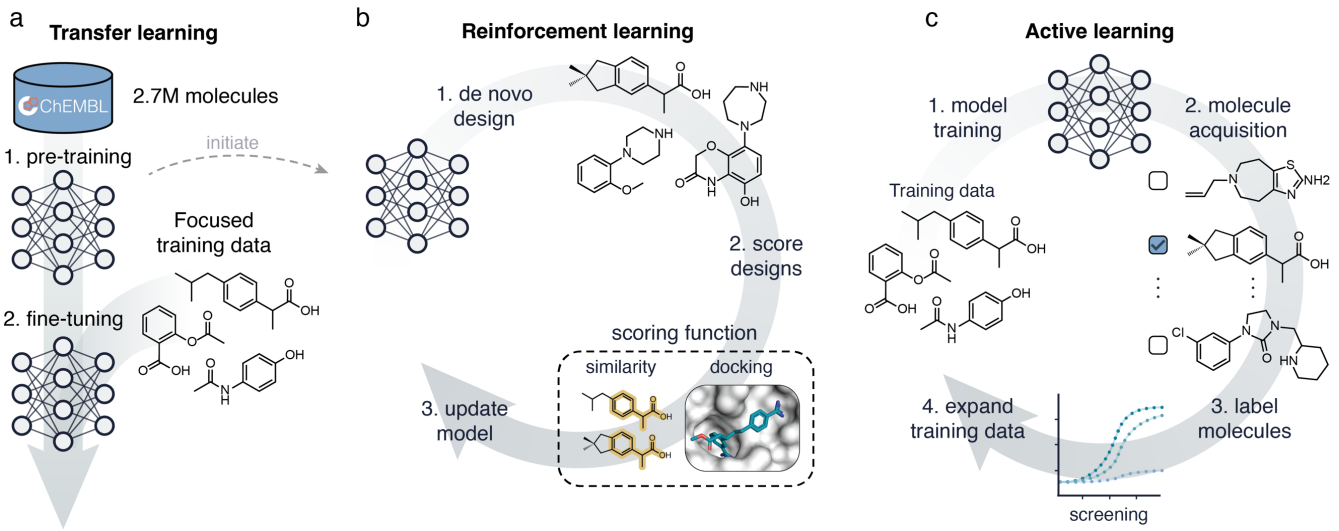

Figure 3. Adaptive training in low-data regimes. a. In transfer learning, a large dataset is used to pre-train a model, which is later fine-tuned (by additional training) on a focused training set for a different, but related, prediction task. **b.** In reinforcement learning, a (pretrained) *de novo* design model is rewarded for favorable designs through an external scoring function, e.g., similarity of the designs to known molecules, or docking scores. **c.** In active learning, a model is trained on a small initial set of molecules and chooses which new molecules from a screening library should be tested and added to the training set over iterations, with the goal of improving the model and reach one's goal faster.

In drug discovery, transfer learning is achieved by 'pre-training' a model on a large dataset (*e.g.*, ChEMBL [36] or ZINC [37]), and then 'fine-tuning' it (by additional training) on a smaller, and task-focused dataset (*e.g.*, bioactivity on a given macromolecular target [22]). The pre-training approach depends on the task to be performed and on the chosen molecular representation [38]. Pre-training enables bypassing the need of numerous labeled training data and leveraging large corpora of unlabeled molecules; this is particularly suited for molecular properties with little experimental annotations available. Transfer learning can improve model performance in small-data regimes, especially for *de novo* drug design (*e.g.*, [39]). However, undesired biases incurred during pre-training may persist after fine-tuning and affect the quality of the *de novo* designs [23]. Overall, the efficacy of transfer learning can depend on how related the pre-training and fine-tuning datasets are (*e.g.*, [22], [40], [41]), and on the chosen molecular representation and related training strategies (*e.g.*, [10]).

