Generate What You Can Make: Achieving in-house synthesizability with readily available resources in de novo drug design

Alan Kai Hassen[1,5*], Martin Šícho[2,3], Yorick J. van Aalst[2], Mirjam C.W. Huizenga[4], Darcy N.R. Reynolds[4], Sohvi Luukkonen[2], Andrius Bernatavicius[1,2], Djork-Arné Clevert[5], Antonius P.A. Janssen[4*], Gerard J.P. van Westen[2*], Mike Preuss[1*]

[1]Leiden Institute of Advanced Computer Science, Leiden University, Leiden, The Netherlands.
[2]Leiden Academic Centre of Drug Research, Leiden University, Leiden, The Netherlands.
[3]CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technolog, University of Chemistry and Technology Prague, Prague, Czech Republic.
[4]Leiden Institute of Chemistry, Leiden University, Leiden, The Netherlands.
[5]Machine Learning Research, Pfizer Research and Development, Berlin, Germany.

*Corresponding author(s). E-mail(s): a.k.hassen@liacs.leidenuniv.nl; a.p.a.janssen@lic.leidenuniv.nl; gerard@lacdr.leidenuniv.nl; m.preuss@liacs.leidenuniv.nl;

Abstract

Molecules generated by Computer-Aided Drug Design often lack synthesizability to be valuable because Computer-Aided Synthesis Planning (CASP) and CASP-based approximated synthesizability scores have rarely been used as generation objectives, despite facilitating the in-silico generation of synthesizable molecules. Published scores approximate a general notion of CASP-based synthesizability with nearly unlimited building block resources. However, this approach is disconnected from the reality of small laboratory drug design, where building block

resources are limited, making a notion of in-house synthesizability that uses already available resources highly desirable. In this work, we show a successful *de novo* drug design workflow generating active and in-house synthesizable ligands of monoglyceride lipase (MGLL). We demonstrate the successful transfer of CASP from 17.4 million commercial building blocks to a small laboratory setting of roughly 6,000 building blocks with only a decrease of –12% in CASP success. Moreover, we present a rapidly retrainable in-house synthesizability score, successfully capturing our in-house synthesizability without relying on external building block resources. We show that including our in-house synthesizability score in a multi-objective *de novo* drug design workflow, alongside a simple QSAR model, provides thousands of potentially active and easily in-house synthesizable molecules. Further, we highlight differences between general and in-house synthesizability scores and demonstrate potential problems with the out-of-distribution predictive performance of synthesizability scores on generated molecules. Finally, we experimentally evaluate the synthesis and biochemical activity of three *de novo* candidates using their CASP-suggested synthesis routes using only in-house building blocks. We find one candidate with evident activity, suggesting potential new ligand ideas for MGLL inhibitors while showcasing the usefulness of our in-house synthesizability score.

Keywords: computer-aided synthesis planning, casp, retrosynthesis, synthesizability, synthesizability score, de novo drug design, virtual screening

1 Introduction

In drug discovery, the traditional Design-Make-Test-Analyze (DMTA) cycle is undergoing substantial changes, driven by the incorporation of novel artificial intelligence approaches [1]. Within the "Design" phase of DMTA, *de novo* drug design methods have emerged that propose novel molecular structures, already demonstrating effectiveness in identifying potential new drug candidates for desired protein targets [2, 3]. In the search process for potential drug candidates, optimization-based *de novo* approaches repeatedly generate a selection of candidate molecules using a chosen method, evaluate these candidate molecules with an objective function, and optimize the method toward generating molecules that satisfy the objective function. This process continues until, hopefully, interesting molecular structures are found [4, 5]. Inherently, this search involves multi-objective optimization, as generated molecules should satisfy various potentially contradicting and, therefore, non-combinable objectives, such as selectivity for the desired protein target, pharmacokinetic properties, or synthetic accessibility [6]. The underlying methods applied for the generation and optimization of molecules include Generative Adversarial Networks, Autoencoders, Genetic Algorithms, Generative Flows, Diffusion Models, and Reinforcement Learning [5], which can generate novel molecular structures in 2D [7–12] and, more recently, 3D [13].

Simultaneously, the "Make" phase of DMTA has also undergone massive changes with the emergence of artificial intelligence approaches, where computer-aided synthesis planning (CASP) determines synthesis routes by deconstructing molecules recursively into molecular precursors until a collection of commercially available molecules, commonly termed "building blocks", is identified [14, 15]. Rather than manually searching for these synthesis routes, contemporary approaches employ neural networks to encapsulate the backward reaction logic and search algorithms to find possible multi-step reaction pathways [16].

One of the existing challenges limiting the broader adoption of *de novo* techniques in the Design phase is the generation of unrealistic, non-synthesizable molecular structures. To combat this, different strategies have become available to include the synthesizability aspect of the Make phase into the Design phase to ensure realistic molecular structures [17]. The most straightforward approach is to directly use synthesis planning, assessing if a synthesis route can be found using one of the available approaches [14, 18–20]. Lately, this approach has been successfully investigated as an objective in *de novo* drug design [21], but has high computational requirements and is time-intensive [4, 17]. In this scenario, each molecule necessitates an entire synthesis planning run, where the duration can range from minutes to several hours depending on the selected retrosynthesis neural network [22, 23].

An alternative to running synthesis planning is the use of a heuristic or learned synthesizability score that provides a computationally inexpensive and fast measure of synthesizability, making them well suited as an objective function for post-hoc virtual screening or as an objective function within de novo drug design [4, 17]. These synthesizability heuristics calculated on the molecular structure can be as simple as the length of the SMILES string [17] or the presence of fragments typical in synthesizable molecules [24]. More advanced metrics, like the frequently used SAScore [25], combine the presence of common structural features of synthesizable molecules with a penalty for structural complexity like rings or stereo-centers. In practice, these heuristic scores are occasionally used as objectives in *de novo* drug design to improve synthesizability (e.g., [17]) or as post-generation filters to identify synthetic accessible molecules (e.g., [11, 24]).

In contrast to heuristic synthesizability scores, CASP-based synthesizability scores approximate synthesis planning results and learn the relationship between a molecule's structure and the successful identification of a synthesis route via synthesis planning [26]. This learning task is either formulated as a classification task of the synthesis planning outcomes [26, 27] or a regression task relying on the resulting synthesis route properties [21, 28]. However, these CASP-based scores are thus far rarely used as an objective in *de novo* drug design and are missing in common *de novo* benchmark frameworks (e.g., [29]). Yet the limited in-silico studies that use these scores show two things: First, they improve synthesizability in terms of the used score itself [28], but lack in-silico evaluation of potential synthesis routes. Second, they show improvements in post-generation synthesis planning evaluations [21], but lack the experimental evaluation of generated structures and synthesis routes.

All of the above ties into a common challenge of the field, where contemporary *de novo* drug design and synthesizability approaches do not take the experimental reality

of drug discovery into account, as most *de novo* approaches are evaluated against synthesizability and activity heuristics (e.g., [29]) instead of synthesizing potential drug candidates and measuring their activity experimentally [30]. This absence of experimental evaluation and focus on computational benchmarking environments is also present in *de novo* methods that explicitly include synthesizability scores to actively enforce realistic and synthetically accessible molecular structures (e.g., [21, 28]), yielding the question of whether suggested approaches also work experimentally regarding the proposed drug candidates and the suggested synthesis routes.

In addition to the lack of experimental evaluation, these general CASP-based synthesizability scores assume near-infinite building block availability. This assumption is, however, far removed from a realistic laboratory setting, where resources are limited regarding budget and lead times for building blocks. Repurposing already available in-house building blocks reduces both costs and experiment lead times in the research process while also reducing the amount of chemical waste at the same time. Naturally, this consideration is especially relevant for universities with limited research budgets, making a specific notion of in-house synthesizability tailored to available resources more valuable than a general notion of synthesizability within the overall drug discovery process.

The transfer of contemporary CASP methods, which rely on millions of commercially available building blocks, to a resource-limited environment might be challenging for two reasons: First, the CASP performance is limited by the quantity and nature of available building blocks, where missing building blocks can lead to unsolvable molecules [26]. Second, current CASP-based synthesizability scores are not building block agnostic as they create their training data to capture a general notion of synthesizability with these millions of commercially available building blocks (e.g., [21, 26, 28]).

This work addresses such challenges in the field of computer-aided *de novo* drug design (see Figure 1):

First, we demonstrate the successful transfer of synthesis planning to an environment with a limited in-house collection of building blocks, revealing that an extensive commercial inventory is unnecessary for identifying potential synthesis routes. Specifically, we show that using only 6,000 in-house building blocks results in merely -12% loss in synthesis planning performance for a large drug-like chemical space, compared to employing a $>$ 1000 times larger library of commercially available building blocks ("Zinc" [18]).

Second, we introduce an in-house CASP-based synthesizability score that can successfully predict if molecules are synthesizable with our in-house building blocks. In addition, we establish that a well-chosen dataset of 10,000 molecules suffices for training this score, allowing rapid retraining to accommodate changes in building blocks through iterative synthesis planning and model training.

Third, we demonstrate the effectiveness and usefulness of both in-house and general CASP-based synthesizability scores within *de novo* drug design. When combined with a MGLL [31] protein target QSAR model as objectives, we show that the in-house synthesizability score facilitates the generation of thousands of in-house, easy-to-synthesize and potentially active drug candidate molecules. In the course of

this, we highlight differences between resulting candidate spaces when using general and in-house synthesizability scores and demonstrate potential problems with the out-of-distribution predictive performance of synthesizability scores on the generated candidate spaces in *de novo* drug design.

Finally, we experimentally evaluate and critically analyze three generated molecules using an in-house synthesizability score after synthesis based on AI-suggested, in-house CASP routes. In the process, we find one candidate with evident activity, suggest potential novel ligand ideas for MGLL inhibitors, and examine differences between our experimentally evaluated molecules, the generated in-house candidate space, and known MGLL ligands.

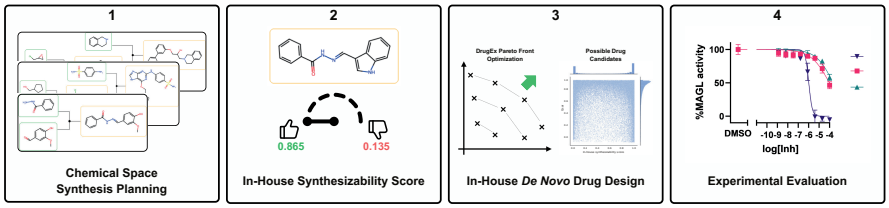

Fig. 1 Schematic overview of the main steps of our study. 1. In-house synthesis is evaluated on a small building block subset from Leiden University for large chemical space. For this, synthesis planning software is used to evaluate in-house synthesizability. 2. An in-house synthesizability score is learned based on the returns of the synthesis planning software estimating the in-house synthesizability of molecules. Its generalizability is successfully evaluated against unseen ChEMBL molecules. 3. This score is used as an objective alongside a MGLL target QSAR model in a multi-objective *de novo* drug design task to successfully provide thousands of potential candidates. 4. Three of these drug candidates are successfully experimentally evaluated using only in-house resources for their synthesis routes, where one shows evident activity.

2 Results & Discussion

2.1 In-house synthesizability

To evaluate the transfer synthesis planning to our real-life, resource-limited university setting, we deployed the open-source synthesis planning toolkit AiZynthFinder [18] with two different building block sets, 5,955 in-house university building blocks ("Led3") and 17.4 million generally available commercial compounds ("Zinc"). The synthesis planning performance was evaluated for two datasets, a set number of centroids of a Butina-clustered [33] subset from Papyrus ("Caspyrus") [32] and a set of 200,000 randomly sampled drug-like ChEMBL [34] molecules.

An overview of the synthesis planning results is presented in Figure 2. This analysis showed that the difference in performance when using only 5,955 Led3 building blocks compared to 17.4 million Zinc building blocks, despite a 3,000-fold increase, is notably small. Using the more limited Led3 building blocks, solvability rates for Caspyrus centroids are around 60%, except when using only 1,000 clusters ("Caspyrus1k") or evaluating on ChEMBL. For the far more extensive Zinc building blocks, solvability

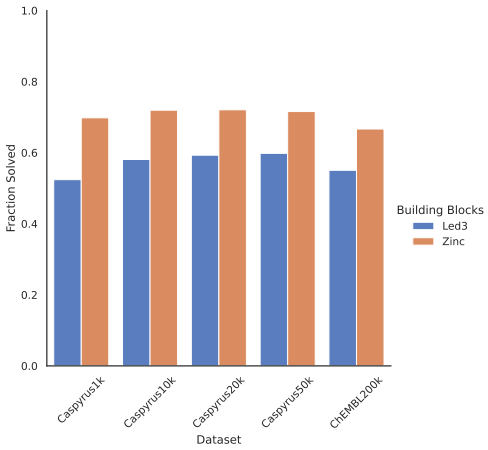

Fig. 2 Synthesis Planning Performance. Evaluation using 5,955 Leiden University in-house ("Led3") or 17.4 million general building blocks ("Zinc"). Percentage of molecules where a complete synthesis route to either building blocks can be found using synthesis planning on different subsets of a Butina-clustered Papyrus [32] ("Caspyrus") or a sample of 200,000 ChEMBL molecules.

rates are around 70% across all datasets. The solvability disparity between both building blocks is around +12% for most datasets except for Caspyrus1k, where roughly +17% more molecules are solved with Zinc building blocks. A notable difference between both building blocks is that the shortest synthesis route found with in-house building blocks is, on average, two reaction steps longer than those using Zinc building blocks, as more building blocks allow shorter synthesis routes across all datasets (see Figure 3).

