Traversing Chemical Space with Active Deep Learning: A Computational Framework for Low-data Drug Discovery

Derek van Tilborg[1,2] and 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.
*Corresponding author: f.grisoni@tue.nl.

Abstract

Deep learning is accelerating drug discovery. However, current approaches are often affected by limitations in the available data, *e.g.*, in terms of size or molecular diversity. Active deep learning has an untapped potential for low-data drug discovery, as it allows to improve a model iteratively during the screening process by acquiring new data, and to adjust its course along the way. However, several *known unknowns* exist when it comes to active learning: (a) what the best computational strategies are for chemical space exploration, (b) how active learning holds up to traditional, non-iterative, approaches, and (c) how it should be used in the low-data scenarios typical of drug discovery. These open questions currently limit the wider adoption of active learning in drug discovery. To provide answers, this study simulates a real-world low-data drug discovery scenario, and systematically analyses six active learning strategies combined with two deep learning architectures, on three large-scale molecular libraries. Not only do we show that active learning can achieve up to a six-fold improvement in hit discovery compared to traditional methods, but we also identify the most important determinants of its success in low-data regimes. This study lays the first-in-time foundations for the prospective use of active deep learning for low-data drug discovery and is expected to accelerate its adoption.

Keywords: *Active learning, Drug Discovery, Deep Learning, Low-data, Virtual Screening*

Introduction

Deep learning is showing increasing promise for drug discovery[1,2]. One of its key applications is virtual screening[3], whereby large commercial libraries (usually consisting of 10^3-10^9 molecular candidates) are prioritized for prospective wet-lab experiments[3–6]. A key bottleneck of deep learning, however, is the need for sufficient training data (preferably 10^3 molecules and above[7–9]). Unfortunately, available ligand-target interaction data are often limited in size and structural diversity – factors that might hamper the usefulness of deep learning models in practice[8,10,11]. Furthermore, the chemical composition of commercial screening libraries is often structurally distinct from the training data[12], resulting in potentially unreliable predictions.

One potential solution to escape the training size and diversity bottlenecks is active learning[13–15]. Active learning is based on the principle that a model can achieve greater accuracy with fewer training data if it is "*allowed to choose the data from which it learns*"[16]. In drug discovery, active learning can be cast into an iterative screening approach (Fig. 1), where instead of performing a single virtual screening experiment, one can test (fewer) molecules across multiple cycles[13–15,17–21]. At each iteration, a selection of molecules is acquired based on model predictions and experimentally tested (*e.g.*,

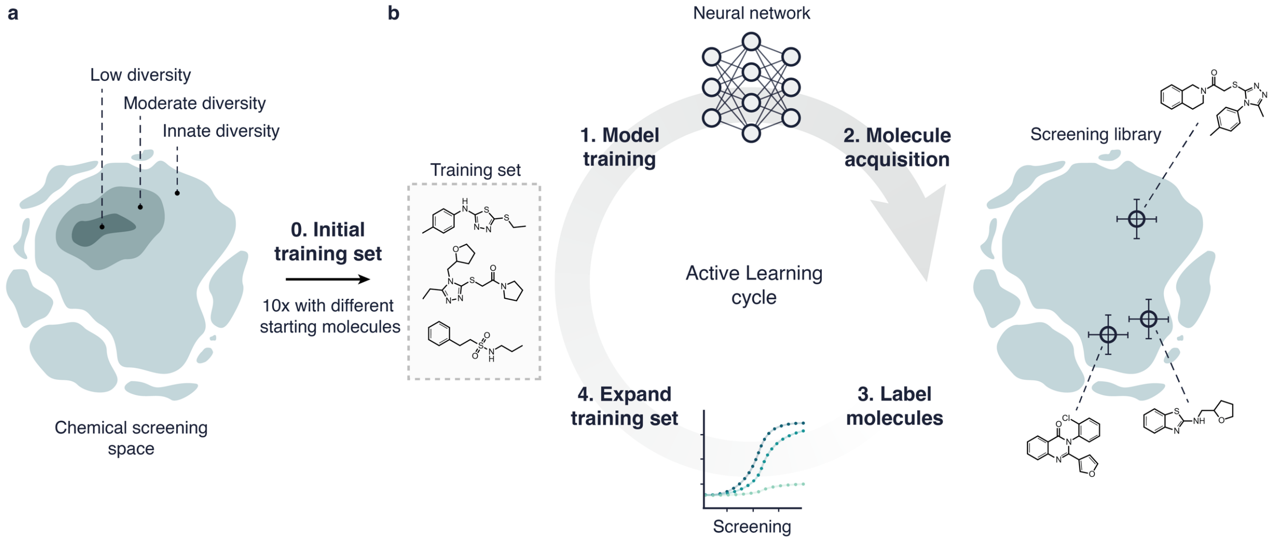

Figure 1: Overview of the experimental setup for low-data active learning. a. A starting set of molecules (64) is selected from the screening library (step 0). This starting set can be sampled from the full library (innate structural diversity) or can be sampled from hierarchical subclusters of molecules with moderate or low structural diversity. **b.** The active deep learning cycle. From the training set, a model is trained and used to predict bioactivity on all unlabelled molecules in the screening library (step 1). Using an acquisition function, molecules are selected for follow-up (step 2), and subsequent labelling (step 3). Finally, the labelled molecules are added to the training set for the next cycle (step 4).

for its biological activity on the intended target). The newly obtained experimental data, together with data previously collected, are then used to update the model, aiming to inform the next iteration. By improving a model over subsequent iterations, active learning bears promise to identify more bioactive molecules using less resources than standard 'one-shot' virtual screening approaches[13].

Although there have been several pioneering active learning studies in the molecular domain[13,22–33], numerous technical and practical aspects remain unaddressed. First, the hypothesised advantage in screening efficiency that active learning offers over traditional methods has not been quantified across different scenarios, and existing studies lack a focus on hit enrichment rather than performance variation. Secondly, existing studies do not address low-data settings that are commonplace in drug discovery, where little training data is leveraged to screen compound libraries many orders of magnitude larger (Supp. Table 1). Finally, how molecular diversity in the initial starting set affects the success of an active learning campaign has not been addressed. Even though active learning is expected to gain relevance in the future, *e.g.*, in the context of automated molecule discovery and self-driving labs[34,35], the current lack of computational guidelines on how to use it in practice might hamper its adoption.

