Computers in Biology and Medicine 175 (2024) 108486

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

DeLA-DrugSelf: Empowering multi-objective de novo design through SELFIES molecular representation

Domenico Alberga [a], Giuseppe Lamanna [a], Giovanni Graziano [b], Pietro Delre [a], Maria Cristina Lomuscio [a], Nicola Corriero [a], Alessia Ligresti [c], Dritan Siliqi [a], Michele Saviano [d], Marialessandra Contino [b], Angela Stefanachi [b], Giuseppe Felice Mangiatordi [a,*]

[a] CNR – Institute of Crystallography, Via Amendola 122/o, 70126, Bari, Italy
[b] Department of Pharmacy - Pharmaceutical Sciences, University of Bari "Aldo Moro", via E. Orabona, 4, I-70125, Bari, Italy
[c] CNR - Institute of Biomolecular Chemistry, Via Campi Flegrei 34, 80078, Pozzuoli, Italy
[d] CNR – Institute of Crystallography, Via Vivaldi 43, 81100, Caserta, Italy

A B S T R A C T

In this paper, we introduce *DeLA-DrugSelf*, an upgraded version of *DeLA-Drug* [J. Chem. Inf. Model. 62 (2022) 1411–1424], which incorporates essential advancements for automated multi-objective de novo design. Unlike its predecessor, which relies on SMILES notation for molecular representation, *DeLA-DrugSelf* employs a novel and robust molecular representation string named SELFIES (SELF-referencing Embedded String). The generation process in *DeLA-DrugSelf* not only involves substitutions to the initial string representing the starting query molecule but also incorporates insertions and deletions. This enhancement makes *DeLA-DrugSelf* significantly more adept at executing data-driven scaffold decoration and lead optimization strategies. Remarkably, *DeLA-DrugSelf* explicitly addresses the SELFIES-related collapse issue, considering only collapse-free compounds during generation. These compounds undergo a rigorous quality metrics evaluation, highlighting substantial advancements in terms of drug-likeness, uniqueness, and novelty compared to the molecules generated by the previous version of the algorithm. To evaluate the potential of *DeLA-DrugSelf* as a mutational operator within a genetic algorithm framework for multi-objective optimization, we employed a fitness function based on Pareto dominance. Our objectives focused on target-oriented properties aimed at optimizing known cannabinoid receptor 2 (CB2R) ligands. The results obtained indicate that *DeLA-DrugSelf,* available as a user-friendly web platform (https://www.ba.ic.cnr.it/softwareic/delaself/), can effectively contribute to the data-driven optimization of starting bioactive molecules based on user-defined parameters.

1. Introduction

1.1. Background

The drug discovery (DD) process can be likened to searching for a needle in a haystack, given that the chemical space is estimated to encompass at least 10^{23} drug-like compounds [1]. When the objective is to identify a small molecule active against a protein target, a conventional approach involves searching a well-established active molecule and endeavoring to enhance its pharmacodynamics and pharmacokinetics properties [2,3]. This is achieved by introducing subtle modifications guided by human intuition [4]. In recent decades, the traditionally costly and time-intensive DD process has accelerated thanks to the emergence of virtual screening and computer-aided drug discovery techniques [5–9]. These advancements allow for in-silico testing of molecular libraries with sizes reaching billions [10,11]. In the ongoing era of Artificial Intelligence (AI), the advent of deep learning [12,13] and, more specifically, generative algorithms has paved the way for complete automation in navigating the vast landscape of chemical space. This capability empowers the identification of novel drug-like compounds in a data-driven manner [14–16]. Specifically, these algorithms undergo a two-step process. Initially, they are trained to acquire a chemical representation from a designated training set. Subsequently, they leverage the acquired syntactical rules to generate novel compounds [17–19]. This dual-step approach allows them to not only learn from existing data but also creatively produce new chemical entities based on the learned patterns and structures. Numerous generative models have been developed based on diverse deep learning architectures including Recurrent Neural Networks (RNNs) [20–23], variational autoencoders (VAEs) [24–28], Generative Adversarial Networks (GANs) [29–33], Reinforcement Learning (RL) [25,34–37], and transformers [38–42]. These models, having a wide range of applications in chemioinformatics tasks such as prediction of protein-protein interactions [13] or protein function annotations [12], are commonly

* Corresponding author.
E-mail address: giuseppefelice.mangiatordi@cnr.it (G.F. Mangiatordi).

https://doi.org/10.1016/j.compbiomed.2024.108486
Received 3 February 2024; Received in revised form 8 April 2024; Accepted 15 April 2024
Available online 16 April 2024

employed for chemical design tasks and typically utilize either molecular graphs [43] or SMILES [44] as the chemical representation. A significant challenge faced by these models is the generation of valid SMILES strings [45]. Specifically, the generated strings must correspond to chemically valid molecular structures. Additionally, they should be easy to synthesize and possess favorable drug-like properties including appropriate molecular weight, lipophilicity, and the absence of potentially toxic chemical groups. However, achieving chemically valid SMILES strings is far from straightforward, as evidenced by the fact that these models are predominantly evaluated based on their ability to produce valid molecules [46].

1.2. The advent of SELFIES molecular representation

A new molecular string representation named Self-referencing Embedded Strings (SELFIES) was recently proposed [47,48]. The main advantage of using SELFIES strings consists of their intrinsic ability to represent a chemically valid molecule whatever token combination is considered [47]. In other words, a validity of 100 % can be easily reached when algorithms based on such a new molecular representation are employed for generation, as testified by several recent papers [49–52]. Additionally, the SELFIES encoding algorithm employs a set of grammar rules designed to be simpler than SMILES, aiming to generate human-readable sequences [47]. Consequently, a generative model using SELFIES can learn its syntax more easily than SMILES, making the training phase faster. However, despite using SELFIES in the context of de novo design is undoubtedly promising, this new molecular representation is not exempt from limitations. While all the generated compounds may be deemed valid, a non-negligible fraction struggles with the so-called "collapse" issue. Specifically, the SELFIES decoding algorithm, in its attempt to self-correct a grammatically incorrect combination of tokens, automatically truncates the string, collapsing it into a valid SELFIES [51–53]. This phenomenon introduces a potential bias and yields false positive feedback during model training, thereby casting doubt on the reliability of the resulting generative models. As far as our knowledge extends, a comprehensive resolution to this issue has been presented in only a handful of papers that introduce novel generative algorithms based on molecular SELFIES [51,52]. In many instances, SELFIES collapse has been overlooked or not addressed at all.

1.3. DeLA-DrugSelf

In a recent work by our group, we introduced a SMILES generative algorithm, called *DeLA-Drug* [22], which is based on Recurrent Neural Networks (RNNs). This algorithm can generate drug-like analogues by altering a query sequence through character substitutions. Subsequently, we harnessed *DeLA-Drug* as a mutational operator within a genetic algorithm-based framework named GENERA [54]. The purpose was to guide the generation of an analogue library possessing desired properties, such as the predicted affinity against a protein target. Building on this background, herein we present a novel RNN-based de novo design algorithm, named *DeLA-DrugSelf,* which can generate novel molecules learning the SELFIES representation. Exploiting the SELFIES encoding robustness, the proposed approach is able to generate a library of analogues of a starting query by manipulating its SELFIES representation. Remarkably, *DeLA-DrugSelf* generates new compounds by introducing character substitutions (as *DeLA-Drug*) and making insertions or deletions at random positions, making it more suitable for performing lead-optimization and scaffold-decoration tasks. Notably, *DeLA-DrugSelf* explicitly addresses the collapse issue as only collapse-free compounds are considered upon generation. Finally, the ability of *DeLA-DrugSelf* to generate new compounds able to bind to the cannabinoid receptor 2 (CB2R), a promising target involved in cancer and neurodegeneration [55–58], was challenged by using it as a mutational operator within a genetic algorithm framework.

2. Materials and methods

2.1. Datasets preparation

The training set (TS) of *DeLA-DrugSelf* was prepared starting from the entire ChEMBL28 database [59] and applying the following data curation steps using KNIME v4.1.4 [60] as data-handling software and employing the CDK [61], RDKit [62], and Open Babel [63] extension nodes for the required chemoinformatics tasks: (i) discard records lacking SMILES notation (ii); remove stereoisomerism; (iii) de-salt and neutralize all of the entries; (iv) remove inorganic and metal atom compounds; (v) remove compounds with elements different from H, C, N, O, F, Br, I, Cl, P, and S; (vi) convert all of the entries into Open Babel neutralized canonical SMILES; (vii) remove duplicates [64]; and (viii) discard SMILES with too low (bottom 5 %) or too high (top 5 %) number of characters. The final TS consists of 1,092,285 compounds whose SMILES were converted into SELFIES notation using an in-house Python script. All the SELFIES characters were tokenized into a single character symbol (e.g. '[C]' were mapped as 'A', and [O] as 'E'). Finally, two additional characters were added at the beginning ("$") and at the end ("~") of each string, respectively. Furthermore, each entry was expanded adding "€" padding characters to standardize the length of the strings (matching the longest ones in the TS). The final alphabet consists of 79 characters and each padded entry is 82 characters long. Each character was represented as a one-hot vector composed of 79 components. Using this representation, each SELFIES was encoded as a binary matrix of dimension 82×79. In addition, a dataset including high affinity CB2R ligands (CB2R-DB), was built extracting affinity entries from ChEMBL25 [59] labelled with the CB2R Target ID (ChEMBL253). The CB2R-DB was filtered retaining entries annotated exclusively with IC_{50} and K_i measures with values < 1 μM, referring to assays conducted on human targets ("target_organism" = "*Homo sapiens*") and marked as direct binding ("assay_type" = "B") and without warnings in the "data_validity_comment" field. The final CB2R-DB contains 1845 compounds after de-salting, neutralizing, and removing all the duplicates.

2.2. Generative model architecture

The model was built resembling the architecture of DeLA-Drug, fully detailed in our previous work [22]. The model consists of an RNN [65] composed of two layers of Long Short-Term Memory (LSTM) units [66] and is trained to predict the probability distribution of the next character given the C previous characters as a context. The training was performed on a node of the pre-exascale Tier-0 EuroHPC supercomputer LEONARDO (supercomputer center CINECA, Italy), in four epochs using the TS with a minibatch size of 256 and using the Adam optimizer [67] (learning rate is 10^{-3}, $\beta_1 = 0.9$ and $\beta_2 = 0.999$) with back-propagation through time [68].

2.3. Sampling with mutations approach

We exploited the LSTM architecture in two generative approaches. The first one, named sampling from scratch (SFS), is the same as described in our previous work [22] and here briefly summarized. Starting from the beginning of the token sequence, we iteratively predict the next character of the sequence by sampling the probability distribution returned in output by the RNN when a context consisting of the previous C characters is given as input. This procedure is iterated till the end of the sequence character (hereinafter referred to as EoS) is sampled. On the other hand, the second strategy, named sampling with mutations (SWM) consists of the following steps. Starting from a query SELFIES string, a number M of tokens in random positions (Ps) is selected from the query. The query string is then scanned token by token starting from the first, when the scanned token is not included in the selected Ps, it is copied to the output string, otherwise it is modified by applying one of the following operations, randomly selected: (i) the token is substituted

with the one predicted by the RNN when the previous output C tokens are given as input (substitution operation); (ii) the token is copied to the output and a new token is added to the output predicting as in (i) (addition operation); (iii) the token is deleted (deletion). The procedure stops when the EoS token is sampled, either because it was copied from the original string or generated by addition and substitution operations. Fig. 1 shows the main steps of the adopted workflow.