- Reinforcement Learning, whereby the actions taken by a model are steered towards promising solutions via a reward function (Fig. 3b). Reinforcement learning waives the need for a labeled starting dataset entirely, at the cost of requiring an 'oracle' (*e.g.*, a machine learning model predicting a specific property) that can accurately reward specific choices. In drug discovery, reinforcement learning has been used mostly for *de novo* design [42], where it consists of the following phases: (1) *de novo* molecule design using a molecule generator (*e.g.*, trained on ChEMBL [36] or ZINC [37]), (2) ranking of the designs via a scoring function (*e.g.*, docking [43], and structural similarity [44]), and (3) use of the top designs as new input to the model, to bias future generations towards desirable properties. However, reinforcement learning is faced with several challenges [24], [25], *e.g.*, related to (a) the difficulty of condensing (multiple) complex chemical properties into single scoring functions, or, (b) possible model shortcuts, where a 'loophole' in the scoring function is exploitatively capitalized upon (*e.g.*, learning to append a single carbon atom to trivially fulfill a novelty criterion). To avoid these failure modes, caution with the used data and reward functions becomes essential [24].
- Active learning, which selects molecules for screening over multiple iterations to expand the current dataset and, correspondingly, improves the model for the next screening round [45]. While many factors can be tuned when performing active learning, how molecules are selected for screening is a key aspect in hit discovery [26]. In low-data bioactivity prediction, active learning can remarkably outperform traditional machine learning approaches, leading to several fold improvement in hit retrieval [26]. In *de novo* design, active learning has (to the best of our knowledge) not been extensively explored yet. Nevertheless, active

learning requires both experimental and computational resources and expertise, which might increase the barriers for its adoption.

Context-enriched training

In this minireview, we define context-enriched training as an umbrella term encompassing approaches that provide additional knowledge to the model ('context') to improve its performance on a given task. Unlike multi-stage learning, in context-enriched training all available information is provided to the model simultaneously. Context-enrichment can be performed in several different ways, such as:

- *Multimodal training*. Multimodal training leverages multiple input types (*e.g.*, different molecular representations) to enhance the model performance on a given task [27]. Combining molecular information from different modalities might allow the model to learn more informative representations for the task at hand, potentially improving performance compared to models that rely on a single representation. In several application domains (*e.g.*, medical image analysis [46] and computer vision [47]) multimodal learning has shown promise to alleviate the limitations of scarce data. In the molecular sciences, it has been applied to several tasks (for instance, to combine molecular graphs and textual data [48], and ligand and proteins graphs [49]), showing promise for zero- and few-shot prediction [50]. Since each molecular representation captures only part of the underlying 'molecular reality', multimodal deep learning is particularly relevant in the molecular sciences. However, it also carries several limitations, *e.g.*, choosing how to effectively combine modalities, and 'modality competition' [28], whereby only a subset of the input modalities is leveraged by the model to make a prediction.
- *Multi-task learning*. In multi-task learning, a model is trained to predict multiple outputs in parallel (*e.g.*, multiple molecular properties [29]). The underlying idea is that the model error is optimized across all tasks (with the possibility to include missing values, if necessary), encouraging learning a shared representation that is beneficial for all tasks [51]. Although multi-task learning does not seem to systematically outperform single-task approaches for bioactivity prediction, it seems beneficial on tasks that have fewer labelled molecules [30]. Multi-task learning also comes with some caveats, for instance: (a) Its effectiveness relies on the assumption that the tasks are related, leading to failure when tasks that are too different from each other [51]; (b) the unavailability of all labels for each datapoint might affect the overall performance; and (c) an 'easy' tasks to model may become dominant during the training process, to the detriment of the more difficult ones. Hence, the complexity-performance trade-off of multi-task learning should be evaluated on a case-by-case basis.

Additional approaches have been used to provide additional context at training time, *e.g.*, by learning associations between a small set of molecules of interest with a larger set of (contextual) molecules [52], [53]. A particularly interesting strategy is meta-learning [54] – whereby the outputs of multiple machine learning algorithms are combined to predict a novel task, which is finding increasing application in drug discovery [55], [56], [57], *e.g.,* to predict bioactivity on novel binding assays and protein targets [57].

Challenges and future opportunities

Deep learning has enabled exciting new avenues in drug discovery. Despite the need for large training datasets being the 'Achilles heel' of deep learning in drug discovery, several advances allow neural networks to be, paradoxically, powerful tools in low-data scenarios. An increasing body of literature shows how strategies like the ones discussed in this minireview can lead to high-performing deep learning models, even with little data. However, many challenges are still lurking in the data shallows.

