ARTICLE

de novo generated combinatorial library design

Simon Viet Johansson,*[a,b] , Morteza Haghir Chehreghani[b], Ola Engkvist[a,b] Alexander Schliep [b,c]

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Artificial intelligence (AI) contributes new methods for designing compounds in drug discovery, ranging from *de novo* design models suggesting new molecular structures or optimizing existing leads to predictive models evaluating their toxicological properties. However, a limiting factor for the effectiveness of AI methods in drug discovery is the lack of access to high-quality data sets leading to a focus on approaches optimizing data generation. Combinatorial library design is a popular approach for bioactivity testing as a large number of molecules can be synthesized from a limited number of building blocks. We propose a framework for designing combinatorial libraries from *de novo* generated building blocks using k-Determinantal Point Processes and Gibbs sampling. We explore optimization of biological activity, Quantitative Estimate of Drug-likeness (QED) and diversity and the trade-offs between them, both in single-objective and in multi-objective library design settings. Using retrosynthesis models to estimate building block availability, the proposed framework is able to explore the prospective benefit from expanding a stock of available building blocks by synthesis or purchase the preferred building blocks before designing a library. In simulation experiments with building block collections from all available commercial vendors near-optimal libraries could be found without synthesis of additional building blocks; in other simulation experiments we showed that even one synthesis step to increase the number of available building blocks could improve library designs when starting with an in-house building block collection of reasonable size.

Introduction

AI and AI-assisted tools have seen rapidly increased popularity in cheminformatics over the past decade. In drug discovery, these tools have impacted bioactivity prediction[1, 2], *de novo* molecular design[3-7], synthesis prediction[8-12] and toxicology prediction[13]. In turn, the demand for high-quality data has increased beyond the extent of existing data sources [14] and there is a need to facilitate a larger number of informative experiments to generate data in a standardized format. Combinatorial chemistry is a popular method for producing large collections of compounds, motivated by material efficiency and more sustainable chemistry [15, 16] since synthesis of 100 molecules using two *building blocks* per synthesis could in the worst case require 200 different building blocks, whereas a library of the same size using combinatorial chemistry would use 20 in a 10×10 design.

Library design has traditionally aimed to optimize the selection of molecules for either molecular diversity[17-19] or molecular properties like high activity towards a target or reduce lipophilicity, i.e. a *focused* library design[20-24]. A diverse library design provides a larger coverage of the chemical space and is often viewed as more 'informative', since similar molecules hypothetically would provide redundancy in the information gained[17, 25]. Focused libraries on the other hand might aim to optimize a selected lead compound[26, 27] by lowering the structural diversity and exploring similar structures to the lead compound to improve a specific property.

The space of synthetically feasible molecules is estimated to be of size 10^{60} [28], whereas traditional High-throughput screening (HTS) has the capability to physically test approximately 10^6compounds. Consequently, virtual compound libraries became the focus as the computational resources became large enough to store their chemical structures[16, 29, 30]. The virtual library CH/PMUNK[31] consists of 95 million compounds by enumerating products using common reactions from combinatorial chemistry. The virtual library REAL[32] has over 6×10^9 molecules for virtual screening that obey Lipinski's rule of 5 [33]. The GDB-17 library of small molecules enumerated by Ruddigkeit et al.[34] contains 160 billion virtual compounds with up to 17 heavy atoms. Additionally, compound suppliers also offer "synthesis on demand" building blocks of which the largest is MADE[35], a catalogue of 770 million building blocks that can be ordered and made with "over 76% success rate".

Generative models for *de novo* design offer an alternative to virtual screening or HTS, by instead generating focused selections with a smaller size[3, 36]. Several deep learning models have been proposed to generate chemical libraries in a focused manner, in particular decorating a *scaffold*[37] by suggesting which building blocks to attach to this scaffold. The Mol-GPT model showed capability to both optimize a lead, as well as decorate a scaffold[38]. STRIFE emphasized pharmacophore

[a] *Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden.*
[b] *Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden.*
[c] *Faculty of Health Sciences, Brandenburg University of Technology Cottbus-Senftenberg, Cottbus, Germany*

† Footnotes relating to the title and/or authors should appear here.
Electronic Supplementary Information (ESI) available: [details of any supplementary information available should be included here]. See DOI: 10.1039/x0xx00000x

information to decorate and optimize proteins[39]. Domenico et al. adapted the *REINVENT*[3] architecture to create focused libraries towards inhibiting NA, AChE and SARS-CoV-2[40]. LibINVENT[41] uses reinforcement learning to generate reaction-constrained decorations to input scaffolds. These methods can generate building blocks for combinatorial library design, but do not inherently offer an optimized combinatorial selection. Given a limited experimental budget, there is motivation to develop workflows for optimizing combinatorial design for novel *de novo* generated building blocks.

Methods that simultaneously optimize both diversity and molecular properties of a library have been used in several previous studies, using for example simulated annealing[42] (SA) or genetic algorithms (GA)[43-45]. These approaches provide optimization over lists of provided building blocks, or virtual libraries but cannot determine whether novel generated building blocks can be acquired or if they are only hypothetical structures impossible to synthesize in practice. As such, a design made by these models on *de novo* generated building blocks is limited by the "synthesis on demand" success rate.

A model that has proven to perform well for modelling the trade-off between quality and diversity is the *Determinantal Point Process* (DPP)[46-48]. DPPs are probabilistic models that have been argued to represent repulsion between items[49]. They are used in other application areas for text summarization[48], pose estimation[47] and diverse image selection[46], but have not yet been investigated for library design. While common methods for selecting diversity are maximizing the sum of pairwise distances[17, 45] or minimizing average pairwise similarity[44], the determinant of the similarities captures the interaction between multiple molecules simultaneously[50]. Additionally, the max-sum or min-average methods scale in time complexity quadratically with the number of building blocks in the optimization space. While the DPP has a cubic scaling, it is instead dependent on the size of the sampled library rather than the number of options.

