

Focused Library Generator: Case of Mdmx Inhibitors

Zhonghua Xia¹, Pavel Karpov^{1,2}, Grzegorz Popowicz¹, and Igor V. Tetko^{1,2}

¹Helmholtz Zentrum München - Research Center for Environmental Health (GmbH), Institute of Structural Biology, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany,

²BigChem GmbH, Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany
zhonghua.xia, pavel.karpov, grzegorz.popowicz, i.tetko "@ " helmholtz-muenchen.de

Abstract

We present a Focused Library Generator that is able to create from scratch new molecules with desired properties. After training the Generator on the ChEMBL database, transfer learning was used to switch the generator to producing new Mdmx inhibitors that are a promising class of anticancer drugs. Lilly medicinal chemistry filters, molecular docking, and a QSAR IC₅₀ model were used to refine the output of the Generator. Pharmacophore screening and molecular dynamics (MD) simulations were then used to further select putative ligands. Finally, we identified five promising hits with equivalent or even better predicted binding free energies and IC₅₀ values than known Mdmx inhibitors. The source code of the project is available on <http://github.com/bigchem/online-chem>.

Introduction

To find molecules that maximize the desirable properties and simultaneously minimize their adverse effects is a significant challenge in the modern chemically-related sciences. This is especially true for drug discovery research where a few atoms have to “encode” specific activity, selectivity, pharmacokinetics etc. The number of chemical compounds is enormous [1, 2], with drug-like space estimated to be 10⁶⁰ molecules, and to even enumerate all of them is an impossible task much less consider synthesis and experimental testing. Several approaches – to name the most beneficial, virtual screening, and *in silico* structure generation – have been developed to enrich the outcome of projects involved in designing new molecules with optimized properties. The former approaches (filters) screen a database of existing organic compounds, (e.g. the ZINC database), or just synthesizable (Enamine REAL [3]), against several prognostic models built within QSAR/QSPR (Quantitative Structure-Activity/Property Relationship) methodology, and select the most promising candidates for further analysis and experimental testing. The latter strategies (generators) can computationally generate new molecules [4] with desired

predicted properties using experimentally validated training sets [5]. Such molecules can be later produced and assayed. The main advantage of the latter procedures is that they can potentially navigate in unexplored areas of chemical space not represented in existing libraries. Such compounds offer potentially better specificity, selectivity, and pharmacokinetic properties. They are also unlikely to be covered by existing intellectual property limitations.

The generators are based on the recent progress with sequence data processing methods. Examples of such data are natural language sentences or music where each subsequent piece of information (words or notes) depends on the previous parts. Since the first work on sequence generation revealed the possibility for recurrent neural networks to model the probability distribution of the next token in a sequence [6], the scientific community immediately realized its power and opportunities. Many works appeared in different data domains, including among others, chemoinformatics [4]. The chemical molecules can be presented as ordinary alphanumeric strings by means of Simplified Molecular Input Line Entry System (SMILES) [7] from which the model can learn the conditional probability distribution of the next symbol with respect to all previous tokens. Once trained, the model can generate a new and diverse set of valid SMILES that are similar to the training dataset [8]. Though the ability to generate an arbitrary compound from "nothing" looks promising and appealing, its practical application is not different from exploiting combinatorial libraries. Therefore, it is crucial to find a way to adapt the model for sampling new molecules from the desired region of chemical space. For this purpose, transfer learning techniques, where the pre-trained model continues training on a task-specific dataset, are widely used [9]. After several epochs, the model converges and begins producing SMILES that are similar to such datasets. Another way to restrict molecular generator is to apply a reinforcement learning paradigm and bias selection of symbols from high probability variants drawn from the predicted distribution [5, 10]. Recurrent neural network architecture shows exceptionally good performance in sequence-related tasks. During sequence processing, the network stores the entire "idea" of the input data in its internal state. This encoding vector can be successfully converted to new SMILES by using it as the starting internal state of another recurrent unit. Thus, other techniques, for example, Generative Adversarial Networks (GAN) [11, 12], and Variational Autoencoders [13], can efficiently navigate in this encoding space and enforce the recurrent network to produce focused SMILES of interest.

In our work we demonstrated the applicability of such approaches to model the mouse double minute x (Mdmx, also known as Mdm4), which is a critical negative regulator of a tumor suppressor protein p53 [14]. Mdmx triggered a wave of research since the discovery of its importance to cancer progression [15]. The mutation of TP53 appears in about half of human cancers [16], and in the remnant, the wild-type p53 is inactivated by its negative regulators, Mdm2 and Mdmx proteins [17]. The former is the first experimentally identified as a major endogenous inhibitor of p53 [18–20], and the latter, which is highly homologous with Mdm2, influences p53 transcriptional activity [14] and is a vital independent regulator of p53 [21]. Until now, however, the studies on small-molecule Mdmx inhibitors have made slow progress. There are nine Mdm2 inhibitors in clinical studies [22, 23], but for Mdmx, there is none. Development of SJ-172550, the first reported inhibitor of Mdmx [24], was put on hold because of stability problems [25], other alternatives, like XI-011 [26], and NSC207895 [27] had difficulties at the preclinical trials.

Even though Mdm2 and Mdmx are structurally closely related, their slight differences lead to Mdm2 inhibitors or even p53 transactivation domain being inferior in terms of binding affinity to Mdmx [28]. The usage of Mdm2 inhibitors in some cases is significantly limited due to its toxicity in normal cells [29]. Recently, it was reported that the loss of Mdmx induces p53 activation but has much less destructive effects *in vivo* than Mdm2 inhibition [30]. Therefore, Mdmx remains an interesting and unexplored target of great potential for cancer treatment.

For the first time, we applied generative neural networks for designing new putative Mdmx inhibitors *in silico* and validated our approach with MD calculations. We showed that the Generator equipped with appropriate constraints, e.g. molecular docking, and a QSAR regression model, can efficiently navigate in chemical space with a high density of promising inhibitors. The overall methodology, as well as the source code provided (<http://github.com/bigchem/online-chem>), can be applied to other drug-design projects with minor changes.

Methods

Datasets

ChEMBL library of SMILES For our primary generator model we selected 1,727,112 compounds from the ChEMBL library [31]. As distinct from all other

works on SMILES generators, we did not perform a pre-processing of the original data, e.g., string size limitation, chirality removing, standardization, and chemistry-relevant tokenization. This was done with a goal of training a universal recurrent model that would be able to cover the whole ChEMBL database and then, finding weights (“embedding”) that could also be used separately from the original generator model. This SMILES-embedding, like word-embeddings which are actively used for construction of language models, could be also applied to QSAR modeling [32]. 77,112 randomly selected SMILES were used as a validation set while the remaining 1,650,000 SMILES were used as a training set.