This first-in-time systematic study on active deep learning in low-data scenarios aims to illuminate such *known unknowns* and lay the foundations for its application in realistic drug screening scenarios. We focus on several facets of the active learning pipeline, such as hit discovery of different deep learning architectures and acquisition functions, the effect of starting set size and molecular diversity, and the relative improvements of active learning compared to non-iterative approaches in low- and very-low data regimes. The proposed frame of reference comprises of recommendations and observations for the prospective use of active deep learning in low-data drug discovery, and it is expected to accelerate its adoption.

Results and Discussion

Active learning setup

In this study, we mimicked a realistic drug discovery scenario. All experiments followed the same setup (Fig. 1), whereby active learning

Table 1: Summary of the data. Subsets of three datasets from LIT-PCBA[36] were used: Pyruvate kinase M2 (PKM2, agonism), Aldehyde dehydrogenase 1 (ALDH1, inhibition), and Vitamin D receptor (VDR, antagonism). Mean and standard deviations of the number of hit molecules in the 10 starting sets are reported.

Dataset	Screening library size	Test set size	Hits in starting sets
PKM2	100,000 (223 hits, 0.2%)	20,000 (44 hits, 0.2%)	1.2 ± 0.4
ALDH1	100,000 (4,986 hits, 5.0%)	20,000 (997 hits, 5.0%)	4.1 ± 3.1
VDR	100,000 (239 hits, 0.2%)	20,000 (48 hits, 0.2%)	1.2 ± 0.4

models were used to iteratively query a screening library of 100,000 molecules with a maximum screening budget of 1,000 molecules. Each series of active learning experiments started with a small set of 64 molecules (step 0) sampled from the screening library, containing at least one bioactive molecule (*see* Materials and Methods). Notably, this procedure led to 80% of the starting sets containing less than three hits, which reflects a realistic drug discovery scenario. The active learning screening procedure then consisted of four iterative steps:

1. *Training*. A machine learning model was trained on all available training data. The trained model was then used to perform bioactivity predictions on all unlabelled molecules in the screening library.
2. *Acquisition*. From these predictions, 64 molecules were selected for follow-up using one out of six pre-determined acquisition functions.
3. *Testing*. The acquired molecules were labelled with their corresponding experimental bioactivity (known in advance but not used for model training), to simulate a wet-lab testing procedure.
4. *Update*. All tested molecules were added to the training set, for the next cycle (step 1).

This four-step cycle was repeated until 1,000 molecules were screened. The acquisition size (step 2) was determined through preliminary experiments where we compared acquiring 16, 32, or 64 molecules per cycle and found no significant differences in performance (see Supplementary Information). Using this setup, we investigated the impact of the following factors on the effectiveness of active learning: (a) the structural diversity and (c) the acquisition function used to (a) the structural diversity of the starting set (comparing three levels of diversity), (b) the chosen deep learning approach (two approaches), and (c) the acquisition function used to decide what molecules to select (six acquisition functions). These systematic analyses were each carried out on three macromolecular targets in replicates of ten with different random starting sets, to ensure the robustness and generalizability of the obtained conclusions. These factors are explained below.

Chosen macromolecular targets. To mimic real drug-screening experiments that rely on large screening libraries, we used three high-throughput screening datasets from LIT-PCBA[36], each for a different biological target. Not only are LIT-PCBA datasets designed to mimic the hit/potency distribution of typical experimental drug screens[36], but they are also large enough to be used as a 'screening library' to simulate prospective active learning campaigns. We selected the three LIT-PCBA datasets containing the most experimentally-validated molecules, which refer to targets of clinical and therapeutic interest[37–40], namely: (a) Pyruvate kinase M2 (PKM2, agonism), (b) Aldehyde dehydrogenase 1 (ALDH1, inhibition), and (c) Vitamin D receptor (VDR, antagonism). For each dataset, 100,000 molecules were randomly extracted, preserving the proportion between active and inactive molecules. These molecules served to construct a screening library, from which the starting training set and the successive molecule picks are drawn (Table 1 and Supp. Fig. 2). An additional set of 20,000 molecules was randomly selected to serve as an external test set for performance monitoring.

Levels of structural diversity. To examine the effect of structural heterogeneity in the starting set, we artificially created subsets of molecules with different degrees of structural diversity. Diversity was defined through molecular similarity by computing the Tanimoto coefficient on Extended Connectivity Fingerprints[41] (ECFPs), which captures the presence of shared substructures.

Using hierarchical clustering, we selected ten clusters in each dataset with moderate structural diversity (average Tanimoto similarity ranging from 0.26 ± 0.03 to 0.28 ± 0.03, Table 2). Within each of these clusters, another subcluster of molecules with low diversity was identified (average similarity higher than 0.36 ± 0.01, Table 2). This gave us a hierarchy of three levels of structural diversity (Fig. 1a): (1) *'innate'* diversity, representing the inherent diversity of the full screening library, (2) *moderate* diversity, and (3) *low* diversity. When constructing the starting set, molecules could be sampled from areas of each level accordingly. Since our approach is hierarchical, this allowed us to vary molecular diversity in the starting set while staying in the same population of molecules for each of the ten experimental replicates.

Table 2: Hierarchy of molecular diversity in starting sets. Molecular diversity is reported as the mean Tanimoto similarity on Extended Connectivity Fingerprints between all molecules in the starting set. For three datasets, the mean similarity of all ten starting sets (*n*=64) is reported with standard deviations. The higher the mean similarity, the lower the diversity. These values were the best achievable for ten hierarchical subsets large enough to sample the starting set from.

		Similarity ↑	
Diversity ↓	PKM2	ALDH1	VDR
Innate	0.14 ± 0.00	0.13 ± 0.00	0.14 ± 0.00
Moderate	0.27 ± 0.03	0.28 ± 0.04	0.26 ± 0.03
Low	0.46 ± 0.03	0.46 ± 0.02	0.36 ± 0.01

Deep learning models. Two deep learning strategies were used to perform bioactivity predictions. Models were trained using either traditional engineered molecular descriptors or learnable molecular representations[29]:

1. Neural networks (multi-layer perceptron) that learn from molecular fingerprints in the form of Extended Connectivity Fingerprints[41] (ECFPs). These molecular fingerprints encode the presence of radial, atom-centred substructures. This method performs comparably to other gold-standard machine learning approaches for molecular property prediction (*e.g.*, random forest with ECFPs)[8].
2. Graph neural networks, which learn directly from the molecular graph[4,42]. Molecular graphs are a direct numeric representation of molecular topology, with nodes and edges representing atoms and chemical bonds respectively.