2.4. Combining DeLA-DrugSelf with a genetic algorithm for multi-objective optimization

DeLA-DrugSelf was inserted into a genetic algorithm framework to test it as a suitable mutational operator for a multi-objective de-novo design pipeline. To achieve this, the following steps were taken into consideration. Given an initial pool of molecules, all the desired scoring functions were calculated for each entry and the Pareto front was derived accordingly [54]. A fixed number of mutations M was randomly chosen within a fixed range and each molecule belonging to the Pareto front modified through the SWM approach. The newly generated molecules were added to the initial pool. The new Pareto front was computed and employed for the further generation, performed using a new value of M. The stopping criterion of the algorithm can be customized case by case. As an example, the generation can terminate when a specified number of molecules with the desired properties are generated. To test our algorithm as a valuable tool for lead-optimization, we started the generation from a single lead molecule (instead of a pool of compounds). If the algorithm generates a molecule with a collapsed SELFIES, a new generation is performed using a new set of random Ps for a maximum of 1000 attempts.

2.5. Docking simulations

All the generated compounds were docked on the recently published crystal structure of CB2R (PDB code: 6KPC) [69] using a protocol validated in a previous work [56] and herein briefly mentioned. Docking simulations were performed using Grid based ligand docking with energetics (GLIDE) [70] as software, available in the Schrödinger suite 2023-4 [71], and, more specifically, the standard precision (SP) protocol. Notice that during the docking process, the protein was held fixed while full flexibility was allowed for the ligands. The interested reader is referred to Intranuovo et al. [56] for methodological details.

2.6. Generation of Protein–Ligand interaction fingerprints

Interaction fingerprints (IFs) were generated using the SIFt tool, available in the Schrödinger Suite 2023-4 [71,72]. It is important to note that IFs are binary one-dimensional (1D) representations that encode the presence or absence of specific interactions occurring between a given compound and the binding site (BS) in the top-scored docking pose. Specifically, nine types of potential interactions were considered for each residue: (i) any contact, (ii) backbone interactions, (iii) side-chain interactions, (iv) contact with polar residues, (v) contact with hydrophobic residues, (vi) formation of hydrogen bonds with H-bond acceptors of the BS, (vii) formation of hydrogen bonds with H-bond donors of the BS, (viii) contact with aromatic residues, and (ix) contact with charged residues. This approach represented each residue within the BS as a nine-bit long string, where '1′ indicates the presence of the corresponding ligand-residue interaction, and '0′ indicates its absence.

3. Results and discussion

The main goal of our RNN algorithm is to learn the grammar of the SELFIES, approximating it to a natural language. The learned grammar can be used to estimate the probability distribution for the ith character given its context (i.e., the previous C characters). When working with SMILES, the validity of the generated string is a crucial aspect to optimize, thus the algorithms are usually designed to maximize the percentage of valid SMILES string generated in a run [22]. On the contrary, the SELFIES decoding algorithm is based on the assumption that every combination of tokens is valid, hence it automatically ignores grammatically invalid sequences [47]. More specifically, incorrect strings collapse into a truncated valid SELFIES string. Although this phenomenon, also known as SELFIES collapse problem, has been addressed only in a few works presenting new generative algorithms, it should be carefully considered as able to strongly alter the performance of the algorithm [51,52]. In the context of analogues generation, it should be noted that including collapsed SELFIES would lead to the generation of molecules that stray too much from the starting query. In other words, the uncollapsed rate (U), defined as the percentage of the generated molecules unaffected by the SELFIES collapse problem, is a quality metric worth considering. Furthermore, to gain more insights into the link between SELFIES collapse and the quality of the generated compounds, we also computed the Levenshtein (D) distance [73] as an

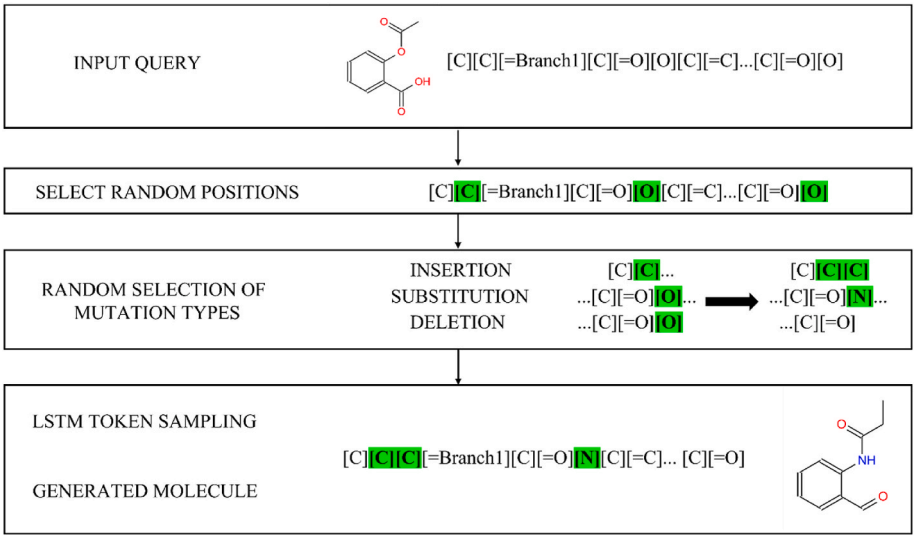


Fig. 1. Flowchart showing the main steps followed by the algorithm of *DeLA-DrugSelf*. Notice that unlike its predecessor DeLA-Drug, which utilized SMILES for molecular representation, it employs SELFIES. Notice that the generation process involves not only substitutions to the initial string representing the starting query molecule but also insertions and deletions and that *DeLA-DrugSelf* explicitly addresses the SELFIES-related collapse issue, considering only collapse-free compounds during generation.

indicator of the degree of collapse of each generated compound. Notice that D is computed using the following procedure: i) The generated SELFIES string is converted into the corresponding SMILES string; ii) The resulting SMILES string is converted back into a SELFIES string; iii) The starting and final SELFIES strings are compared, and D is computed as the counts of SELFIES character edits (insertions, deletions, and substitutions) required to transform one string into the other one. Furthermore, to evaluate the quality of the molecules generated by means of both the SFS and SWM approaches, a set of quality metrics typically employed in the context of de novo design [74,75] were computed: (i) unicity (Uni) defined as the percentage of unique generated molecules; (ii) internal diversity (ID) [76] defined as the mean of the Tanimoto distances (measured using a Morgan fingerprint [77] with radius 2) between each molecule and all the others belonging to the considered set; (iii) novelty (Nov) defined as the percentage of the generated molecules not present in the TS; (iv) the quantitative estimate of drug-likeness (QED) score [78], an estimation of the compound to resemble a drug based on computed physicochemical properties; (v) the synthetic accessibility (SA) score [79], that quantify the difficulty of the chemical synthesis of the compound, ranging from 1 (easy) to 10 (hard); (vi) the percentage of generated compounds without structural alerts known to be responsible for false positives in *in vitro* assays (PAINS) [80]. Regarding the SWM approach, we also measured the query similarity (QS), defined as the mean Tanimoto similarity (using a Morgan fingerprint [77] with radius 2) computed between each generated molecule and the corresponding query.

3.1. SFS approach evaluation

To evaluate the quality of the compounds proposed by the SFS approach, we generated a set of 100,000 SELFIES strings using different values of C (*i.e.;* C = 1, 2, 4, 6, 8, 10, 12 and 14). The above-mentioned quality metrics were computed for each set (Table 1) considering: all the generated SELFIES strings (D ≥ 0), the uncollapsed ones only (D = 0) and the collapsed ones only (D > 0). Satisfactorily, the uncollapsed rate U reached values > 75 % with a maximum value exceeding 80 % when a very high value of context is considered (C = 12). Exceptions are represented by the pool generated using as C very low values (C = 1 and C = 2, both returning U values < 70 %). Importantly, these data suggest that increasing the context, a parameter recently introduced by our group [22], might be a winning strategy to minimize the collapsing of the SELFIES string. Remarkably, for all the considered sets, high Uni (>99 %), Nov (>99 %) and ID values (>0.83) were obtained demonstrating the ability of the SFS approach to generate compounds spanning a large area of the chemical space. It is worth noting that significantly lower values (always <89 %) of Uni are returned when the SFS approach is applied on the same RNN architecture but using SMILES strings instead of SELFIES ones [22]. Furthermore, the obtained data indicate that both QED and SA are affected by C, as a significant improvement is observed in going from C = 1 (QED = 0.51 ± 0.22, SA = 3.29 ± 0.96) and C = 2 (QED = 0.51 ± 0.22, SA = 3.21 ± 0.94) to C ≥ 4 where QED ranges from 0.58 ± 0.21 to 0.60 ± 0.20 (except for C = 14) and SA from 2.95 ± 0.83 to 3.11 ± 0.93. Remarkably, this improvement seems to be the consequence of a reduced collapse rate (*i.e.;* higher U values). This is evident looking at the values of the quality metrics returned by the subsets including only uncollapsed (D = 0) and collapsed (D > 0) SELFIES strings. As an example, when a C value equal to 12 is employed, the returned averaged QED score improves from 0.50 ± 0.21 (D > 0) to 0.60 ± 0.20 (D > 0), hence exceeding the values obtained when SMILES strings are used [22]. Such a trend is confirmed even when the SA score is considered (improving from 3.52 ± 0.90 to 2.81 ± 0.75). To delve deeper into this point, we plotted the dependence of both the averaged QED and SA with respect to the degree of collapse (C = 12). In other words, these metrics were computed for subsets returning specific D values. As evident from Fig. 2A, the higher the degree of collapse, the lower the quality of the generated compounds. This behavior can be

Table 1
Quality Metrics of the set of SELFIES generated with the Sampling from Scratch (SFS) approach at different values of context (C). Unicity (Uni), internal diversity (ID), novelty (Nov), the quantitative estimate of drug-likeness (QED) score, the synthetic accessibility (SA) score and the percentage of generated compounds without structural alerts known to be responsible for false positives in vitro assays (PAINS) are reported. For each set the metrics are computed considering all the molecules, only the uncollapsed SELFIES with Levenshtein distance D = 0 and only the collapsed SELFIES (D > 0).