We propose a library optimization workflow for *de novo* generated building blocks in a combinatorial fashion applying recombination[51, 52]. Using LibINVENT[41], we generate and filter building blocks that can attach to an example scaffold using specified reactions. We then use the Computer Aided Synthesis Prediction (CASP) tool AiZynthFinder[12] to evaluate all generated building blocks and their availability in the eMolecules building block platform[53] of purchasable building blocks, or estimate the number of reaction steps needed to synthesize them using template-based retrosynthesis prediction[8, 9]. We simultaneously explore and optimize the library selection for Quantitative Estimate of Drug-likeness (QED)[54], Quantitative Structure-Activity Relationship (QSAR)[1, 36, 55, 56] and Structural diversity (ECFP6)[57] using Gibbs sampling[58], conditioned on a constant size, thus sampling from a determinantal point process of constant size *k* (*k-DPP*)[59]. The workflow is model-agnostic and can be applied to any list of building blocks and any CASP tool that break down the building blocks into stock-available precursors. We apply this workflow to optimize a library from all available building blocks from eMolecules [53]. We also simulate an in-house building block store by optimizing over a subset of the available building blocks and explore the differences in optimized libraries between using available building blocks and commercially available building blocks.

The main contributions of this framework are as follows. we

- extend combinatorial library design to score *de novo* designed building blocks,
- propose the use of DPPs, in particular *k*-DPPs, to sample libraries that optimize the trade-off between quality and diversity, and
- estimate the difference in score between libraries using available building blocks and total pool of generated reactants, and estimate the potential gain from expanding the available building blocks.

Methods

The framework (see Figure 1) consists of the generation of building blocks, followed by use of retrosynthesis prediction models to estimate if the building blocks are available in a defined stock data set, or if they could be produced from this stock through synthesis. While the implementation here [https://github.com/SeemonJ/combinatorial-library-design-dpp] is specifically made to work with the open source versions of LibINVENT[60] and AiZynthfinder[61], the framework itself can be adapted to work with any metrics.

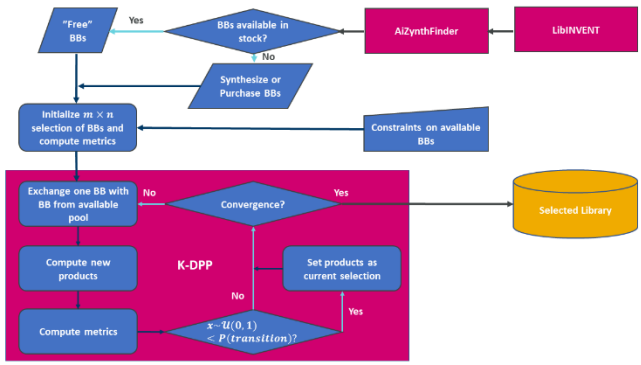

Figure 1. Flowchart of methods used for the combinatorial library design.

Application example

The scaffold displayed in Figure 2 is adapted from the original LibINVENT publication[41]. The reactions used are Buchwald-Hartwig[62] for the left attachment point and primary amide coupling[63] for the right one. We will refer to these reactions as BH and AC respectively in the following.

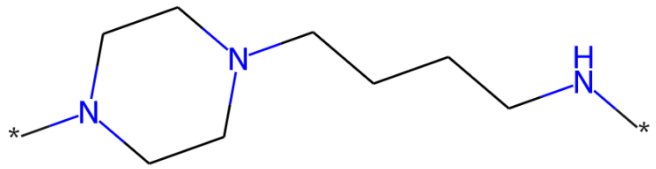

Figure 2. Scaffold used as input for the generation of building blocks. This figure is adapted from [1].

Target activity model

The QSAR model is a random forest model[64] built using Scikit-learn 0.21.3[65] with 50 estimators. The training data used is all DRD2 data available in ExcapeDB[66], with a threshold for active/inactive pXC50 of 6. Compounds from HTS assays from ChEMBL[67] without pXC50 data were assigned as inactive. With these definitions for activity, the data set had 6,304 active compounds and 344,905 inactive compounds. The compounds were represented by the extended connectivity fingerprint with 2,048 bits and radius 3 (ECFP6). The model was trained using an 80%/20% training/test data split. The data is imbalanced with most of training points labelled as inactive compounds, resulting in AUC-ROC score of 0.995 by having a pessimistic bias. This model was used both as part of the LibINVENT reinforcement learning run and during Library selection.

Building block generation using LibINVENT

The building blocks were generated using the pre-trained prior model of LibINVENT[60]. The reinforcement learning was run for 1,000 epochs with a batch size of 128 and a learning rate of 5×10^{-6}. The default diversity filter, which penalizes previously sampled building blocks, and the custom alerts for non-druglike groups were included during training. Reaction filters for the BH and AC reactions were applied, which penalize building blocks that do not match the reaction SMARTS[68].

A total of 104,991 unique molecules (82%) were generated, of which 94,808 (74%) matched the reaction filters. All molecules for which QSAR model assigned a probability of being active lower than 0.8 were removed in post-processing. This yielded 45,928 remaining products, from which the building blocks were extracted. 32,159 unique carboxylic acids and 2,084 unique aromatic halides were identified, corresponding to AC and BH reactions, respectively. The runtime was approximately 2 hours using a Nvidia 2080Ti.

Building block availability

The public version of AiZynthFinder[61] was used to check which building blocks were available directly 'in stock', and which building blocks would require synthesis to be available. The baseline stock consists of purchasable building blocks from eMolecules[53], and consists of approximately 1.5 million building blocks (including 227K carboxylic acids and 444K aromatic halides). AiZynthFinder was set to a maximum search time of 5 minutes, and maximum 10 reaction steps for identifying a synthetic route. AiZynthFinder was run in batches across multiple CPU's of varying models as performing the analysis on $\sim$34K building blocks for up to 5 minutes each would, in the worst case, require $\sim$2,800 CPU hours, in the scenario that no building blocks were available directly in stock. This analysis was performed both for the baseline stock and for five limited availability subsets, used to simulate internal stock. The limited availability subsets were sampled uniformly without replacement from the baseline stock and were chosen to be 3% of the size of the baseline size ($\sim$45k building blocks).