Both approaches share the same multi-layer perceptron, which either learns directly from molecular fingerprints, or from the output of several layers of graph convolutions[43]. This enables a direct comparison between ECFPs and molecular graphs. Additionally, these methods enable robust uncertainty estimation through approximate Bayesian modelling. Both methods were implemented as such via anchored ensembling[44], which produces predictive posterior distributions that closely approximate exact Bayesian methods (*see* Materials and Methods).

Acquisition functions. An important contributor to hit discovery is the so-called acquisition function. The acquisition function determines which molecules are selected for screening and how the training set is expanded, governing hit retrieval and influencing future iterations. Six acquisition functions were investigated as a strategy to select molecules for the next cycle, and for their effect on the active learning performance:

1. *Molecular similarity*. Molecules in the screening library with the highest structural similarity (Tanimoto similarity on ECFPs) to any previously found hit are selected.
2. *Exploitation*. Molecules with the best model predictions are selected (*Eq.* 6).
3. *Exploration*. Molecules with the most uncertain predictions are selected (*Eq.* 7), with the goal of 'patching knowledge gaps' in the model.
4. *Mutual Information*[45]. Based on Bayesian Active Learning by Disagreement (BALD)[45], molecules with the lowest mutual information (*Eq.* 8) are selected from the screening library. Mutual information is low when there are many possible ways of predicting the data with high certainty, given the same model.

5. *Exploitation without retraining*. Molecules are selected with exploitation using a model trained on just the start dataset (*Eq.* 6). This method resembles traditional 'one-shot' virtual screening where a single model prioritizes the molecules for the whole screening experiment.
6. *Random acquisition*. Molecules are randomly selected from the screening library. This method serves as a control.

Evaluation of active deep learning. The factors investigated in this study were evaluated for their effect on hit enrichment across active learning cycles. This was quantified using the enrichment factor (EF)[46], which captures the ratio between the number of hits found among all acquired molecules and the number of hits expected in selecting the same number of molecules at random from the screening library. Enrichment factor values larger than 1 indicate methods that can enrich a selection of hits more than a random pick (the higher, the better), while enrichment factor values lower than 1 perform worse than random at hit retrieval.

Structural diversity: Nature *versus* nurture

The active learning process is 'seeded' by the starting data that are available on a given target of interest (and their structural diversity), with usually little or no control over what molecules are available for training. Here were tested the effect of the structural diversity of the starting set on subsequent active learning cycles. We compared starting sets with different degrees of structural diversity: low, moderate, and the innate diversity of the screening libraries.

We found that, in general, the starting molecular diversity (its *'nature'*) has little influence on the diversity of the molecules acquired later (Fig. 2) or hit retrieval (Supp. Fig. 3) – owed to active-learning selections (*'nurture'*). The initial structural bias is quickly compensated for in the first 1-5 cycles for most methods (corresponding to 64-320 acquired molecules, Fig. 2) and converges to the levels of innate diversity of the screening library (Table 2). Only for similarity-based acquisition, by definition, the structural bias of the starting set lingers. In a few cases (mainly for exploitative and mutual information-based acquisition functions) fewer hits seem to be found when starting with highly similar molecules, although this proved to not be statistically significant in most cases (Supp. Fig. 3). This behaviour is observed regardless of the chosen deep learning approach and datasets. These results indicate that the structural diversity of the starting set does not play a big role in hit identification when screening for several cycles, hence showing the usefulness of iterative train-test-update cycles in traversing chemical space effectively.

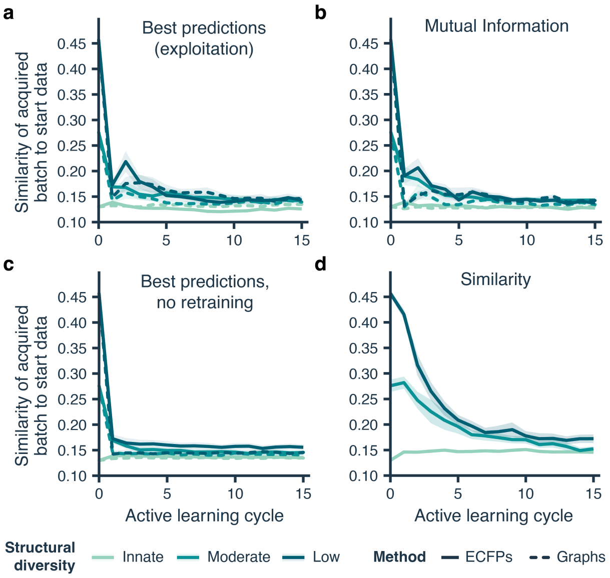

Figure 2: The effect of structural diversity in the starting data. Using the ALDH1 dataset, 64 molecules are acquired each cycle. Shaded areas represent the standard error of 10 experiments with different random starting molecules. Structural diversity is defined as mean Tanimoto similarity within the starting set (Table 2). **a**. Molecule acquisition based on best predictions (exploitation). **b**. Molecule acquisition based on mutual information. **c**. Molecule acquisition based on best predictions (exploitation) without updating models. **d**. Similarity-based molecule acquisition. Since molecules are not acquired based on model predictions, but through molecular similarity, results for ECFP- and graph-based models are identical.

Choosing an acquisition function

Choosing how to acquire the next iteration of molecules is arguably one of the most crucial factors in active learning, since it will determine hit allocation and what structure-activity