	C = 1			C = 2			C = 4		
	D ≥ 0	D = 0	D > 0	D ≥ 0	D = 0	D > 0	D ≥ 0	D = 0	D > 0
U (%)	66.58	–	–	68.83	–	–	79.01	–	–
Uni (%)	99.93	99.92	99.96	99.97	99.97	99.96	99.89	99.87	99.95
ID (Td)	0.85	0.85	0.86	0.84	0.84	0.85	0.83	0.83	0.86
Nov (%)	99.97	99.96	99.98	99.97	99.97	99.98	99.95	99.94	99.98
QED	0.51 ± 0.22	0.54 ± 0.21	0.44 ± 0.21	0.51 ± 0.22	0.54 ± 0.21	0.43 ± 0.21	0.58 ± 0.21	0.60 ± 0.20	0.49 ± 0.21
SA	3.29 ± 0.96	3.05 ± 0.83	3.78 ± 1.01	3.21 ± 0.94	2.95 ± 0.79	3.78 ± 1.00	3.00 ± 0.85	2.85 ± 0.75	3.58 ± 0.94
PAINS (%)	95.30	95.28	95.33	95.21	95.34	94.92	95.23	95.26	95.19
	C=6			**C=8**			**C=10**		
	D≥0	**D=0**	**D>0**	**D≥0**	**D=0**	**D>0**	**D≥0**	**D=0**	**D>0**
U (%)	78.04	–	–	76.87	–	–	79.19	–	–
Uni (%)	99.81	99.76	99.97	99.89	99.87	99.96	99.81	99.76	99.94
ID (Td)	0.84	0.84	0.86	0.85	0.84	0.87	0.85	0.86	0.87
Nov (%)	99.94	99.93	99.99	99.96	99.95	99.98	99.94	99.93	99.97
QED	0.57 ± 0.21	0.60 ± 0.20	0.49 ± 0.21	0.60 ± 0.21	0.62 ± 0.20	0.51 ± 0.21	0.60 ± 0.20	0.62 ± 0.19	0.50 ± 0.21
SA	3.01 ± 0.89	2.83 ± 0.77	3.63 ± 0.99	3.11 ± 0.93	2.92 ± 0.82	3.72 ± 1.00	3.00 ± 0.88	2.84 ± 0.78	3.65 ± 0.93
PAINS (%)	94.77	94.71	95.03	95.48	95.49	95.47	95.22	95.29	94.95
	C=12			**C=14**			**RANDOM**		
	D≥0	**D=0**	**D>0**	**D≥0**	**D=0**	**D>0**	**D≥0**	**D=0**	**D>0**
U (%)	**80.22**	–	–	76.05	–	–	2.5	–	–
Uni (%)	99.83	99.80	99.95	99.90	99.88	99.95	60.34	43.08	61.33
ID (Td)	0.84	0.84	0.86	0.84	0.84	0.86	0.98	0.98	0.98
Nov (%)	99.95	99.95	99.98	99.95	99.95	99.98	100.00	100.00	100.00
QED	0.58 ± 0.20	0.60 ± 0.20	0.50 ± 0.21	0.52 ± 0.21	0.55 ± 0.20	0.43 ± 0.21	0.25 ± 0.11	0.31 ± 0.80	0.31 ± 0.80
SA	2.95 ± 0.83	2.81 ± 0.75	3.52 ± 0.90	3.04 ± 0.87	2.84 ± 0.73	3.67 ± 0.97	7.04 ± 0.90	6.75 ± 1.17	6.75 ± 1.17
PAINS (%)	95.28	95.30	95.22	94.68	94.74	94.49	90.77	96.11	96.11

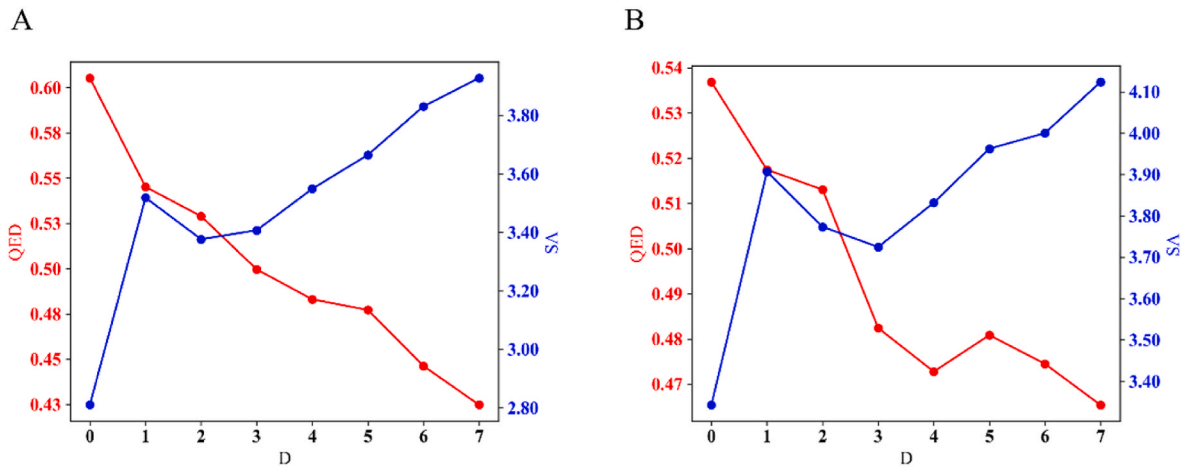


Fig. 2. Average of Quantitative Estimate of Drug-likeness (QED) and Synthetic Accessibility (SA) scores calculated for different values of SELFIES degree of collapse (D) for the set generated using: *i)* the SFS approach and a C value equal to 12 (A); ii) the SWM approach and with mutations number equal to 1 (B).

explained considering that the RNN learns to reproduce the probability distribution of the SELFIES token in the TS. When truncation occurs in the case of collapsed SELFIES, the resulting strings might correspond to molecules that are far from those of the TS in terms of drug-likeness and synthetic accessibility. These results further support the statement that the collapse phenomenon in generative algorithms using SELFIES strings should be properly considered by the scientific community, being able to significantly compromise the quality of the designed compounds. In other words, efforts to reduce the occurrence of this phenomenon are highly desirable, similarly to what has been done in the last years to maximize the validity rate of the molecules generated using SMILES strings for molecular representation. We compared the performance of our generative algorithm with that returned by a SELFIES generator producing strings by randomly concatenating the tokens. Significantly worse performances were obtained with random concatenation, as evident looking at the low U (2.5 %), averaged QED (0.25 ± 0.11) and averaged SA (7.04 ± 0.90) values, hence further confirming that *DeLA-DrugSelf* properly learned the SELFIES syntax during the training phase. Furthermore, the algorithm's reliability is strengthened by its ability to generate molecules that replicate the properties of the training set [75]. Indeed, the average SA and QED values of the molecules used for training are 3.02 ± 0.90 and 0.54 ± 0.20, respectively. As shown in Table 1, the employed SFS approach successfully reproduces these metrics while also ensuring a slight improvement compared to the training. Additionally, it is noteworthy that, based on the data reported in Table 1, *Dela-DrugSelf* is capable of generating compounds with slightly better QED and SA values compared to the majority of generators available in the literature, as evidenced by the results reported by Wang et al. in a recent benchmark study [81].

3.2. SWM approach evaluation

The SWM approach is crafted to generate analogues based on a given query. We tested our algorithm by generating a set of molecules for each number of mutations (M) ranging from 1 to 5. This involved creating molecules from 1000 randomly selected queries from CB2R-DB, generating 100 molecules for each query. We employed the RNN model trained with C = 12, as it exhibited the highest value of U in the SFS approach (Table 1). For each set of generated molecules, we computed the predefined quality metrics, and the results are presented in Table 2.

Table 2
Quality Metrics of the set of SELFIES generated with the Sampling with Mutations algorithm (SWM) at different values of mutations number (M). Unicity (Uni), internal diversity (ID), novelty (Nov), the quantitative estimate of drug-likeness (QED) score, the synthetic accessibility (SA) score and the percentage of generated compounds without structural alerts known to be responsible for false positives in vitro assays (PAINS) and the average Query Similarity (QS) are reported. For each set, the metrics are computed considering all the molecules, only the uncollapsed SELFIES with Levenshtain distance D = 0 and only the collapsed SELFIES (D > 0).

	M = 1			M = 2			M = 3		
	D ≥ 0	D = 0	D > 0	D ≥ 0	D = 0	D > 0	D ≥ 0	D = 0	D > 0
U (%)	26.95	–	–	15.27	–	–	8.90	–	–
Uni (%)	62.68	65.06	62.78	89.84	90.41	90.17	95.02	97.28	94.90
ID (Td)	0.87	0.84	0.88	0.88	0.85	0.88	0.88	0.86	0.88
Nov (%)	99.95	99.85	99.99	99.96	99.80	99.98	99.98	99.83	99.99
QED	0.50 ± 0.20	0.53 ± 0.19	0.49 ± 0.20	0.49 ± 0.20	0.52 ± 0.19	0.48 ± 0.20	0.48 ± 0.20	0.51 ± 0.20	0.47 ± 0.20
SA	3.76 ± 0.86	3.38 ± 0.79	3.90 ± 0.84	4.03 ± 0.89	3.55 ± 0.84	4.11 ± 0.88	4.23 ± 0.90	3.70 ± 0.88	4.28 ± 0.89
PAINS (%)	97.50	96.21	97.99	97.73	96.41	97.96	97.87	96.34	98.02
QS	0.49 ± 0.21	0.70 ± 0.17	0.41 ± 0.17	0.39 ± 0.19	0.63 ± 0.18	0.35 ± 0.16	0.32 ± 0.16	0.56 ± 0.18	0.30 ± 0.14

	M = 4			M = 5		
	D ≥ 0	D = 0	D > 0	D ≥ 0	D = 0	D > 0
U (%)	5.52	–	–	3.46	–	–
Uni (%)	96.00	99.06	95.85	96.17	99.48	96.06
ID (Td)	0.89	0.86	0.89	0.89	0.87	0.89
Nov (%)	99.99	99.83	100.00	99.99	99.85	100.00
QED	0.47 ± 0.20	0.50 ± 0.20	0.46 ± 0.20	0.46 ± 0.92	0.49 ± 0.21	0.45 ± 0.20
SA	4.39 ± 0.91	3.85 ± 0.92	4.43 ± 0.90	4.51 ± 0.92	4.00 ± 0.95	4.53 ± 0.91
PAINS (%)	97.79	96.42	97.87	97.85	95.99	97.10
QS	0.27 ± 0.14	0.49 ± 0.18	0.26 ± 0.13	0.24 ± 0.12	0.44 ± 0.17	0.23 ± 0.12

Like earlier evaluations, all metrics were calculated for: i) all molecules; ii) only the uncollapsed ones (D = 0); iii) only the collapsed ones (D > 0). Notably, the metric U decreases with an increase in the number of mutations (M), dropping from 26.95 % for M = 1 to 3.46 % for M = 5. This is an anticipated outcome, as elevating the number of perturbations to the original SELFIES sequences heightens the risk of grammatically incorrect token combinations. As our tool was designed for data-driven lead optimization, we also computed the Query Similarity (QS) for each set of generated compounds. QS is defined as the averaged similarity between the generated molecules and their respective query. Significantly, when transitioning from M = 1 to M = 5, QS decreases from 0.49 ± 0.21 to 0.24 ± 0.12 (Table 2). Notably, these values increase when considering only the uncollapsed SELFIES (D = 0), with QS ranging from 0.44 ± 0.17 (M = 5) to 0.70 ± 0.17 (M = 1). The substantial differences observed can be attributed to the collapse phase, where generating a grammatically valid sequence of tokens may result in molecules uncontrollably different from the original query. The need for control over the distance from the query is another fact advocating for the critical importance of minimizing collapse when designing generative algorithms for automated lead optimization procedures. As seen when the SFS approach is employed, clear correlations were observed between two quality metrics (QED and SA) and the degree of collapse (D) (Table 2), further supporting the idea that lower D values correspond to higher-quality generated compounds. This trend is visually represented in Fig. 2B, with data generated when M = 1. Importantly, the same trend is evident when considering subsets generated with M > 1 (Table 2). Notably, generations conducted using the SWM approach with two well-known drugs as queries (i.e., aspirin and paracetamol) further indicate that *DeLA-DrugSelf* outperforms *DeLA-Drug* in terms of both QED and SA scores (refer to Tables S1 and S2 in the supporting information for details), as observed when inspecting the molecules generated using the SFS approach. Finally, it is worth noting that when employing the SWM approach, the drug-likeness of the generated compounds also depends on that of the initial molecule used as a query. In other words, using a molecule with a low QED score may result in analogues with less favorable QED values compared to those produced by starting from molecules with high drug-likeness.