Mdmx inhibitors The training set of inhibitors (total 293) with experimental IC_{50} values [33–44], which have been tested with human Mdmx, were selected from the ChEMBL and BindingDB [45] databases. Their IC_{50} values were measured by several kinds of methods,

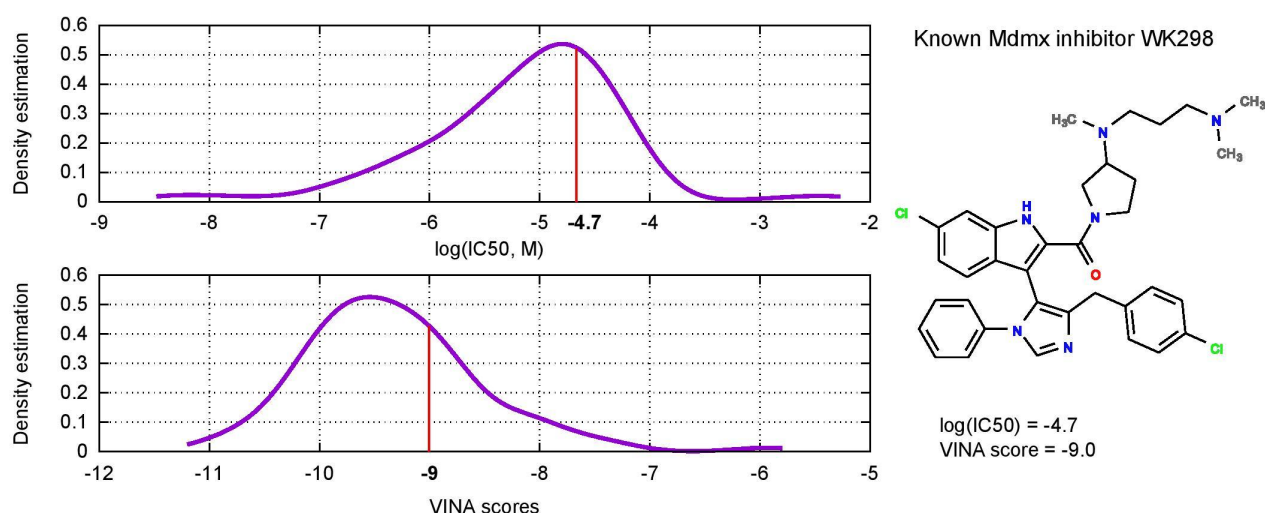
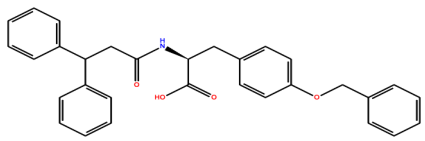
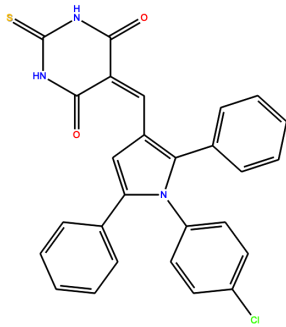
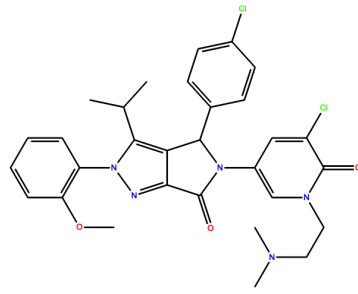


Figure 1. Distribution of $\log(IC_{50})$ experimental values, and also VINA scores for the known Mdmx inhibitor.

such as the enzyme-linked immunosorbent assay (ELISA) (21 compounds), the dose-response confirmation of Inhibitors of Mdm2/MdmX interaction in luminescent format (19 compounds), the time resolved fluorescence energy transfer (TR-FRET) assay (237 compounds), fluorescence polarization assay (FPA) (2 compounds), the quantitative sandwich immune-enzymatic assay (14 compounds). Most of the inhibitors had $\log(IC_{50})$ values less than -4.0, Figure 1.

The average Tanimoto similarity within the dataset is 0.56. One can cluster all compounds in three different scaffolds, Table. 1, though only one cluster contains the majority of inhibitors (247 compounds).

Table 1. Diversity of the original inhibitors. Numbers mean the number of compounds in a respective cluster.

		
30 compounds	16 compounds	247 compounds

SMILES generator neural network (SGNN)

The generator neural network model consists of 2 LSTM (Long-Short Term Memory) [46] layers stacked upon each other with the number of internal units equal to 512 and hyperbolic tangent as a nonlinearity activation function. Stacked recurrent layers are commonly used for generator architectures but vary in the number of layers and the number of processing units in each layer, Table 2.

Table 2. Neural network parameters from the literature.

Deep-learning framework	Number and type of layers	Number of units	Reference
Keras [49]	3 LSTM	64	[8]
Keras	3 LSTM	1024	[4]
Keras	3 LSTM	512	[9]
PyTorch [50]	3 GRU [51]	512	[47]
Keras	2 LSTM	256	[48]

The last model in the Table 2 contains the smallest number of layers. We chose this architecture but with a bigger number of units equal to 512, which resulted in 3,306,043 parameters. SGNN vocabulary is bigger (stereochemistry and inorganic ions) than in [48], so more training parameters are needed to encode the information. Following the last LSTM layer, we added an ordinary dense layer with a softmax output, Figure 2. To facilitate batch computation, we applied

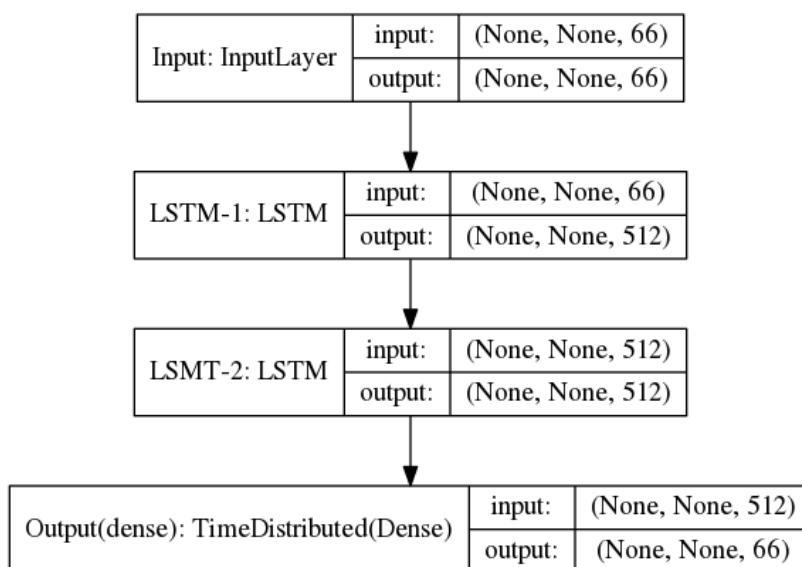


Figure 2. The SGNN model architecture.

masking to differentiate which positions in the input sequence are valid symbols. By design, our model can work with batches of different lengths, also consisting of strings of various sizes.

The input vocabulary was fixed to 66 letters including characters for the start (^) and the end (\$) of a string

\$#%()+-./0123456789=@ABCFGHIKLMNOPRSTVXZ[\]abcdegilnoprstu^

Besides drug-like compounds, the proposed vocabulary can additionally encode inorganic ions often co-crystallized with the core fragment. It also can work with common reagents in organic synthesis, thus, facilitating the use of the same molecular representation in different models, being a universal SGNN model.

Each letter was one-hot encoded and then passed to recurrent units. We did not employ any tokenization scheme for the SMILES generation model because, as stated above, the primary goal was to train a universal model covering the whole ChEMBL database. Therefore, any restriction, even if it may seem chemistry-relevant, could lead to an under-representation of some classes of tokens in the training dataset, thus, worsening the quality of the entire model. Our preliminary analysis showed that using the popular tokenization scheme [49] over simple character-based treatment did not improve the accuracy of the model. The output of the SGNN is a probability distribution of the next symbol conditioned on all previous symbols in a string over the vocabulary. During decoding, we randomly pick a symbol from the vocabulary according to this distribution. Sampling at higher temperatures results in a smoother probability

landscape and, therefore, can produce more diverse but also invalid SMILES. To adjust the softmax output to different temperatures, the following formula is used (see details in Supporting Information):

$$p_i = \frac{e^{\frac{\ln(y_i)}{T}}}{\sum_i e^{\frac{\ln(y_i)}{T}}} \quad (1)$$

where y_i is the output of the SGNN, T is the temperature, and p_i is the corrected probability.

The SGNN was implemented in Lasagne [50] with Theano backend [51]. SMILES validation was done using RDKit library [52]. SGNN was trained for 100 epochs with the Adagrad optimization algorithm [53] (constant learning rate 0.01, gradient clipping 100, batch size 2048). To avoid overfitting, we used the early-stopping technique with a validation dataset. However, we did not see the overfitting and stopped training after the maximum number of epochs.