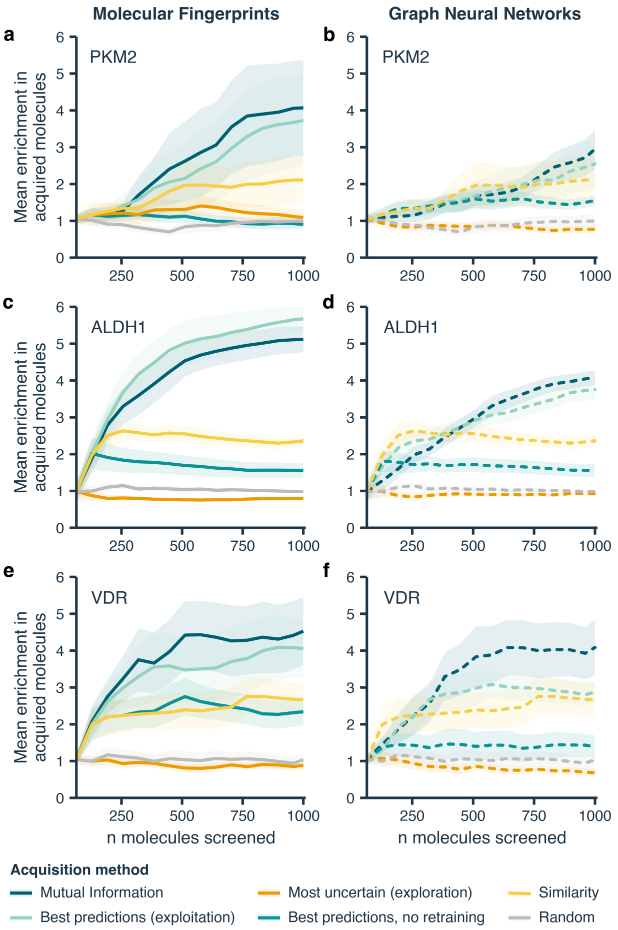

Figure 3: Effect of acquisition functions on hit discovery. Enrichment factor of all acquired molecules with models trained using ECFPs and molecular graphs, reported across three different datasets. Lines represent the average values and shaded areas represent the standard error, computed over 10 experiments with different starting molecules. Molecular fingerprints (ECFPs) are reported on the left-hand column (solid line), and graph neural networks on the right-hand column (dashed line). Every row represents one dataset: **a, b**. PKM2, **c, d**. ALDH1, **e, f**. VDR.

information will be fed back into the model for future screening cycles. Here, we found that exploitation (selection of the best predicted molecules) and acquisition based on mutual information outperform all other acquisition functions, regardless of the dataset and the deep learning approach (Fig. 3). Active learning achieves a remarkable increase in the number of hits compared to screening the same number of molecules in just one go ('one-shot'). For instance, active learning with exploitation leads to a two- to four-fold enrichment compared the corresponding 'one-shot' approach.

Across all datasets and most molecule acquisition functions, ECFP-based models outperform graph-based models in hit enrichment. Graph neural networks, however, achieve a slightly higher prediction accuracy on the PKM2 and ALDH1 datasets (Supp. Table 2 and Supp. Fig. 4). Although graph-based methods tend to find less hits, they exhibit very similar behaviour to the ECFP-based methods in the relative performance of acquisition functions. Therefore, from here on, we will report mostly on the results of ECFP-based models (while those of graph neural networks can be found as Supplementary Information).

Strikingly, iterative similarity-based acquisition (not requiring any machine learning in principle) also yields higher enrichment than 'one-shot' virtual screening across the board, although lower than deep learning. Especially in early iterations, similarity-based acquisition seems to be highly effective, even when compared to the best performing deep learning methods. This suggests that the choice of the acquisition function (*e.g.*, deep learning *vs* similarity-based) also needs to consider the total 'budget' of molecules to be tested experimentally. Finally, exploration (selecting the most uncertain molecules) yields the fewest hits, in line with previous work[13].

Chemical space exploration

Finding structurally novel hits is often one of the objectives of virtual screening since it might increase the chances of success of hit-to-lead optimisation and provide access to unexplored regions in the chemical space. Hence, we analysed the hits acquired (in terms of physicochemical and structural features) at the end of the iterative procedure, compared to the respective starting set. Based on qualitative projections of chemical space, active deep learning resulted to be able to explore broader regions of the chemical space (Fig. 4a), whereas a traditional 'one-shot' (Fig. 4b) or similarity-based (Fig. 4c) approaches stayed in 'narrower' regions.

To get more nuanced insights into the impact of the acquisition functions in the capacity