The parameters chosen both for generative modelling and retrosynthesis let both models run for a longer time, 1000 epochs compared to 100 during generation and 5 minutes instead of 2 for retrosynthesis evaluation, than previous uses of the same architectures[12, 41]. This yields more output building blocks and solves more routes than previous use in demonstrated studies, and potentially include LibINVENT output that could be a result of over-exploiting the QSAR model. This was done intentionally to increase the size of the search space and provide a larger diversity of building blocks with respect to quality properties to showcase the effect of the different strategies.

Determinantal Point Processes

In library design, diversity is often computed between compounds through the matrix of pairwise distances. When optimizing the library, the most common approaches maximize the sum of distances, maximize the minimum distance, or maximize the average distance to the nearest neighbour[17, 44, 45]. This captures the distance between a pair of two molecules well, but does not capture the relationships between multiple molecules simultaneously[50].

Discrete DPPs are probability distributions first used by Odile to model fermions[69], and have been increasingly popular within machine learning for capturing the trade-off between diversity and quality[46]. Let $L \in \mathbb{R}^{n \times n}$ be a positive semi-definite (PSD) matrix. A discrete DPP with *kernel* L is a probability distribution $\mu: 2^{[n]} \to \mathbb{R}_+$ defined by

$$\mu(S) \propto Det(L_S), \forall S \subseteq [n]. \qquad (1)$$

where L_S is the principal submatrix of L indexed by the elements of S. Consider that if each row of the matrix is a feature vector that represents an item, then the probability of a set of items is proportional to the volume of the hull spanned by the vectors. A diverse selection in the given features will correspond to a larger volume. For this study, the feature representation used to describe the products of the selection is the ECFP6 similar to the QSAR model, and the similarity measure described with the Tanimoto index[70] (also known as

the Jaccard index). This is well suited for application into DPPs, as the pairwise similarities L is a typical kernel[46].

Kulesza and Taskar [46] demonstrate that the quality of terms can be incorporated into DPPs by decomposing the kernel into

$$L_{i,j} = q_i \phi_i{}^T \phi_j q_j, \qquad (2)$$

where $\phi_i{}^T \phi_j$ represents the similarity between items i, j and q_i is a measure of the quality of the item. This applies to multiple quality measures and inserting equation 2 into the definition of DPP thus yields the probability for observing the set Y while sampling the DPP

$$P_L(Y) \propto (\prod_{i \in Y} q_i{}^2) Det(S_Y). \quad (3)$$

Sampling process

Evaluating the determinant of all possible products at once may introduce practical problems, since the naive implementation of determinant calculations are $O(n^3)$. This naïve implementation is used in most libraries. Due to parallelization in smaller blocks of submatrices across multiple threads, it is possible to compute determinants of matrices with $n > 10{,}000$ in minutes. For the sampled number of possible products, $32{,}159 \times 6{,}213 = 199{,}803{,}867$, it is computationally infeasible to evaluate all subsets, let alone optimize across all possible selections. For scenarios such as ours, however, the only selections of relevance are sets of practical size, such as the same sizes as screening plates, i.e., 96, 384 or 1536. *K*-DPPs are an extension of general DPPs that are conditioned to selected sets of size exactly k. Gharan and Rezaei[71] introduced a computationally efficient method for sampling *k*-DPPs using a Gibbs sampling scheme shown to have fast mixing properties. Here, the proposal distribution samples suggestions only from exchange operations between one element and one non-element of the current k-set. This ensures that the size of selection always remains constant. Moreover, at time step *t* during sampling, it requires only computation of the transition probability

$$P_L(Y_{t+1}) \propto \left(\prod_{i \in Y_{t+1}, j \in Y_t, l=G} \left(\frac{q_i}{q_j} \right)^l \right) \left(\frac{Det(S_{Y_{t+1}})}{Det(S_{Y_t})} \right)^{\omega_{div}}, \qquad (4)$$

where G is the set of quality parameters included and $\omega_{(\cdot)}$ are the respective weights for each parameter. These weights are tuneable. To give equal importance to QSAR value, QED score and diversity, we set $\omega_{QSAR} = \omega_{QED} = \omega_{div} = 0.33$ as constant. At each point t, this results in two computations of complexity $O(k^3)$ for the two determinant calculations. The following sampling scheme was implemented for selecting u and v number of building blocks from the respective sets A, B of available building blocks for two attachment points:

Algorithm 1.

1. Initialize selection with u and v building blocks at random from A, B respectively
2. Create $u \times v$ matrix of products Y_0, denote this matrix as the active set Q
3. Compute the quality values, q_{Y_0} and the matrix of pairwise similarities, S_{Y_0}
4. Compute $P_L(Y_0) \propto \left(\sum_{i \in Y_0, l \in G} \omega_l \log\left(q_i{}^2\right)\right) + \omega_{div} logDet\left(S_{Y_0}\right)$
5. Select a new building block from either A or B uniformly
6. Compute the new matrix Y_1, and the corresponding values, q_{Y_1} S_{Y_1}
7. Calculate the transition probability

$$\log\left(P_L(Y_{t+1})\right) = f\Big(\left(\sum_{i \in Y_{t+1}, l \in G} \omega_l \log\left(q_i{}^2\right)\right) + \omega_{div} logDet\left(S_{Y_{t+1}}\right) - \left(\sum_{j \in Q, l \in G} \omega_l \log\left(q_j{}^2\right)\right) - \omega_{div} logDet\left(S_Q\right)\Big), \qquad (4)$$

where,

$$f(x) = \begin{cases} 0, if\ x > 0 \\ \alpha x, otherwise \end{cases}$$

and α is a tunable parameter on the *acceptance probability*,

8. Move to the new state $Q = Y_1$ with probability $P_L(Y_{t+1})$ or stay with $Q = Y_0$ with probability $1 - P_L(Y_{t+1})$
9. Repeat steps 5-8 until termination.