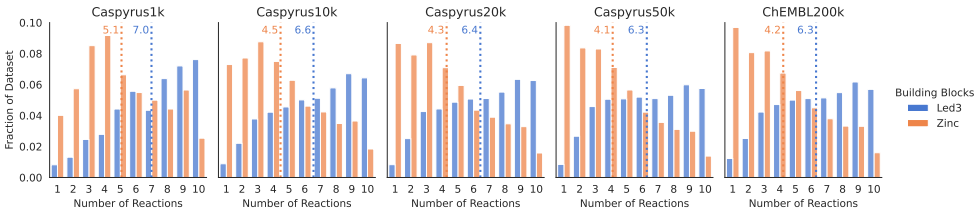

Fig. 3 Distribution of the shortest synthesis route found. Evaluation using synthesis planning with 5,955 building blocks (Led3) and 17.4 million building blocks (Zinc) on the Caspyrus and 200,000 ChEMBL molecules datasets. The dotted line indicates the average route length for both building block sets.

Overall, these results suggest that storing a large commercially sized stock of building blocks is unnecessary to run synthesis planning, as a small building set loses only –12% solvability when accepting slightly longer synthesis routes. These results open the possibility of planning the synthesis of desired compounds in-house instead of buying new building blocks from a vendor and potentially allowing the prioritization of interesting drug discovery candidates according to available in-house resources.

2.2 In-house synthesizability score

After discovering that in-house building blocks are sufficient for performing synthesis planning, we trained a CASP-based synthesizability score for assessing the in-house synthesizability of molecules without requiring resource-intensive synthesis planning. In short, we trained an XGBoost model [35], following the methodology suggested by RaScore [26], to predict if a complete synthesis route can be found for a molecule using synthesis planning. Here, we used the previously generated routes for the in-house Led3 and Zinc building blocks as training data. Afterward, we evaluated the models on respective independent test sets (10% of the data - "IND-Test") and 200,000 newly sampled ChEMBL molecules not present in any training datasets ("ChEMBL-Test") to further evaluate generalizability, for which we additionally conducted synthesis planning with both building block sets (see Figure 4).

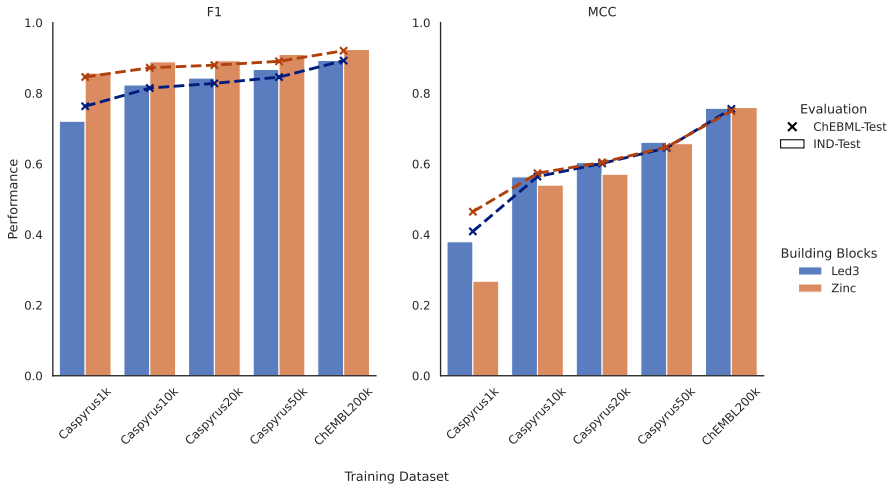

Fig. 4 Benchmarking in-house and general synthesizability scores. Performance comparison of CASP-based synthesizability scores predicting the synthesizability using in-house ("Led3") and general ("Zinc") building blocks in contrast to finding a synthesis route using synthesis planning. Scores are evaluated by measuring the F1 and MCC scores on independent test sets of the respective training datasets ("IND-TEST") and 200,000 newly sampled and, to all models, unknown ChEMBL molecules ("ChEMBL-TEST").

On both evaluation tasks, our trained in-house models achieved excellent results in both F1 and Matthews Correlation Coefficient (MCC) [36, 37] classification scores, which were used to assess the predictivity of synthetic accessibility by the trained scorer. For datasets with at least 10,000 molecules, the F1 performance on the respective test sets surpassed 0.8, proving competitive with the results from larger training datasets. The MCC performance generally improved with more training data, reaching acceptable levels with at least 10,000 molecules, likely because more data enhances the discernment of non-synthesizable molecules. When employing the same training data but using routes based on Zinc building blocks instead, the resulting classifiers performed comparably to those trained with in-house building blocks. Like the Led3

building blocks, classifiers based on Zinc building blocks achieved acceptable F1 and MCC performance when trained on datasets of at least 10,000 molecules. The performance differences in F1 and MCC between the respective dataset test sets and the additionally sampled and unseen 200,000 ChEMBL molecules were minor (except for Caspyrus1k).

These results indicate that our models can accurately estimate in-house synthesizability on a large drug-like chemical space and generalize beyond their respective test sets, allowing us to assess in-house synthesizability for our laboratory in the drug discovery process.

2.3 In-house synthesizability of generated molecules

Since we can successfully predict if a molecule is in-house synthesizable, we wanted to investigate if these scores can be used in a *de novo* drug design setting to generate in-house synthesizable drug candidates.

For this purpose, we combined our in-house synthesizability scores with an MGLL QSAR model to train a multi-objective DrugEx [11] molecular generator to find potent and readily synthesizable compounds for this target (compare training details in methods 4.3). We deployed a novel DrugEx training strategy that helped our generator to learn the desired chemical spaces by guiding it from a general drug-like chemical space towards our target space with both a fine-tuned target-specific generator model, capturing the known ligand distribution, and a QSAR model, capturing the scaffold specific information. As we wanted to evaluate the effect of different synthesizability scores, we trained multiple molecular generators with different QSAR and synthesizability model combinations. We used the QSAR model without any synthesizability score or in combination with either the SAScore [25] or our in-house and general synthesizability scores trained on 10,000 and 200,000 molecules (Caspyrus10k & ChEMBL200k). To evaluate the trained molecular generators, we sampled 100,000 molecules for each trained generator and assessed how many are synthesizable with either building blocks using synthesis planning ("solved") and are seen as active by the QSAR model with a probability larger than 0.8 ("active").

The performance of different synthesizability scores in combination with our QSAR model is presented in Figure 5. The compounds generated with only a QSAR model as an objective have a very low yield of solvable and active structures. This is true when solving with the in-house and general building blocks. In contrast, SAScore produces a lot of solvable molecules that are, however, not active. Regarding synthesizability scores trained using synthesis planning, all CASP-based synthesizability scores perform well and produce between 20,000 and 30,000 predicted active and synthesizable candidates using either the in-house or general building blocks. Surprisingly, scores trained on Caspyrus10k produce the most solved and active molecules, whereas CASP-based synthesizability scores trained on 200,000 ChEMBL molecules produce more solved molecules but not more active ones. It is worth noting that the solvability of the generated molecules is expectably lower than the ChEMBL test sets (compare Figure 4) as molecules are generated along the Pareto front between the QSAR model and the respective used synthesizability score (compare Supplementary: Figure C5 for an example of the generated objective space).

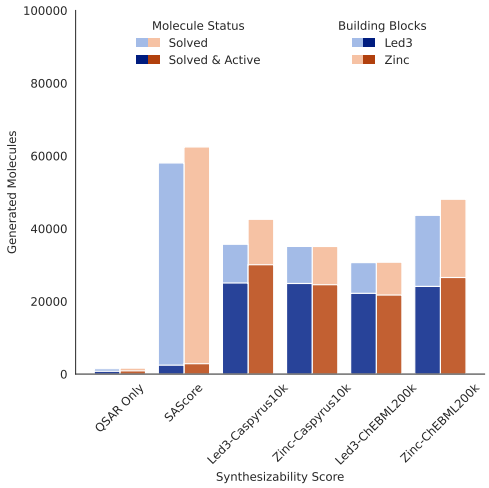

Fig. 5 Generated synthesizable and potentially active molecules using in-house synthesizability scores. Evaluation of 100,000 molecules generated per selected QSAR model and CASP-based synthesizability score combination. "Solved" denotes the successful identification of a synthesis route for a particular molecule with the respective building blocks (in-house Led3 and Zinc), while "Active" is measured by the QSAR model with a probability threshold of greater than 0.8.

Quantitatively, our experiment shows that using in-house synthesizability scores within a *de novo* generator can produce thousands of in-house synthesizable molecules, which can function as a starting point for experimental in-house evaluation.

2.4 Synthesizability score impact on generated molecules

After we showed that CASP-based synthesizability scores facilitate the generation of synthesizable molecules, we set out to investigate their impact on the generated candidates and potential problems with their predictive performance in the desired candidate space.

First, given that we tested in-house and general synthesizability scores alongside our QSAR model, an obvious question is whether these different scores target separate chemical spaces and generate, consequently, distinct candidates. Our primary motivation stems from the fact that the number of solved *de novo* candidate molecules from the in-house and general Caspyrus10k synthesizability scores are comparable when using in-house building blocks within synthesis planning. This yields the question of whether one can use a general synthesizability score in *de novo* design first and solve with in-house building blocks afterward to receive the same candidates. For this purpose, we created a joint UMAP projection [38] of all the solved and potentially active candidate molecules from both the in-house and general synthesizability scores trained with Caspyrus10k, making the synthesizability score results comparable as they are trained on the same dataset. Here, molecules generated with these two scores prioritize different chemical sub-spaces, showing that utilizing only a general synthesizability score and running synthesis planning with in-house building blocks afterward is problematic as the generated results can differ (see Figure 6, Supplementary: Figure C1

for ChEMBL200k). In detail, the usage of only a general score produces sparse results in areas prioritized by the in-house score and, while still partially recovering the same key scaffolds, creates different molecules. Between both candidate spaces, only 1,124 unique molecules, solved with in-house building blocks and seen as active by the QSAR model, are shared (based on InChI comparisons).

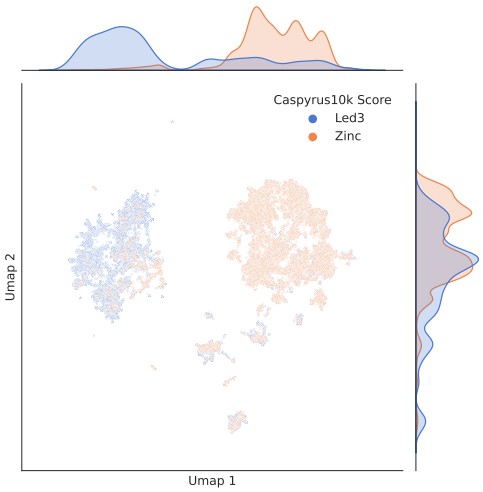

Fig. 6 Contrasting the shared generated chemical space of in-house and general synthesizability scores. UMAP visualization of the solved and potentially active molecular space derived from combining the molecules generated from both in-house and general synthesizability scores trained on the same dataset ("Caspyrus10k"). In both instances, in-house building blocks are used for synthesis planning to evaluate solvability. UMAP is calculated using Morgan Fingerprints (Radius 3, Size 2048).

Second, CASP-based synthesizability scores are trained on a specific drug-like chemical space, in our case 200,000 ChEMBL or up to 50,000 Caspyrus molecules, for which synthesis planning is conducted and that is consequently known to the model. However, a specific target chemical space explored by our *de novo* generation might fall outside of this known model scope and produce unreliable predictions. To analyze if this happens in our generation process, we evaluated if our CASP-based scores correctly predict the route planning results for the 100,000 generated molecules and compared the performance to the independent ChEMBL 200k test set (compare Figure 4). Naturally, we could only compare scores used during the generation with their respective building blocks, meaning that a score trained using synthesis planning results from Zinc building blocks is now also evaluated against Zinc building blocks. Across all models, the performance on generated molecules decreases and performs worse than on the ChEMBL test set, showing a clear domain shift away from the training data (see Figure 7). However, the overall performance for most scores is still acceptable, with around 0.7 F1 and an MCC of around 0.5. For the worst performing Caspyrus10k score based on Zinc building blocks, it is questionable if an MCC of 0.26 is still sufficient to be reliably used.