3.3. An example of DeLA-DrugSelf application

We tested the ability of *DeLA-DrugSelf* to be used for the optimization of the recently published compound **1** depicted in Fig. 3 (compound **26** in Ref. [56]) responsible for an interesting experimental affinity (IC_{50} = 11 nM) towards the cannabinoid receptor II (CB2R), a target of interest for cancer and neurodegeneration [55,56].

To offer a tangible illustration of how users can generate molecules based on different criteria according to their needs (*e.g.,* preserving a structural of binding mode similarity to a reference molecule), we conducted three distinct generations, referred to as GEN1, GEN2, and GEN3, each configured with different parameters and constraints. For GEN1, the number of mutations (M) ranged from 1 to 5, and the parameters used for Pareto front calculation were: i) the docking score (DS); ii) the Tanimoto similarity (sim) with respect to the starting molecule, calculated by computing Morgan circular fingerprints with a radius of 2; and iii) SA. Additionally, any compounds returning SA values exceeding one unit compared to the starting lead compound were discarded during the generation process. For GEN2, we utilized the same parameters as GEN1, with the only exception being M, which ranged from 1 to 3. Additionally, an additional filter was applied: all generated compounds were discarded if their sim values were less than 0.5. GEN3 differed from GEN1 as it incorporated interaction fingerprints (IF) Tanimoto similarity (IFsim) in the Pareto front. IFsim was calculated by comparing the IF computed on the top-scored docking poses with those returned by the cognate ligand [82], as defined in the experimental section. Using IFsim, it was possible to reward molecules that exhibited a predicted binding mode like the one experimentally observed for the cognate ligand. It is noteworthy that considering this parameter during the selection of molecules has been observed to significantly increase the success rate in virtual screening procedures [82]. All generations were terminated when at least 1500 molecules were generated. In Figs. 4–6 the following metrics are reported for the GEN1, GEN2, and GEN3 sets, respectively: (i) the number of total generated molecules owning DSs higher than the query molecule, labelled as HDS, plotted against the number of generation loops; (ii) the distributions of the optimized parameters along with the respective values owned by the query lead compound represented as vertical red lines; (iii) the joint distributions of the DS with the other parameters.

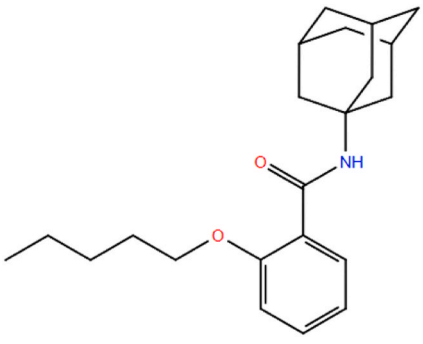


Fig. 3. Chemical structure of compound **1**, a recently published CB2R ligand [56] used as starting query to test the ability of *DeLA-DrugSelf* to be used for lead optimization.

Fig. 4 clearly shows that GEN1 molecules have good DS and SA values. As reported in Table 3, the set has a good average SA (3.04 ± 0.77) and 34.08 % of the generated molecules have a DS better than that returned by the query (HDS molecules). The DS/SA joint distribution further emphasizes that these properties are simultaneously possessed by a substantial number of compounds. However, sim values are frequently below 0.5, and the majority of HDS molecules exhibit low molecular similarity compared to the query. This may not be problematic if a broad exploration of the chemical space around the lead compound is desired, as also testified by the high ID value (0.74). Alternatively, if the objective is to evaluate molecules with high sim values, additional constraints during the generation process are essential. The parameters chosen for GEN2 are specifically tailored to address this goal, as evident from Fig. 5. By reducing the maximum number of mutations to 3 and implementing a similarity cutoff of 0.5, GEN2 compels the generation of molecules exhibiting both high similarity to the query and high docking scores, reflected in the HDS percentage value of 54.32 %. Consequently, a low ID value is observed (0.54), indicating that chemical space exploration is confined around the lead molecule. As anticipated, due to the high similarity to the query, the average SA (3.54 ± 0.48) of GEN2 compounds closely approximates the value of the query (3.31). The GEN3 set shares similarities with GEN1 in terms of optimized parameters, with the addition of the IFsim Tanimoto similarity as the fourth parameter in the Pareto front. In this case, the heightened complexity of the Pareto front introduces a trade-off in the optimization of individual parameters, as the genetic algorithm seeks to find a balance among a larger number of factors. This may elucidate the reduction of the HDS percentage value (22.27 %). Nevertheless, a favorable ID (0.71) and average SA are observed (3.12 ± 0.75). To further evaluate the potential of the generated molecules as CB2R ligands, we subjected the HDS molecules to ALPACA [58], a machine learning-based platform for predicting CB2R affinity. ALPACA, exploiting a consolidated machine learning protocol [83], has demonstrated high accuracy in assessing the potential CB2R affinity of query molecules (AUC >0.90). Table 3 presents the percentage of HDS compounds predicted by ALPACA as high-affinity CB2R binders (CB2Rsel). GEN1 demonstrates a noteworthy CB2Rsel value of 35.92 %. Importantly, in GEN2, by emphasizing similarity to the lead compound, there is a

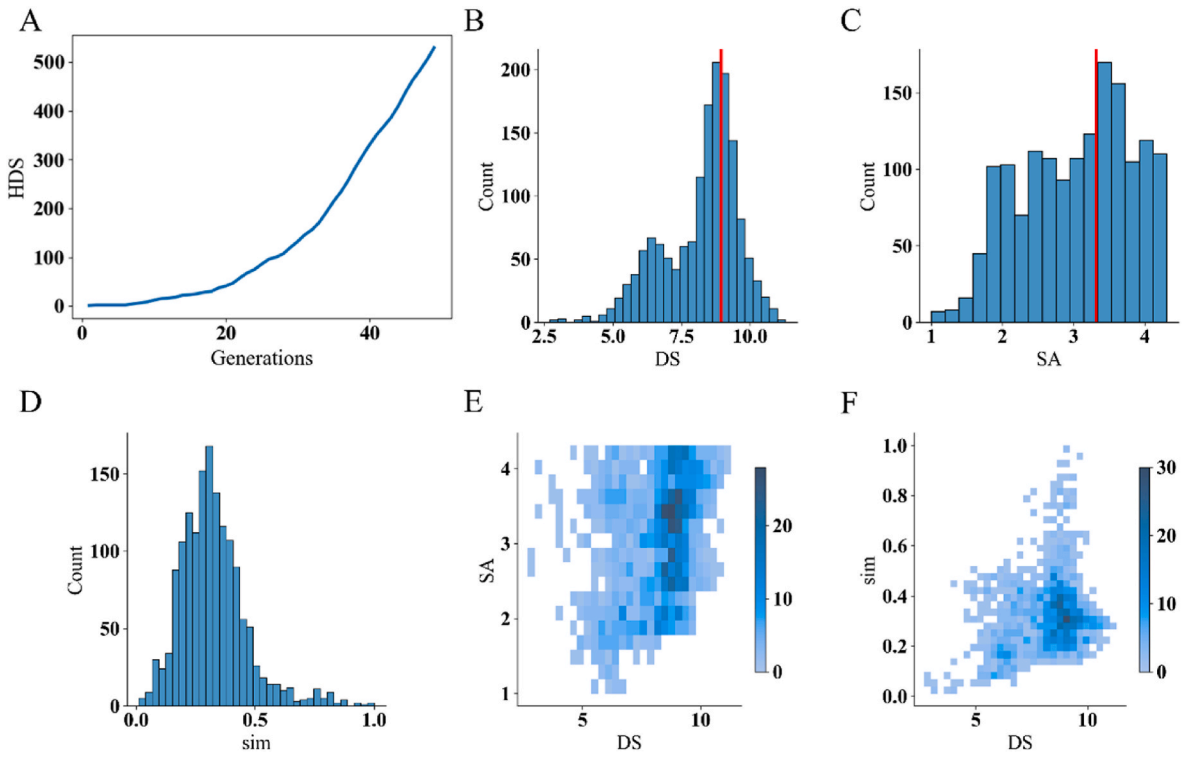


Fig. 4. Number of total generated molecules owning DSs higher than the lead molecule (HDS) plotted against the number of generation loops for the GEN1 set (A). Docking scores (reported as absolute values - DSs) (B), synthetic accessibility (SA) score (C) and the Tanimoto similarity (sim) with respect to the starting molecule, calculated by computing Morgan circular fingerprints with a radius of 2 (D) distributions returned by all the compounds belonging to the GEN1 set along with the respective values owned by the lead compound represented as vertical red lines. Joint distributions of DS-SA (E) and DS-sim (F) are also provided.

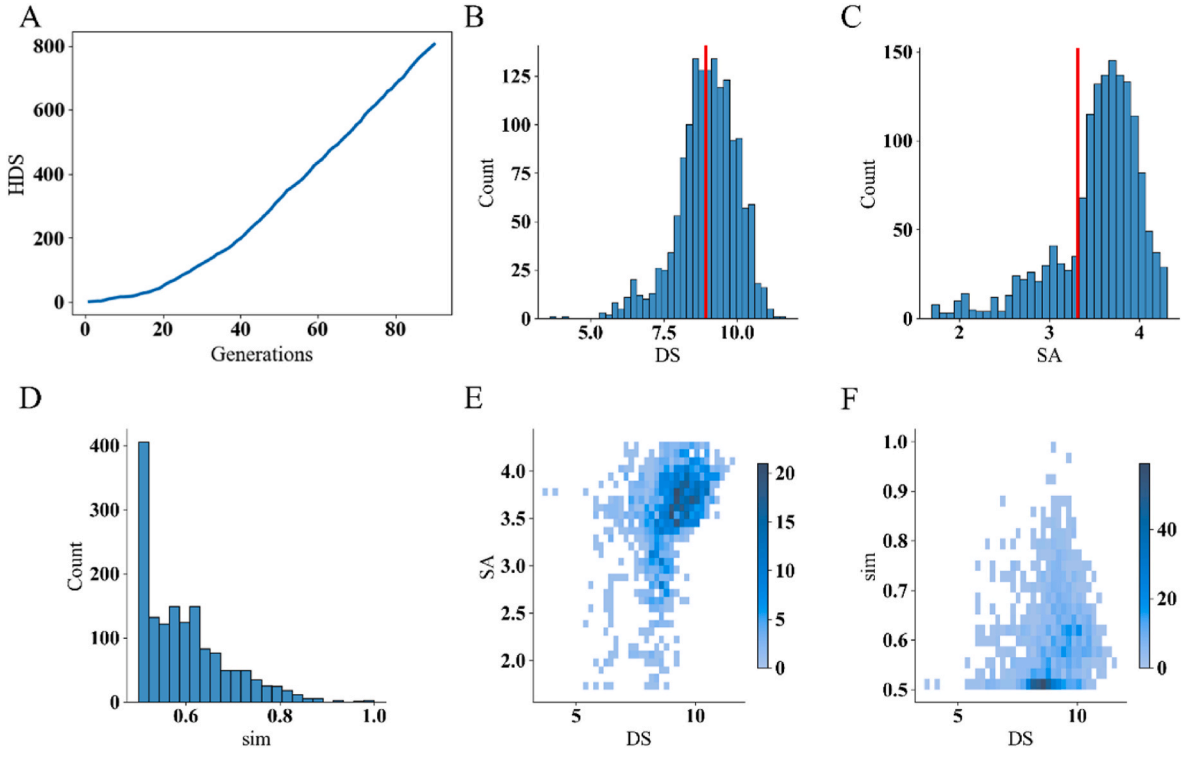


Fig. 5. Number of total generated molecules owning DSs higher than the lead molecule (HDS) plotted against the number of generation loops for the GEN2 set (A). Docking scores (reported as absolute values - DSs) (B), synthetic accessibility (SA) score (C) and the Tanimoto similarity (sim) with respect to the starting molecule, calculated by computing Morgan circular fingerprints with a radius of 2 (D) distributions returned by all the compounds belonging to the GEN2 set along with the respective values owned by the lead compound represented as vertical red lines. Joint distributions of DS-SA (E) and DS-sim (F) are also provided.