Molecular docking

The crystal structure of Mdmx in complex with p53 (PDB ID: 3dab [54]) at a resolution of 1.9 Å was used for the docking studies. The receptor was protonated at neutral pH with only polar hydrogens by AutoDockTools-1.5.6 and Gasteiger charges were added. By centering on p53, the docking grid was created. In particular, a grid box size of 24×18×22 and centered at 0.514 (x), -21.838 (y), and 8.047 (z) was set with spacing of 1.0 Å between the grid points. The ligands were docked using the AutoDock VINA [55] program (hereafter simply called VINA). Ten predicted conformations were generated for each ligand. The rest of the parameters were set to the default values. From the 10 binding poses, we chose the optimal one in terms of the docking score.

QSAR models for IC₅₀ and solubility estimation

To estimate IC₅₀ values of new Mdmx inhibitors we built a regression model using the OCHEM platform. OCHEM platform [56] automatically estimates the applicability domain (AD) [57] that is of crucial need for the current project. The generator treats molecules as simple sequences of characters with predefined grammar rules. Though the final sequence may be a valid SMILES, it can result in a molecule, which is chemically non-synthesizable or is very

different from the training set for which IC_{50} estimation may be unreliable. The QSAR model was used for additional filtering of the generated compounds selecting only those molecules that are in its AD. We explored different machine learning methods and descriptors available on OCHEM website and found that the best model was based on a Transformer-CNN method [58]. This model's parameters are coefficient of determination $r^2 = 0.69$, root mean squared error RMSE=0.51, and mean absolute error MAE = 0.35 (n=293). The model is publicly available on <http://ochem.eu/model/785>. We used all 293 compounds for modeling though some values were measured with different methods because the number of points on individual assays were not enough for building and validating a model (see Datasets section).

Though our primary interest was to find new molecules that can bind to Mdmx we also ran the full generation cycle with solubility as one of the constraints besides IC_{50} and VINA endpoints. The solubility model used in this study was built with the same Transformer-CNN approach and resulted in $r^2 = 0.92$, RMSE=0.58, and MAE = 0.41 (n=1311) and is available on <http://ochem.eu/model/784>.

Tuning the generator

After training the generator on the ChEMBL database, we tuned it with the Mdmx dataset according to the following algorithm:

1. Retraining the SGNN model solely on known Mdmx inhibitors, which formed the transfer set. The number of unique molecules in the transfer set was small compared to the initial ChEMBL training dataset. Therefore, we augmented (maximum 10 times) this set with non-canonical SMILES [59, 60], which were shown to increase the accuracy of the models. For example, toluene canonical SMILES is Cc1ccccc1, and non-canonical variants are: c1cccc(C)c1, c1ccc(C)c1, c1c(cccc1)C, etc. Both canonical and non-canonical SMILES were used in the training set for the SGNN.
2. New SMILES were generated by SGNN. The compounds which passed a Lipinski-like filter (MW > 100 and MW < 700, HBD < 5, and HBA < 10) and a collection of Lilly filters [61] were then docked with VINA [55]. An IC_{50} QSAR model was also used to estimate the IC_{50} values for the molecules generated. Both docking scores and IC_{50} values were saved in the temporary database.

3. The generated compounds were ranked by considering both IC_{50} and VINA scores simultaneously. According to Fig. 1, most active compounds have $-8 \leq \log(IC_{50}) \leq -4$ and $-11 \leq \text{VINA score} \leq -8$. The Spearman correlation coefficient between VINA scores and $\log(IC_{50})$ is equal to -0.16 thus showing that both endpoints presumably capture different aspects of structural features allowing compounds to be active. Therefore, we normalized values of both endpoints to the same interval and used the average of the normalized values as the final score for the molecules. A new training dataset was formed from the original data and highly-scored putative ligands from Step 2, keeping the ratio of known and generated ligands 1:1. As in step 1, all SMILES were augmented.
4. The SGNN model was retrained on this new dataset.
5. The sampling temperature was increased according to the formula, as discussed in the Results and Discussion section:

$$T_{cycle} = 1.0 + 0.5 * (cycle - 1) \quad (2)$$

Steps 2-5 were repeated by generating new SMILES at the end of each cycle. In this study, we performed two separate runs for tuning the generator. The first run (run-1) used the VINA score and IC_{50} values to select promising inhibitors. The second run (run-2) utilized additional solubility values. Note, additional filters decrease the number of generated molecules, and for chemical space exploration, one needs to increase the number of attempts while generating compounds. In run-2, we generated ten times more SMILES compared to run-1.

Virtual screening

To perform virtual screening of the output of the Generator (run-1), a pharmacophore model was built by Schrödinger Phase [62, 63] based on the known small-molecule Mdmx inhibitor WK298 [64], which formed a complex with Mdmx. The structure of this complex was obtained by docking WK298 into Mdmx. The initial conformations of WK298 and Mdmx for docking were extracted respectively from crystal structures with PDB IDs of 3l bj [65] and 3dab [54]. The top-scored binding pose was well superimposed with the crystal structure of WK298-Mdmx complex and when it was aligned with the crystal structure of p53-Mdmx complex, we can see WK298 binds to Mdmx in a way that mimics parts of the binding of the native p53 peptide. The Trp23 pocket is filled with the 6-chloroindole substituent and the 4-chlorobenzyl ring protrudes the Leu26 pocket.

Finally the third key substituent of WK298, the phenyl ring, occupies the Phe19 pocket, although the plane of this ring is nearly perpendicular to the plane of the Phe19 side chain of p53. The pharmacophore model, Figure 3, was based on the aforementioned critical residues of p53, (Phe19, Trp23, and Leu26) [65]. The model was validated internally in Schrödinger suite. Hereby, 204 of 274 active compounds have been successfully retrieved and the parameters are Area Under the Curve (AUC) 0.72, EF1% 10.62. We used this pharmacophore model to screen all the stereoisomers from the Schrödinger LipPrep. Finally, we analyzed molecules with Molecular Dynamic (MD) simulations.

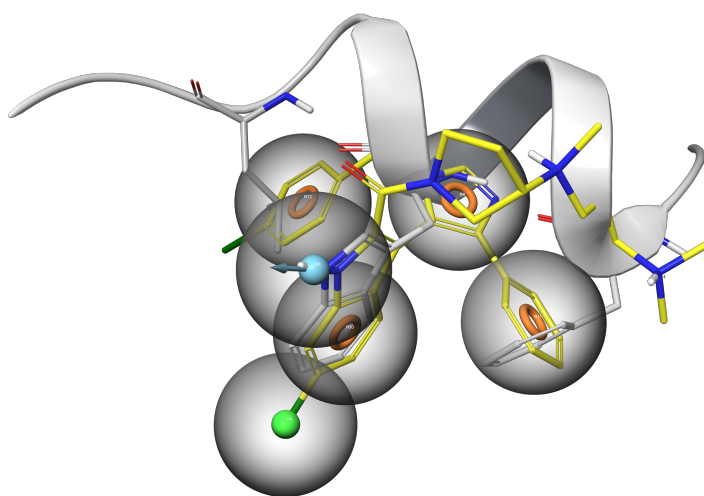


Figure 3 The pharmacophore model of Mdmx inhibitor used in this study (yellow: WK298, white: p53)

MD simulations were performed with AmberTools18 [66], including ligand reduction with Reduce; atomic charge calculation with Antechamber and am1bcc as the backend; topology and coordinate files generation with tleap. Each MD simulation ran for 10 ns of production time under the NPT ensemble referring to the constant 0.987 atm with isotropic position scaling and 300 K using the weak-coupling and SHAKE algorithm; maximum time step of 2 fs was used. The default value of the nonbonded cutoff was used properly along with the Particle Mesh Ewald (PME) calculations. For the stable ones which have a small standard deviation of RMSD value (“sd” in Table 3), we further calculated the binding free energies between the protein and ligands using the MMPBSA.py Python script [67] provided by AmberTools18. It is known that post-processing trajectories according to Molecular Mechanics/ Poisson Boltzmann (or Generalized Born) Surface Area (MM/PB(GB)SA) approximation allows a better evaluation of the binding patterns of these ligands to Mdmx and, therefore, can

help to identify the most potent hits [68]. We selected 100 frames from the last nanosecond of simulation with 10 ps intervals for calculation of ΔG values. The ΔG calculation was not refined by the entropy contribution in order to reduce the computational time. Therefore, the calculated ΔG was only an approximation of true free energy but could be used to compare the generated ligands with similar Mdmx binding pattern.