Since the pairwise similarity values of S_X are all in $[0,1]$, the determinants may become too small for double precision with relevant choices of *k*. For numerical stability, the logarithm of the right hand side of equation 3 is used in step 7. The logarithm of the determinant become negative, where a greater value represents a more diverse set. In the numerical experiments we let $m = 12, n = 8$, corresponding to the generated building blocks of carboxylic acids and aromatic halides respectively, and used $k = 96$ as it is a common plate size.

We chose to conduct experiments for $\alpha = 0$ such that we only accept strict improvements (hill climbing, which is a greedy search). The selections of the model for different optimization strategies were examined, see Table 1. To explore the mixing time, the termination criteria were set as a patience parameter, sampling the distribution until 10,000 samples were drawn without finding a better solution. We compare the results against the average result of 100 random selections and the top 96 cherry-picked compounds by QSAR values from the LibINVENT run.

Results

In this section, we first show the results of processing the generated building blocks from LibINVENT through AiZynthFinder, to give a measure of the selection space for the framework. We then present the average results of each optimization strategy for different levels of availability related to required number of reaction steps. Next we show optimization results for a simulated scenario of limited stock building block availability. Finally, we discuss the computational

performance of the model when scaling up to larger selection space.

The 32,159 unique carboxylic acids and 2,084 unique aromatic halides generated through LibINVENT were analysed using AiZynthFinder. The retrosynthetic prediction found that 88.7% of the generated carboxylic acids and 98.3% of the aromatic halides could be synthesized within 2 steps of reactions from the base eMolecules stock. Of the building blocks, 6,203 carboxylic acids (19.3% of the generated building blocks) and 763 aromatic halides (36.6%) were directly available in stock; i.e., required no synthesis. The full distribution of reaction availability can be seen in Figure 3.

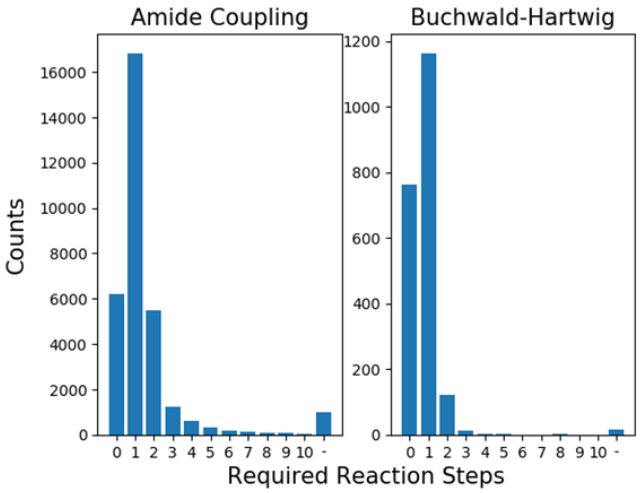

Figure 3. Distribution of number of reaction steps needed for the generated building blocks from the entire eMolecules stock. The building blocks for which a retrosynthetic route could not be found are denoted with '-'

The compound selection was performed on the criteria of only QSAR, only QED, only Diversity and all the metrics simultaneously with equal weight. For the rest of this section, we will refer to the strategy of optimizing the metrics simultaneously with Simultaneous Optimization (SO). The single-objective strategies were performed by setting the weights ω_k in **Algorithm 1** for the ignored metrics to 0. This was performed for building blocks available from 0-4 reaction steps, as extending the search to the remaining compounds added few additional options (see Figure 3). At each step, the new building blocks were added to the existing pool of available blocks to model the marginal gain for the chemist to perform synthesis for acquisition of new building blocks. We repeated 10 runs for each level of reaction step for **Algorithm 1** from different randomized initializations.

The results for single-objective search, cf. Table 1, show that the average QSAR values while optimizing for the other objectives tended to stay between 0.6-0.7, indicating that an arbitrary re-combination of building blocks from LibINVENT compounds of high QSAR values does not always result in a product that also has a high QSAR value.

Expanding the search to building blocks available by 1-4 reaction steps resulted in samples of slightly lower diversity as average QSAR value went from very close to 1.0 to selections that had each compound with a value of exactly 1.0. Optimizing for diversity maintained the average QSAR value in the observed selections. The results of SO did not improve as the number of available building blocks increased. This indicates that the set of purchasable building blocks that is already available covers optimal solutions given our scoring parameters. For the single-objective optimization strategies, the QED value tended to decrease as the size of the search space increased. A possible explanation could be that the building blocks corresponding to several steps of reactions are more complex, which tend to have a negative effect on the QED value[54]. The difference between the selections from baseline available building blocks and selections of building blocks one reaction step away represent the largest change in QED score, while further expansions of the building block availability resulted in much smaller or no changes for all metrics. This observation is likely explained by the distribution of building blocks we previously observed in Figure 3; one reaction step represents a change from a space of $6{,}203 \times 763$ products to a space of $23{,}034 \times 1{,}926$, almost ten times larger. The next reaction steps increase the size of the product space relative to the previous step by 31.7% and 4.9%, respectively. The sampling process thus selects building blocks from a pool that is very similar between these three selections, and as such the distributions are similar.

The top 96 compounds by QSAR value generated by LibINVENT had an average QSAR value of 1.0, average QED of 0.43. While these compounds are more diverse than any selection found in our combinatorial selection, they achieve this by breaking the combinatorial constraint. The selection had 96 different carboxylic acids and 3 different aromatic halides. 95 carboxylic were evaluated by AiZynthfinder to be synthesizable, in at most four reaction steps. The 3 aromatic halides were all available directly in stock.

To compare these results against random selection, we sampled 100 combinatorial selections of size 12×8, where each building block for the respective AC and BH reactions was sampled with equal probability. This was repeated for building block availability from each level of reaction steps up to 4 reaction steps from the stock. The random selections consistently had worse QSAR values and QED values than SO, while having diversity values that were not noticeably different from the optimized selections. The average QED value among the random selections is <0.25, which is significantly lower than the average of an "attractive drug" [54]. In addition, the average QSAR value is lower than 0.8, which means many products in the selection are not very likely to be bioactive. This validates the need for optimizing these selections.