One of the central trials models face in a low-data setting is out-of-domain generalization. Since deep learning models are typically trained on a specific set of molecules (and the corresponding structure-activity relationships), they might be challenged in generalizing to new, unseen molecules that may come from a different distribution (*e.g.*, novel molecular scaffolds, structural motifs, or binding modes). Although this aspect is relevant for deep learning in general, low-data regimes intrinsically put additional strains on the model's ability to generalize out of the (limited) training distribution. Awareness of prediction uncertainty and out-of-distribution performance are expected to become crucial guides in the future for prospective decision making, especially in low-data scenarios. We also expect causal [58] and explainable deep learning [59] to become instrumental tools in low-data scenarios and for out-of-domain generalization, by shedding light on causal relationships, spurious correlations, and potential model shortcuts.

Geometric deep learning – which incorporates and processes symmetry information [60] – is also getting increasing attention in the molecular sciences, especially in the context of complex, three-dimensional molecular systems [5]. Incorporating symmetry

information, such as invariance or equivariance to roto-translations into neural network architectures, bears promise to learn sophisticated molecular information, which might be especially relevant when little training data is available. However, little is currently known on the performance of geometric deep learning in low-data scenarios, and the incorporation of molecular symmetry might not necessarily lead to better performing approaches [61].

Structure-based drug discovery also bears a great potential in the low-data setting [62]. These approaches, in fact, can leverage large corpora of protein-ligand affinity annotations, and can apply them to targets for which little (or no) ligand affinity information is available. Current structure-based approaches do not necessarily outperform approaches based on ligand information only [63], and hence we encourage the cheminformatics community to explore novel strategies to combine protein structure and ligand information with deep learning. Finally, multimodal learning [48], [49], [50] and meta-learning [55], [56], [57] strategies are getting increasing traction, and we expect them to become commonplace in drug discovery with low-data.

A current 'known unknown' in the field is the minimal data requirement for deep learning in drug discovery. Only a limited number of studies systematically examine the effect of dataset size and diversity on the model performance and out-of-domain generalization [17], [64], [65]. The same holds for knowledge on what deep learning strategy to choose based on the task and data at hand. In this context, FS-mol [41] – which provides the first-in-kind benchmark and set of baselines for low-data training – is a notable effort to further propel deep learning approaches in drug discovery. We expect the development of metrics and datasets tailored to low-data training to be key to harmonize the evaluation, choice, and development of novel approaches with an increased potential for drug discovery.

Author Contributions

Conceptualization: all authors. *Investigation:* all authors. *Visualization:* D.v.T., L.R., F.G., with contributions from all authors. *Writing – brainstorming*: all authors. *Writing – first draft:* H.B., D.v.T., and F.G. *Writing – reviewing:* all authors. All authors have given approval to the final version of the manuscript.

Acknowledgements

This study was co-funded by the European Union (ERC, ReMINDER, 101077879). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. The authors also acknowledge support from the Irene Curie Fellowship and the Centre for Living Technologies.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning", *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, doi: 10.1038/nature14539.

[2] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold", *Nature*, vol. 596, no. 7873, pp. 583–589, 2021, doi: 10.1038/s41586-021-03819-2.

[3] M. Segler, M. Preuss, and M. Waller, "Planning chemical syntheses with deep neural networks and symbolic AI", *Nature*, vol. 555, no. 7698, pp. 604–610, 2018, doi: 10.1038/nature25978.

[4] 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, doi: 10.1021/ci00057a005.

[5] K. Atz, F. Grisoni, and G. Schneider, "Geometric deep learning on molecular representations", *Nat. Mach. Intell.*, vol. 3, no. 12, pp. 1023–1032, 2021, doi: 10.1038/s42256-021-00418-8.

[6] R. Kurczab, S. Smusz, and A. Bojarski, "The influence of negative training set size on machine learning-based virtual screening", *J. Cheminformatics*, vol. 6, no. 1, p. 32, 2014, doi: 10.1186/1758-2946-6-32.

[7] F. Imrie, T. Hadfield, A. Bradley, and C. Deane, "Deep generative design with 3D pharmacophoric constraints", *Chem. Sci.*, vol. 12, no. 43, pp. 14577–14589, 2021, doi: 10.1039/D1SC02436A.

[8] L. David, A. Thakkar, R. Mercado, and O. Engkvist, "Molecular representations in AI-driven drug discovery: a review and practical guide", *J. Cheminformatics*, vol. 12, no. 1, p. 56, 2020, doi: 10.1186/s13321-020-00460-5.

[9] 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, 2022, doi: 10.1021/acsomega.2c02738.

[10] 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, 2022, doi: 10.1021/acs.jcim.2c01073.

[11] R. Gymez-Bombarelli *et al.*, "Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules", *ACS Cent. Sci.*, vol. 4, no. 2, pp. 268–276, 2018, doi: 10.1021/acscentsci.7b00572.