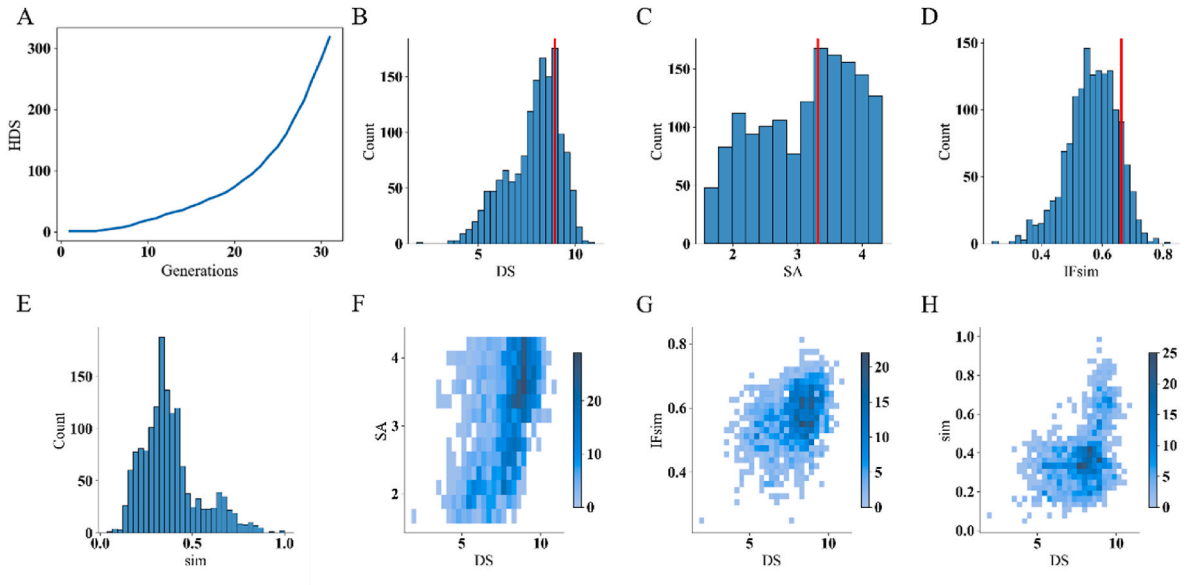


Fig. 6. Number of total generated molecules owning DSs higher than the lead molecule (HDS) plotted against the number of generation loops for the GEN3 set (A). Docking scores (reported as absolute values - DSs) (B), synthetic accessibility (SA) score (C), interaction fingerprints Tanimoto similarity (IFsim) score (E) and the Tanimoto similarity (sim) with respect to the starting molecule, calculated by computing Morgan circular fingerprints with a radius of 2 (E) distributions returned by all the compounds belonging to the GEN3 set along with the respective values owned by the lead compound represented as vertical red lines. Joint distributions of DS-SA (F), DS-IFsim (G) and DS-sim (H) are also provided.

Table 3
The quantitative estimate of drug-likeness (QED) score, the synthetic accessibility (SA) score, the internal diversity (ID) and the percentage of generated compounds without structural alerts known to be responsible for false positives in vitro assays (PAINS) for the GEN1, GEN2 and GEN3 sets along with the percentage of the generated molecules having a docking score against CB2R higher than that returned by the query (HDS) and the percentage of HDS compounds predicted by ALPACA [58] as high-affinity CB2R binders (CB2Rsel).

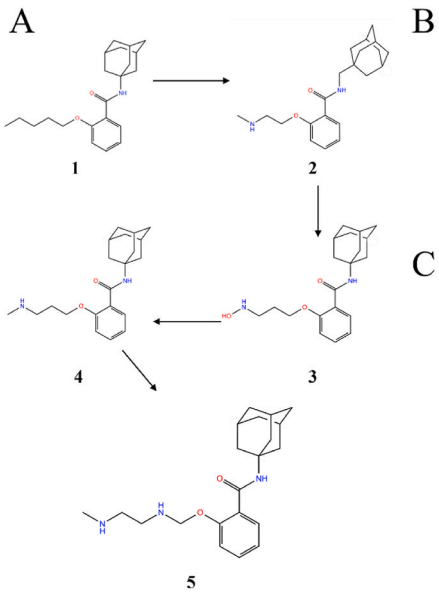
Generation	QED	SA	ID	PAINS (%)	HDS (%)	CB2Rsel (%)
GEN1	0.50 ± 0.17	3.04 ± 0.77	0.74	99.41	34.08	35.92
GEN2	0.57 ± 0.15	3.54 ± 0.48	0.54	99.87	54.32	92.34
GEN3	0.58 ± 0.17	3.12 ± 0.75	0.71	99.13	22.27	54.22

corresponding high percentage of HDS molecules predicted as CB2R ligands by ALPACA (92.34 %). Furthermore, comparing GEN3 to GEN1 reveals that the inclusion of the IFsim optimization parameter results in an increased CB2Rsel value of 54.22 %, hence again supporting the notion that the IF inclusion enhances the probability of identifying molecules active on the target of interest, aligning with findings in the existing literature [82].

Based on this latest data, we conducted a thorough analysis of the GEN3 dataset to identify a concrete example of how *DeLADrug-Self* can enable the optimization of user-set parameters. Fig. 7 illustrates an example of the evolution of the starting molecule (**1**) during the performed generations from a structural perspective (Fig. 7A) and in relation to the parameters used during the multi-objective optimization (Fig. 7B). As highlighted in the figure, the algorithm significantly improves DS (from −8.93 to −10.19 kcal/mol), maintaining a good similarity to the starting molecule and, more importantly, a binding mode similar to that experimentally observed for the co-crystallized ligand (IFsim always >0.6). Interestingly, the modifications introduced by *DeLA-Drug-Self* allow for a reduction in the lipophilicity, highly desirable in the context of the design of CB2 modulators [84], as the logP value decreases from 4.95 to 2.53, (Fig. 7A and B). The proposed changes (specifically the insertion of nitrogen atoms and the elongation of the alkyl chain attached to the phenyl group) enable the establishment of new interactions for these ligands, particularly with T114 and W194, as highlighted in Fig. 7C, representing the top-scored docking pose of the final molecule **5.** Remarkably, compound **5** represents an example of data-driven lead optimization. Indeed, our algorithm herein suggests how to improve both pharmacokinetic and originality of the lead compound. Additionally, the synthesis of **5** can be easily carried out, thanks to the commercial availability of required reagents or documented synthetic procedures (as detailed in reference [85]). This further underscores the capability of *DeLA-DrugSelf* to generate compounds with practical applicability in real-life scenarios.

3.4. *DeLA-DrugSelf: A user-friendly web tool*

DeLA-DrugSelf is a user-friendly web tool accessible at https://www.ba.ic.cnr.it/softwareic/delaself/. Users can input a query molecule in two ways: either by drawing its 2D structure using the JSME canvas applet [86] or by entering a SMILES string into the provided text field. Additionally, JSME allows for the direct importation of .mol or .SDF files into the system. Users can configure the tool according to their preferences by adjusting parameters such as the desired number of compounds (ranging from 10 to 100, default setting: 10) and the number of mutations (ranging from 1 to 5, default setting: 1). It is important to note that a context C = 12 is utilized for generation. Noteworthy, upon inserting the query molecule, the web portal provides a QED value for that molecule. If this score falls below a certain threshold (<0.35), it is flagged to the user as a warning, and she/he has the option to decide whether to proceed with generating analogues or not. Upon completion of the generation process, the tool presents an interactive list of uncollapsed SELFIES (converted in SMILES format for the sake of simplicity) with SA values not exceeding one unit in comparison to the starting query. Users can explore this list in several ways: firstly, compounds are ranked based on their QED, SA, and Tanimoto similarity to the query. Secondly, users have the option to download the ranking as SMILES (in .txt format) or .SDF files. Furthermore, the 2D structures of the generated compounds are readily accessible via the JSME editor by clicking on the corresponding SMILES string. At the user's request, the platform also



mol	DS	SA	IFsim	sim	logP
1	8.93	3.31	0.66	1.00	4.95
2	9.43	3.46	0.63	0.62	3.23
3	9.51	3.59	0.49	0.80	3.13
4	9.78	3.43	0.65	0.82	3.37
5	10.19	3.72	0.64	0.71	2.53

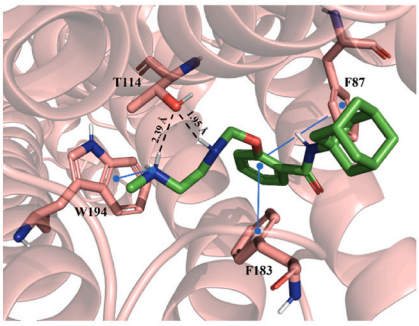


Fig. 7. An example of structural evolution of the starting query (**1**) performed by *DeLA-DrugSelf* during the generation of the GEN3 set (A); values returned by all the parameters used as objectives during the generation of the GEN3 set, docking scores (reported as absolute values - DS), synthetic accessibility (SA) score, interaction fingerprints Tanimoto similarity (IFsim) score, the Tanimoto similarity (sim) with respect to the starting molecule, calculated by computing Morgan circular fingerprints with a radius of 2 and the logarithm of the partition coefficient (logP), for the compounds **1–5** (B); top-scored docking pose returned by the final compound **5** (C). Notice that the ligand and important residues are rendered as sticks, the protein as cartoon and the main intermolecular (ligand-protein) interactions according to the following scheme: H-bond interactions (dotted black lines), pi-pi and cation-pi interactions (blue lines). For the sake of clarity, only polar hydrogen atoms are shown.

provides predictions on whether the generated compounds will interact with specific cytochrome P450 isoforms, including CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP2D6, CYP2E1, and CYP3A4. This prediction is performed by the software program CypReact, as detailed in the publication by Tian et al. [87]. Links to download the resulting data are sent to the user's provided email address. Additionally, the "History" page stores a record of all user executions, containing details such as input SMILES and generated output. Fig. 8 illustrates an example of output page returned by the tool after a generation performed using the default settings.