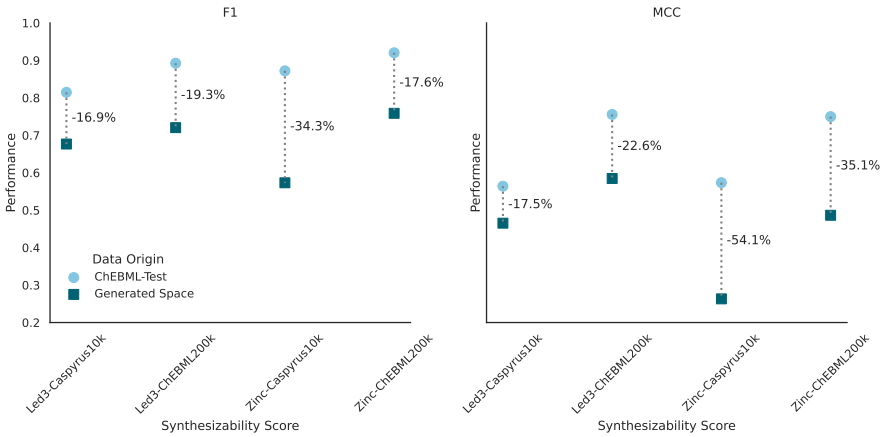

Fig. 7 Out-of-distribution predictive performance of synthesizability scores on the explored chemical space. Evaluation of the predictive performance of CASP-based synthesizability scores on *de novo* generated molecules contrasted with the performance on the ChEMBL-Test set (compare Figure 4). The predictive performance of each score is evaluated by synthesis planning using the building blocks specific to each score's training.

Overall, these results suggest that synthesizability scores, in-house or general, can be used to generate desired candidates, but it is necessary to be careful when using such scores as they might produce different candidate distributions, and the reliability of the individual score predictions might differ.

2.5 Experimental candidate and synthesis route evaluation

Next, we experimentally evaluated our methodology regarding the predicted activity and their suggested in-house synthesis routes. For this purpose, we first deployed a virtual screening approach to reduce the candidate set to a manageable size. In detail, we filtered the molecules generated with the in-house Caspyrus10k synthesizability score, requiring that molecules be perceived as active and synthesizable by their respective objective function using a probability filter threshold of 0.8 (32,907 candidates). Next, we reduced the resulting molecules by the requirement that a synthesis route with our in-house building blocks could be found, resulting in 20,055 potential candidate molecules (compare Supplementary: Table C5 for the other scores). It is noteworthy that we relied here on a virtual screening setting rather than directly using the solved candidates from the prior experiments (compare Figure 5) since this setting reflects a more realistic application of our synthesizability scores in the future, reducing resource-intensive synthesis planning. To decrease the resulting large number of synthesis candidates further, we first analyzed the entire candidate set regarding the Tanimoto similarity for each molecule to the known ligands of MGLL (see Supplementary: Figure C2). We then applied further filtering in that a found synthesis route cannot be longer than five reaction steps to focus on easy-to-make candidates (4,675), required drug-likeness by satisfying the Lipinski rule of 5 [39] (950), and enforced novelty by having a Tanimoto similarity to known ligands of smaller than 0.7 (609). From

these 609 candidates, domain experts selected three candidates for experimental validation based on diversity, potential activity ("chemical eye"), and the presence of a short synthesis route (1 or 2 steps). These three candidates were made using the suggested synthesis routes by the synthesis planning algorithm and evaluated in a natural substrate assay for MGLL inhibition.

The experimental inhibition results of our candidates and their respective in-house synthesis routes are presented in Figure 8. Compound **1** showed clear activity with an IC_{50} of 1 µM, and compounds **2** and **3** show slight activity of around 100 µM IC_{50}.

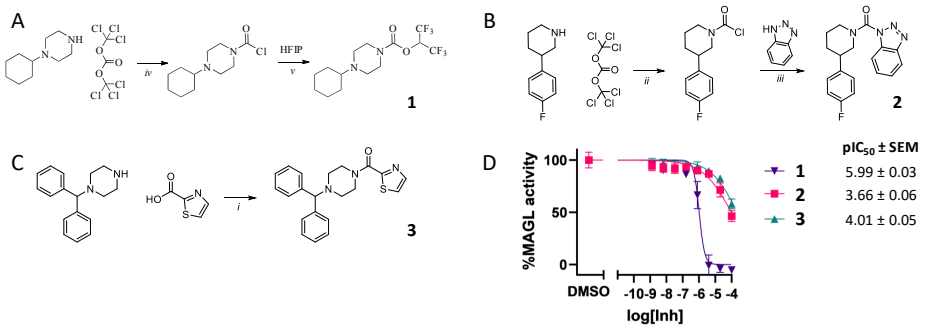

Fig. 8 Selected *de novo* generated candidates, synthesis routes based on in-house building blocks, and their experimentally validated activity. A, B, C) Selected candidates **1**, **2**, **3** for experimental evaluation and their respective in-house synthesis routes. D) Residual MGLL enzyme activity after treatment with varying concentrations of inhibitor as measured by natural substrate assay (compare Supplementary: Experimental Evaluation D for details).

Although all three tested molecules showed some level of inhibitory activity, a stricter boundary of ≤ 10 µM, generally used for hit finding, only leaves one candidate that can be classified as active. This lower potency is unsurprising, given that the selection of molecules to synthesize was based on conducting at most two synthesis steps. Nevertheless, from these experimental results, we can conclude that we can generate in-house synthesizable and active drug candidates that rely on CASP routes using our limited building blocks.

2.6 Critical analysis of *de novo* generated candidates

Given that most *de novo* methods only do an in-silico evaluation of their drug candidates [30], it is vital to critically analyze our experimentally evaluated and active molecules stemming from a *de novo* drug design approach to provide further inside.

For this purpose, we first contrasted our synthesized candidates with known ligands to analyze their novelty. When directly inspecting our selected candidates, even though active and in-house synthesizable, their novelty in key scaffolds is limited. Looking at the closest known ligands, as determined by a Tanimoto similarity threshold, for the respective candidate structures, **2** and **3** are variations of the closest ligand. However, candidate **1**, which was also the most active one in our experiments, deviates more from the closest known ligands in the training dataset and seems to combine distinct

motives found in previously explored analogs using the same key scaffold (see Figure 9, Supplementary: Figure C3 for candidate **2** & Figure C4 for candidate **3**), akin to what a medicinal chemist would think of trying in the various Design cycles of a candidate.

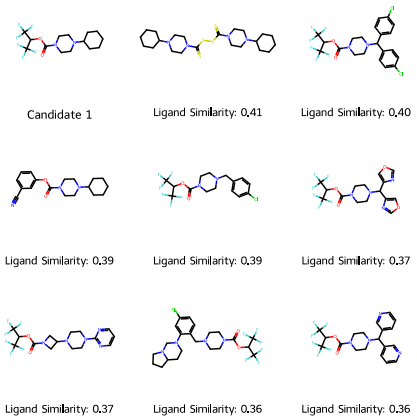

Fig. 9 Closest known ligands compared to most active candidate 1. Measured by Tanimoto similarity on Morgan Fingerprints (Radius 3, Size 2048).

In the second step, we compared our solved candidate space to the known ligands to understand what constitutes our generated space and how our objective functions influence the generation of potential candidates and the presence of key scaffolds. For this purpose, we created a joint UMAP projection of all the solved generated candidate molecules, our three synthesized candidates, and all known ligands for the target. For the known ligands, we annotated which molecules are active or inactive in terms of our QSAR model (compare methods 4.3) and for which of the active ligands a synthesis route could be found with our in-house building blocks. When analyzing the joint UMAP projection of the generated candidate molecules and known ligands (see Figure 10), candidate molecules are generated in areas where active ligands that are synthesizable with our in-house building blocks are present. From this, we can conclude that the QSAR model works as intended, which is supported by the direct rediscovery of 145 known active ligands in our candidate space (based on InChI comparisons) that the QSAR model also classified as active and, in comparison, the rediscovery of 0 inactive ligands. This, however, also explains the usage of key scaffolds in our generated candidates, as the QSAR model operates on the structures of known ligands for MGLL and does not generalize well beyond that. Inactive known ligands, in comparison, tend to be in areas of low candidate density. They can, however, also be close to active ligands with higher density, especially when analogs to known actives are tested.

We can conclude further that the applied in-house synthesizability score works as intended as a generation objective, as unsolvable active ligands are outside areas with high candidate density. Intriguingly, the model generates two major clusters of molecules with little to no known molecules tested for MGLL. These areas could hold

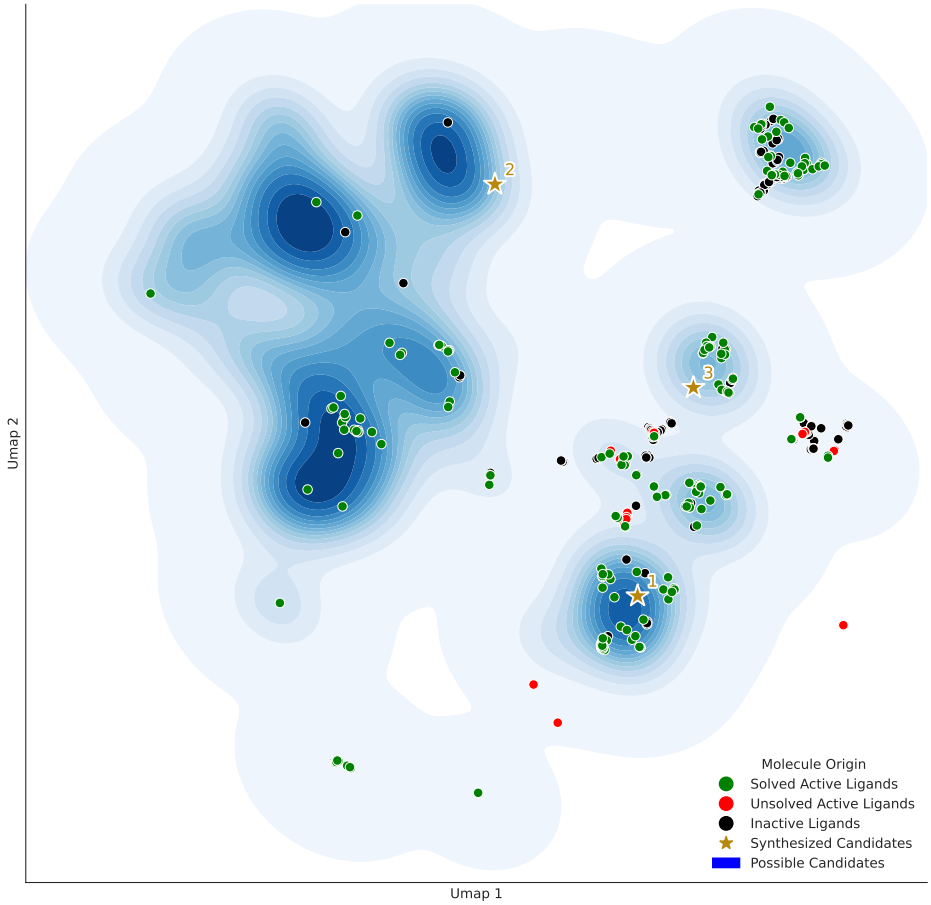

Fig. 10 Contrasting the generated drug candidate space with the known MGLL ligand space. UMAP visualization of the solved molecular candidate space of 20,055 molecules generated with in-house synthesizability score ("Caspyrus10k") and target QSAR model as training objectives and known MGLL ligands. Known MGLL ligands are marked as either inactive ("black circle") or active. Active ligands are differentiated between synthesizable using in-house building blocks ("green circle") and those that are not ("red circle"). Experimentally tested candidates are denoted with a star. UMAP is calculated using Morgan Fingerprints (Radius 3, Size 2048).

more 'creative' ligands, which was also illustrated by their lengthier synthetic routes. For synthetic reasons, these were outside of the scope of this research.

3 Conclusion

In this work, we have introduced an end-to-end and experimentally evaluated in-house *de novo* drug design approach that provides active drug candidates and their in-house synthesis routes by repurposing already available chemicals to reduce costs, lead times and potentially chemical waste in the drug discovery process.

We have demonstrated that synthesis planning can be successfully conducted by using only a small set of roughly 6,000 in-house available building blocks, making it unnecessary to have a commercially vendor-sized stock of building blocks available. With this, we demonstrated the possibility of conducting potential synthesis in-house while repurposing already available resources. Compared to utilizing general vendor building blocks, this in-house approach yields only a -12% decrease in synthesis planning success rate when accepting the resulting, on average, two reactions longer synthesis routes. Next, we leveraged our in-house synthesis planning approach to create an in-house machine learning synthesizability score to predict if a molecule is synthesizable with our in-house building blocks. We further showed that it is possible to train such a score on a small, selected subset of molecules, allowing the recreation of our score within a day in case of changes in our available building blocks, reactions, or the general adaptation to a new laboratory environment by the broader research community. Finally, we showed the successful application of this score in *de novo* drug design by generating molecules that are both active against our selected MGLL target and in-house synthesizable. We further demonstrated that combining synthesis planning and *de novo* drug design is viable and valuable in a small laboratory setting by providing a large set of in-house accessible candidate molecules to our chemists, showing that including such a synthesizability score increased the number of in-house synthetically accessible molecules manifold. Out of this candidate pool, we validated three selected candidates not only in silico but experimentally, finding an active molecule with new disconnection ideas for our target and additionally verifying that the algorithmically proposed in-house synthesis routes are feasible in our laboratory setting.