[12] B. Dou *et al.*, "Machine Learning Methods for Small Data Challenges in Molecular Science", *Chem. Rev.*, vol. 123, no. 13, pp. 8736–8780, 2023, doi: 10.1021/acs.chemrev.3c00189.

[13] E. Bjerrum, "SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules", arXiv, 2017, doi: 10.48550/arXiv.1703.07076.

[14] 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, doi: 10.1038/s42256-020-0160-y.

[15] J. Arús-Pous *et al.*, "Randomized SMILES strings improve the quality of molecular generative models", *J. Cheminformatics*, vol. 11, no. 1, p. 71, 2019, doi: 10.1186/s13321-019-0393-0.
[16] C. Shorten, T. M. Khoshgoftaar, and B. Furht, "Text Data Augmentation for Deep Learning", *J. Big Data*, vol. 8, no. 1, p. 101, 2021, doi: 10.1186/s40537-021-00492-0.
[17] M. Skinnider, R. Stacey, D. Wishart, and L. Foster, "Chemical language models enable navigation in sparsely populated chemical space", *Nat. Mach. Intell.*, vol. 3, no. 9, pp. 759–770, 2021, doi: 10.1038/s42256-021-00368-1.
[18] J. Hemmerich, E. Asilar, and G. Ecker, "COVER: conformational oversampling as data augmentation for molecules", *J. Cheminformatics*, vol. 12, no. 1, p. 18, 2020, doi: 10.1186/s13321-020-00420-z.
[19] M. Hechinger, K. Leonhard, and W. Marquardt, "What is Wrong with Quantitative Structure–Property Relations Models Based on Three-Dimensional Descriptors?", *J. Chem. Inf. Model.*, vol. 52, no. 8, pp. 1984–1993, 2012, doi: 10.1021/ci300246m.
[20] D. Zankov *et al.*, "QSAR Modeling Based on Conformation Ensembles Using a Multi-Instance Learning Approach", *J. Chem. Inf. Model.*, vol. 61, no. 10, pp. 4913–4923, 2021, doi: 10.1021/acs.jcim.1c00692.
[21] P. Kyaw Zin, A. Borrel, and D. Fourches, "Benchmarking 2D/3D/MD-QSAR Models for Imatinib Derivatives: How Far Can We Predict?", *J. Chem. Inf. Model.*, vol. 60, no. 7, pp. 3342–3360, 2020, doi: 10.1021/acs.jcim.0c00200.
[22] C. Cai *et al.*, "Transfer Learning for Drug Discovery", *J. Med. Chem.*, vol. 63, no. 16, pp. 8683–8694, 2020, doi: 10.1021/acs.jmedchem.9b02147.
[23] M. Moret, F. Grisoni, P. Katzberger, and G. Schneider, "Perplexity-Based Molecule Ranking and Bias Estimation of Chemical Language Models", *J. Chem. Inf. Model.*, vol. 62, no. 5, pp. 1199–1206, 2022, doi: 10.1021/acs.jcim.2c00079.
[24] 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, 2022, doi: 10.1186/s13321-022-00601-y.
[25] P. Renz, D. Van Rompaey, J. Wegner, S. Hochreiter, and G. Klambauer, "On failure modes in molecule generation and optimization", *Drug Discov. Today Technol.*, vol. 32–33, pp. 55–63, 2019, doi: 10.1016/j.ddtec.2020.09.003.
[26] D. van Tilborg and F. Grisoni, "Traversing Chemical Space with Active Deep Learning", ChemRxiv, 2023, doi: 10.26434/chemrxiv-2023-wgl32.
[27] J. Gao, P. Li, Z. Chen, and J. Zhang, "A Survey on Deep Learning for Multimodal Data Fusion", *Neural Comput.*, vol. 32, no. 5, pp. 829–864, 2020, doi: 10.1162/neco_a_01273.
[28] Y. Huang, J. Lin, C. Zhou, H. Yang, and L. Huang, "Modality Competition: What Makes Joint Training of Multi-modal Network Fail in Deep Learning? (Provably)", arXiv, 2022, doi: 10.48550/arXiv.2203.12221.