4. Conclusions

In this paper, we introduce *DeLA-DrugSelf*, a novel deep-learning algorithm designed for generating analogues of drug-like compounds. This algorithm serves as a mutational operator within a genetic algorithm framework, enabling multi-objective optimizations based on user-defined parameters. While sharing the general architecture (RNN model composed of two layers of LSTM cells) with its predecessor (e.g.; DeLA-Drug) [22] *DeLA-DrugSelf* incorporates crucial advancements. Primarily, it employs a more robust molecular representation string, named SELFIES (SELF-referencing Embedded String) [47]. Furthermore, it can perform generation not only through token substitutions but also insertions and deletions. This versatility makes it particularly well-suited for data-driven scaffold decoration and lead optimization tasks. Our analysis of the generated libraries highlights significant improvements in terms of drug-likeness, uniqueness, and novelty of the compounds. It emphasizes that minimizing collapse in generative algorithms for automated optimization procedures is critical, as evident from clear correlations between collapse rates and two quality metrics (*e.g.;* QED and SA). *DeLA-DrugSelf* was further evaluated as a tool for optimizing a compound known to efficiently bind CB2R, a target implicated in various pathological conditions such as cancer and neurodegeneration [84,88]. The obtained data underscore the ability of the algorithm, now available as a user-friendly web platform (http://www.ba.ic.cnr.it/softwareic/delaself/), to automatically enhance both pharmacokinetics and predicted binding affinity of the initial lead compound while maintaining high synthetic accessibility. While further improvements, already planned by our group, are desirable (*e.g.;* using alternative architectures other than RNNs), the discussed data supports *DeLA-DrugSelf* as a valuable tool for the de novo design of compounds with practical applicability in real-life scenarios.

Data and software availability

The following data are made available in the supporting information:

List of the compounds generated via SFS approach setting C = 1, including the corresponding degree of collapse D (c1.csv);
List of the compounds generated via SFS approach setting C = 2, including the corresponding degree of collapse D (c2.csv);
List of the compounds generated via SFS approach setting C = 4, including the corresponding degree of collapse D (c4.csv);
List of the compounds generated via SFS approach setting C = 6, including the corresponding degree of collapse D (c6.csv);
List of the compounds generated via SFS approach setting C = 8, including the corresponding degree of collapse D (c8.csv);
List of the compounds generated via SFS approach setting C = 10, including the corresponding degree of collapse D (c10.csv);
List of the compounds generated via SFS approach setting C = 12, including the corresponding degree of collapse D (c12.csv);
List of the compounds generated via SFS approach setting C = 14, including the corresponding degree of collapse D (c14.csv);

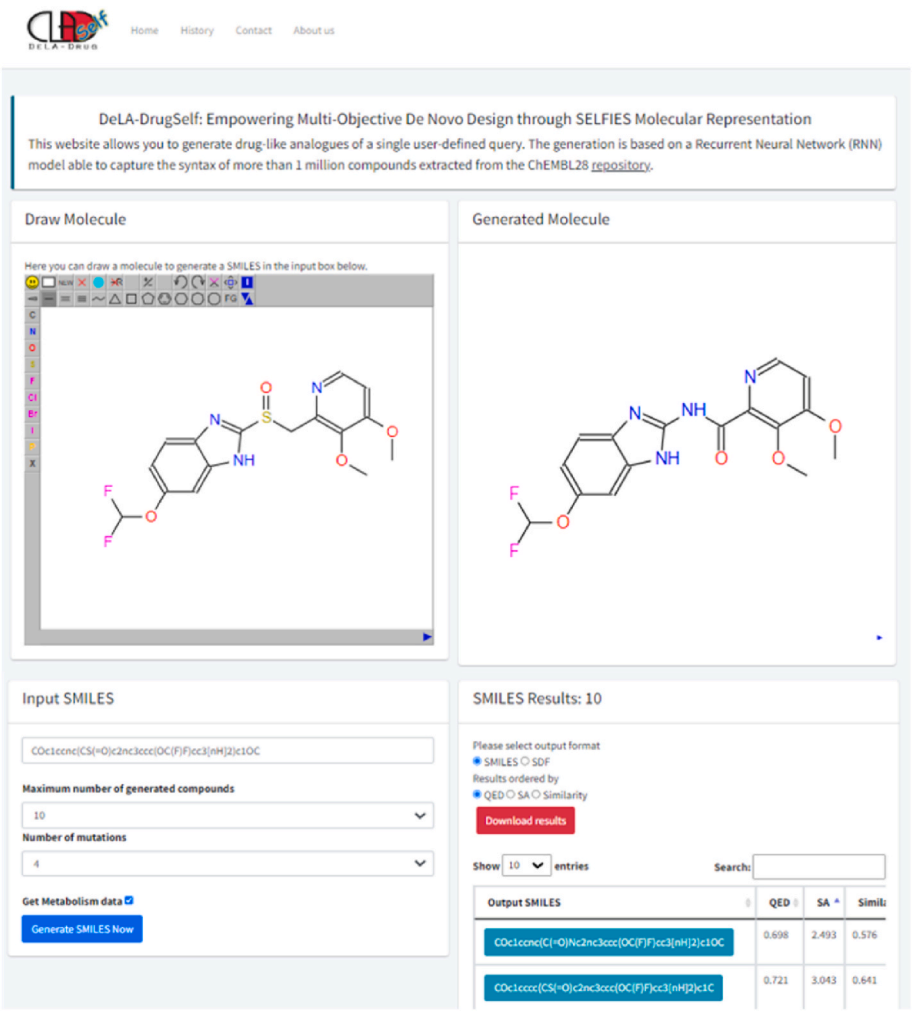


Fig. 8. Example of output page returned by the DeLA-DrugSelf web platform after using the SMILES string of a known proton pump inhibitor (Pantoprazole) as input.

List of the compounds generated via SWM approach setting C = 12 and M = 1, including the corresponding degree of collapse D (c12_m1.csv);
List of the compounds generated via SWM approach setting C = 12 and M = 2, including the corresponding degree of collapse D (c12_m2.csv);
List of the compounds generated via SWM approach setting C = 12 and M = 3, including the corresponding degree of collapse D (c12_m3.csv);
List of the compounds generated via SWM approach setting C = 12 and M = 4, including the corresponding degree of collapse D (c12_m4.csv);
List of the compounds generated via SWM approach setting C = 12 and M = 5, including the corresponding degree of collapse D (c12_m5.csv);
List of the compounds belonging to the GEN1 dataset, including the corresponding docking scores, similarity respect to the query compound and SA score (GEN1.csv).
List of the compounds belonging to the GEN2 dataset, including the corresponding docking scores, similarity respect to the query compound and SA score (GEN2.csv).
List of the compounds belonging to the GEN3 dataset, including the corresponding docking scores, interaction fingerprints similarity, similarity respect to the query compound and SA score (GEN3.csv).
Quality Metrics (QED, SA and QS) computed over 50 molecules generated using the structure of aspirin (QED = 0.55, SA = 1.58) as query and using both *DeLA-Drug* (1 and 5 substitutions) and *DeLA-DrugSelf* (1 and 5 mutations) (Table S1 in the Supporting Information).
Quality Metrics (QED, SA and QS) computed over 50 molecules generated using the structure of paracetamol (QED = 0.59, SA = 1.41). as query and using both *DeLA-Drug* (1 and 5 substitutions) and *DeLA-DrugSelf* (1 and 5 mutations) (Table S2 in the supporting information).
The code of DeLADrug-Self can be freely downloaded at https://github.com/alberdom88/DeLA-DrugSelf

CRediT authorship contribution statement

Domenico Alberga: Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. **Giuseppe Lamanna:** Writing – review & editing, Software, Methodology. **Giovanni Graziano:** Writing – review & editing, Formal analysis. **Pietro Delre:** Writing – review & editing, Software, Methodology, Formal analysis. **Maria Cristina Lomuscio:** Writing – review & editing, Methodology. **Nicola Corriero:** Software. **Alessia Ligresti:** Writing – review & editing, Formal analysis. **Dritan Siliqi:** Writing – review & editing, Methodology, Investigation, Formal analysis. **Michele Saviano:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Marialessandra Contino:** Writing – review & editing, Supervision, Formal analysis. **Angela Stefanachi:** Writing – review & editing,

Methodology, Formal analysis. **Giuseppe Felice Mangiatordi:** Writing – review & editing, Writing – original draft, Supervision, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Domenico Alberga reports financial support was provided by European Union. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was performed within the project "Potentiating the Italian Capacity for Structural Biology Services in Instruct Eric (ITACA.SB)" (Project n° IR0000009), call MUR 3264/2021 PNRR M4/C2/L3.1.1, and the Mission 4, component 2 "From Research to Business"; Investment 3.3 "Introduction of innovative doctorates that respond to the innovation needs of businesses and promote the recruitment of researchers by the companies" PNRR funded by the European Union NextGenerationEU. We acknowledge the CINECA awards no. HP10C0C37T under the ISCRA initiative for the availability of high-performance computing resources.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2024.108486.