The parameters and the environment for MD simulations worked well on the WK298-Mdmx complex. We were able to get stable trajectories by running 10ns MD simulations on this complex, according to the RMSD (see Supporting information).

Results and Discussion

The generator

We examined the ability of the Generator to produce chemical structures following a comparison of distributions for molecular weight, lipophilicity, topological surface area, and synthetic accessibility introduced in [8]. For this purpose, we sampled approximately ten millions SMILES from the SGNN, and 79% of them were valid unique SMILES, see Table 3. The results in Figure 4 (blue dotted and solid lines labeled with "ChEMBL" and "ChEMBL (generated)") showed a good correlation between distribution shapes for different properties.

Table 3. Parameters of generated molecules.

Strings	All	Valid SMILES	Unique SMILES
All	10 495 701	9 377 274 (89.3%)	8 297 705 (79.1%)
Stereo (with @)	1 677 294	1 388 454 (13.2%)	1 325 236 (12.6%)
Cis/trans (with / or \)	1 153 284	998 506 (9.5%)	905 536 (8.6%)

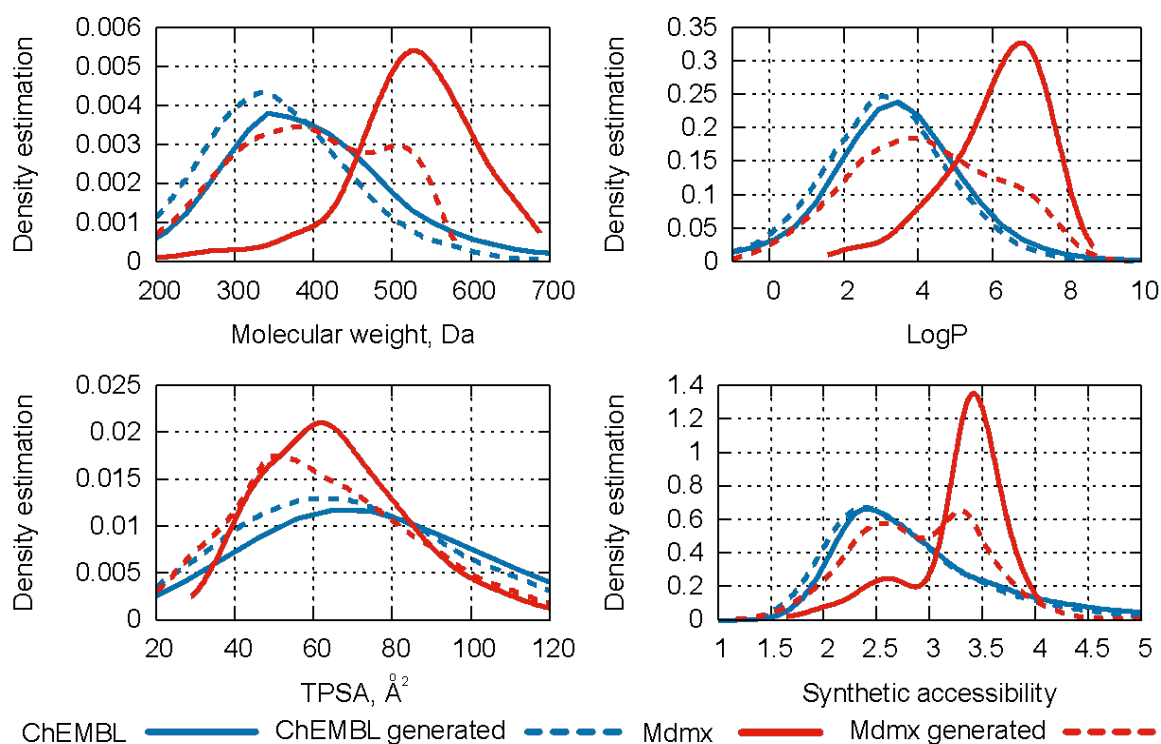


Figure 4 Distribution of molecular weight, logP, TPSA, and synthetic accessibility for original ChEMBL database, generated molecules before and after tuning SGNN model with Mdmx inhibitors.

Notably, our generator sampled 13.2% valid molecules with chiral centers, and 9.5% geometric isomers with cis/trans configurations.

The Mdmx transfer dataset of known inhibitors consisted of 293 molecules. It was augmented to 2916 SMILES and then used for retraining the original ChEMBL model. While an early-stopping technique was not required for training using ChEMBL, it was essential on this step because the number of parameters of the model was huge compared to the amount of training data, and the model could easily overfit [69]. In our experiments, 3-4 epochs were sufficient for transfer of the information. Total 5803 (run-1) valid SMILES that passed drug-like Lilly filters were generated during all 10 cycles of tuning and 778 with VINA score less than -7.5 were selected for further virtual screening.

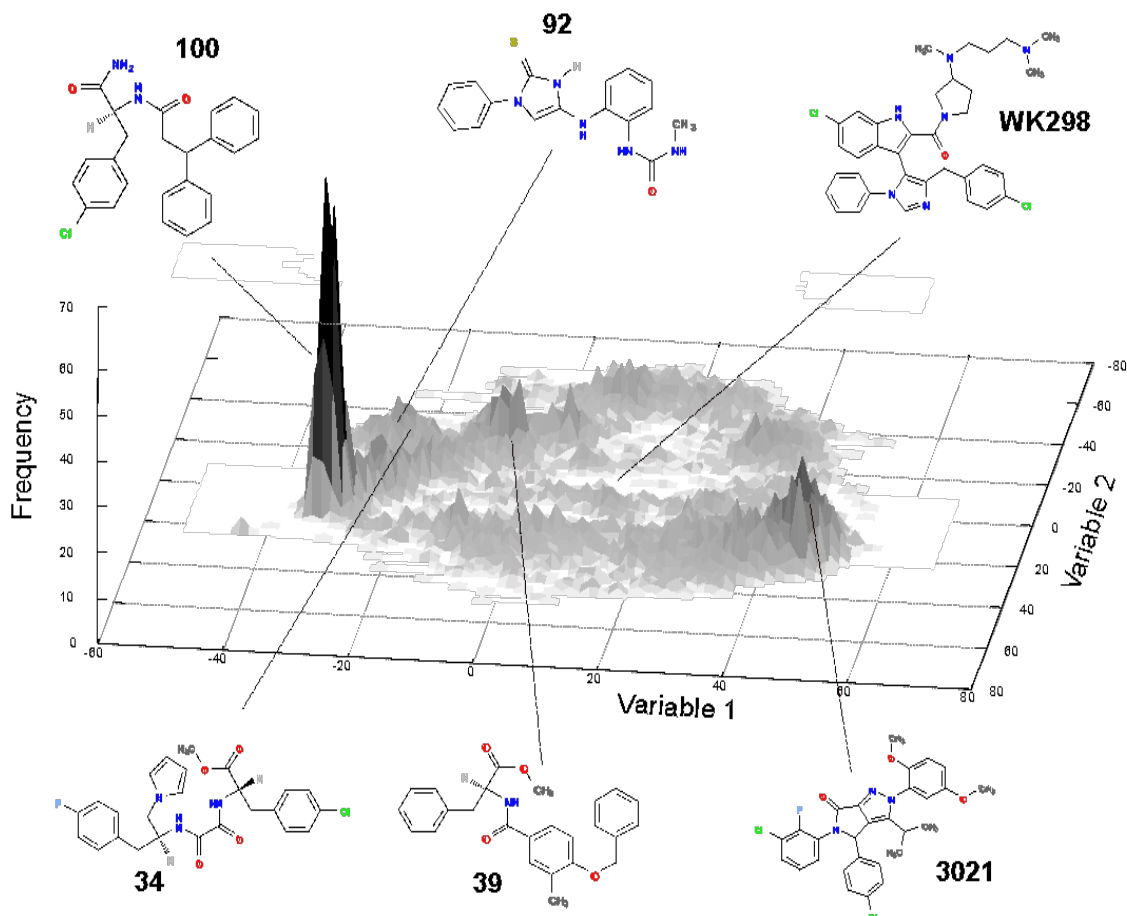


Figure 5. Visualization of chemical space of Mdmx inhibitors.