Selection strategy	N Reaction steps	Avg QSAR	Avg QED	Avg logDet
QSAR	$\rangle$	0.999	0.297	-196.0
	0	0.993	0.370	-206.8
	1	1.000	0.281	-192.9
	2	1.000	0.278	-195.9
	3	1.000*	0.281	-193.1
	4	1.000*	0.277	-192.1
QED	$\sum_0^4$	0.679	0.782	-154.9
	0	0.676	0.785	-155.9
	1	0.677	0.782	-154.4
	2	0.685	0.781	-155.5
	3	0.682	0.782	-153.5
	4	0.675	0.781	-155.1
Diversity	$\sum_0^4$	0.692	0.139	-95.81
	0	0.698	0.244	-101.3
	1	0.699	0.138	-95.88
	2	0.688	0.110	-94.12
	3	0.687	0.103	-94.13
	4	0.686	0.099	-93.65
Simultaneous Optimization	$\sum_0^4$	0.848	0.701	-126.8
	0	0.852	0.703	-126.8
	1	0.848	0.701	-126.8
	2	0.845	0.704	-127.3
	3	0.843	0.699	-126.2
	4	0.851	0.699	-127.2
Random selection	$\sum_0^4$	0.777	0.247	-126.7
	0	0.765	0.354	-128.3
	1	0.781	0.231	-126.7
	2	0.778	0.213	-125.6
	3	0.779	0.215	-126.3
	4	0.781	0.213	-125.9
LibINVENT top 96	-	1.000	0.43	-88.44

Table 1. Summary of average metrics across all selection strategies used. LogDet is the logarithm of determinant of the kernel matrix, or matrix of all pairwise Tanimoto similarities in the current selection. A value closer to 0 is more diverse. Random selection is the average values of 100 combinations selected for each reaction step availability. For each optimization strategy, we show the results of stock-available building blocks (0 reaction steps) and building blocks up to 4 reaction steps away. The overall average results are denoted by $\sum_0^4$.

The selected products of the single-objective optimizations as well as the SO were also compared visually. Figure 4 shows a small sample of 2×2 combinatorial examples from the different selections for visual clarity. The single-objective selections leave plenty of room for improvement. QED-optimized and diversity-optimized selections both have QSAR values around 0.7, but while the QED-optimized compounds are small, the diversity optimized compounds promote larger building blocks with several rings and side chains. QSAR-optimized selections have the lowest diversity and cover a range of low QED-scores, favouring building blocks with 1-2 rings each and are generally too large still for being druglike. It is likely that the QSAR score of 1.0 indicates that LibINVENT finds exactly which bits in the fingerprint representation that exploit the QSAR model. SO yielded a balanced selection of smaller building blocks that still yielded a high average QSAR value of ~ 0.848.

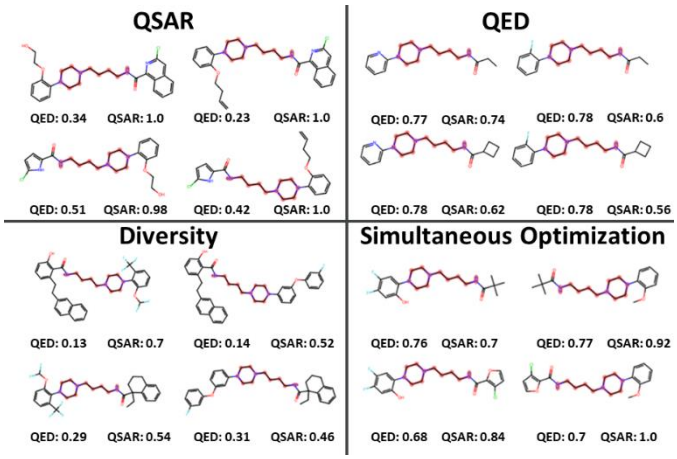

Figure 4. Sampled compounds using the selection strategies of Max QSAR, Max QED, Max diversity and Simultaneous Optimization of all three criteria. The shown examples are using building blocks available in the eMolecules stock.

To evaluate the selection strategies in a more practically relevant setting, we restricted our building block stock availability to a subset of 3% of the original size (~45k building blocks) simulating an approximate availability of building blocks available for a pharmaceutical company. The distribution of solved retrosynthesis routes for the building block subsets are shown in Figure 5. The unsolved routes on average were 26,504 with a standard deviation of 526.6 and 1,072 with a standard deviation of 132.9 for AC and BH reactions, respectively.

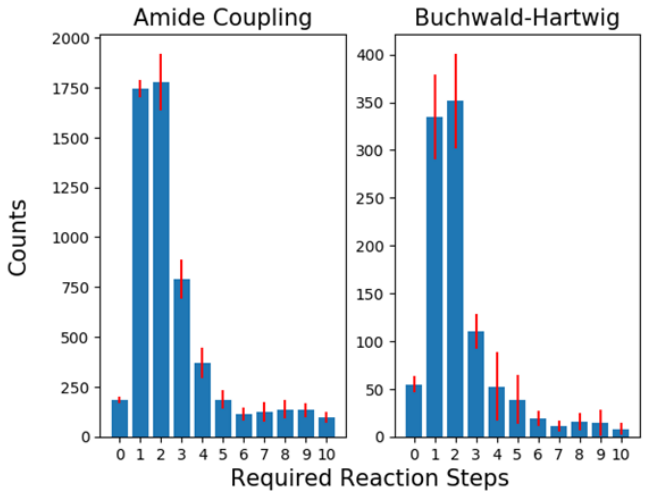

Figure 5. Distribution of average number of reaction steps needed for the generated building blocks while using a 3% subset of the stock. The error bars show the standard deviation across the 5 splits. The number of unsolved routes is omitted from the figure for visual clarity.

It is noteworthy that the proportion of building blocks added per reaction step relative to the current available size is larger for these limited availability subsets, i.e., as 1,745 and 385 building blocks are added for AC and BH after one reaction, compared to 16,831 and 1,163 building blocks added for the full stock. The general trend continues as the selection space is expanded to more reaction steps and in the first four reaction steps almost half of the total number of aromatic halides and more than half of the carboxylic acids become available.