Even though the proof-of-concept for in-house synthesizability of generated structures is the main focus of this study, a primary limitation relates to the novelty of the generated structures. Generally, we see in our candidates one of the current problems in *de novo* drug design, where key scaffolds for the target are re-used, and the sidechains are algorithmically altered (e.g., [3]). In our work, we do not explore potentially more active candidates with more complex side chains and, consequently, longer synthesis routes, as we find novel ideas for a possible MGLL inhibitor, even when looking only at fairly undecorated molecules. Still, the re-usage of key scaffolds is also present in our work. Even though we do not enforce or fix any scaffolds for the target, our trained molecular generator re-discovers active and in-house synthesizable molecules with known scaffolds on its own.

A natural future improvement is to replace the target QSAR model, potentially limiting the diversity of generated key scaffolds, with other methods for assessing protein-ligand activity like a shape-based pharmacophore [40] or docking [41, 42]. Since both synthesis planning and synthesizability scores are active research fields, improving the synthesis planning performance with more complex neural networks capturing that capture the reaction logic [22, 23] or better approximation models for synthesizability [27, 28] that combine more synthesis route criteria beyond binary CASP-synthesizability [21]. Along the same lines, optimizing the right in-house building blocks to open synthetically accessible chemical spaces might be of further interest. Here, the presence of the right mix of small laboratory building blocks could allow the synthesis of a broader chemical space with as few as possible reactions. Beyond

focusing only on in-house synthesizability, merging in-house with cheap and easy-to-acquire vendor building blocks could be of practical interest to maximize cost-efficient synthesis.

Finally, our in-house synthesizability score is regularly used in our university setting for *de novo* drug design and virtual screening to streamline the overall drug discovery process. Its internal usage and the application of similar scores in other institutions will hopefully facilitate a change for a more efficient and sustainable drug discovery process and a further combination of contemporary artificial intelligence methods with real-world laboratory experimentation going forward. For this purpose, we provide all relevant code that relies solely on open-source software and all data to reproduce the results presented in this work, allowing easy and cost-free creation of other in-house synthesizability scores.

4 Methods

4.1 Synthesis planning

For all synthesis planning in this study, we used the publicly available open-source AiZynthfinder [18] synthesis planning framework. Specifically, we relied on the AiZynthfinder-provided NeuralSym reaction network [43] that is trained on publicly available USPTO reactions [44] and Monte-Carlo Tree Search [14] as the respective search algorithm. The search settings were limited to a search time of 900 seconds per molecule, 1000 search iterations, and a synthesis route depth of 8. Further, we added 50 possible reactions to the tree search per reaction model call (compare Supplementary: Table A2 for details). The building blocks used, i.e., search targets in the tree search, were 17,422,831 Zinc building blocks provided by AiZynthinder [18], used for the general evaluation of synthesizability, and 5,955 building blocks provided by the Leiden University Early Drug Discovery & Development department [45], used for in-house synthesizability.

We utilized two datasets to evaluate synthesizability using the respective building blocks: First, we created a representative subset of the synthesizable drug-like molecules space that allows fast evaluation and retraining of potential synthesis scores named Caspyrus. The creation process mimicked our work evaluating different model architectures in synthesis planning with 10,000 molecules [23]. We selected the high-quality Papyrus dataset [32] of 1,238,835 molecules and cleaned them with the Guacamol cleaning strategy [29] to ensure drug-like molecules. We further removed known building blocks stemming from Zinc [18], Enamine [46], MolPort [47] and eMolecules [48]. We then clustered the remaining molecules using Butina clustering [33] with a cut-off of 0.6 using Morgan fingerprints [49] (radius of 2, fingerprint size of 1024), which resulted in 137,963 cluster centroids. From these centroids, we removed 19 centroids that are directly in clinical study phases 1-3 [50] as we wanted to prevent later molecular generation towards intellectual property spaces. Finally, we took centroids of the n largest clusters to create the different Caspyrus versions (see Table 1).

Table 1 Different Caspyrus versions. Overview of the selected cluster centroids per Caspyrus dataset and their overall represented molecules.

Name	Centroids	Molecules
Caspyrus1k	1,000	82,352
Caspyrus10k [23]	10,000	280,956
Caspyrus20k	20,000	371,231
Caspyrus50k	50,000	491,422

Second, we sampled 200,000 molecules from ChEBML, following the evaluation framework of RaScore [26], and cleaned them with the same Guacamol cleaning strategy. Compared to the clustered Caspyrus dataset, this dataset is more likely to contain noisy data, duplicates, and potential building blocks.

We measured the number of molecules for which at least one complete synthesis route with the respective building block sets could be found on both evaluation datasets. Furthermore, we used the shortest found route of all found synthesis routes to evaluate the minimum route length.

4.2 Synthesizability scores

We leveraged the results of the synthesis planning to train our general and in-house synthesizability scores. To approximate synthesis planning, we used XGBoost [35] as a binary classifier to learn the relationship between the selected molecules and their synthesis planning result (synthesis route found/not found). We selected the rather "simplistic" XGBoost, following the well-working RaScore [26], as we were more interested in the general applicability of our approach and because more complex Graph Neural Network architectures showed only slight performance improvements [27, 28]. The input into all XGBoost models were Morgan fingerprints (radius of 3, size of 2048) using additional selected chemical properties following DrugEx [11].

All classifiers were trained and evaluated with the following scheme: Initially, we split away 10% of the respective data as a test set following the process of RaScore [26], where we used the ability to find a synthesis route with Led3 building blocks as a stratifying criterion. On the remaining 90% of the data, the training dataset, we conducted a 5-fold cross-validation to evaluate different hyperparameter settings. Our hyperparameter optimization scheme consists of 1000 rounds of Bayesian Optimization for every classifier using Bayesian Optimization and Hyperband [51] - in total, multiple days of runtime per classifier. Here, the selected hyperparameters were the learning rate (0.05-0.4), maximum depth of a tree (1-50), minimum loss reduction required for further partition of a tree (0-10), and number of trees (5-250). The final score is then trained on the entire training dataset using the best hyperparameters.

The final performance of each score is evaluated on two datasets: First, the respective 10% test data for each dataset not used during training. Second, we sampled an additional 200,000 cleaned molecules from ChEMBL [34] and conducted synthesis planning to create a new test to measure the generalizability of the trained scores on a

large chemical space (compare Supplementary: Table B3 for optimal hyperparameter settings and results). Noteworthy, we ensured that the molecules from this ChEMBL test set are neither in the Caspyrus nor the Chembl200k datasets used to train our CASP-based synthesizability scores.

4.3 De novo molecular generation

The trained CASP-based synthesizability scores were evaluated in a *de novo* drug design setting, where the goal was to generate active and in-house synthesizable molecules for our selected MGLL protein target [31], evaluated by in silico synthesis planning and experimental evaluation.

For this purpose, we used our molecular generator DrugEx [11] alongside a set of desirable generation objectives, in our case, a trained target QSAR model and multiple different synthesizability scores. We selected DrugEx v3 as the molecular generator for two reasons: First, DrugEx is currently the only Reinforcement Learning (RL) approach that uses a reward based on the Pareto front instead of a single or a scalarized objective [52], which allows the model to more accurately learn the trade-offs between different objectives and produce more diverse solutions. This is especially important in our setting as the biological activity predicted by the QSAR model and synthesizability are non-consumable without losing information about the trade-offs between both objectives, meaning that a synthesizable molecule is not necessarily active and vice versa. Second, we hope for the adoption of our approach in the future, as DrugEx is open source, well-maintained with high code quality [53] and allowed for all the data and methods used to create this work to be publicly available. Given that the DrugEx framework offers several generative model architectures, we decided to use the latest graph-based transformer model operating on fragments in this work [11], where the goal is to learn the generation of novel and valid molecules from a predetermined chemical space given a set of starting fragments – substructures smaller than known key scaffolds. The version 3.4.0.dev1 of the DrugEx software was used throughout this work.

In our case, the training process of DrugEx consisted of three steps:

(1) A pretrained model was obtained, that captures the general drug-like chemical space by learning the mapping between fragments and their respective molecules. Here, we used a pre-trained model based on Papyrus 05.5 [32] that was trained by applying BRICS fragmentation [54] on the molecules in Papyrus to achieve the aforementioned goal.

(2) A fine-tuned DrugEx model was created by conducting transfer learning on the pre-trained model with the chemical space related to MGLL. For this purpose, we extracted 700 structures related to MGLL from Papyrus 05.5 [32] using the MGLL Uniprot ID Q99685 (Supporting information: Q99685.tsv) and utilized them to fine-tune the pre-trained model. These 700 ligands in the fine-tuning set were also fragmented with the BRICS method following the same protocol as the pre-trained model (1). Out of the resulting data set of fragment-molecule pairs, 10% were used for validation and implementation of the early stopping strategy. The training process ran for 200 epochs with a batch size of 512 until no improvement in loss could be observed after 50 epochs (compare Supplementary: Figure C6).

(3) In the final step, we used RL to steer our model towards generating active and synthesizable molecules by repeatedly generating a set of molecules, evaluating the generated molecules with our objectives, and retraining the model based on the Pareto-front of both active and synthesizable molecules. Here, the general pre-trained model (1) was used as the actively trained network (G_ϑ) and the fine-tuned model (2) as the fixed network (G_φ) in the DrugEx RL exploration strategy [11]. To train the model, the same set of training and validation fragment-molecule pairs was used as in the fine-tuning step (2). Given that we wanted to evaluate the effect of different synthesizability scores, we trained multiple models that each combined a different synthesizability score with our QSAR model (see Table 2). Further, several values for the exploration parameter epsilon were explored that controlled the fraction of data originating from the fixed fine-tuned ligand space model during training (compare Supplementary: Figure C7). For all objectives, modifier settings were set according to values recommended in the literature or based on a suitable classification threshold to support smooth model training (compare Supplementary: Table C7). For each trained model, the training was set to continue for at most 500 epochs, with early stopping being triggered once the overall desirability on the validation set stopped improving. Based on the epsilon trade-off data obtained (compare Supplementary: Figure C7), the final set of 100,000 compounds was generated with models with an exploration parameter epsilon of 0.2 as they offered the best trade-off between objective optimization (desirability) and structural diversity. All models built are made available in the public domain as part of the provided data.

Table 2 Trained DrugEx models. Models are trained using a combination of the QSAR model alongside a synthesizability score, relying in the case of CASP-based synthesizability scores on a unique set of training data and building blocks.

Synthesizability Score	Training Data	Building Blocks
QSAR Only	-	-
SAScore	-	-
Led3Caspyrus10k	Caspyrus10k	In-house
Led3ChEMBL200k	ChEMBL200k	In-house
ZincCasyprus10k	Casyprus10k	General
ZincChEMBL200k	ChEMBL200k	General

The QSAR model used for the MGLL [31] activity objective was trained by using the QSPRPred library [55], which directly interfaces with DrugEx to facilitate QSAR model scoring. The same set of 700 MGLL ligands from Papyrus, as described in the fine-tuning step (2), was used to obtain bioactivity data for this model. For model evaluation and selection, we divided the ligands into training and test sets using both a scaffold split (80% training, 20% test) and a time split (pre-2018 training, since 2018 test), comparing the results obtained from different models under both evaluation strategies. Here, we opted for a classification task instead of a regression task for the QSAR modeling as, from our experience, classification works better in DrugEx during

RL optimization. The labels to distinguish active and inactive molecules were taken from the pChEMBL values in Papyrus, where molecules with at least 6.5 pChEMBL were treated as active. For both scaffold- and time-splits, we applied hyperparameter optimization using grid-search with a 5-fold cross-validation on the training data (compare Supplementary: Table C8) to find the optimal hyperparameters and selected the best model algorithm based on the overall test-set performance across both evaluation strategies. Out of the nine evaluated models via QSPRPred (Random Forrest, Extra Tree Classifier, XGBoost, Multi-Layer Perceptron, Gradient Boosting Classifier, AdaBoost, k-nearest neighbors, Support Vector Classification, and Gaussian Naïve Bayes) [35, 56], we picked XGBoost for our QSAR model as it performed consistently well across both the scaffold and time split benchmarks (see Figure 11) and provided fast inference speeds required for our RL training. Due to data scarcity, we retrained the selected XGBoost classifier afterward with all known bioactivity data for our target. The optimal hyperparameters for this final model were chosen from the prior scaffold-split optimization workflow, as the resulting model showed the best performance both during cross-validation and on the external test set.

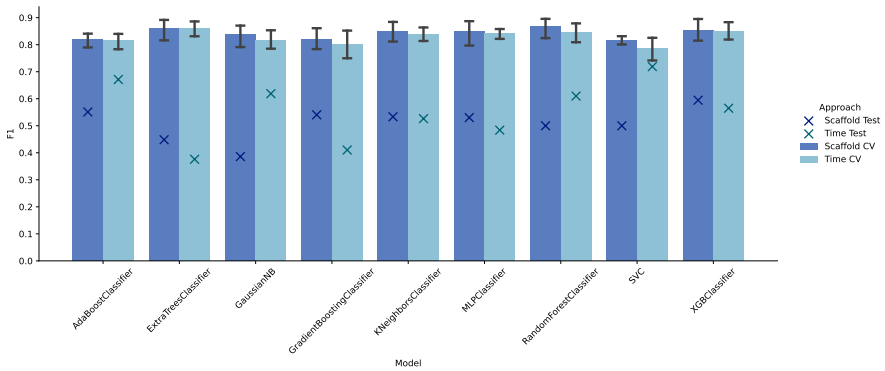

Fig. 11 Performance of the QSAR model evaluated on known MGLL ligands. Performance is measured using 5-fold cross-validation on the training data ("CV") and an independent test dataset ("Test") while employing both scaffold- and time-splits.