The visualization of the chemical space of generated inhibitors, Figure 5, based on the deep neural network model [70] reveals an interesting landscape consisting of a valley with the known WK298 inhibitor in the middle and two ridges. Most of compounds selected by MD simulations are located at the highest and most occupied ridge to the left of the valley. The landscape evidently shows that the generator explored particular directions in the vast chemical space, and it was not a pure random walk.

The temperature, Equation 1, can control the diversity of the generated SMILES. At low T , the system can produce valid compounds similar to the training set and with excellent quality. But to design new creative molecules, we need to add curiosity to the generator by increasing T . The increase of temperature decreased the number of valid SMILES which was 66% for $T = 1.45$, see Supporting Information. This higher temperature would cause the generator to produce garbage more in more than 1/3 of the outputs. In this study we made 10 such cycles and the corresponding maximum temperature was $T = 1.45$. The authors [48] used 5-12 cycles for a similar procedure using uniqueness of

generated SMILES and their distance to the original dataset as a stopping criteria.

Mdmx inhibitors

The 5803 molecules, which passed advanced drug-likeness filters, were produced by the Generator and were ranked according to the sum of normalized VINA scores and $\log(\text{IC}_{50})$ values (Fig. 6). Then pharmacophore model filtered the top 102 molecules to 96. We simulated Mdmx complexes with these ligands and 50 trajectories were obtained.

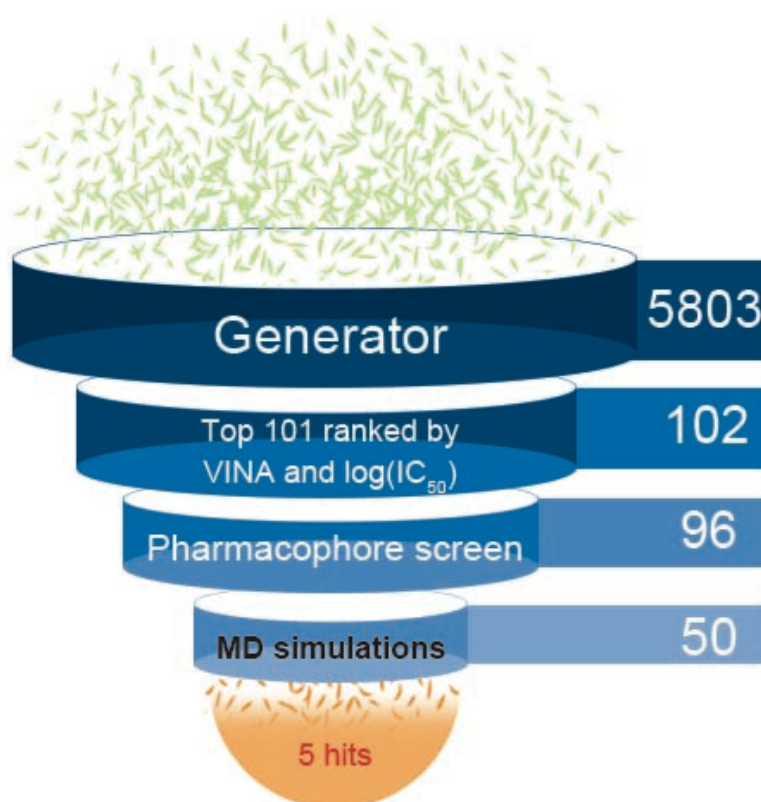


Figure 6. The scheme of the virtual screening workflow.

To investigate the stability of the ligands' binding pose in the receptor pocket, we analyzed the RMSD values for the ligand of the Mdmx complex. For the known inhibitor WK298, which was subjected to the same process, the standard deviations of RMSD are in the range of 0.317 to 2.758. We set the threshold to 2.758 to filter the relatively unstable trajectories. This filter further decreased the number of molecules to 49, for which we also estimated the binding free energies. Theoretical ΔG_{exp} can be estimated by experimental IC_{50}

values with formula $\Delta G_{\text{exp}} \approx RT \ln(\text{IC}_{50})$ [68, 71]. For WK298 with IC_{50} value of 19.7 μM [65], its ΔG_{exp} is around -6.5 kcal/mol set. Below we analyze 5 promising hits (Table 4) with $\Delta G_{\text{exp}} < -6.5$, which could be thus expected to have a higher activity than WK298.

Table 4. RMSD and the binding free energy (ΔG) of the representative compounds and WK298 ascending by ΔG

Compound	RMSD				$-\log(\text{IC}_{50})$	ΔG (kcal/mol)
	avg ^a	std ^b	min ^c	max ^d		
WK298	1.5	0.3	0.52	2.9	-4.7	-9.4
3021	4.7	0.4	0.63	5.9	-5.2	-13.0
92	1.6	0.5	0.51	3.7	-7.7	-10.8
100	1.7	0.7	0.48	5.8	-7.9	-6.9
34	2.8	0.7	0.67	5.2	-7.9	-6.7
39	4	1	0.78	8.0	-7.6	-6.7

a. avg = the average; b. std = standard deviation; c. min = the minimum; d. max = the maximum

Compound 92 has a unique scaffold that is 3,5-disubstituted-2-thioxo-2,3-dihydro-1*H*-imidazole. We aligned its Mdmx complex with the crystal structure of p53-Mdmx complex (PDB ID: 3dab). As shown in Fig. 8c, its urea group points towards the indole ring of the p53 residue Trp23 and both of its imino groups have H-bond interaction with Met32. 3-phenyl is fairly close to Phe19 pocket. The three aromatic rings of it lay on the top of helix $\alpha 2$ of Mdmx and the imino group has H-bond interaction with His33. Moreover, Tyr78 of Mdmx kept the “closed” conformation, which was specially studied in [65].

3,3-diphenylpropanamide is one of the common substructures among generated molecules, such as compound 100. 4-chlorophenyl is located in the Trp23 pocket. One of the two-phenyl rings occupies the space of Lys24 of p53 and the other one is exposed to the solvent. In the case of compound 34, one phenyl ring is substituted by its bioisosteres like pyrrole and methyl ester. The latter appears more frequently so as to become another common fragment. Here, 4-fluorophenyl mimics Trp23 of p53 and 4-chlorophenyl mimics Phe19. Besides, 1*H*-pyrrole toward the opposite direction is aligned with Lys24.

Outside the range of filtered molecules, compound 39 with ΔG of -6.7 kcal/mol also has an ester-substituted two-phenyl “umbrella”. Its long straight chain lies across the cleft of the protein and the hydrophobic pocket shrinks significantly to fit around it. In the crystal structure of Mdmx complex with Mdmx

selective inhibitor (PDB ID: 6q9y; compound 16, IC_{50} of 3.7 μ M), we can observe the similar binding conformation [72]. Apart from filling the Trp23 and Leu26 pockets, the remaining phenyl group nicely overlaps with Ser20 of p53.