Selection strategy	N Reaction steps	Avg QSAR	Avg QED	Avg logDet
QSAR	$\sum_0^4$	0.974	0.357	-186.1
	0	0.909	0.373	-152.3
	1	0.984	0.386	-189.5
	2	0.992	0.349	-196.3
	3	0.992	0.341	-195.4
	4	0.993	0.335	-196.9
QED	$\sum_0^4$	0.697	0.764	-152.0
	0	0.734	0.701	-143.2
	1	0.685	0.775	-151.4
	2	0.691	0.781	-155.1
	3	0.687	0.782	-155.2
	4	0.686	0.772	-155.3
Diversity	$\sum_0^4$	0.712	0.223	-102.8
	0	0.722	0.305	-108.3
	1	0.704	0.237	-102.7
	2	0.707	0.200	-100.8
	3	0.708	0.186	-100.3
	4	0.716	0.187	-101.8
Simultaneous Optimization	$\sum_0^4$	0.832	0.691	-127.1
	0	0.789	0.650	-127.9
	1	0.836	0.700	-127.1
	2	0.842	0.700	-126.2
	3	0.846	0.703	-127.2
	4	0.846	0.703	-127.3

Table 2. Summarization of average metrics across all selection strategies used for optimizing over the smaller (3%) subsets of available building blocks. LogDet is the logarithm of determinant of the kernel matrix, or matrix of all pairwise Tanimoto similarities in the current selection. A value closer to 0 is more diverse. For each optimization strategy, we show the results of stock-available building blocks (0 reaction steps) and building blocks up to 4 reaction steps away. The overall average results are denoted by $\sum_0^4$.

The same four selection strategies were used for building blocks available from 0-4 reaction steps with ten starting randomized initializations each. Here, the selection from stock-available (zero reaction steps), seen in Table 2, shows that the highest achievable values are drastically lower than after acquiring more building blocks by synthesis. For this smaller space the algorithm is likely to result in the same optimum for the given stock with multiple initializations.

The results show that optimized selections approach their respective values from the full eMolecules availability already after extending the selection space to building blocks available within one reaction, and that the stock-available selections score similar in average QSAR and diversity to the random selection of previous experiment. There are smaller improvements in selections with building blocks available within two reaction steps and no improvements with further reactions. We can draw parallels with the distribution of available building blocks in Figure 4 to the distribution of the previous experiment, and note that the improvements occur when a relatively large number of new building blocks are added to the selection space. When the relative expansion of the space is low the probability of finding a new improved solution is also low.

Unlike the previous experiment, however, the QED score remains at a similar level or, in some cases, improves as the number of reaction steps increase. It is likely that the number of added building blocks through reactions that are "too complex" are lower in this experiment.

The methodology of comparing the optimization results between two different stocks of availability might be useful to estimate the prospective gain from synthesizing new building blocks compared to buying available compounds or simply using the current stock by comparing the optimization results with different selection spaces. This can assist the decision-maker in designing efficient libraries in a combinatorial manner. The number of building blocks estimated to be available through synthesis shows a substantial/relevant increase in search space as the number of reaction steps increases. In practice, only stock-available building blocks or building blocks that can be synthesized in one reaction step will often be used. Alternatively, one could introduce a constraint on the total number of reaction steps used for the selected library, which could be accounted for using e.g., reaction sampling.

Computational time

During selection, we opted for relatively small selection dimensions to limit the computational time to less than ten hours per run, since we performed 12 optimizations, for 10 splits and 5 different building block availabilities, for a total of 600 selections. The observed runs would perform for approximately 20,000-100,000 samples depending on selection space, initialization and number of metrics, which could take between 20 minutes and 4 hours on a single CPU with the QSAR model being the biggest bottleneck. However, since the evaluation of a random forest model is linear in the number of new products between two samples (12 or 8 depending on the exchanged building block) and determinant calculations have the time complexity of $O(\mathrm{k}^3)$ with total number of products, the method will eventually be limited by evaluations of diversity rather than QSAR. This appears feasible with size 1,536 as here $n = products^2 = (u \times v)^2$. The termination criterion for 10,000 samples without improvement was chosen after some

initial experimentation. For larger library dimensions, it is possible that more samples are more suitable to find convergence. The increase in number of building blocks to choose results in more decision variables to determine for an optimal solution. Additionally, larger dimensions generally mean the marginal change of exchanging one building block on the average values in the selection is smaller, which implies the acceptance ratio becomes closer to 1. On an Intel Xeon W-2125 CPU @ 4.00GHz machine with 8 threads the 12×8 configuration required approximately 0.11s for the QSAR computations compared to 0.04s for computing diversity for each sample, while a 48×32 configuration required 0.14s for the QSAR and 4.0s for computing the diversity. A full exhaustive search was never considered even for the smallest subsets as e.g., the size of the average 3% subset at stock-availability in a 12×8 configuration results in $\sim 2 \times 10^{27}$ different possible combinations. For the same reasons, hyperparameter optimization of α and ω was not performed, as this scaffold is hypothetical and that a marginally better selection would not lead to generalizable guidelines for these parameters.

Conclusions

We present a framework for combinatorial library design evaluated using available public data and open source software to allow reproducibility. The framework can be controlled by specifying both importance of different evaluation metrics and the acceptance ratio α. Our experimental results show that it is possible to perform the multi-objective optimization towards both quality and diversity for our example library. The results show that our framework can navigate the search space around combinatorial library design and find selections of high (>0.8) QSAR values while retaining good (>0.7) QED values and high diversity. The trade-offs between the different objectives were investigated and it was found that the multi-objective optimization maintained a QED relatively close to the maximum possible while optimizing QSAR and diversity. Building blocks that were selected at random showed on average low (<0.25) QED values and lower QSAR value (~ 0.78) than the quality-focused optimization strategies. Our experiments indicate that the set of all available purchasable building blocks require minimal extra synthesis to reach the highest observed scores, while simulated scenarios of limited stock greatly benefit—to comparable score levels—from single-step synthesis of building blocks. The latter scenario might be useful in practise in a larger company with a sizable building block store. It might be faster and cheaper to synthesize the needed building blocks for the combinatorial library design in one step compared to purchasing additional building blocks. It was also shown that synthesizing building blocks in more than one step was not attractive given the size of the internal building block store. For an institution with a very small internal building block store, it might be favourable to synthesize the needed building blocks for the libraries in more than one step.