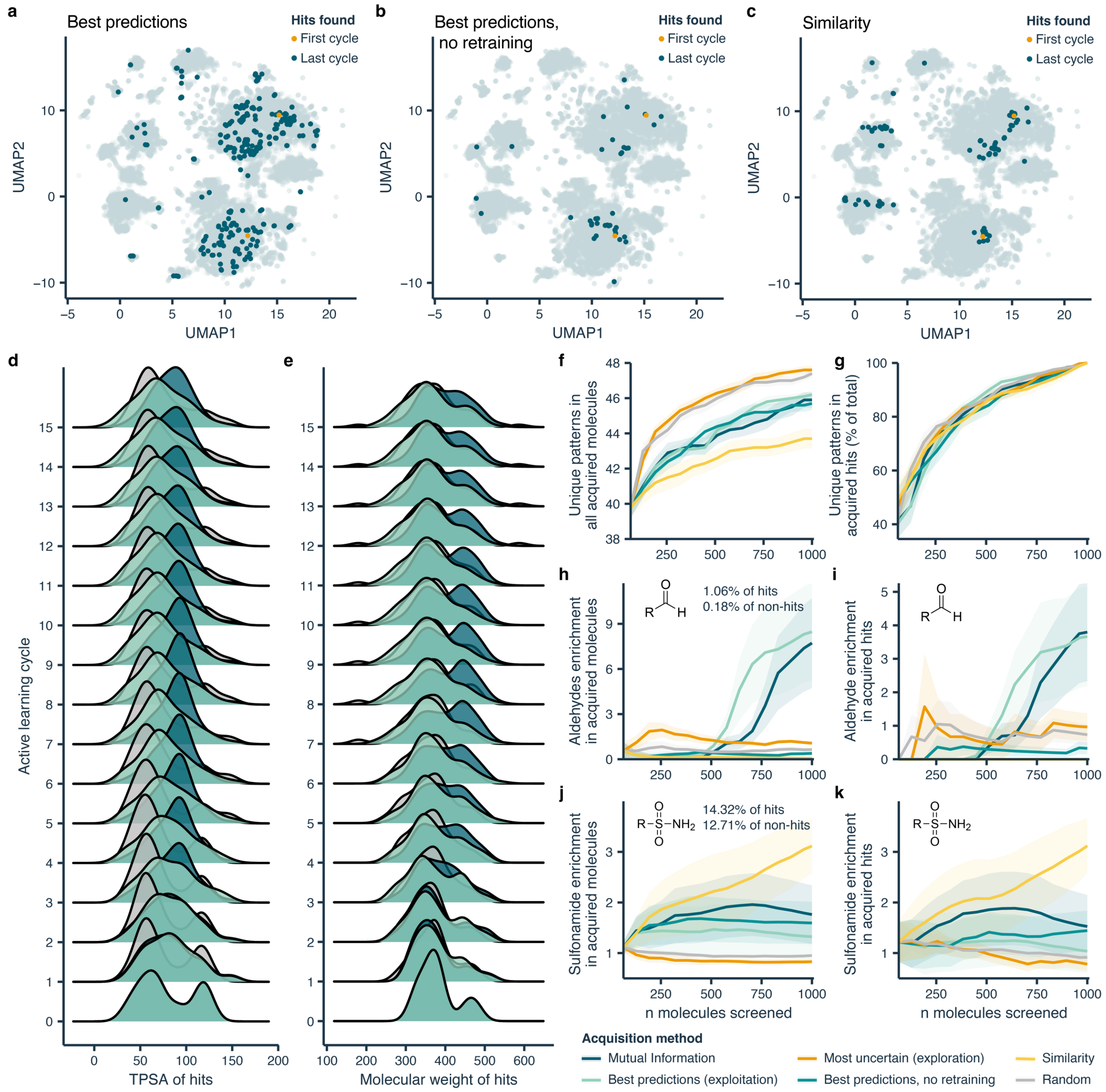

Figure 4: The acquisition function determines the voyage through chemical space. Using the ALDH1 dataset, 64 molecules were acquired each cycle with models trained on ECFPs. **a-c**. Example case of all hit molecules found in the first (orange) and last (dark blue) active learning cycle using three acquisition methods. Here, all molecules in the screening library (n=100,000, depicted in light grey) are visualized in two dimensions using UMAP[47] on RDKit[48] molecular descriptors (see *Materials and Methods*). Every point represents a molecule in the dataset, with the starting and final hits highlighted. The closer the molecules in the projection, the more similar their physicochemical properties. **d.** The distribution of total polar surface area (TPSA) of hit molecules throughout active learning cycles. **e**. Molecular weight of hit molecules. **f**. The number of unique chemically relevant molecular substructures found across all acquired molecules. Shaded areas represent the standard error of 10 experiments with different random starting molecules. **g**. The share of unique molecular substructures found in hit molecules per acquisition method. **h**. The enrichment of aldehydes in acquired molecules. **i**. Enrichment of aldehydes among found hits. **j**. The enrichment of sulphonamides in acquired molecules. **k**. Enrichment of sulphonamides among found hits.

to traverse the chemical space, we analysed the physicochemical properties of the acquired hits over iterations. Different acquisition functions steer a screening experiment towards populations of molecules with different physicochemical properties than the starting set. An example experiment is shown (Fig. 4d, e), where, starting from the same set of molecules, the physicochemical properties of the acquired hits (*e.g.*, total polar surface area and molecular weight) will progressively move towards different distributions over screening cycles based on the chosen acquisition function. This indicates that different acquisition functions can discover distinct chemistry, as captured by the analysed physicochemical properties.

Furthermore, we curated a list of 50 chemically relevant molecular substructures (Supp. Table 3), to monitor chemical space exploration across active learning cycles. We found that exploration (selection of most uncertain molecules) and random acquisition are the best acquisition methods in terms of (global) early enrichment of new substructures (Fig. 4f). Similarity-based acquisition, by definition, is the slowest at finding novel chemistry. When looking at hit molecules only, exploitation and mutual information-based acquisition yield the highest number of new (unique) substructures (Supp. Fig. 5). Hence, in general, mutual information is the acquisition method that overall discovers more hits and provides a high degree of novel substructure exploration (both overall and in the retrieved hits). When normalizing these statistics by the number of hits found by each method (Fig. 4g), no single method finds novel substructures faster.

Interestingly, active deep learning showed an ‘adaptive behaviour’ in the molecular substructures that are acquired during screening. For example, aldehydes – overrepresented in ALDH1 hits (1.06%) compared to all molecules (0.18%) – were, at some point, prioritized by exploitative and mutual information-based acquisition functions (Fig. 4h). In turn, the enrichment of aldehydes found among hits increased accordingly (Fig. 4i). This indicates that relevant structure-activity relationships can be picked up along the way based on newly discovered information. This is not true for methods not based on active learning. In fact, irrelevant substructures might get prioritized purely by their presence in the starting set. For example, in the ALDH1 dataset, similarity-based acquisition started to prioritize sulphonamides (Fig. 4j, h), which are common in this dataset, but not overrepresented in hit molecules (occurring in 14.32% of hits and 12.71% of non-hits). The adaptive nature of active learning allows it to change course during screening and might prevent models to get stuck in specific areas of the chemical space. This motivates once more the adoption of active learning to explore novel chemistry, especially in combination with exploitative and mutual-information-based acquisition functions.

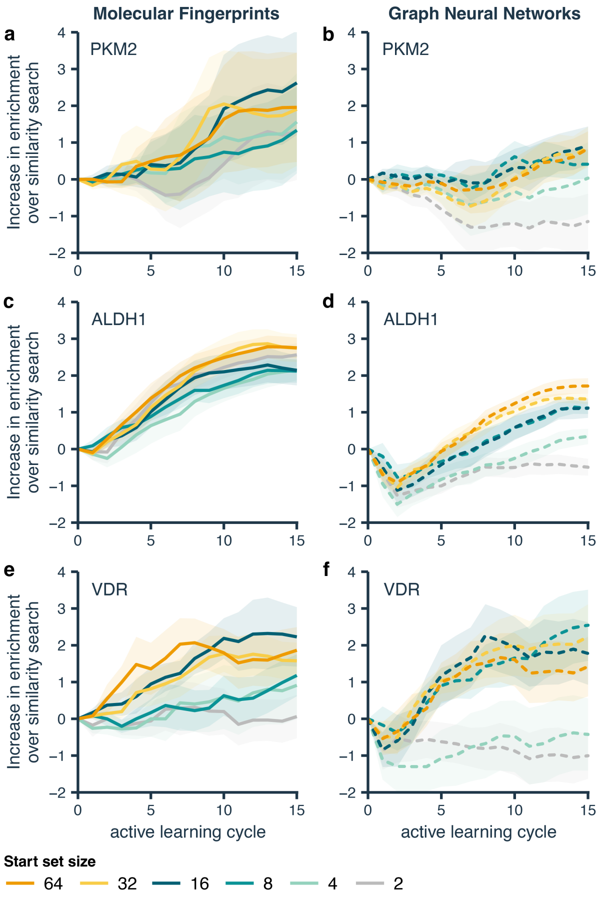

Figure 5: Effect of the starting set size on hit discovery by active learning, compared to similarity search. The increase in enrichment factor of models acquiring molecules using mutual information over similarity search across active learning cycles for different starting set sizes. Lines represent the average values and shaded areas represent the standard error, computed over 10 experiments with different starting molecules. Molecular fingerprints (ECFPs) are reported on the left-hand column, and graph neural networks on the right-hand column. Every row represents one dataset: **a, b**. PKM2, **c, d**. ALDH1, **e, f**. VDR.

Low-data active learning: How low can you go?

To stress-test active learning in low-data regimes, we further decreased the number of starting molecules available for training. We used differently sized random starting sets, containing as little as 2 molecules in the most extreme scenario, up to 64 molecules (Fig. 5, Supp. Fig. 6). The best performing acquisition function (mutual information) was employed with both ECFP-based and graph-based models.