References

[1] P.G. Polishchuk, T.I. Madzhidov, A. Varnek, Estimation of the size of drug-like chemical space based on GDB-17 data, J. Comput. Aided Mol. Des. 27 (2013) 675–679, https://doi.org/10.1007/s10822-013-9672-4.
[2] G.N. Kumar, S. Surapaneni, Role of drug metabolism in drug discovery and development, Med. Res. Rev. 21 (2001) 397–411, https://doi.org/10.1002/med.1016.
[3] H. Derendorf, L.J. Lesko, P. Chaikin, W.A. Colburn, P. Lee, R. Miller, R. Powell, G. Rhodes, D. Stanski, J. Venitz, Pharmacokinetic/pharmacodynamic modeling in drug Research and development, J. Clin. Pharmacol. 40 (2000) 1399–1418, https://doi.org/10.1177/009127000004001211.
[4] L. Gomez, Decision making in medicinal chemistry: the power of our intuition, ACS Med. Chem. Lett. 9 (2018) 956–958, https://doi.org/10.1021/acsmedchemlett.8b00359.
[5] E.H.B. Maia, L.C. Assis, T.A. de Oliveira, A.M. da Silva, A.G. Taranto, Structure-based virtual screening: from classical to artificial intelligence, Front. Chem. 8 (2020). https://www.frontiersin.org/articles/10.3389/fchem.2020.00343. (Accessed 11 January 2024).
[6] B. Shaker, S. Ahmad, J. Lee, C. Jung, D. Na, In silico methods and tools for drug discovery, Comput. Biol. Med. 137 (2021) 104851, https://doi.org/10.1016/j.compbiomed.2021.104851.
[7] W. Xue, F. Yang, P. Wang, G. Zheng, Y. Chen, X. Yao, F. Zhu, What contributes to serotonin-norepinephrine reuptake inhibitors' dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation, ACS Chem. Neurosci. 9 (2018) 1128–1140, https://doi.org/10.1021/acschemneuro.7b00490.
[8] W. Xue, T. Fu, S. Deng, F. Yang, J. Yang, F. Zhu, Molecular mechanism for the allosteric inhibition of the human serotonin transporter by antidepressant escitalopram, ACS Chem. Neurosci. 13 (2022) 340–351, https://doi.org/10.1021/acschemneuro.1c00694.
[9] J. Yin, H. Zhang, X. Sun, N. You, M. Mou, M. Lu, Z. Pan, F. Li, H. Li, S. Zeng, F. Zhu, Decoding drug response with structurized gridding map-based cell representation, IEEE J. Biomed. Health Informatics (2023) 1–12, https://doi.org/10.1109/JBHI.2023.3342280.
[10] W.P. Walters, R. Wang, New trends in virtual screening, J. Chem. Inf. Model. 60 (2020) 4109–4111, https://doi.org/10.1021/acs.jcim.0c01009.
[11] Y. Luo, P. Wang, M. Mou, H. Zheng, J. Hong, L. Tao, F. Zhu, A novel strategy for designing the magic shotguns for distantly related target pairs, Briefings Bioinf. 24 (2023) bbac621, https://doi.org/10.1093/bib/bbac621.
[12] L. Zheng, S. Shi, M. Lu, P. Fang, Z. Pan, H. Zhang, Z. Zhou, H. Zhang, M. Mou, S. Huang, L. Tao, W. Xia, H. Li, Z. Zeng, S. Zhang, Y. Chen, Z. Li, F. Zhu, AnnoPRO: a strategy for protein function annotation based on multi-scale protein representation and a hybrid deep learning of dual-path encoding, Genome Biol. 25 (2024) 41, https://doi.org/10.1186/s13059-024-03166-1.
[13] M. Mou, Z. Pan, Z. Zhou, L. Zheng, H. Zhang, S. Shi, F. Li, X. Sun, F. Zhu, A transformer-based ensemble framework for the prediction of protein–protein interaction sites, Research 6 (2023), https://doi.org/10.34133/research.0240, 0240.
[14] Y. Cheng, Y. Gong, Y. Liu, B. Song, Q. Zou, Molecular design in drug discovery: a comprehensive review of deep generative models, Briefings Bioinf. 22 (2021) bbab344, https://doi.org/10.1093/bib/bbab344.
[15] Y. Bian, X.-Q. Xie, Generative chemistry: drug discovery with deep learning generative models, J. Mol. Model. 27 (2021) 71, https://doi.org/10.1007/s00894-021-04674-8.
[16] X. Zeng, F. Wang, Y. Luo, S. Kang, J. Tang, F.C. Lightstone, E.F. Fang, W. Cornell, R. Nussinov, F. Cheng, Deep generative molecular design reshapes drug discovery, Cell Reports Med. 3 (2022) 100794, https://doi.org/10.1016/j.xcrm.2022.100794.
[17] T. Blaschke, J. Bajorath, Fine-tuning of a generative neural network for designing multi-target compounds, J. Comput. Aided Mol. Des. 36 (2022) 363–371, https://doi.org/10.1007/s10822-021-00392-8.
[18] M. Wang, Z. Wang, H. Sun, J. Wang, C. Shen, G. Weng, X. Chai, H. Li, D. Cao, T. Hou, Deep learning approaches for de novo drug design: An overview, Curr. Opin. Struct. Biol. 72 (2022) 135–144, https://doi.org/10.1016/j.sbi.2021.10.001.
[19] D.D. Martinelli, Generative machine learning for de novo drug discovery: A systematic review, Comput. Biol. Med. 145 (2022) 105403, https://doi.org/10.1016/j.compbiomed.2022.105403.
[20] A. Gupta, A.T. Müller, B.J.H. Huisman, J.A. Fuchs, P. Schneider, G. Schneider, Generative recurrent networks for de novo drug design, Molecul. Informat. 37 (2018) 1700111, https://doi.org/10.1002/minf.201700111.
[21] F. Grisoni, M. Moret, R. Lingwood, G. Schneider, Bidirectional molecule generation with recurrent neural networks, J. Chem. Inf. Model. 60 (2020) 1175–1183, https://doi.org/10.1021/acs.jcim.9b00943.
[22] T.M. Creanza, G. Lamanna, P. Delre, M. Contino, N. Corriero, M. Saviano, G. F. Mangiatordi, N. Ancona, DeLA-drug: a deep learning algorithm for automated design of druglike analogues, J. Chem. Inf. Model. (2022), https://doi.org/10.1021/acs.jcim.2c00205.
[23] J. Zou, L. Zhao, S. Shi, Generation of focused drug molecule library using recurrent neural network, J. Mol. Model. 29 (2023) 361, https://doi.org/10.1007/s00894-023-05772-5.
[24] 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, A. Aspuru-Guzik, Automatic chemical design using a data-driven continuous representation of molecules, ACS Cent. Sci. 4 (2018) 268–276, https://doi.org/10.1021/acscentsci.7b00572.
[25] A. Zhavoronkov, Y.A. Ivanenkov, A. Aliper, M.S. Veselov, V.A. Aladinskiy, A. V. Aladinskaya, V.A. Terentiev, D.A. Polykovskiy, M.D. Kuznetsov, A. Asadulaev, Y. Volkov, A. Zholus, R.R. Shayakhmetov, A. Zhebrak, L.I. Minaeva, B. A. Zagribelnyy, L.H. Lee, R. Soll, D. Madge, L. Xing, T. Guo, A. Aspuru-Guzik, Deep learning enables rapid identification of potent DDR1 kinase inhibitors, Nat. Biotechnol. 37 (2019) 1038–1040, https://doi.org/10.1038/s41587-019-0224-x.
[26] M. Lee, K. Min, MGCVAE: multi-objective inverse design via molecular graph conditional variational autoencoder, J. Chem. Inf. Model. 62 (2022) 2943–2950, https://doi.org/10.1021/acs.jcim.2c00487.
[27] H. Iwata, T. Nakai, T. Koyama, S. Matsumoto, R. Kojima, Y. Okuno, VGAE-MCTS: a new molecular generative model combining the variational graph auto-encoder and Monte Carlo tree search, J. Chem. Inf. Model. 63 (2023) 7392–7400, https://doi.org/10.1021/acs.jcim.3c01220.
[28] C. Hu, S. Li, C. Yang, J. Chen, Y. Xiong, G. Fan, H. Liu, L. Hong, ScaffoldGVAE: scaffold generation and hopping of drug molecules via a variational autoencoder based on multi-view graph neural networks, J. Cheminf. 15 (2023) 91, https://doi.org/10.1186/s13321-023-00766-0.
[29] A. Kadurin, S. Nikolenko, K. Khrabrov, A. Aliper, A. Zhavoronkov, druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico, Mol. Pharm. 14 (2017) 3098–3104, https://doi.org/10.1021/acs.molpharmaceut.7b00346.
[30] Ł. Maziarka, A. Pocha, J. Kaczmarczyk, K. Rataj, T. Danel, M. Warchoł, Mol-CycleGAN: a generative model for molecular optimization, J. Cheminf. 12 (2020) 2, https://doi.org/10.1186/s13321-019-0404-1.
[31] Y.J. Lee, H. Kahng, S.B. Kim, Generative adversarial networks for de novo molecular design, Molecul. Informat. 40 (2021) 2100045, https://doi.org/10.1002/minf.202100045.
[32] M. Abbasi, B.P. Santos, T.C. Pereira, R. Sofia, N.R.C. Monteiro, C.J.V. Simões, R.M. M. Brito, B. Ribeiro, J.L. Oliveira, J.P. Arrais, Designing optimized drug candidates with generative adversarial network, J. Cheminf. 14 (2022) 40, https://doi.org/10.1186/s13321-022-00623-6.
[33] J. Zou, J. Yu, P. Hu, L. Zhao, S. Shi, STAGAN: an approach for improve the stability of molecular graph generation based on generative adversarial networks, Comput. Biol. Med. 167 (2023) 107691, https://doi.org/10.1016/j.compbiomed.2023.107691.
[34] M. Olivecrona, T. Blaschke, O. Engkvist, H. Chen, Molecular de-novo design through deep reinforcement learning, J. Cheminf. 9 (2017) 48, https://doi.org/10.1186/s13321-017-0235-x.
[35] S.R. Atance, J.V. Diez, O. Engkvist, S. Olsson, R. Mercado, De novo drug design using reinforcement learning with graph-based deep generative models, J. Chem. Inf. Model. 62 (2022) 4863–4872, https://doi.org/10.1021/acs.jcim.2c00838.
[36] Q. Wang, Z. Wei, X. Hu, Z. Wang, Y. Dong, H. Liu, Molecular generation strategy and optimization based on A2C reinforcement learning in de novo drug design, Bioinformatics 39 (2023) btad693, https://doi.org/10.1093/bioinformatics/btad693.

[37] A. Domenico, G. Nicola, T. Daniela, C. Fulvio, A. Nicola, N. Orazio, De novo drug design of targeted chemical libraries based on artificial intelligence and pair-based multiobjective optimization, J. Chem. Inf. Model. 60 (2020) 4582–4593, https://doi.org/10.1021/acs.jcim.0c00517.

[38] V. Bagal, R. Aggarwal, P.K. Vinod, U.D. Priyakumar, MolGPT: molecular generation using a transformer-decoder model, J. Chem. Inf. Model. 62 (2022) 2064–2076, https://doi.org/10.1021/acs.jcim.1c00600.

[39] X. Liu, K. Ye, H.W.T. van Vlijmen, A.P. Ijzerman, G.J.P. van Westen, DrugEx v3: scaffold-constrained drug design with graph transformer-based reinforcement learning, J. Cheminf. 15 (2023) 24, https://doi.org/10.1186/s13321-023-00694-z.

[40] E. Mazuz, G. Shtar, B. Shapira, L. Rokach, Molecule generation using transformers and policy gradient reinforcement learning, Sci. Rep. 13 (2023) 8799, https://doi.org/10.1038/s41598-023-35648-w.

[41] Y. Matsukiyo, C. Yamanaka, Y. Yamanishi, De novo generation of chemical structures of inhibitor and activator candidates for therapeutic target proteins by a transformer-based variational autoencoder and bayesian optimization, J. Chem. Inf. Model. (2023), https://doi.org/10.1021/acs.jcim.3c00824.

[42] N.R.C. Monteiro, T.O. Pereira, A.C.D. Machado, J.L. Oliveira, M. Abbasi, J. P. Arrais, FSM-DDTR: end-to-end feedback strategy for multi-objective De Novo drug design using transformers, Comput. Biol. Med. 164 (2023) 107285, https://doi.org/10.1016/j.compbiomed.2023.107285.

[43] L. David, A. Thakkar, R. Mercado, O. Engkvist, Molecular representations in AI-driven drug discovery: a review and practical guide, J. Cheminf. 12 (2020) 56, https://doi.org/10.1186/s13321-020-00460-5.

[44] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36, https://doi.org/10.1021/ci00057a005.

[45] J. Meyers, B. Fabian, N. Brown, De novo molecular design and generative models, Drug Discov. Today 26 (2021) 2707–2715, https://doi.org/10.1016/j.drudis.2021.05.019.

[46] K. Handa, M.C. Thomas, M. Kageyama, T. Iijima, A. Bender, On the difficulty of validating molecular generative models realistically: a case study on public and proprietary data, J. Cheminf. 15 (2023) 112, https://doi.org/10.1186/s13321-023-00781-1.

[47] M. Krenn, F. Häse, A. Nigam, P. Friederich, A. Aspuru-Guzik, Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation, Mach. Learn, Sci. Technol. 1 (2020) 045024, https://doi.org/10.1088/2632-2153/aba947.

[48] M. Krenn, Q. Ai, S. Barthel, N. Carson, A. Frei, N.C. Frey, P. Friederich, T. Gaudin, A.A. Gayle, K.M. Jablonka, R.F. Lameiro, D. Lemm, A. Lo, S.M. Moosavi, J. M. Nápoles-Duarte, A. Nigam, R. Pollice, K. Rajan, U. Schatzschneider, P. Schwaller, M. Skreta, B. Smit, F. Strieth-Kalthoff, C. Sun, G. Tom, G. Falk von Rudorff, A. Wang, A.D. White, A. Young, R. Yu, A. Aspuru-Guzik, SELFIES and the future of molecular string representations, Patterns 3 (2022) 100588, https://doi.org/10.1016/j.patter.2022.100588.