Author Contributions

SVJ jointly with MHC, OE and AS conceptualized the approach and developed the methodology. SVJ performed the data curation, formal analysis, and investigation, implemented the software, and performed validation, and visualization; he also wrote the original draft. MHC, OE and AS supervised and administered the project. All authors reviewed and edited the submitted manuscript.

Conflicts of interest

The authors declare no conflict of interest

Acknowledgements

The authors would like to thank Dr. Samuel Genheden for help with AiZynthFinder. and discussions regarding building blocks, Dr. Thierry Kogej for discussions on filtering stock files for building block types and different entry formats. The authors also thank Mr. Hampus Gummesson Svensson for scientific discussions and for reviewing this manuscript. Additionally, the authors want to thank the Molecular AI department at AstraZeneca and the Department of Computer Science and Engineering at Chalmers for many discussions and support. Finally, the authors thank the Knut and Alice Wallenberg foundation and the WASP program for financial support.

Notes and references

1. E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. A. Winkler, D. Agrafiotis, A. Cherkasov and A. Tropsha, *Chemical Society Reviews*, 2020, **49**, 3525-3564.
2. M. Withnall, E. Lindelöf, O. Engkvist and H. Chen, *Journal of Cheminformatics*, 2020, **12**, 1-1.
3. M. Olivecrona, T. Blaschke, O. Engkvist and H. Chen, *Journal of Cheminformatics*, 2017, **9**, 48-48.
4. O. Prykhodko, S. V. Johansson, P. C. Kotsias, J. Arús-Pous, E. J. Bjerrum, O. Engkvist and H. Chen, *Journal of Cheminformatics*, 2019, **11**.
5. R. Mercado, T. Rastemo, E. Lindelöf, G. Klambauer, O. Engkvist, H. Chen and E. J. Bjerrum, *Machine Learning: Science and Technology*, 2020.
6. R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Central Science*, 2018, **4**, 268-276.
7. M. H. S. Segler, T. Kogej, C. Tyrchan and M. P. Waller, *ACS Central Science*, 2018, **4**, 120-131.
8. M. H. S. Segler and M. P. Waller, *Chemistry - A European Journal*, 2017, **23**, 5966-5971.
9. M. H. S. Segler, M. Preuss and M. P. Waller, *Nature*, 2018, **555**, 604-610.
10. C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green and K. F. Jensen, *ACS Central Science*, 2017, **3**, 434-443.
11. H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green and K. F. Jensen, *ACS Central Science*, 2018, **4**, 1465-1476.