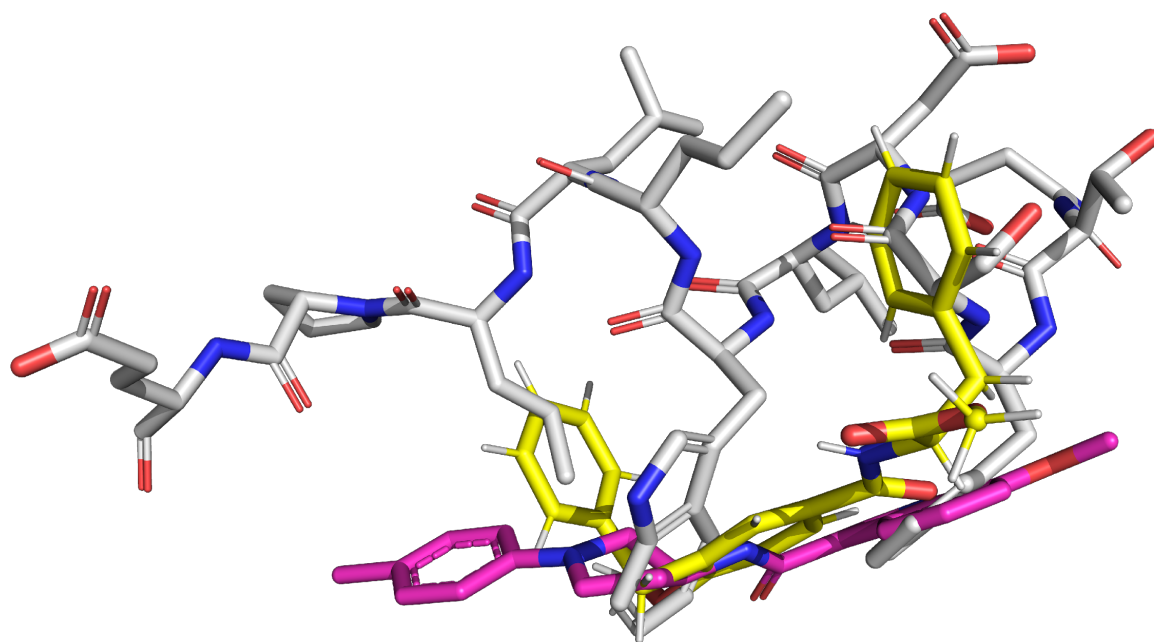


Figure 7 Compound 39 and Mdmx complex aligned with known inhibitor in complex with Mdmx (PDB ID: 6q9y). (white: p53, magenta: 6q9y, yellow: compound 39).

Comparing the binding pattern of these complexes, the molecule's extremities can be well aligned to Phe19, Trp23 and Leu26 of p53. It is frequently seen that fragments of ligands insert into the Trp23 hydrophobic pocket. Simultaneously, the pocket shrinks to accommodate the size of small molecules. Interestingly, almost every compound analyzed can mimic the three key interacting residues. And some of them can better align with p53 in more residues beyond those three. Especially the surface over the helix α 2 is a hot spot for fragments to occupy like Ser20 of p53. As for Tyr78, in most cases, it remains in a "closed conformation" pointing to the ligand. Therefore, we assume it tends to interact with the ligand and flips only with the movement of the ligand. To select some of them that are typical and very different from known Mdmx inhibitors and test them by bioassay in the future would be a worthwhile endeavor. It is noted that some compounds with good ΔG values are not similarly ordered at the top of the results from Generator. This is due to their

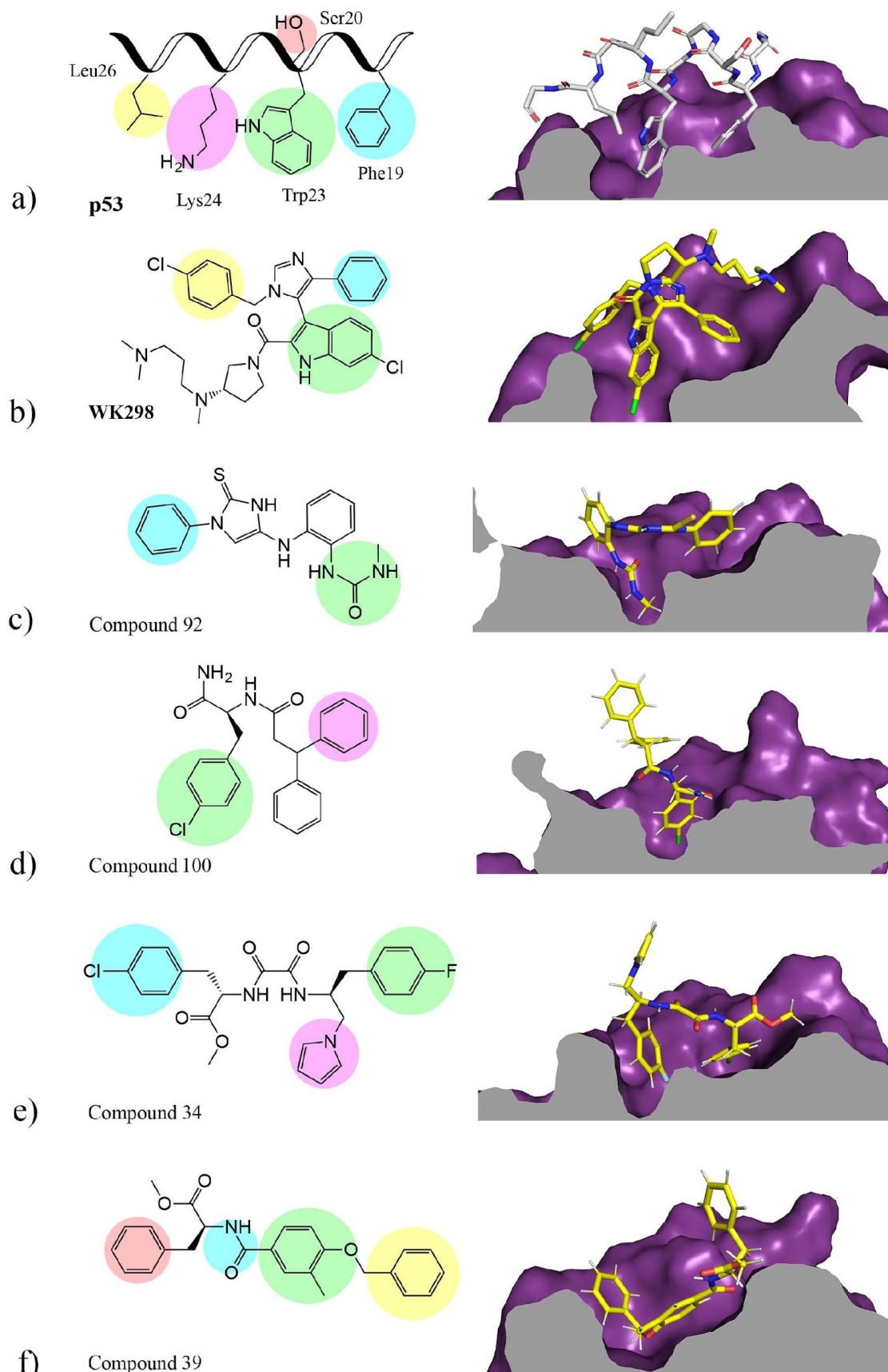


Figure 8 a) p53-Mdmx complex (PDB ID: 3dab). White color indicates p53 and the surface of Mdmx is shown in purple colour; b) WK298-Mdmx complex (PDB ID: 3lbj). Yellow color indicates WK298; c) –f) The binding patterns of respective compounds (shown in yellow) in complex with Mdmx are shown. Compounds 3021 and 92 had similar binding modes and thus only complex with compound 92 is shown.

different computational principles. The scoring function integrates both empirical information and experimental affinity measurements. Relative ΔG in this work is calculated based on molecular mechanics and Poisson-Boltzmann Surface Area (MM/PBSA) in the absence of entropic contribution. And the evaluation criteria of the Generator also equivalently take the predicted IC_{50} values into account.