[49] A. Nigam, R. Pollice, M. Krenn, G. dos Passos Gomes, A. Aspuru-Guzik, Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES, Chem. Sci. 12 (2021) 7079–7090, https://doi.org/10.1039/D1SC00231G.

[50] L. Chen, Q. Shen, J. Lou, Magicmol: a light-weighted pipeline for drug-like molecule evolution and quick chemical space exploration, BMC Bioinf. 24 (2023) 173, https://doi.org/10.1186/s12859-023-05286-0.

[51] S. Piao, J. Choi, S. Seo, S. Park, SELF-EdiT: structure-constrained molecular optimisation using SELFIES editing transformer, Appl. Intell. 53 (2023) 25868–25880, https://doi.org/10.1007/s10489-023-04915-8.

[52] J. Choi, S. Seo, S. Choi, S. Piao, C. Park, S.J. Ryu, B.J. Kim, S. Park, ReBADD-SE: multi-objective molecular optimisation using SELFIES fragment and off-policy self-critical sequence training, Comput. Biol. Med. 157 (2023) 106721, https://doi.org/10.1016/j.compbiomed.2023.106721.

[53] F. Grisoni, Chemical language models for de novo drug design: Challenges and opportunities, Curr. Opin. Struct. Biol. 79 (2023) 102527, https://doi.org/10.1016/j.sbi.2023.102527.

[54] G. Lamanna, P. Delre, G. Marcou, M. Saviano, A. Varnek, D. Horvath, G. F. Mangiatordi, GENERA: a combined genetic/deep-learning algorithm for multiobjective target-oriented de novo design, J. Chem. Inf. Model. 63 (2023) 5107–5119, https://doi.org/10.1021/acs.jcim.3c00963.

[55] G.F. Mangiatordi, F. Intranuovo, P. Delre, F.S. Abatematteo, C. Abate, M. Niso, T. M. Creanza, N. Ancona, A. Stefanachi, M. Contino, Cannabinoid receptor subtype 2 (CB2R) in a multitarget approach: perspective of an innovative strategy in cancer and neurodegeneration, J. Med. Chem. 63 (2020) 14448–14469, https://doi.org/10.1021/acs.jmedchem.0c01357.

[56] F. Intranuovo, L. Brunetti, P. DelRe, G.F. Mangiatordi, A. Stefanachi, A. Laghezza, M. Niso, F. Leonetti, F. Loiodice, A. Ligresti, M. Kostrzewa, J. Brea, M.I. Loza, E. Sotelo, M. Saviano, N.A. Colabufo, C. Riganti, C. Abate, M. Contino, Development of N-(1-Adamantyl)benzamides as novel anti-inflammatory multitarget agents acting as dual modulators of the cannabinoid CB2 receptor and fatty acid amide hydrolase, J. Med. Chem. 66 (2023) 235–250, https://doi.org/10.1021/acs.jmedchem.2c01084.

[57] G.F. Mangiatordi, M.M. Cavalluzzi, P. Delre, G. Lamanna, M.C. Lumuscio, M. Saviano, J.-P. Majoral, S. Mignani, A. Duranti, G. Lentini, Endocannabinoid degradation enzyme inhibitors as potential antipsychotics: a medicinal chemistry perspective, Biomedicines 11 (2023) 469, https://doi.org/10.3390/biomedicines11020469.

[58] P. Delre, M. Contino, D. Alberga, M. Saviano, N. Corriero, G.F. Mangiatordi, ALPACA: a machine Learning Platform for Affinity and selectivity profiling of CAnnabinoids receptors modulators, Comput. Biol. Med. 164 (2023) 107314, https://doi.org/10.1016/j.compbiomed.2023.107314.

[59] A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, J.P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery, Nucleic Acids Res. 40 (2012) D1100–D1107, https://doi.org/10.1093/nar/gkr777.

[60] M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, B. Wiswedel, KNIME - the Konstanz information miner: version 2.0 and beyond, SIGKDD Explor. Newsl 11 (2009) 26–31, https://doi.org/10.1145/1656274.1656280.

[61] S. Beisken, T. Meinl, B. Wiswedel, L.F. de Figueiredo, M. Berthold, C. Steinbeck, KNIME-CDK: workflow-driven cheminformatics, BMC Bioinf. 14 (2013) 257, https://doi.org/10.1186/1471-2105-14-257.

[62] G. Landrum, P. Tosco, B. KelleyRic, D. Cosgrovesriniker, gedeck, R. Vianello, NadineSchneider, E. Kawashima, G. Jones, D. N, A. Dalke, B. Cole, M. Swain, S. Turk, AlexanderSavelyev, A. Vaucher, M. Wójcikowski, I. Take, V.F. Scalfani, D. Probst, K. Ujihara, G. godin, A. Pahl, R. Walker, J. Lehtivarjo, F. Berenger, jasondbiggs, strets123, rdkit/rdkit: 2023_09_4 (Q3 2023) Release. https://doi.org/10.5281/zenodo.10460537, 2024.

[63] N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, Open Babel: an open chemical toolbox, J. Cheminf. 3 (2011) 33, https://doi.org/10.1186/1758-2946-3-33.

[64] D. Gadaleta, A. Lombardo, C. Toma, E. Benfenati, A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications, J. Cheminf. 10 (2018) 60, https://doi.org/10.1186/s13321-018-0315-6.

[65] J.L. Elman, Finding structure in time, Cognit. Sci. 14 (1990) 179–211, https://doi.org/10.1207/s15516709cog1402_1.

[66] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735.

[67] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv.Org, https://arxiv.org/abs/1412.6980v9, 2014. (Accessed 12 January 2024).

[68] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back-propagating errors, Nature 323 (1986) 533–536, https://doi.org/10.1038/323533a0.

[69] T. Hua, X. Li, L. Wu, C. Iliopoulos-Tsoutsouvas, Y. Wang, M. Wu, L. Shen, C. A. Brust, S.P. Nikas, F. Song, X. Song, S. Yuan, Q. Sun, Y. Wu, S. Jiang, T.W. Grim, O. Benchama, E.L. Stahl, N. Zvonok, S. Zhao, L.M. Bohn, A. Makriyannis, Z.-J. Liu, Activation and signaling mechanism revealed by cannabinoid receptor-gi complex structures, Cell 180 (2020) 655–665.e18, https://doi.org/10.1016/j.cell.2020.01.008.

[70] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M. P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P. Francis, P.S. Shenkin, Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy, J. Med. Chem. 47 (2004) 1739–1749, https://doi.org/10.1021/jm0306430.

[71] Schrödinger Release 2023-4, Schrödinger, LLC, New York, NY, 2023 (n.d.).

[72] Z. Deng, C. Chuaqui, J. Singh, Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional Protein−Ligand binding interactions, J. Med. Chem. 47 (2004) 337–344, https://doi.org/10.1021/jm030331x.

[73] V.I. Levenshtein, et al., Binary codes capable of correcting deletions, insertions, and reversals, Sov. Phys. Dokl. 10 (1966) 707–710.

[74] D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, S. Golovanov, O. Tatanov, S. Belyaev, R. Kurbanov, A. Artamonov, V. Aladinskiy, M. Veselov, A. Kadurin, S. Johansson, H. Chen, S. Nikolenko, A. Aspuru-Guzik, A. Zhavoronkov, Molecular sets (moses): a benchmarking platform for molecular generation models, Front. Pharmacol. 11 (2020). https://www.frontiersin.org/articles/10.3389/fphar.2020.565644. (Accessed 12 January 2024).

[75] N. Brown, M. Fiscato, M.H.S. Segler, A.C. Vaucher, GuacaMol: Benchmarking Models for de Novo Molecular Design, J. Chem. Inf. Model. 59 (2019) 1096–1108, https://doi.org/10.1021/acs.jcim.8b00839.

[76] M. Levandowsky, D. Winter, Distance between sets, Nature 234 (1971) 34–35, https://doi.org/10.1038/234034a0.

[77] H.L. Morgan, The generation of a unique machine description for chemical structures-A technique developed at chemical abstracts service, J. Chem. Doc. 5 (1965) 107–113, https://doi.org/10.1021/c160017a018.

[78] G.R. Bickerton, G.V. Paolini, J. Besnard, S. Muresan, A.L. Hopkins, Quantifying the chemical beauty of drugs, Nat. Chem. 4 (2012) 90–98, https://doi.org/10.1038/nchem.1243.

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

[80] J.L. Dahlin, J.W.M. Nissink, J.M. Strasser, S. Francis, L. Higgins, H. Zhou, Z. Zhang, M.A. Walters, PAINS in the assay: chemical mechanisms of assay interference and promiscuous enzymatic inhibition observed during a sulfhydryl-scavenging HTS, J. Med. Chem. 58 (2015) 2091–2113, https://doi.org/10.1021/jm5019093.

[81] G. Weng, H. Zhao, D. Nie, H. Zhang, L. Liu, T. Hou, Y. Kang, RediscMol: benchmarking molecular generation models in biological properties, J. Med. Chem. 67 (2024) 1533–1543, https://doi.org/10.1021/acs.jmedchem.3c02051.

[82] T.M. Creanza, P. Delre, N. Ancona, G. Lentini, M. Saviano, G.F. Mangiatordi, Structure-based prediction of hERG-related cardiotoxicity: a benchmark study, J. Chem. Inf. Model. 61 (2021) 4758–4770, https://doi.org/10.1021/acs.jcim.1c00744.

[83] M.C. Lomuscio, C. Abate, D. Alberga, A. Laghezza, N. Corriero, N.A. Colabufo, M. Saviano, P. Delre, G.F. Mangiatordi, AMALPHI: a machine learning platform for predicting drug-induced PhospholIpidosis, Mol. Pharm. (2023), https://doi.org/10.1021/acs.molpharmaceut.3c00964.

[84] Z.M. Whiting, J. Yin, S.M. de la Harpe, A.J. Vernall, N.L. Grimsey, Developing the Cannabinoid Receptor 2 (CB2) pharmacopoeia: past, present, and future, Trends Pharmacol. Sci. 43 (2022) 754–771, https://doi.org/10.1016/j.tips.2022.06.010.
[85] S. Huang, P. Huang, L. Wang, J. Han, Y. Chen, H. Zhong, Halogenated-methylammonium based 3D halide perovskites, Adv. Mater. 31 (2019) 1903830, https://doi.org/10.1002/adma.201903830.
[86] B. Bienfait, P. Ertl, JSME: a free molecule editor in JavaScript, J. Cheminf. 5 (2013) 24, https://doi.org/10.1186/1758-2946-5-24.
[87] S. Tian, Y. Djoumbou-Feunang, R. Greiner, D.S. Wishart, CypReact: a software tool for in silico reactant prediction for human cytochrome P450 enzymes, J. Chem. Inf. Model. 58 (2018) 1282–1291, https://doi.org/10.1021/acs.jcim.8b00035.
[88] G. Graziano, P. Delre, F. Carofiglio, J. Brea, A. Ligresti, M. Kostrzewa, C. Riganti, C. Gioè-Gallo, M. Majellaro, O. Nicolotti, N.A. Colabufo, C. Abate, M.I. Loza, E. Sotelo, G.F. Mangiatordi, M. Contino, A. Stefanachi, F. Leonetti, N-adamantyl-anthranil amide derivatives: new selective ligands for the cannabinoid receptor subtype 2 (CB2R), Eur. J. Med. Chem. 248 (2023) 115109, https://doi.org/10.1016/j.ejmech.2023.115109.