12. S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist and E. Bjerrum, *Journal of Cheminformatics*, 2020, **12**, 70.
13. M. Garcia de Lomana, F. Svensson, A. Volkamer, M. Mathea and J. Kirchmair, *Digital Discovery*, 2022, **1**, 158-172.
14. S. Johansson, A. Thakkar, T. Kogej, E. Bjerrum, S. Genheden, T. Bastys, C. Kannas, A. Schliep, H. Chen and O. Engkvist, *Drug Discovery Today: Technologies*, 2019, **32-33**, 65-72.
15. K. H. Bleicher, H.-J. Böhm, K. Müller and A. I. Alanine, *Nature Reviews Drug Discovery*, 2003, **2**, 369-378.
16. T. Kodadek, *Chemical Communications*, 2011, **47**, 9757-9763.
17. R. Pascual, J. Borrell Ji Fau - Teixidó and J. Teixidó.
18. E. A. Jamois, M. Hassan and M. Waldman, *Journal of Chemical Information and Computer Sciences*, 2000, **40**, 63-70.
19. B. R. Beno and J. S. Mason, *Drug Discovery Today*, 2001, **6**, 251-258.
20. D. C. Spellmeyer and P. D. J. Grootenhuis, in *Annual Reports in Medicinal Chemistry*, ed. A. M. Doherty, Academic Press, 1999, vol. 34, pp. 287-296.
21. F. L. Stahura, L. Xue, J. W. Godden and J. Bajorath, *Journal of Molecular Graphics and Modelling*, 1999, **17**, 1-52.
22. D. K. Agrafiotis and V. S. Lobanov, *Journal of Chemical Information and Computer Sciences*, 2000, **40**, 1030-1038.
23. R. P. Sheridan, S. G. SanFeliciano and S. K. Kearsley, *Journal of Molecular Graphics and Modelling*, 2000, **18**, 320-334.
24. E. A. Jamois, C. T. Lin and M. Waldman, *Journal of Molecular Graphics and Modelling*, 2003, **22**, 141-149.
25. *Concepts and applications of molecular similarity*, John Wiley & Sons, Nashville, TN, 1990.
26. S. D. Pickett, I. M. McLay and D. E. Clark, *Journal of Chemical Information and Computer Sciences*, 2000, **40**, 263-272.
27. E. Jacoby, B. Wroblowski, C. Buyck, J.-M. Neefs, C. Meyer, M. D. Cummings and H. van Vlijmen, *Molecular Informatics*, 2018, **37**, 1700119.
28. G. Schneider and U. Fechner, *Nature Reviews Drug Discovery*, 2005, **4**, 649-663.
29. N. van Hilten, F. Chevillard and P. Kolb, *Journal of Chemical Information and Modeling*, 2019, **59**, 644-651.
30. W. P. Walters, *Journal of Medicinal Chemistry*, 2019, **62**, 1116-1124.
31. L. Humbeck, S. Weigang, T. Schäfer, P. Mutzel and O. Koch, *ChemMedChem*, 2018, **13**, 532-539.
32. Enamine, REAL Building Blocks, https://enamine.net/compound-collections/real-compounds/real-database, (accessed 2023-04-12).
33. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Advanced Drug Delivery Reviews*, 2001, **46**, 3-26.
34. L. Ruddigkeit, R. van Deursen, L. C. Blum and J.-L. Reymond, *Journal of Chemical Information and Modeling*, 2012, **52**, 2864-2875.
35. Enamine, MADE Building Blocks, https://enamine.net/building-blocks/made-building-blocks, (accessed 2023-04-12).
36. P. Schneider, W. P. Walters, A. T. Plowright, N. Sieroka, J. Listgarten, R. A. Goodnow, J. Fisher, J. M. Jansen, J. S. Duca, T. S. Rush, M. Zentgraf, J. E. Hill, E. Krutoholow, M. Kohler, J. Blaney, K. Funatsu, C. Luebkemann and G. Schneider, *Nature Reviews Drug Discovery*, 2020, **19**, 353-364.
37. J. Arús-Pous, A. Patronov, E. J. Bjerrum, C. Tyrchan, J.-L. Reymond, H. Chen and O. Engkvist, *Journal of Cheminformatics*, 2020, **12**, 38.
38. V. Bagal, R. Aggarwal, P. K. Vinod and U. D. Priyakumar, *Journal of Chemical Information and Modeling*, 2022, **62**, 2064-2076.
39. T. E. Hadfield, F. Imrie, A. Merritt, K. Birchall and C. M. Deane, *Journal of Chemical Information and Modeling*, 2022, **62**, 2280-2292.
40. A. Domenico, G. Nicola, T. Daniela, C. Fulvio, A. Nicola and N. Orazio, *Journal of Chemical Information and Modeling*, 2020, **60**, 4582-4593.
41. V. Fialková, J. Zhao, K. Papadopoulos, O. Engkvist, E. J. Bjerrum, T. Kogej and A. Patronov, *Journal of Chemical Information and Modeling*, 2022, **62**, 2046-2063.
42. D. K. Agrafiotis, *Molecular Diversity*, 2000, **5**, 209-230.
43. V. J. Gillet, W. Khatib, P. Willett, P. J. Fleming and D. V. S. Green, *Journal of Chemical Information and Computer Sciences*, 2002, **42**, 375-385.
44. H. Chen, U. Börjesson, O. Engkvist, T. Kogej, M. A. Svensson, N. Blomberg, D. Weigelt, J. N. Burrows and T. Lange, *Journal of Chemical Information and Modeling*, 2009, **49**, 603-614.
45. T. Meinl, C. Ostermann and M. R. Berthold, *Journal of Chemical Information and Modeling*, 2011, **51**, 237-247.
46. A. Kulesza, B. J. F. Taskar and T. i. M. Learning, 2012, **5**, 123-286.
47. A. Kulesza and B. Taskar, presented in part at the Advances in Neural Information Processing Systems, 2010, 2010.
48. J. Gillenwater, A. Kulesza and B. Taskar, presented in part at the Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012/7, 2012.
49. N. Miyoshi and T. Shirai, *Advances in Applied Probability*, 2014, **46**, 832-845.
50. T. Nakamura, S. Sakaue, K. Fujii, Y. Harabuchi, S. Maeda and S. Iwata, *Scientific Reports*, 2022, **12**, 1124.
51. D. Sydow, P. Schmiel, J. Mortier and A. Volkamer, *Journal of Chemical Information and Modeling*, 2020, **60**, 6081-6094.
52. G. V. Andrianov, W. J. Gabriel Ong, I. Serebriiskii and J. Karanicolas, *Journal of Chemical Information and Modeling*, 2021, **61**, 5967-5987.
53. Emolecules, Emolecules, https://downloads.emolecules.com/free/, (accessed 28-02-2023).
54. G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan and A. L. Hopkins, *Nature Chemistry*, 2012, **4**, 90-90.
55. J. Gasteiger, *ChemPhysChem*, 2020, **21**, 2233-2242.
56. C. Hansch, P. P. Maloney, T. Fujita and R. M. Muir, *Nature*, 1962, **194**, 178-180.
57. M. Hassan, S. Brown Rd Fau - Varma-O'brien, D. Varma-O'brien S Fau - Rogers and D. Rogers.
58. S. Geman and D. Geman, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1984, **PAMI-6**, 721-741.
59. A. Kulesza and B. Taskar, presented in part at the Proceedings of the 28th International Conference on International Conference on Machine Learning, 2011, 2011.

60. V. Fialková, J. Zhao, K. Papadopoulos, O. Engkvist, E. J. Bjerrum, T. Kogej and A. Patronov, Implementation of the Lib-INVENT Decorator model, https://github.com/MolecularAI/Lib-INVENT, (accessed 28-02-2023).
61. S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist and E. Bjerrum, AiZynthFinder, https://github.com/MolecularAI/aizynthfinder, (accessed 28-02-2023).
62. M. B. Smith and J. March, John Wiley & Sons, Somerset, 7th edn., 2013, pp. 751-755.
63. B. Mahjour, Y. Shen, W. Liu and T. Cernak, *Nature*, 2020, **580**, 71-75.
64. T. K. Ho, USA, 1995.
65. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825-2830.
66. J. Sun, N. Jeliazkova, V. Chupakhin, J.-F. Golib-Dzib, O. Engkvist, L. Carlsson, J. Wegner, H. Ceulemans, I. Georgiev, V. Jeliazkov, N. Kochev, T. J. Ashby and H. Chen, *Journal of Cheminformatics*, 2017, **9**, 17.
67. A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic Acids Research*, 2012, **40**, D1100-D1107.
68. Daylight, SMARTS - A Language for Molecular Patterns, https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, (accessed 2023-02-28, 2023).
69. O. Macchi, *Advances in Applied Probability*, 1975, **7**, 83-122.
70. T. T. Tanimoto, *An Elementary Mathematical Theory of Classification and Prediction*, International Business Machines Corporation, 1958.
71. S. O. Gharan and A. J. a. p. a. Rezaei, 2018.