[29] S. Sosnin, M. Vashurina, M. Withnall, P. Karpov, M. Fedorov, and I. Tetko, "A Survey of Multi-task Learning Methods in Chemoinformatics", *Mol. Inform.*, vol. 38, no. 4, p. 1800108, 2019, doi: 10.1002/minf.201800108.
[30] C. Valsecchi, M. Collarile, F. Grisoni, R. Todeschini, D. Ballabio, and V. Consonni, "Predicting molecular activity on nuclear receptors by multitask neural networks", *J. Chemom.*, vol. 36, no. 2, p. e3325, 2022, doi: 10.1002/cem.3325.
[31] Y. Zhu *et al.*, "Learning Over Molecular Conformer Ensembles: Datasets and Benchmarks", arXiv, 2023, doi: 10.48550/arXiv.2310.00115.
[32] R. Magar *et al.*, "AugLiChem: data augmentation library of chemical structures for machine learning", *Mach. Learn. Sci. Technol.*, vol. 3, no. 4, p. 045015, 2022, doi: 10.1088/2632-2153/ac9c84.
[33] J. Jiang *et al.*, "NoiseMol: A noise-robusted data augmentation via perturbing noise for molecular property prediction", *J. Mol. Graph. Model.*, vol. 121, p. 108454, 2023, doi: 10.1016/j.jmgm.2023.108454.
[34] H. Altae-Tran, B. Ramsundar, A. Pappu, and V. Pande, "Low Data Drug Discovery with One-Shot Learning", *ACS Cent. Sci.*, vol. 3, no. 4, pp. 283–293, 2017, doi: 10.1021/acscentsci.6b00367.
[35] S. Honda, S. Shi, and H. Ueda, "SMILES Transformer: Pre-trained Molecular Fingerprint for Low Data Drug Discovery", arXiv, 2019, doi: 10.48550/arXiv.1911.04738.
[36] B. Zdrazil *et al.*, "The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods", *Nucleic Acids Res.*, p. gkad1004, 2023, doi: 10.1093/nar/gkad1004.
[37] J. Irwin and B. Shoichet, "ZINC -- A Free Database of Commercially Available Compounds for Virtual Screening", *J. Chem. Inf. Model.*, vol. 45, no. 1, pp. 177–182, 2005, doi: 10.1021/ci049714+.
[38] L. Yu, Y. Su, Y. Liu, and X. Zeng, "Review of unsupervised pretraining strategies for molecules representation", *Brief. Funct. Genomics*, vol. 20, no. 5, pp. 323–332, 2021, doi: 10.1093/bfgp/elab036.
[39] M. Ballarotto *et al.*, "*De Novo* Design of Nurr1 Agonists *via* Fragment-Augmented Generative Deep Learning in Low-Data Regime", *J. Med. Chem.*, vol. 66, no. 12, pp. 8170–8177, 2023, doi: 10.1021/acs.jmedchem.3c00485.
[40] D. Merk, F. Grisoni, L. Friedrich, and G. Schneider, "Tuning artificial intelligence on the de novo design of natural-product-inspired retinoid X receptor modulators", *Commun. Chem.*, vol. 1, no. 1, p. 68, 2018, doi: 10.1038/s42004-018-0068-1.
[41] M. Stanley *et al.*, "FS-Mol: A Few-Shot Learning Dataset of Molecules", 35th Conference on Neural Information Processing Systems (NeurIPS), 2021, https://openreview.net/forum?id=701FtuyLlAd.
[42] 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, 2017, doi: 10.1186/s13321-017-0235-x.
[43] J. Boitreaud, V. Mallet, C. Oliver, and J. Waldispühl, "OptiMol: Optimization of Binding Affinities in Chemical Space for Drug Discovery", *J. Chem. Inf. Model.*, vol. 60, no. 12, pp. 5658–5666, 2020, doi: 10.1021/acs.jcim.0c00833.
[44] 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, 2020, doi: 10.1021/acs.jcim.0c00915.
[45] D. Reker and G. Schneider, "Active-learning strategies in computer-assisted drug discovery", *Drug Discov. Today*, vol. 20, no. 4, pp. 458–465, 2015, doi: 10.1016/j.drudis.2014.12.004.