Constraints during tuning

The current implementation of the tuning workflow for the generator allows introducing several constraints to take into account while generating new molecules. In this study, we used the IC_{50} model to estimate the inhibition potential of a compound, and VINA scores to measure its binding affinity. However, most of the generated compounds have solubility issues. We investigated is it possible to create new inhibitors considering solubility as well as other endpoints. We started another cycle of generation but with an additional solubility filter (run-2).

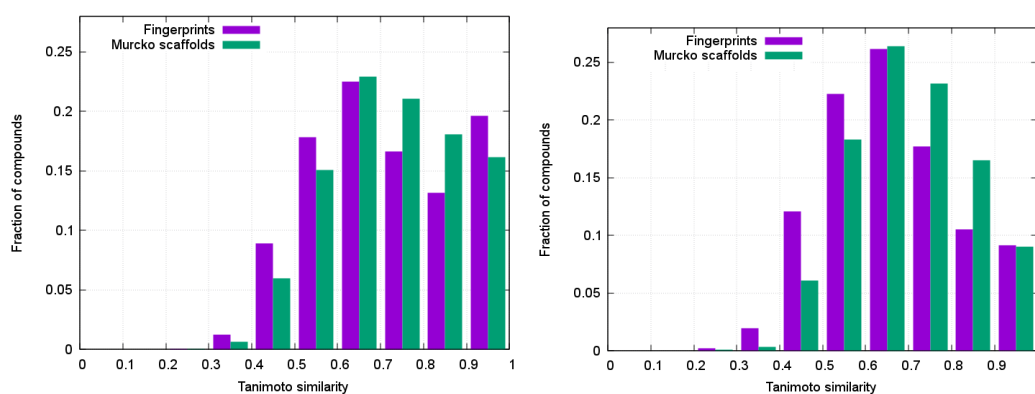


Figure 9. Comparison of similarity distributions for run-1 of generation (without solubility, left) and run-2 (with additional solubility parameter, right).

Distributions of the Tanimoto similarity of generated molecules for entire compounds and Murcko scaffolds [73] are depicted in Fig. 9. The first TOP-500 compounds from the second run are also available in the Supporting Information. It is clear that adding another filter shifts the generator to produce a more diverse set of compounds as well as chemical scaffolds. Utilizing more QSAR filters at this stage should benefit the entire de-novo design pipeline and we will further investigate the mutual influences of different filters to the overall scoring procedure.

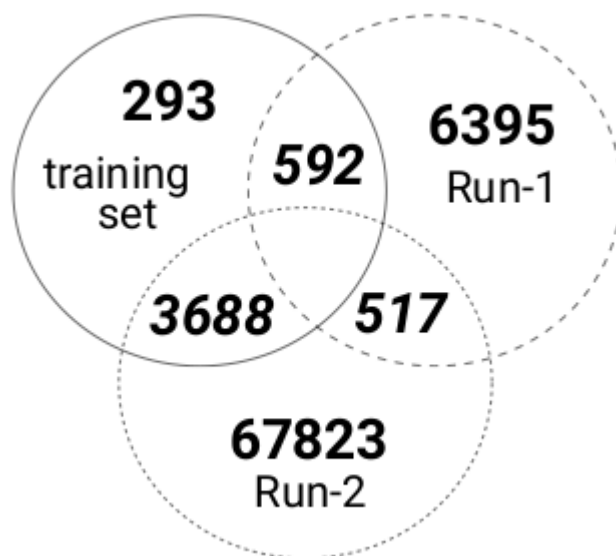


Figure 10. Number of compounds generated during run-1 and run-2 as compared to the original dataset.

As shown in Fig. 10, the overlap between generated compounds during two different runs is less than 10% (517/6395). Note, the total number of generated compounds also includes those structures that overlap with the training dataset. For example, for run-1 the generator produced in total 6395 structures, but 592 of them were already in the training dataset (the generator can produce the same molecule more than one time). Thus, different filters guide exploration of different regions of chemical space and can be used complementary to each other. But the question of how to account for all these filters in one workflow remains open.

Conclusions

Designing putative inhibitors for Mdmx is a challenging task, which so far has not resulted in a new drug candidate. In this work, we exploited the SGNN guided sampling, with docking scores to design new, putative inhibitors of Mdmx. The validation of the focused library generated by the computer through molecular dynamics simulation identified several compounds with binding energies similar or higher than the reference inhibitor. The overall approach being entirely artificial intelligence driven until the final inspection of the binding poses proved its ability, at least virtually, to design new compounds with machine learning methods and validate them with traditional molecular dynamics simulations in an automated manner. Despite several proposed generators [8–13, 48, 74–76], our study is the first one to describe and implement the full process of virtual drug design of molecules using “generators”. The public

availability of its code and whole workflow allows other users to adapt and use it for their own projects. Though the number and quality of filters applied during the generation phase are to be optimized further, the whole algorithmic scheme, we believe, will guide future drug-development process, bring new drug candidates for challenging targets and will contribute to the diversification of medicinal chemistry methodology.

Acknowledgments

This study was partially supported by the China Scholarship Council (CSC) for providing the fellowship for Miss. Xia (201706880010) and ERA-CVD (<http://era-cvd.eu>) "Cardio-Oncology" project, BMBF 01KL1710. The authors also express gratitude to NVIDIA Corporation for donating Titan V, Xp and Quadro P6000 graphics cards for this research.

References

1. Ruddigkeit L, van Deursen R, Blum LC, Reymond JL (2012) Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J Chem Inf Model* 52 (11):2864-2875. doi:10.1021/ci300415d
2. Reymond JL (2015) The chemical space project. *Acc Chem Res* 48 (3):722-730. doi:10.1021/ar500432k
3. Shivanyuk A, Ryabukhin S, Bogolyubsky AV, Mykytenko DM, Chuprina A, Heilman W, Kostyuk AN, Tolmachev A (2007) Enamine real database: Making chemical diversity real. *Chimica Oggi* 25:58-59
4. Segler MHS, Kogej T, Tyrchan C, Waller MP (2018) Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent Sci* 4 (1):120-131. doi:10.1021/acscentsci.7b00512
5. Olivecrona M, Blaschke T, Engkvist O, Chen H (2017) Molecular de-novo design through deep reinforcement learning. *J Cheminform* 9 (1):48. doi:10.1186/s13321-017-0235-x
6. Graves A (2013) Generating Sequences With Recurrent Neural Networks. arXiv e-prints:arXiv:1308.0850
7. Weininger D (1988) Smiles, a Chemical Language and Information-System .1. Introduction to Methodology and Encoding Rules. *J Chem Inf Comput Sci* 28 (1):31-36. doi:Doi 10.1021/Ci00057a005
8. Ertl P, Lewis R, Martin E, Polyakov V (2017) In silico generation of novel, drug-like chemical matter using the LSTM neural network. arXiv e-prints:arXiv:1712.07449