Like in previous experiments (Fig. 3), ECFP-based models outperform graph-based models in hit enrichment (Fig. 5, Supp. Fig. 6), with the latter struggling particularly in extremely low-data starts (2 to 4 molecules). As expected, larger starting sets tend to be better, with the magnitude of the effects of data size reduction depending on the target, apart from some serendipity-driven anomalies. When faced with little initial data (*e.g.*, less than 10 molecules), active learning still showed remarkable hit enrichment. In early screening cycles, similarity search (Fig. 5), which would be a natural choice for virtual screening[49] in extremely-low data scenarios, performs on par with active learning. Yet, active learning progressively showed a substantial improvement over similarity search throughout subsequent cycles (Fig. 5). This motivates the adoption of active learning instead of similarity search for projects that are more time- and cost-permissive and that would allow a larger number of iterations (*e.g.*, more than five).

Conclusions and recommendations

In this systematic study, we presented a reference framework for active deep learning in low-data drug discovery. We examined the behaviour of active deep learning models across a range of different real-world scenarios. We outlined some guidelines for the prospective application of active deep learning:

1. Active learning can find more bioactive molecules in (extreme) low-data settings (up to a six-fold hit enrichment in our case) compared to traditional screening methods like similarity-search or 'one-shot' virtual screening. However, starting out with more data is generally better.
2. Active learning is more effective when screening for several cycles, *e.g.*, five or more as per our experimental setup. With a restricted screening budget (*e.g.*, in the order of dozens of molecules), a similarity-based approach might prove a better solution.
3. The biggest factor driving hit retrieval is the chosen acquisition function, that controls what molecules are selected for follow-up at each cycle. Using best predictions (exploitation) or a mutual information-based acquisition method yields the most bioactive molecules. These acquisition methods present no downside in their ability to discover diverse molecules.
4. A low structural diversity of the initial training set does not seem to affect hit discovery in later active learning cycles. We showed that a lack of diversity is rapidly 'corrected' for within the next 1-5 screening cycles.
5. A fingerprint-based deep learning approach has a better capacity to identify novel hits and is more apt to low-data settings compared to graph neural networks.

Our results corroborate active learning as an effective tool for drug discovery, especially in low-data scenarios, in terms of dataset size, structural bias, and class imbalance. Since the method of acquiring molecules proved to be the main performance driver, there is ample room for improved acquisition functions. Acquisition functions that consider a more chemistry-centred approach, on top of model predictions, bear particular promise to improve molecular novelty among found hits. In this study, we use an approximate Bayesian approach for uncertainty prediction. Different methods of uncertainty quantification might also play a big role in molecule acquisition and could be object of future research[50].

We hope that this research will encourage the adoption of active deep learning in prospective studies. We envision that such adoption will accelerate the shift towards fully automated, deep learning-guided molecule screening, to ultimately discover better drugs with fewer resources.

Materials and Methods

Data pre-processing and analysis

Data pre-processing. ALDH1, PKM2 and VDR were downloaded from LIT-PCBA[36] (accessed on August 2023 at https://drugdesign.unistra.fr/LIT-PCBA). SMILES strings were canonicalized using RDkit[48] v. 2022.09.5, after which non-unique SMILES strings and molecules that could not be "kekulized" or featurized (*see Molecule featurization*) were omitted. All molecules were randomly shuffled. Data was randomly split into a screening library (training set) containing 100,000 molecules and test set, containing 20,000 molecules (Table 1).

Molecule featurization. ECFPs were computed with the following settings: length = 1024 bis, radius = 2. For the featurization of molecular graphs, the following atom features were one-hot-encoded: atom type (C, N, O, S, F, Cl, Br, I, P, Si, B, and Se), implicit atom degree, total atom degree (including hydrogens), explicit valence, implicit valence, total valence, implicit hydrogens, total hydrogens, formal charge, and hybridization. Additionally, atom membership of a set of molecular substructures was binary encoded (Supplementary Table 3). ECFPs and all atom features were computed from canonicalized SMILES strings using RDkit[48] v. 2022.09.5.

Clustering. Structurally diverse groups of molecules in the training set were identified using average distance agglomerative clustering, implemented with scikit-learn v. 1.2.1. A distance matrix of all molecules was used based on the 'Tanimoto distance' (computed as 1 – Tanimoto similarity) on ECFPs. The hierarchical clustering dendrogram was cut at a Tanimoto distance (T_d) of T_d = 0.8 to find moderately diverse subclusters and at T_d = 0.61, T_d = 0.70, and T_d = 0.70 for PKM2, VDR, and ALDH1 respectively to find smaller, low diversity, sub-subclusters. These specific cut-off values were chosen so that each dataset contained 10 clusters. All subclusters had a minimal size of 128 molecules and contained a single sub-subcluster of 64 molecules. Additionally, clusters should contain a minimal number of hits, comparable to the proportion of hits in the full dataset, as determined by:

$$n_{hits} > n_{min}\mathbb{E}[Y_{screen}], \qquad (1)$$

where n_{hits} is the number of hits in the subcluster, n_{min} is the minimum size of the subcluster, and Y_{screen} is the binary class vector of the full screening library.

Molecular substructure assignation. A set of 50 unique SMILES Arbitrary Target Specification (SMARTS) patterns were used to identify substructures (Supplementary Table 3) using RDkit[48] v. 2022.09.5.

Chemical space visualization. UMAP (Uniform Manifold Approximation and Projection)[47] was applied to the set of all available molecular descriptors in RDkit[48] v. 2022.09.5 (see Supplementary Information) using the Python library umap[47] (v. 0.5.3). Molecular descriptors were scaled using max scaling and default parameters were used, except for *n_neighbors*=10 and *min_dist*=0.25.

Active learning

For each active learning experiment, 64 molecules were selected from the screening library. These molecules are then removed from the library. Every set starts with a random hit molecule, after which the remaining molecules are randomly sampled using a uniform distribution. Until the budget of 1,000 molecules was depleted, a model M was trained on the training set, after which bioactivity predictions were made on the screening library. Using one of five acquisition methods (see Acquisition methods), n molecules are labelled and moved to the training set.