To investigate the effect of our synthesizability scores on generated molecules, we used different synthesizability scores as a second objective alongside the QSAR model (see Table 2). In our baseline setting, we only used the QSAR model without any synthesizability score ("QSAR Only") or combined SAScore [25] with the QSAR model ("SAScore"). We picked SAScore as a heuristic synthesizability baseline as it is a widely adopted measure to evaluate molecules (e.g., [29]) and differs substantially from our CASP-based synthesizability scores as it measures the topological complexity of a molecule instead of approximating the ability to find a synthesis route using synthesis planning. As SAScore does not provide a probability for synthetic complexity, we transformed the scores using a smoothed-clipped score function (compare Supplementary: Table C7). For our non-baseline setting, we selected four different CASP-based synthesizability scores alongside our QSAR model, where two measured the in-house

synthesizability and the other two measured general synthesizability. For our in-house synthesizability scores, we used models trained on the Caspyrus10k and ChEMBL200k datasets using in-house building blocks. The rationale behind this selection was two-fold: First, we wanted to know how much data is required to train a synthesizability score. Second, a synthesizability score based on 10,000 molecules is easily retrainable in case of available building blocks or reaction changes, as the computational requirements of running synthesis planning differ substantially between 10,000 and 200,000 molecules. For the general synthesizability scores, we selected models based on the same Caspyrus10k and ChEMBL200k datasets, as this allowed a direct comparison on the same training dataset between our sparse locally available in-house building blocks and generally available building blocks. Noteworthy, the ChEMBL200k score mimics the RaScore [26], as it is trained with the same amount of data and comparable building blocks.

To evaluate different combinations of the QSAR model and synthesizability score, we generated 100,000 molecules for each uniquely trained DrugEx model. We evaluated the synthesizability of our generated molecules by conducting synthesis planning using in-house and general building blocks on the generated molecules with the same settings as in the prior synthesis planning step. Given that we can sample indefinitely from our trained models, we sampled 100,000 molecules for each trained model, assuming that a denser population of candidates generated along the Pareto front should increase our hit probabilities (e.g., [52]) and provide us with enough examples to evaluate each score profusely.

Supplementary information. Additional figures, tables and experimental details are provided in the supplementary information.

Declarations

Availability of data and materials. All source code, models and relevant data of this work can be found at https://github.com/AlanHassen/led3score.

Competing interests. The authors declare no competing interests.

Funding. AKH was supported by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Innovative Training Network European Industrial Doctorate grant agreement No. 956832 "Advanced machine learning for Innovative Drug Discovery". MŠ was supported by Czech Science Foundation Grant No. 22-17367O and by the Ministry of Education, Youth and Sports of the Czech Republic (project number LM2023052). SL was supported by funding from the Dutch Research Council (NWO) in the framework of the Science PPP Fund for the top sectors and acknowledges the Dutch Research Council (NWO ENPPS.LIFT.019.010).

Authors' contributions. **AKH:** Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration. **MS:** Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration. **YJvA:** Software, Formal Analysis, Investigation, Data

Curation, Visualization. **MCWH:** Formal Analysis, Investigation, Data Curation, Visualization. **DNRR:** Formal Analysis, Investigation, Data Curation, Visualization. **SL:** Conceptualization, Methodology, Writing - Review & Editing, Supervision. **AB:** Conceptualization, Methodology, Validation, Writing - Review & Editing. **DAC:** Validation, Resources, Writing - Review & Editing, Supervision, Funding acquisition. **APAJ:** Conceptualization, Methodology, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Supervision, Project administration. **GJPvW:** Conceptualization, Methodology, Validation, Resources, Writing - Review & Editing, Supervision, Project administration. **MP:** Conceptualization, Methodology, Validation, Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Acknowledgements. Parts of this work were performed using the ALICE compute resources provided by Leiden University. Large Language Models (LLMs) were used throughout the creation of this manuscript to improve spelling mistakes, grammar, and the overall reading flow. All LLM suggestions were profusely checked for correctness and refined by the authors of this work. No research was conducted throughout this work by the LLM.

Authors' information (optional). Does not apply.

Appendix A In-house synthesis planning

Table A1 Synthesis Planning Performance. Evaluation using 5,955 Leiden University in-house ("Led3") or 17.4 million general building blocks ("Zinc"). Percentage of molecules where a complete synthesis route to either building blocks can be found using synthesis planning on different subsets of a Butina-clustered subset from Papyrus [32] ("Caspyrus") or a sample of 200,000 ChEMBL molecules.

	Dataset				
Building Blocks	Caspyrus1k	Caspyrus10k	Caspyrus20k	Caspyrus50k	CHEMBL200k
Led3	52.4%	58.1%	59.3%	59.8%	55.0%
Zinc	69.8%	72.0%	72.1%	71.6%	66.7%

Table A2 Synthesis Planning search settings. AiZynthFinder [18] search settings used throughout this work.

Parameter	Value
Search Algorithm	Mcts
C	1.4
cutoff_cumulative	0.995
cutoff_number	50
Max_transforms	9
Iteration_limit	1000
Time_limit	900
Use_prior	True
Return_first	True
Exclude_target_from_stock	True
Prune_cycles_in_search	True

Appendix B In-house synthesizability score

Table B3 Hyperparameter search results for synthesizability scores. Optimal hyperparameters (Learning Rate, Max Tree Depth, Minimum Loss, Numbers of Trees) found for the used XGBoost models and their performance measured by F1 and MCC scores on independent test sets of the respective training datasets ("IND-Test") and 200,000 newly sampled and, to all models, unknown ChEMBL molecules ("ChEMBL-Test").

		Hyperparameter				IND-Test		ChEMBL-Test	
Building Blocks	Dataset	LR	Max. Tree Depth	Min. Loss	Number of Trees	F1	MCC	F1	MCC
Led3	Caspyrus1k	0.272	1	4.950	205	0.721	0.379	0.763	0.409
	Caspyrus10k	0.269	18	0.009	244	0.823	0.563	0.815	0.564
	Caspyrus20k	0.196	17	0.039	244	0.843	0.604	0.828	0.601
	Caspyrus50k	0.200	17	0.016	223	0.867	0.661	0.846	0.645
	ChEMBL200k	0.228	18	0.093	241	0.893	0.758	0.892	0.756
Zinc	Caspyrus1k	0.156	34	0.220	247	0.857	0.268	0.846	0.465
	Caspyrus10k	0.112	14	0.061	221	0.889	0.540	0.872	0.573
	Caspyrus20k	0.097	14	0.266	221	0.892	0.571	0.880	0.604
	Caspyrus50k	0.141	16	0.056	243	0.910	0.657	0.890	0.648
	ChEMBL200k	0.167	18	0.090	245	0.924	0.760	0.920	0.750

Appendix C In-house de novo drug design

Table C4 Generated synthesizable and potentially active molecules using in-house synthesizability scores. Evaluation of 100,000 molecules generated per selected QSAR model and CASP-based synthesizability score combination. "Solved" denotes the successful identification of a synthesis route for a particular molecule with the respective building blocks (in-house Led3 and Zinc), while "Active" is measured by the QSAR model with a probability threshold of greater than 0.8.

Synthesizability Score	Building Blocks	Solved	Solved & Active
QSAR Only	Led3	1,468	762
QSAR Only	Zinc	1,635	883
SAScore	Led3	58,052	2,447
SAScore	Zinc	62,452	2,851
Led3-Caspyrus10k	Led3	35,697	25,044
Led3-Caspyrus10k	Zinc	42,564	30,071
Zinc-Caspyrus10k	Led3	35,102	24,912
Zinc-Caspyrus10k	Zinc	35,084	24,575
Led3-ChEBML200k	Led3	30,650	22,202
Led3-ChEBML200k	Zinc	30,765	21,732
Zinc-ChEBML200k	Led3	43,655	24,109
Zinc-ChEBML200k	Zinc	48,078	26,554

Table C5 Virtual Screening results with post-generation filtering. Virtual Screening outcomes for the 100,000 generated molecules from each trained DrugEx model, employing training objectives as post-generation filters. The 100,000 generated molecules are filtered by the QSAR model, the respective synthesizability score, and both combined (desired molecules). These desired molecules are evaluated with synthesis planning using both Led3 and Zinc building blocks. Filter thresholds are set at > 0.8 for both the QSAR and synthesizability models and ≤ 4.5 for the SAScore.

	QSAR Model Filter			Desired Solved	
RL Training Objectives	QSAR Only	SAScore	Both (Desired)	Led3 BB	Zinc BB
QSAR Only	76,828	-	-	762	883
SAScore	7,689	95,331	6,089	2,420	2,808
Led3-Caspyrus10k	67,407	48,093	32,907	20,055	23,054
Led3-ChEMBL200k	64,913	26,229	18,307	14,077	13,597
Zinc-Caspyrus10k	66,628	66,664	48,338	22,419	22,058
Zinc-ChEMBL200k	51,310	54,668	27,775	19,780	21,440

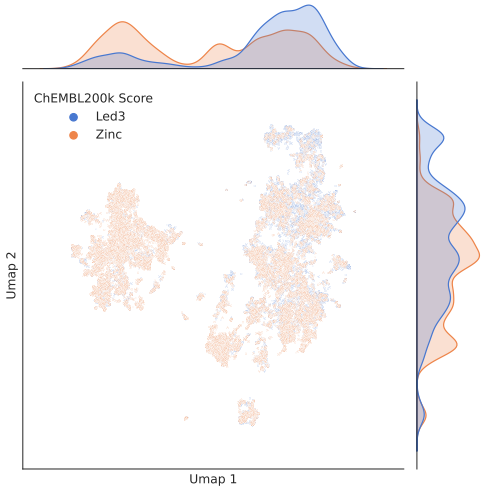

Fig. C1 Contrasting the shared generated chemical space of in-house and general synthesizability scores using ChEMBL200k. UMAP visualization of the solved and potentially active molecular space derived from combining the molecules generated from both in-house and general synthesizability scores trained on the same dataset ("ChEMBL200k"). In both instances, in-house building blocks are used for synthesis planning to evaluate solvability. UMAP is calculated using Morgan Fingerprints (Radius 3, Size 2048).

Table C6 Out-of-distribution predictive performance of synthesizability scores on the explored chemical space. Evaluation of the predictive performance of CASP-based synthesizability scores on de novo generated molecules. The predictive performance of each score is evaluated by synthesis planning using the building blocks specific to each score's training.

Synthesizability Score	Accuracy	Precision	Recall	F1	MCC
Led3-Caspyrus10k	0.696	0.545	0.892	0.677	0.465
Led3-ChEMBL200k	0.809	0.653	0.804	0.720	0.585
Zinc-Caspyrus10k	0.496	0.408	0.965	0.573	0.263
Zinc-ChEMBL200k	0.716	0.641	0.929	0.759	0.486

Table C7 DrugEx RL objective modifier functions and class decision thresholds. This table details the parameters of different modifier functions used within DrugEx, including their lower and upper bounds (Lower_x, Upper_x) and class decision thresholds. Detailed descriptions of modifier functions are available in Table S2 [53].

DrugEx Objective	Modifier Function	Lower_x	Upper_x	Class Decision Threshold
QSAR Classifier	ClippedScore	0.2	0.8	0.5
SAScore	SmoothClippedScore	7	4	0.5
Led3-based Score	ClippedScore	0.2	0.8	0.5
Zinc-based Score	ClippedScore	0.2	0.8	0.5

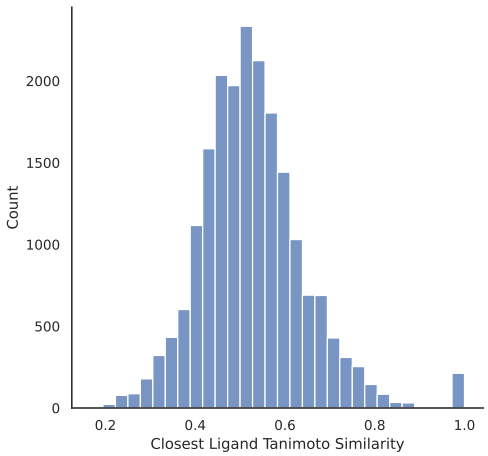

Fig. C2 Tanimoto Similarity to the closest known ligands. Evaluated on the solved virtual screening results using the in-house Caspyrus10k model, determined using Morgan Fingerprints (Radius 3, Size 2048).