[46] F. Aydin, M. Zhang, M. Ananda-Rajah, and G. Haffari, "Medical Multimodal Classifiers Under Scarce Data Condition", arXiv, 2019, doi: 10.48550/arXiv.1902.08888.
[47] F. Pahde, M. Puscas, T. Klein, and M. Nabi, "Multimodal Prototypical Networks for Few-Shot Learning", Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2644–2653, 2021, https://openaccess.thecvf.com/content/WACV2021/html/Pahde_Multimodal_Prototypical_Networks_for_Few-Shot_Learning_WACV_2021_paper.html
[48] B. Su *et al.*, "A Molecular Multimodal Foundation Model Associating Molecule Graphs with Natural Language", arXiv, 2022, doi: 10.48550/arXiv.2209.05481.
[49] P. Wang *et al.*, "Structure-Aware Multimodal Deep Learning for Drug–Protein Interaction Prediction", *J. Chem. Inf. Model.*, vol. 62, no. 5, pp. 1308–1317, 2022, doi: 10.1021/acs.jcim.2c00060.
[50] P. Seidl, A. Vall, S. Hochreiter, and G. Klambauer, "Enhancing Activity Prediction Models in Drug Discovery with the Ability to Understand Human Language", arXiv, 2023, doi: 10.48550/arXiv.2303.03363.
[51] S. Wu, H. Zhang, and C. Ré, "Understanding and Improving Information Transfer in Multi-Task Learning", 2020, doi: 10.48550/ARXIV.2005.00944.
[52] C. Fifty, J. Leskovec, and S. Thrun, "In-Context Learning for Few-Shot Molecular Property Prediction", arXiv, 2023, doi: 10.48550/arXiv.2310.08863.
[53] J. Schimunek *et al.*, "Context-enriched molecule representations improve few-shot drug discovery", arXiv, 2023, doi: 10.48550/arXiv.2305.09481.
[54] J. Vanschoren, "Meta-learning", *Autom. Mach. Learn. Methods Syst. Chall.*, pp. 35–61, 2019, doi: 10.1007/978-3-030-05318-5_2.
[55] Q. Lv, G. Chen, Z. Yang, W. Zhong, and C. Chen, "Meta Learning With Graph Attention Networks for Low-Data Drug Discovery", *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–13, 2023, doi: 10.1109/TNNLS.2023.3250324.
[56] C. Nguyen, C. Kreatsoulas, and K. Branson, "Meta-Learning Initializations for Low-Resource Drug Discovery", ChemRxiv, 2020, doi: 10.26434/chemrxiv.11981622.v1.
[57] H. Yao, Y. Wei, L. Huang, D. Xue, J. Huang, and Z. Li, "Functionally Regionalized Knowledge Transfer for Low-resource Drug Discovery", in *Advances in Neural Information Processing Systems*, vol. 34, pp. 8256–8268, 2021, https://proceedings.neurips.cc/paper_files/paper/2021/hash/459a4ddcb586f24efd9395aa7662bc7c-bstract.html
[58] J. Berrevoets, K. Kacprzyk, Z. Qian, and M. van der Schaar, "Causal Deep Learning", arXiv, 2023, doi: 10.48550/arXiv.2303.02186.
[59] J. Jiménez-Luna, F. Grisoni, and G. Schneider, "Drug discovery with explainable artificial intelligence", *Nat. Mach. Intell.*, vol. 2, no. 10, pp. 573–584, 2020, doi: 10.1038/s42256-020-00236-4.
[60] M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric Deep Learning: Going beyond Euclidean data", *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 18–42, 2017, doi: 10.1109/MSP.2017.2693418.
[61] Y. Wang, A. Elhag, N. Jaitly, J. Susskind, and M. Bautista, "Generating Molecular Conformer Fields", arXiv, 2023, doi: 10.48550/ARXIV.2311.17932.
[62] R. Özçelik, D. van Tilborg, J. Jiménez-Luna, and F. Grisoni, "Structure-Based Drug Discovery with Deep Learning", *ChemBioChem*, vol. 24, no. 13, p. e202200776, 2023, doi: 10.1002/cbic.202200776.
[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, 2022, doi: 10.1021/acs.jmedchem.2c00487.
[64] J. Deng, Z. Yang, H. Wang, I. Ojima, D. Samaras, and F. Wang, "A systematic study of key elements underlying molecular property prediction", *Nat. Commun.*, vol. 14, no. 1, Art. no. 1, 2023, doi: 10.1038/s41467-023-41948-6.
[65] Y. Ji *et al.*, "DrugOOD: Out-of-Distribution (OOD) Dataset Curator and Benchmark for AI-aided Drug Discovery -- A Focus on Affinity Prediction Problems with Noise Annotations", arXiv, 2022, doi: 10.48550/ARXIV.2201.09637.