Deep learning

Neural network implementation and architectures. All models were implemented using PyTorch[51] v. 1.12.1 and PyTorch Geometric[52] v. 2.2.0. Graph Convolutional Networks (GCN) consisted of an atom embedding layer, followed by three layers of graph convolutions[43] with batch normalization[53]. Global pooling by summing was then used to get molecular embeddings from atom embeddings. Finally, a multi-layer perceptron (MLP) was used with three fully connected layers with batch normalization[53]. The same MLP architecture was used for models trained on ECFPs. All models used a hidden size of 1024 neurons and were optimized for 50 epochs using the Adam algorithm, with a learning rate of 3 x 10-4 with mixed precision and a minibatch size of 64. During training, minibatches were resampled based on their class with:

$$P_c = 1 - \frac{n_c}{N}, \quad (2)$$

where P_c is the sampling probability for a class c, n_c is the number of samples of class c, and N is the total number of samples.

Uncertainty estimation. For uncertainty estimation, anchored ensembling was used[44]. We used an ensemble of $M = 10$ models. For each model, $m \in \{1 \dots M\}$, its parameters θ_m are regularized with a set of anchored parameters $\theta_{anchor,m}$. Each model is initiated with distinct θ_{anchor}, controlled by random seeding. The classification loss in our implementation is defined as:

$$\mathcal{L}oss = -\frac{1}{N}\sum_{i=1}^{N}\log\big(p_m(y_i|x_i)\big) + \frac{\lambda}{N}\left\|\theta_m - \theta_{anchor,m}\right\|^2, \quad (3)$$

where λ is a regularization coefficient (set to 3×10^{-4}). For estimating the expected value $\mathbb{E}$ of a molecule x_i, we take the mean prediction across all models in the ensemble, as follows:

$$\mathbb{E}(y_i|x_i) = \frac{1}{M}\sum_{m=1}^{M} p_m(y_i|x_i). \quad (4)$$

Similarly, the prediction uncertainty for a molecule x_i is defined as the mean entropy $\mathbb{H}$ over the ensemble:

$$\mathbb{H}(y_i|x_i) = -\frac{1}{M}\sum_{m=1}^{M} p_m(y_i|x_i)\log\big(p_m(y_i|x_i)\big). \quad (5)$$

Acquisition functions

Five acquisition functions were used to select follow-up molecules at each iteration. Each acquisition function (a) greedily selects n molecules based on model prediction on the screening library.

1. *Similarity-based*: samples are selected based on their highest Tanimoto similarity (computed with ECFPs; with 1024 bits and a radius of 2) to any previously acquired hit compound).

2. *Exploitative*: the best predicted samples are selected with:

$$a_{exploit} = \operatorname{argmax}_n\big(\mathbb{E}(y|x)\big). \quad (6)$$

3. *Explorative*: most uncertain samples are selected with:

$$a_{explore} = \operatorname{argmax}_n\big(\mathbb{H}(y|x)\big). \quad (7)$$

4. *Mutual Information*: selects samples with low mutual information ($\mathbb{I}$) with:

$$a_{\mathbb{I}} = \operatorname{argmin}_n(\mathbb{H}(y|x) - \mathbb{E}_M[\mathbb{H}(y|x,\theta)]). \quad (8)$$

   Based on Bayesian Active Learning by Disagreement (BALD)[45], the left term represents the entropy of the model predictions (uncertainty) and the right term represents the expected value of the entropy of the model predictions for each draw of the model parameters (*i.e.*, 'disagreement' between the different models in the ensemble)[54]. To have low mutual information, the model must have many ways of explaining the data with high certainty, *i.e.*, low uncertainty and high disagreement.

5. *Random*: samples are selected from a uniform probability distribution.

Compute

All simulated active learning experiments were performed on a Lenovo ThinkSystem SD650-N v2 server equipped with Intel Xeon Platinum 8360Y CPUs and NVIDIA A100 (40GB) GPUs, using approximately 8,000 hours of compute time.

Acknowledgements

This research 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, the Centre for Living Technologies, and SURF (NWO grant EINF-5379). The authors would like to thank A. Ortiz-Рйrez for proofreading the manuscript, L. Rossen for early discussion on active learning, and R. Цzзelik for his suggestion of agglomerative clustering.

Data and code availability

The curated datasets and the Python code to replicate and extend our study are freely available on GitHub at the following URL: https://github.com/molML/traversing_chem_space.

Author Contributions

Conceptualization: D.v.T and F.G. Data curation: D.v.T. Formal analysis: D.v.T. Methodology: D.v.T and F.G. Software: D.v.T. Visualisation: D.v.T. Writing – original draft: D.v.T. Writing – review and editing: D.v.T and F.G. Both authors have given approval to the final version of the manuscript.

References

1. Qureshi, R. *et al.* AI in drug discovery and its clinical relevance. *Heliyon* **9**, e17575 (2023).
2. Jiménez-Luna, J., Grisoni, F., Weskamp, N. & Schneider, G. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opinion on Drug Discovery* **16**, 949–959 (2021).
3. Walters, W. P., Stahl, M. T. & Murcko, M. A. Virtual screening—an overview. *Drug Discovery Today* **3**, 160–178 (1998).
4. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **180**, 688-702.e13 (2020).
5. Liu, G. *et al.* Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii. *Nat Chem Biol* **19**, 1342-1350 (2023).
6. Neves, B. J. *et al.* Discovery of New Anti-Schistosomal Hits by Integration of QSAR-Based Virtual Screening and High Content Screening. *J. Med. Chem.* **59**, 7075–7088 (2016).
7. Deng, J. *et al.* A systematic study of key elements underlying molecular property prediction. *Nat Commun* **14**, 6395 (2023).
8. Van Tilborg, D., Alenicheva, A. & Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *J. Chem. Inf. Model.* **62**, 5938–5951 (2022).
9. Van Tilborg, D. et al. *Deep Learning for Low-Data Drug Discovery: Hurdles and Opportunities*. ChemRxiv, (2024).
10. Volkov, M. *et al.* On the Frustration to Predict Binding Affinities from Protein–Ligand Structures with Deep Neural Networks. *J. Med. Chem.* **65**, 7946–7958 (2022).
11. Wallach, I. & Heifets, A. Most Ligand-Based Classification Benchmarks Reward Memorization Rather than Generalization. *J. Chem. Inf. Model.* **58**, 916–932 (2018).
12. Zabolotna, Y. *et al.* Chemspace Atlas: Multiscale Chemography of Ultralarge Libraries for Drug Discovery. *J. Chem. Inf. Model.* **62**, 4537–4548 (2022).
13. Reker, D. & Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug Discovery Today* **20**, 458–465 (2015).
14. Graff, D. E., Shakhnovich, E. I. & Coley, C. W. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chem. Sci.* **12**, 7866–7881 (2021).
15. Reker, D. Practical considerations for active machine learning in drug discovery. *Drug Discovery Today: Technologies* **32–33**, 73–79 (2019).
16. Settles, B. *Active Learning Literature Survey*. https://minds.wisconsin.edu/handle/1793/60660 (2009).
17. Murphy, R. F. An active role for machine learning in drug development. *Nat Chem Biol* **7**, 327–330 (2011).
18. Fujiwara, Y. *et al.* Virtual Screening System for Finding Structurally Diverse Hits by Active Learning. *J. Chem. Inf. Model.* **48**, 930–940 (2008).