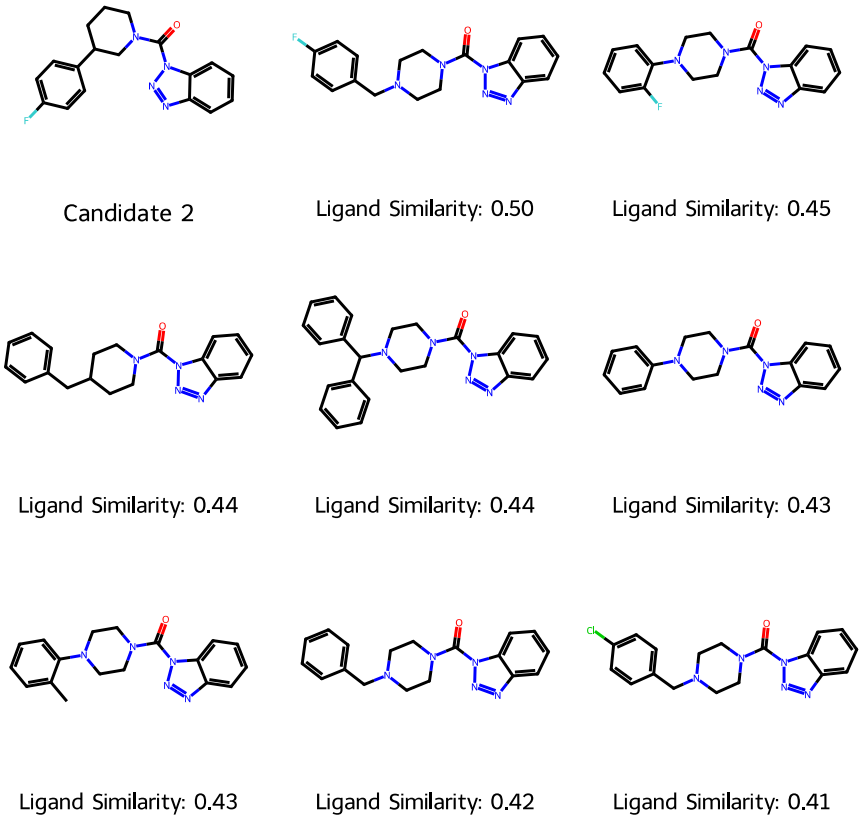

Fig. C3 Closest known ligands compared to most active candidate 2. Measured by Tanimoto similarity on Morgan Fingerprints (Radius 3, Size 2048).

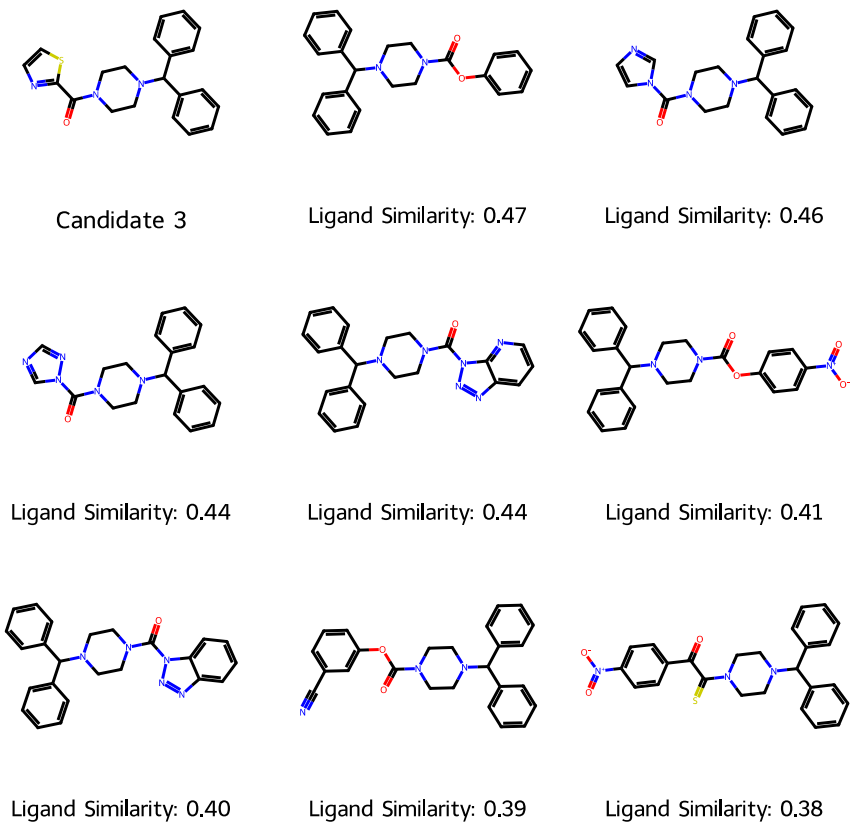

Fig. C4 Closest known ligands compared to most active candidate 3. Measured by Tanimoto similarity on Morgan Fingerprints (Radius 3, Size 2048).

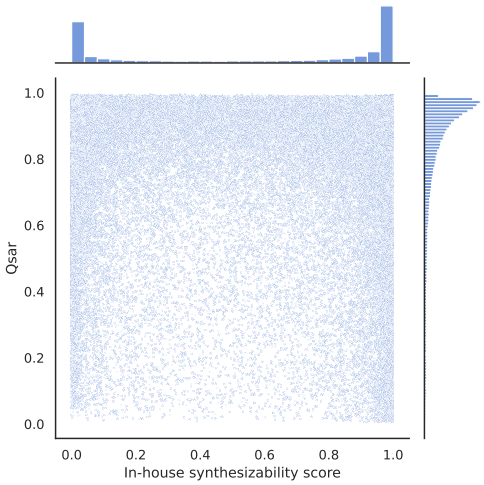

Fig. C5 Pareto Front of the generated molecules using DrugEx. QSAR indicates the perceived activity with respect to our protein target, in-house synthesizability score indicates the synthesizability perceived by our in-house Caspyrus10k synthesizability score.

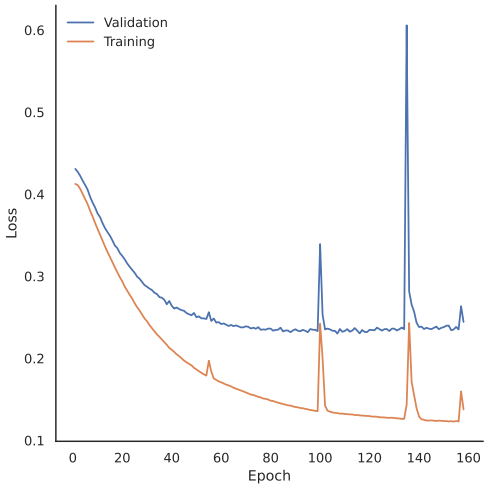

Fig. C6 DrugEx Fine-Tuning Loss. Validation and Training loss during DrugEx fine-tuning of the domain-specific ligand space model.

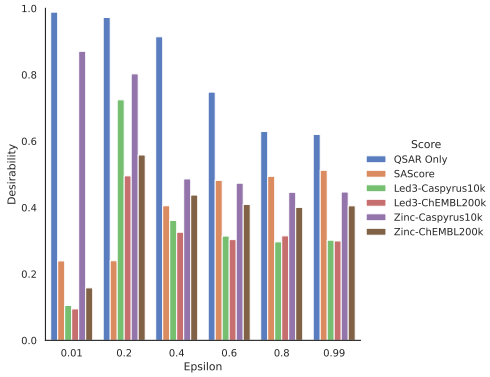

Fig. C7 DrugEx epsilon parameter trade-offs. Maximum Desirability achieved by each model during RL for varying values of the exploration parameter epsilon. The parameter epsilon captures the fraction of data that is sampled during each training iteration from the fixed fine-tuned ligand space model instead of the actively trained model.

Table C8 QSAR model hyperparameter optimization. Overview of tested classifiers, different hyperparameters, possible grid search values and the best-found values.

Algorithm	Hyperparameter	Possible Values	Best
RandomForest	n_estimators	50, 200, 1000	1000
	criterion	gini, entropy, log_loss	log_loss
	min_samples_split	30, 2, 0.1, 0.05	2
	min_samples_leaf	30, 1, 0.1, 0.05	1
	max_features	sqrt, log2	sqrt
	class_weight	balanced, balanced_subsample, None	balanced
	max_samples	0.3, 0.7, 1.0	1.0
	ccp_alpha	0.005, 0.05, 0.1	0.005
ExtraTrees	n_estimators	50, 200, 1000	200
	criterion	gini, entropy, log_loss	log_loss
	min_samples_split	30, 2, 0.1, 0.05	2
	min_samples_leaf	30, 1, 0.1, 0.05	1
	max_features	sqrt, log2	sqrt
	class_weight	balanced, balanced_subsample, None	balanced
	max_samples	0.3, 0.7, 1.0	1.0
	ccp_alpha	0.005, 0.05, 0.1	0.005
XGBoost	learning_rate	0.001, 0.01, 0.1, 0.3, 1.0	0.1
	max_depth	5, 10, 50, 100	50
	n_estimators	50, 200, 1000	50
	colsample_bytree	0.3, 0.7, 1.0	0.3
	colsample_bylevel	0.3, 0.7, 1.0	0.3
	colsample_bynode	0.3, 0.7, 1.0	1.0
	lambda	0.0, 0.5, 1, 2, 5	0
	alpha	0.0, 0.5, 1, 2, 5	0
Multi-Layer Perceptron	hidden_layer_sizes	(50,), (50, 50), (100,), (100, 100), (500,), (500, 500), (500, 100), (100, 500)	(100, 500)
	alpha	0.0001, 0.001, 0.01	0.001
	early_stopping	True, False	False
	max_iter	50, 100, 200, 1000	50
GradientBoosting	n_estimators	50, 200, 1000	200
	min_samples_split	30, 2, 0.1, 0.05	0.1
	min_samples_leaf	30, 1, 0.1, 0.05	1
	max_features	sqrt, log2	sqrt
	ccp_alpha	0.005, 0.05, 0.1	0.005
	loss	log_loss, exponential	exponential
	learning_rate	0.001, 0.01, 0.1, 0.3, 1.0	0.1
	subsample	0.3, 0.7, 1.0	0.7
	n_iter_no_change	1, 5	5
	tol	0.0001, 0.001, 0.01, 0.1	0.0001
AdaBoost	n_estimators	50, 200, 1000	1000
	learning_rate	0.1, 1.0, 2.0, 5.0	0.1
KNN	n_neighbors	1, 3, 5, 10	5
	weights	uniform, distance	distance
	metric	cityblock, manhattan, euclidean, cosine	cityblock
SVC	C	0.5, 1.0, 5.0	5.0
	kernel	linear, poly, rbf, sigmoid	linear
Gaussian Naive Bayes	var_smoothing	1e-9, 1e-6	1e-6

Appendix D Experimental Evaluation

D.1 Biochemistry experimental

D.1.1 Cloning, overexperssion and membrane preparation

Full-length cDNA encoding human MGLL (GenBank ID: BC006230.2; obtained from Source Bioscience) was amplified by PCR and cloned into expression vector pcDNA3.1 in frame with a C-terminal FLAG-tag. All plasmids were isolated from transformed XL10-Gold competent cells (prepared using E. coli transformation buffer set; Zymo Research) using plasmid isolation kits following the supplier's protocol (Qiagen). Constructs were verified by Sanger sequencing (Macrogen).

HEK293T (human embryonic kidney) cells were obtained from ATCC and tested on regular basis for mycoplasma contamination. Cultures were discarded after 2-3 months of use. Cells were cultured at 37 °C under 7% CO_2 in high-glucose DMEM containing phenol red, stable glutamine, 10% (v/v) high iron newborn calf serum (Seradigm), penicillin and streptomycin (200 μg/mL each; Duchefa). Medium was refreshed every 2-3 days and cells were passaged two times a week at 80-90% confluence. One day prior to transfection, HEK293T cells were transferred from confluent 10 cm dishes to 15 cm dishes. Before transfection, medium was refreshed (13 mL). A 3:1 mixture of polyethyleneimine (PEI; 60 μg/dish) and plasmid DNA (20 μg/dish) was prepared in serum-free medium (2 mL) and incubated for 15 min at RT. The mixture was then added dropwise to the cells, after which the cells were grown to confluence in 72 h. Cells were then harvested by suspension in PBS, followed by centrifugation (200 g, 5 min). Cell pellets were flash-frozen in liquid nitrogen and stored at -80 °C.

Cell pellets were thawed on ice and resuspended in lysis buffer A (20 mM HEPES (pH 7.2), 2 mM DTT, 250 mM sucrose, 1 mM $MgCl_2$, and 25 U/ml benzonase). Suspensions were homogenized by polytron (3 × 7 s, 20,000 rpm, SilentCrusher S; Heidolph, Schwabach, Germany), incubated on ice for 30 min, and subsequently centrifuged at 93,000 g for 30 min at 4°C (Ti70 or Ti70.1 rotor; Beckman Coulter, Woerden, The Netherlands). Pellet was resuspended in storage buffer B (20 mM HEPES (pH 7.2), 2 mM DTT)]. Suspension was homogenized by polytron (1 × 10 s, 20,000 rpm). Protein concentrations were determined with Quick Start Bradford reagent (Bio-Rad, Hilversum, The Netherlands) or Qubit fluorometric quantitation (Life Technologies, Breda, The Netherlands). Membranes were diluted with storage buffer B to the desired concentration, aliquoted, frozen in liquid nitrogen, and stored at -80°C.

D.1.2 Biochemical evaluation of MGLL inhibitors

Assays were performed in HEMNB buffer (50 mM HEPES pH 7.4, 1 mM EDTA, 5 mM $MgCl_2$, 100 mM NaCl, 0.5% (w/w) BSA) in black, flat-bottom 96-well plates (Greiner). Inhibitors were added from 40x concentrated stock solution in DMSO. MGLL-overexpressing membrane preparations (0.3 μg per well) were incubated with inhibitor for 20 min at RT in a total volume of 100 μL. Next, 100 μL assay mix containing glycerol kinase (GK), glycerol-3-phosphate oxidase (GPO), horse radish peroxidase (HRP), adenosine triphosphate (ATP), Amplifu™Red and 2-arachidonoylglycerol (2-AG) was added. Fluorescence (λ_{ex} = 535 nm, λ_{em} = 595 nm) was measured at RT in 5

min intervals for 60 min on a Clariostar (BMG Labtech) plate reader. Final assay concentrations: 1.5 ng/μL MGLL-overexpressing membranes, 0.2 U/mL GK, GPO and HRP, 125 μM ATP, 10 μM Amplifu™Red, 25 μM 2-AG, 5% DMSO, 0.5% ACN in a total volume of 200 μL. For IC_{50} determinations, the assay was performed as described above, but with variable inhibitor concentrations. All measurements were performed in N = 2 (individual plates), n = 2 (technical replicates on same plate) or N = 2, n = 4 for controls. Fluorescence values were corrected for the average fluorescence of the negative control (mock-membranes + vehicle). Slopes of the corrected data were determined in the linear interval. The Z'-factor for each assay plate was calculated using the formula $Z' = 1 - 3(\sigma_{pc} + \sigma_{nc})/(\mu_{pc} - \mu_{nc})$ with σ = standard deviation, μ = mean, pc = positive control and nc = negative control, and plates with $Z' \geq 0.6$ were accepted for further analysis. For IC_{50} determination, slopes were normalized to the positive control and analysed in a non-linear dose-response analysis with variable slope (GraphPad Prism 9.0).

Appendix E Chemistry Experimental

E.1 General chemistry

All used glassware was oven dried. Reagents were either acquired from Sigma-Aldrich, Acros and Merck and used without further purification unless specified otherwise. Moisture sensitive reactions were performed under a nitrogen atmosphere using anhydrous solvents dried over activated molecular sieves (4 Å). Traces of water were removed from starting materials through co-evaporation with toluene. Thin layer chromatography (TLC) was performed using TLC Silica gel 60 F_{245} on aluminums sheets (Merck). Compounds were visualized using an ultraviolet lamp (λ_{max} = 254 nm), $KMnO_4$ staining (K_2CO_3 (40 g), MnO_4 (6 g), H_2O (600 mL) and 10% NaOH (5 mL)) or ninhydrin staining (ninhydrine (200 mg), AcOH (5 mL) and EtOH (100 mL). The crude compounds were purified by either flash column chromatography using Screening Devices silica gel 60, or automated flash column chromatography using Biotage Isolera One or Four Flash Chromatography Systems and pre-packed cartridges of Screening Devices UltraPure Irregular Silica Gel (40 – 63 μm, 60 Å). LC-MS measurements were performed on a Thermo Finnigan LCQ Advantage Max ion-trap mass spectrometer (ESI+), coupled to a Surveyor HPLC system (Thermo Finnigan) or a Thermo Vanquish Focused UHPLC$^+$ system, equipped with a C18 column and coupled to a Thermo LCQ Fleet ion-trap mass-spectrometer. Both LC-MS systems were equipped with a standard C18 (Gemini, 4.6 mmD × 50 mmL, 5 μm particle size, Phenomenex) analytical column. Eluents A: H_2O, B: ACN, C: 1% aq. TFA, gradients: 10 – 90% or a 0 – 50% gradient of ACN in water with 0.1% TFA. A Bruker AV-400 Cryomagnet was used to obtain proton(^{1}H)-NMR and carbon(^{13}C)-NMR. Chemical shifts (δ) are reported are reported in parts per million (ppm) downfield of tetramethylsilane (TMS) or solvent resonance as the internal standard ($CDCl_3$: δ 7.26 for ^{1}H, δ 77.16 for ^{13}C, CD_3OD: δ 3.31 for ^{1}H, δ 49.00 for ^{13}C). Splitting patterns reported in an abbreviated manner (s = singlet, d = doublet, t = triplet, q = quartet, and m = multiplet), coupling constants (J) are quoted in Hertz (Hz). Peak assignments were aided by 2D COSY, HSQC, and HMBC experiments. MestReNova software (version 14.0.1-23559)

was used for the analysis of the NMR spectra. PerkinElmer ChemDraw Professional (version 22.2) was used to draw molecular structures presented in this work.

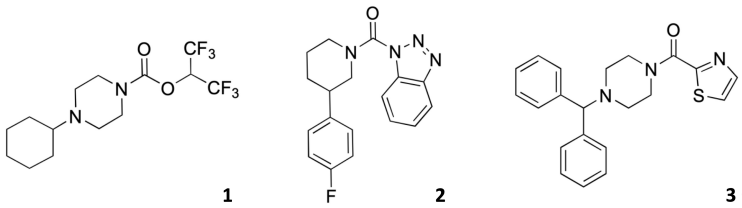

Fig. E8 Experimentally evaluated candidates.

E.1.1 1,1,1,3,3,3-hexafluoropropan-2-yl 4-cyclohexylpiperazine-1-carboxylate (1)

Triphosgene (0.5 eq., 89 mg, 0.30 mmol) and Na_2CO_3 (1 eq., 63 mg, 0.60 mmol) were dissolved in DCM (3.0 mL, 0.1M) under argon and stirred on ice. Subsequently 1-cyclohexylpiperazine (1 eq., 100 mg, 0.60 mmol) in DCM (3 mL) was added and the obtained mixture was stirred at 0 ºC for 1 h. Upon full conversion to the carbamoyl chloride intermediate, the reaction mixture was filtered, rinsed with DCM (10 mL) and concentrated under reduced pressure. Hexafluoroisopropanol (1 eq., 100 mg, 0.60 mmol) and DiPEA (2 eq., 155 mg, 1.20 mmol) were dissolved in DCM (3 mL, 0.2 M) and added drop wisely to the carbamoyl chloride in DCM (3 mL). The obtained solution was stirred under argon atmosphere for 19 h. Upon full conversion of the starting materials, the reaction mixture was quenched with sat. NH_4Cl (2 mL) and brought to pH 7 using 1M NaOH (1 mL). Subsequently, the neutral solution was extracted with DCM (3x 10 mL), after which the combined organic layers were dried ($MgSO_4$) and concentrated under reduced pressure. After purification with column chromatography eluting with isocratic DCM, the title compound was partially isolated as a white solid (43.0 mg, 0.12 mmol, 20%). ^{1}H NMR (400 MHz, $CDCl_3$) δ 5.75 (p, J = 6.2 Hz, 1H), 3.95 – 3.25 (m, 4H), 2.83 – 2.49 (m, 5H), 2.52 – 2.24 (m, 1H), 2.03 – 1.79 (m, 5H), 1.65 (dd, J = 13.0, 3.6 Hz, 1H), 1.48 – 1.15 (m, 5H), 1.16 – 0.99 (m, 1H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 151.41, 120.80 (t, J = 285.0 Hz), 68.15 (hept, J = 136 Hz), 64.02, 49.07, 48.99, 48.89, 48.66, 48.42, 46.63, 44.93, 44.56, 28.71, 26.20, 25.83.

E.1.2 (1H-benzo[d][1,2,3]triazol-1-yl) (3-(4-fluorophenyl)piperidin-1-yl)methanone (2)

3-(4-Fluorophenyl)piperidine (1 eq., 100 mg, 0.56 mmol) and Na_2CO_3 (1 eq., 46.3 mg, 0.56 mmol) were dissolved in DCM (5.6 mL, 0.1) under argon and stirred on ice. Subsequently, triphosgene (0.5 eq., 82.8 mg, 0.28 mmol) was added and the obtained mixture was stirred at 0 ºC for 1 h. Upon full conversion to the carbamoyl chloride intermediate, the reaction mixture was filtered, rinsed with DCM (10 mL) and concentrated under reduced pressure. To the obtained solid the benzotriazole 1H-benzo[d][1,2,3]triazole (1 eq., 66.5 mg, 0.56 mmol) were dissolved in DCM (2.8 mL,

0.2 M), after which DiPEA (2 eq., 144 mg, 1.12 mmol) was added drop wisely and the obtained solution was stirred under argon atmosphere for 19 h. Upon full conversion of the starting materials, the reaction mixture was quenched with sat. NH_4Cl (2 mL) and brought to pH 7 using 1M NaOH (1 mL). Subsequently, the neutral solution was extracted with DCM (3x 10 mL), after which the combined organic layers were dried ($MgSO_4$) and concentrated under reduced pressure. After purification with column chromatography eluting with isocratic DCM, the title compound was partially isolated as a translucent oil (42.0 mg, 0.13 mmol, 23%). ^{1}H NMR (400 MHz, CD_2Cl_2) δ 8.04 (dd, J = 42.5, 8.3 Hz, 2H), 7.52 (dt, J = 60.0, 7.6 Hz, 2H), 7.29 – 7.21 (m, 2H), 7.01 (t, J = 8.2 Hz, 2H), 4.62 (d, J = 15.9 Hz, 2H), 3.47 – 2.74 (m, 3H), 2.17 (d, J = 13.7 Hz, 1H), 1.98 – 1.78 (m, 2H). ^{13}C NMR (101 MHz, CD_2Cl_2) δ 161.83 (d, J = 245.0 Hz), 149.47, 145.46, 138.06, 133.28, 129.42, 128.62 (d, J = 7.9 Hz), 125.29, 119.91, 115.56 (d, J = 21.1 Hz), 113.56, 42.12, 31.59, 25.80.

E.1.3 (4-benzhydrylpiperazin-1-yl)(thiazol-2-yl)methanone (3)

The free amine 1-benzhydrylpiperazine (1 eq., 200 mg, 0.79 mmol), PyAOP (2 eq., 597 mg, 1.59 mmol) and the benzoic acid thiazole-2-carboxylic acid (1 eq., 102 mg, 0.79 mmol) were dissolved in DMF (0.4 M), after which DiPEA (4 eq., 0.56 ml, 1.59 mmol) was added dropwisely and the mixture was stirred for 21 h. Upon reaction completion, the solution was dissolved in EtOAc (10 mL) and washed with brine (2 x 10 mL), after which the aqueous layer was extracted with EtOAc (3x 10 mL). The combined organic layers were dried ($MgSO_4$), filtered and concentrated under reduced pressure and the remaining oil was further purified by silica gel column chromatography eluting with a gradient of 0-40% ether in pentane to obtain the title compound as a pale-yellow solid (259 mg, 712 mmol, 90%). ^{1}H NMR (400 MHz, $CDCl_3$) δ 7.77 (d, J = 3.2 Hz, 1H), 7.45 – 7.39 (m, 5H), 7.31 – 7.22 (m, 4H), 7.21 – 7.13 (m, 2H), 4.39 (t, J = 5.0 Hz, 2H), 4.25 (s, 1H), 3.80 (t, J = 5.1 Hz, 2H), 2.48 (dt, J = 11.8, 5.0 Hz, 4H). ^{13}C NMR (101 MHz, $CDCl_3$) δ 165.21, 159.05, 143.03, 142.13, 128.62, 127.89, 127.16, 123.91, 75.96, 52.40, 51.73, 46.55, 43.62.