19. Pyzer-Knapp, E. O. Bayesian optimization for accelerated drug discovery. *IBM J. Res. & Dev.* **62**, 2:1-2:7 (2018).

20. Gentile, F. *et al.* Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Cent. Sci.* **6**, 939–949 (2020).

21. Gusev, F., Gutkin, E., Kurnikova, M. G. & Isayev, O. Active Learning Guided Drug Design Lead Optimization Based on Relative Binding Free Energy Modeling. *J. Chem. Inf. Model.* **63**, 583–594 (2023).

22. Bellamy, H., Rehim, A. A., Orhobor, O. I. & King, R. Batched Bayesian Optimization for Drug Design in Noisy Environments. *J. Chem. Inf. Model.* **62**, 3970–3981 (2022).

23. Graff, D. E. *et al.* Self-Focusing Virtual Screening with Active Design Space Pruning. *J. Chem. Inf. Model.* **62**, 3854–3862 (2022).

24. Desai, B. *et al.* Rapid Discovery of a Novel Series of Abl Kinase Inhibitors by Application of an Integrated Microfluidic Synthesis and Screening Platform. *J. Med. Chem.* **56**, 3033–3047 (2013).

25. Xue, D. *et al.* Accelerated search for materials with targeted properties by adaptive design. *Nat Commun* **7**, 11241 (2016).

26. Reker, D., Schneider, P. & Schneider, G. Multi-objective active machine learning rapidly improves structure–activity models and reveals new protein–protein interaction inhibitors. *Chem. Sci.* **7**, 3919–3927 (2016).

27. Yuan, R. *et al.* Accelerated Discovery of Large Electrostrains in $BaTiO_3$ -Based Piezoelectrics Using Active Learning. *Advanced Materials* **30**, 1702884 (2018).

28. Mekki-Berrada, F. *et al.* Two-step machine learning enables optimized nanoparticle synthesis. *npj Comput Mater* **7**, 55 (2021).

29. Zhang, Y. & Lee, A. A. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chem. Sci.* **10**, 8154–8163 (2019).

30. Ortiz-Perez, A., Van Tilborg, D., Van Der Meel, R., Grisoni, F. & Albertazzi, L. Machine learning-guided high throughput nanoparticle design. ChemRxiv (2023).

31. Besnard, J. *et al.* Automated design of ligands to polypharmacological profiles. *Nature* **492**, 215–220 (2012).

32. Borkowski, O. *et al.* Large scale active-learning-guided exploration for in vitro protein production optimization. *Nat Commun* **11**, 1872 (2020).

33. Bertin, P. *et al.* RECOVER identifies synergistic drug combinations in vitro through sequential model optimization. *Cell Reports Methods* **3**, 100599 (2023).

34. Abolhasani, M. & Kumacheva, E. The rise of self-driving labs in chemical and materials sciences. *Nat. Synth* **2**, 483–492 (2023).

35. Seifrid, M. *et al.* Autonomous Chemical Experiments: Challenges and Perspectives on Establishing a Self-Driving Lab. *Acc. Chem. Res.* **55**, 2454–2466 (2022).

36. Tran-Nguyen, V.-K., Jacquemard, C. & Rognan, D. LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening. *J. Chem. Inf. Model.* **60**, 4263–4273 (2020).

37. Jiang, F. *et al.* Aldehyde Dehydrogenase 1 Is a Tumor Stem Cell-Associated Marker in Lung Cancer. *Molecular Cancer Research* **7**, 330–338 (2009).

38. Yang, C. *et al.* Aldehyde dehydrogenase 1 (ALDH1) isoform expression and potential clinical implications in hepatocellular carcinoma. *PLoS ONE* **12**, e0182208 (2017).

39. Palsson-McDermott, E. M. *et al.* Pyruvate Kinase M2 Is Required for the Expression of the Immune Checkpoint PD-L1 in Immune Cells and Tumors. *Front. Immunol.* **8**, 1300 (2017).

40. Plum, L. A. & DeLuca, H. F. Vitamin D, disease and therapeutic opportunities. *Nat Rev Drug Discov* **9**, 941–955 (2010).

41. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

42. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J Comput Aided Mol Des* **30**, 595–608 (2016).

43. Kipf, T. N. & Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. ArXiv (2017).

44. Pearce, T., Leibfried, F. & Brintrup, A. Uncertainty in Neural Networks: Approximately Bayesian Ensembling. in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics* 234–244 (PMLR, 2020).

45. Houlsby, N., Huszár, F., Ghahramani, Z. & Lengyel, M. Bayesian Active Learning for Classification and Preference Learning. ArXiv (2011).

46. Truchon, J.-F. & Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **47**, 488–508 (2007).

47. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv (2020).

48. Landrum, G. RDKit: Open-source cheminformatics. (2006).

49. Cereto-Massagué, A. *et al.* Molecular fingerprint similarity search in virtual screening. *Methods* **71**, 58–63 (2015).

50. Yin, T., Panapitiya, G., Coda, E. D. & Saldanha, E. G. Evaluating uncertainty-based active learning for accelerating the generalization of molecular property prediction. *J Cheminform* **15**, 105 (2023).

51. Paszke, A. *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems* vol. 32 (2019).

52. Fey, M. & Lenssen, J. E. Fast Graph Representation Learning with PyTorch Geometric. ArXiv (2019).

53. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. ArXiv (2015).

54. Kirsch, A., van Amersfoort, J. & Gal, Y. BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning. *Advances in Neural Information Processing Systems* vol. 32 (2019).

55. Li, B. & Rangarajan, S. A diversity maximizing active learning strategy for graph neural network models of chemical properties. *Mol. Syst. Des. Eng.* **7**, 1697–1706 (2022).

56. Huggins, D. J., Venkitaraman, A. R. & Spring, D. R. Rational Methods for the Selection of Diverse Screening Compounds. *ACS Chem. Biol.* **6**, 208–217 (2011).

57. Paricharak, S. *et al.* Data-driven approaches used for compound library design, hit triage and bioactivity modeling in high-throughput screening. *Brief Bioinform* bbw105 (2016).