References

[1] Vijayan, RSK, Kihlberg, J, Cross, JB, Poongavanam, V (2022) Enhancing preclinical drug discovery with artificial intelligence. Drug Discovery Today **27**(4), 967–984 https://doi.org/10.1016/j.drudis.2021.11.023

[2] Moret, M, Pachon Angona, I, Cotos, L, Yan, S, Atz, K, Brunner, C, Baumgartner, M, Grisoni, F, Schneider, G (2023) Leveraging molecular structure and bioactivity with chemical language models for de novo drug design. Nature Communications **14**(1), 114 https://doi.org/10.1038/s41467-022-35692-6

[3] Ballarotto, M, Willems, S, Stiller, T, Nawa, F, Marschner, JA, Grisoni, F, Merk, D (2023) De Novo Design of Nurr1 Agonists via Fragment-Augmented Generative Deep Learning in Low-Data Regime. Journal of Medicinal Chemistry **66**(12), 8170–8177 https://doi.org/10.1021/acs.jmedchem.3c00485

[4] Stanley, M, Segler, M (2023) Fake it until you make it? Generative de novo design and virtual screening of synthesizable molecules. Current Opinion in Structural Biology **82**, 102658 https://doi.org/10.1016/j.sbi.2023.102658

[5] Anstine, DM, Isayev, O (2023) Generative Models as an Emerging Paradigm in the Chemical Sciences. Journal of the American Chemical Society **145**(16), 8736–8750 https://doi.org/10.1021/jacs.2c13467

[6] Nicolaou, CA, Brown, N (2013) Multi-objective optimization methods in drug design. Drug Discovery Today: Technologies **10**(3), 427–435 https://doi.org/10.1016/j.ddtec.2013.02.001

[7] Gómez-Bombarelli, R, Wei, JN, Duvenaud, D, Hernández-Lobato, JM, Sánchez-Lengeling, B, Sheberla, D, Aguilera-Iparraguirre, J, Hirzel, TD, Adams, RP, Aspuru-Guzik, A (2018) Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Central Science **4**(2), 268–276 https://doi.org/10.1021/acscentsci.7b00572

[8] Blaschke, T, Arús-Pous, J, Chen, H, Margreitter, C, Tyrchan, C, Engkvist, O, Papadopoulos, K, Patronov, A (2020) REINVENT 2.0: An AI Tool for De Novo Drug Design. Journal of Chemical Information and Modeling **60**(12), 5918–5922 https://doi.org/10.1021/acs.jcim.0c00915

[9] Winter, R, Montanari, F, Steffen, A, Briem, H, Noé, F, Clevert, DqA (2019) Efficient multi-objective molecular optimization in a continuous latent space. Chem. Sci. **10**(34), 8016–8024 https://doi.org/10.1039/C9SC01928F

[10] Liu, X, Ye, K, van Vlijmen, HWT, Emmerich, MTM, IJzerman, AP, van Westen, GJP (2021) DrugEx v2: De novo design of drug molecules by Pareto-based multi-objective reinforcement learning in polypharmacology. Journal of Cheminformatics **13**(1), 85 https://doi.org/10.1186/s13321-021-00561-9

[11] Liu, X, Ye, K, van Vlijmen, HWT, IJzerman, AP, van Westen, GJP (2023) DrugEx v3: Scaffold-constrained drug design with graph transformer-based reinforcement learning. Journal of Cheminformatics **15**(1), 24 https://doi.org/10.1186/s13321-023-00694-z

[12] Méndez-Lucio, O, Baillif, B, Clevert, DqA, Rouquié, D, Wichard, J (2020) De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. Nature Communications **11**(1), 10 https://doi.org/10.1038/s41467-019-13807-w

[13] Hoogeboom, E, Satorras, VG, Vignac, C, Welling, M (2022) Equivariant Diffusion for Molecule Generation in 3D. In: Chaudhuri, K, Jegelka, S, Song, L, Szepesvari, C, Niu, G, Sabato, S (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 8867–8887. PMLR, Baltimore, Maryland, USA

[14] Segler, MHS, Preuss, M, Waller, MP (2018) Planning chemical syntheses with deep neural networks and symbolic AI. Nature **555**(7698), 604–610 https://doi.org/10.1038/nature25978

[15] Corey, EJ, Cheng, XqM (1989) The Logic of Chemical Synthesis. John Wiley & Sons, Ltd, New York

[16] Schwaller, P, Vaucher, AC, Laplaza, R, Bunne, C, Krause, A, Corminboeuf, C, Laino, T (2022) Machine intelligence for chemical reaction space. WIREs Computational Molecular Science **12**(5), 1604 https://doi.org/10.1002/wcms.1604

[17] Gao, W, Coley, CW (2020) The Synthesizability of Molecules Proposed by Generative Models. Journal of Chemical Information and Modeling **60**(12), 5714–5723 https://doi.org/10.1021/acs.jcim.0c00174

[18] Genheden, S, Thakkar, A, Chadimová, V, Reymond, JL, Engkvist, O, Bjerrum, E (2020) AiZynthFinder: A fast, robust and flexible open-source software for retrosynthetic planning. Journal of Cheminformatics **12**(1), 70 https://doi.org/10.1186/s13321-020-00472-1

[19] Chen, B, Li, C, Dai, H, Song, L (2020) Retro*: Learning Retrosynthetic Planning with Neural Guided A* Search. In: III, HD, Singh, A (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 1608–1616. PMLR, Virtual

[20] Yu, Y, Wei, Y, Kuang, K, Huang, Z, Yao, H, Wu, F, Koyejo, S, Mohamed, S, Agarwal, A, Belgrave, D, Cho, K, Oh, A (2022) GRASP: Navigating Retrosynthetic Planning with Goal-driven Policy. In: Advances in Neural Information Processing Systems, vol. 35, pp. 10257–10268. Curran Associates, Inc., New Orleans, Louisiana, USA

[21] Parrot, M, Tajmouati, H, da Silva, VBR, Atwood, BR, Fourcade, R, Gaston-Mathé, Y, Do Huu, N, Perron, Q (2023) Integrating synthetic accessibility with AI-based generative drug design. Journal of Cheminformatics **15**(1), 83 https://doi.org/10.1186/s13321-023-00742-8

[22] Hassen, AK, Torren-Peraire, P, Genheden, S, Verhoeven, J, Preuss, M, Tetko, I (2022) Mind the Retrosynthesis Gap: Bridging the divide between Single-step and Multi-step Retrosynthesis Prediction. In: NeurIPS 2022 AI for Science: Progress and Promises

[23] Torren Peraire, P, Hassen, AK, Genheden, S, Verhoeven, J, Clevert, DqA, Preuss, M, Tetko, IV (2024) Models Matter: The Impact of Single-Step Retrosynthesis on Synthesis Planning. Digital Discovery https://doi.org/10.1039/D3DD00252G

[24] Urbina, F, Lowden, CT, Culberson, JC, Ekins, S (2022) MegaSyn: Integrating Generative Molecular Design, Automated Analog Designer, and Synthetic Viability Prediction. ACS Omega **7**(22), 18699–18713 https://doi.org/10.1021/acsomega.2c01404

[25] Ertl, P, Schuffenhauer, A (2009) Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. Journal of Cheminformatics **1**(1), 8 https://doi.org/10.1186/1758-2946-1-8

[26] Thakkar, A, Chadimová, V, Bjerrum, EJ, Engkvist, O, Reymond, JL (2021) Retrosynthetic accessibility score (RAscore)-rapid machine learned synthesizability classification from AI driven retrosynthetic planning. Chem. Sci. **12**(9), 3339–3349 https://doi.org/10.1039/D0SC05401A

[27] Yu, J, Wang, J, Zhao, H, Gao, J, Kang, Y, Cao, D, Wang, Z, Hou, T (2022) Organic Compound Synthetic Accessibility Prediction Based on the Graph Attention Mechanism. Journal of Chemical Information and Modeling **62**(12), 2973–2986 https://doi.org/10.1021/acs.jcim.2c00038

[28] Liu, CqH, Korablyov, M, Jastrzebski, S, Włodarczyk-Pruszyński, P, Bengio, Y, Segler, M (2022) RetroGNN: Fast Estimation of Synthesizability for Virtual Screening and De Novo Design by Learning from Slow Retrosynthesis Software. Journal of Chemical Information and Modeling **62**(10), 2293–2300 https://doi.org/10.1021/acs.jcim.1c01476

[29] Brown, N, Fiscato, M, Segler, MHS, Vaucher, AC (2019) GuacaMol: Benchmarking Models for de Novo Molecular Design. Journal of Chemical Information and Modeling **59**(3), 1096–1108 https://doi.org/10.1021/acs.jcim.8b00839

[30] Luukkonen, S, Van Den Maagdenberg, HW, Emmerich, MTM, Van Westen, GJP (2023) Artificial intelligence in multi-objective drug design. Current Opinion in Structural Biology **79**, 102537 https://doi.org/10.1016/j.sbi.2023.102537

[31] The UniProt Consortium Q99685 | MGLL | Monoglyceride Lipase | Homo Sapiens (Human) | UniProt (2023). https://www.uniprot.org/uniprotkb/Q99685/entry Accessed 2023-10-24

[32] Béquignon, OJM, Bongers, BJ, Jespers, W, IJzerman, AP, van der Water, B, van Westen, GJP (2023) Papyrus: A large-scale curated dataset aimed at bioactivity predictions. Journal of Cheminformatics **15**(1), 3 https://doi.org/10.1186/s13321-022-00672-x

[33] Butina, D (1999) Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. Journal of Chemical Information and Computer Sciences **39**(4), 747–750 https://doi.org/10.1021/ci9803381

[34] Mendez, D, Gaulton, A, Bento, AP, Chambers, J, De Veij, M, Félix, E, Magariños, MP, Mosquera, JF, Mutowo, P, Nowotka, M, Gordillo-Marañón, M, Hunter, F, Junco, L, Mugumbate, G, Rodriguez-Lopez, M, Atkinson, F, Bosc, N, Radoux, CJ, Segura-Cabrera, A, Hersey, A, Leach, AR (2018) ChEMBL: Towards direct deposition of bioassay data. Nucleic Acids Research **47**(D1), 930–940 https://doi.org/10.1093/nar/gky1075

[35] Chen, T, Guestrin, C (2016) XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16, pp. 785–794. Association for Computing Machinery, San Francisco, California, USA. https://doi.org/10.1145/2939672.2939785

[36] Matthews, BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure **405**(2), 442–451 https://doi.org/10.1016/0005-2795(75)90109-9

[37] Chicco, D, Jurman, G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics **21**(1), 6 https://doi.org/10.1186/s12864-019-6413-7

[38] McInnes, L, Healy, J, Saul, N, Großberger, L (2018) UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software **3**(29), 861 https://doi.org/10.21105/joss.00861

[39] Lipinski, CA, Lombardo, F, Dominy, BW, Feeney, PJ (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settingsq. Advanced Drug Delivery Reviews **23**(1), 3–25 https://doi.org/10.1016/S0169-409X(96)00423-1

[40] Papadopoulos, K, Giblin, KA, Janet, JP, Patronov, A, Engkvist, O (2021) De novo design with deep generative models based on 3D similarity scoring. Bioorganic & Medicinal Chemistry **44**, 116308 https://doi.org/10.1016/j.bmc.2021.

116308

[41] Eberhardt, J, Santos-Martins, D, Tillack, AF, Forli, S (2021) AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. Journal of Chemical Information and Modeling **61**(8), 3891–3898 https://doi.org/10.1021/acs.jcim.1c00203

[42] Corso, G, Stark, H, Jing, B, Barzilay, R, Jaakkola, TS (2023) DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. In: The Eleventh International Conference on Learning Representations

[43] Segler, MHS, Waller, MP (2017) Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. Chemistry - A European Journal **23**(25), 5966–5971 https://doi.org/10.1002/chem.201605499

[44] Lowe, DM (2012) Extraction of chemical structures and reactions from the literature. Thesis, University of Cambridge

[45] Universiteit Leiden Leiden Early Drug Discovery & Development (2023). https://www.universiteitleiden.nl/en/science/led3 Accessed 2023-10-25

[46] Enamine Ltd. Enamine Building Blocks Catalog (2023). https://enamine.net/building-blocks/building-blocks-catalog Accessed 2023-05-15

[47] Molport SIA Molport Compound Sourcing, Selling and Purchasing Platform (2023). https://www.molport.com/shop/index Accessed 2023-05-15

[48] eMolecules, Inc. eMolecules Chemical Building Blocks (2023). https://www.emolecules.com/products/building-blocks Accessed 2023-05-15

[49] Rogers, D, Hahn, M (2010) Extended-Connectivity Fingerprints. Journal of Chemical Information and Modeling **50**(5), 742–754 https://doi.org/10.1021/ci100050t

[50] Corsello, SM, Bittker, JA, Liu, Z, Gould, J, McCarren, P, Hirschman, JE, Johnston, SE, Vrcic, A, Wong, B, Khan, M, Asiedu, J, Narayan, R, Mader, CC, Subramanian, A, Golub, TR (2017) The Drug Repurposing Hub: A next-generation drug library and information resource. Nature Medicine **23**(4), 405–408 https://doi.org/10.1038/nm.4306

[51] Falkner, S, Klein, A, Hutter, F (2018) BOHB: Robust and Efficient Hyperparameter Optimization at Scale. In: Dy, JG, Krause, A (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 1437–1446. PLMR, Stockholmsmässan, Stockholm, Sweden

[52] Fromer, JC, Coley, CW (2023) Computer-aided multi-objective optimization in small molecule discovery. Patterns **4**(2) https://doi.org/10.1016/j.patter.2023.100678

[53] Šícho, M, Luukkonen, S, van Den Maagdenberg, HW, Schoenmaker, L, Béquignon, OJM, Van Westen, GJP (2023) DrugEx: Deep Learning Models and Tools for Exploration of Drug-Like Chemical Space. Journal of Chemical Information and Modeling **63**(12), 3629–3636 https://doi.org/10.1021/acs.jcim.3c00434

[54] Degen, J, Wegscheid-Gerlach, C, Zaliani, A, Rarey, M (2008) On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. ChemMedChem **3**(10), 1503–1507 https://doi.org/10.1002/cmdc.200800178

[55] van den Maagdenberg, H, Sicho, M, Schoenmaker, L, Bequignon, OJM, Luukkonen, S, Gorosiola González, M, Araripe, D QSPRPred: A Tool for Creating Quantitative Structure Property Relationship (QSPR) Models (2023). https://github.com/CDDLeiden/QSPRPred Accessed 2023-06-06

[56] Pedregosa, F, Varoquaux, G, Gramfort, A, Michel, V, Thirion, B, Grisel, O, Blondel, M, Prettenhofer, P, Weiss, R, Dubourg, V, Vanderplas, J, Passos, A, Cournapeau, D, Brucher, M, Perrot, M, Duchesnay, E (2011) Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**, 2825–2830