9. Awale M, Sirockin F, Stiefl N, Reymond JL (2019) Drug Analogs from Fragment-Based Long Short-Term Memory Generative Neural Networks. *J Chem Inf Model* 59 (4):1347-1356. doi:10.1021/acs.jcim.8b00902
10. Yang X, Zhang J, Yoshizoe K, Terayama K, Tsuda K (2017) ChemTS: an efficient python library for de novo molecular generation. *Sci Technol Adv Mater* 18 (1):972-976. doi:10.1080/14686996.2017.1401424
11. Kadurin A, Nikolenko S, Khrabrov K, Aliper A, Zhavoronkov A (2017) druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. *Mol Pharm* 14 (9):3098-3104. doi:10.1021/acs.molpharmaceut.7b00346
12. Gomez-Bombarelli R, Wei JN, Duvenaud D, Hernandez-Lobato JM, Sanchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A (2018) Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent Sci* 4 (2):268-276. doi:10.1021/acscentsci.7b00572
13. Kusner MJ, Paige B, Hernández-Lobato JM (2017) Grammar Variational Autoencoder. *ArXiv e-prints 1703:arXiv:1703.01925*
14. Shvarts A, Steegenga WT, Riteco N, van Laar T, Dekker P, Bazuine M, van Ham RC, van der Houven van Oordt W, Hateboer G, van der Eb AJ, Jochemsen AG (1996) MDMX: a novel p53-binding protein with some functional properties of MDM2. *The EMBO journal* 15 (19):5349-5357
15. Vogelstein B, Lane D, Levine AJ (2000) Surfing the p53 network. *Nature* 408 (6810):307-310
16. Vousden KH, Lu X (2002) Live or let die: the cell's response to p53. *Nature Reviews Cancer* 2 (8):594-604
17. Toledo F, Wahl GM (2006) Regulating the p53 pathway: in vitro hypotheses, in vivo veritas. *Nature Reviews Cancer* 6 (12):909-923
18. Momand J, Zambetti GP, Olson DC, George D, Levine AJ (1992) The mdm-2 oncogene product forms a complex with the p53 protein and inhibits p53-mediated transactivation. *Cell* 69 (7):1237 - 1245
19. Picksley SM, Lane DP (1993) What the papers say: The p53-mdm2 autoregulatory feedback loop: A paradigm for the regulation of growth control by p53? *BioEssays* 15 (10):689-690
20. Haupt Y, Maya R, Kazaz A, Oren M (1997) Mdm2 promotes the rapid degradation of p53. *Nature* 387 (6630):296-299
21. Haupt S, Mejía-Hernández JO, Vijayakumaran R, Keam SP, Haupt Y (2019) The long and the short of it: the MDM4 tail so far. *Journal of Molecular Cell Biology* 11 (3):231-244
22. Espadinha M, Barcherini V, Santos* EAL, Maria M. M (2018) An Update on MDMX and Dual MDM2/X Inhibitors. *Curr Top Med Chem* 18 (8):647-660

23. Gupta A, Shah K, Oza MJ, Behl T (2019) Reactivation of p53 gene by MDM2 inhibitors: A novel therapy for cancer treatment. *Biomedicine & Pharmacotherapy* 109:484 - 492
24. Reed D, Shen Y, Shelat AA, Arnold LA, Ferreira AM, Zhu F, Mills N, Smithson DC, Regni CA, Bashford D, Cicero SA, Schulman BA, Jochemsen AG, Guy RK, Dyer MA (2010) Identification and Characterization of the First Small Molecule Inhibitor of MDMX. *J Biol Chem* 285 (14):10786-10796
25. Bista M, Smithson D, Pecak A, Salinas G, Pustelny K, Min J, Pirog A, Finch K, Zdzalik M, Waddell B, Wladyka B, Kedracka-Krok S, Dyer MA, Dubin G, Guy RK (2012) On the Mechanism of Action of SJ-172550 in Inhibiting the Interaction of MDM4 and p53. *PLOS ONE* 7 (6):1-9
26. Roh J-L, Park JY, Kim EH (2014) XI-011 enhances cisplatin-induced apoptosis by functional restoration of p53 in head and neck cancer. *Apoptosis* 19 (11):1594-1602
27. Wang H, Ma X, Ren S, Buolamwini JK, Yan C (2011) A Small-Molecule Inhibitor of MDMX Activates p53 and Induces Apoptosis. *Molecular Cancer Therapeutics* 10 (1):69-79
28. Joseph TL, Madhumalar A, Brown CJ, Lane DP, Verma CS (2010) Differential binding of p53 and nutlin to MDM2 and MDMX: Computational studies. *Cell Cycle* 9 (6):1167-1181
29. Li Q, Lozano G (2013) Molecular pathways: targeting Mdm2 and Mdm4 in cancer therapy. *Clinical cancer research : an official journal of the American Association for Cancer Research* 19 (1):34-41
30. Kon N, Gu W (2018) Is MDMX the better target? *Aging* 10 (6):1184-1185
31. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 40 (Database issue):D1100-1107. doi:10.1093/nar/gkr777
32. Ramsundar B, Eastman P, Walters P, Pande V (2019) *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More.* O'Reilly Media,
33. Hayashi R, Wang D, Hara T, Iera JA, Durell SR, Appella DH (2009) N-Acylpolyamine inhibitors of HDM2 and HDMX binding to p53. *Bioorg Med Chem* 17 (23):7884-7893. doi:<https://doi.org/10.1016/j.bmc.2009.10.032>
34. Berghausen J, Buschmann N, Furet P, Gessier F, Lisztwan JH, Holzer P, Jacoby E, Kallen J, Masuya K, Soldermann CP (2015) Substituted isoquinolinones and quinazolinones. United States of America Patent US9051279B2, Jun. 9, 2015
35. Furet P, Guagnano V, Holzer P, Mah R, Masuya K, Schlapbach A, Stutz S, Vaupel A (2015) Pyrazolopyrrolidine compounds. United States of America Patent US8969341B2, Mar. 3, 2015
36. Furet P, Guagnano V, Holzer P, Kallen J, Masuya K, Stutz S (2014) Cyclohexyl isoquinolinone compounds. United States of America Patent US8859586B2, Oct. 14, 2014

37. Furet P, Guagnano V, Holzer P, Kallen J, Lv L, Mah R, Mao L, Masuya K, Schlapbach A, Stutz S (2014) Substituted pyrrolo [3, 4-D] imidazoles for the treatment of MDM2/4 mediated diseases. United States of America Patent US8815926B2, Aug. 26, 2014
38. Madden MM, Muppidi A, Li Z, Li X, Chen J, Lin Q (2011) Synthesis of cell-permeable stapled peptide dual inhibitors of the p53-Mdm2/Mdmx interactions via photoinduced cycloaddition. *Bioorg Med Chem Lett* 21 (5):1472-1475. doi:<https://doi.org/10.1016/j.bmcl.2011.01.004>
39. Noguchi T, Oishi S, Honda K, Kondoh Y, Saito T, Kubo T, Kaneda M, Ohno H, Osada H, Fujii N (2013) Affinity-based screening of MDM2/MDMX-p53 interaction inhibitors by chemical array: Identification of novel peptidic inhibitors. *Bioorg Med Chem Lett* 23 (13):3802-3805. doi:<https://doi.org/10.1016/j.bmcl.2013.04.094>
40. Blackburn TJ, Ahmed S, Coxon CR, Liu J, Lu X, Golding BT, Griffin RJ, Hutton C, Newell DR, Ojo S, Watson AF, Zaytzev A, Zhao Y, Lunec J, Hardcastle IR (2013) Diaryl- and triaryl-pyrrole derivatives: inhibitors of the MDM2-p53 and MDMX-p53 protein-protein interactions. *MedChemComm* 4 (9):1297-1304. doi:10.1039/C3MD00161J
41. Macchiarulo A, Giacchè N, Carotti A, Moretti F, Pellicciari R (2011) Expanding the horizon of chemotherapeutic targets: From MDM2 to MDMX (MDM4). *MedChemComm* 2 (6):455-465. doi:10.1039/C0MD00238K
42. Holzer P, Masuya K, Furet P, Kallen J, Valat-Stachyra T, Ferretti S, Berghausen J, Bouisset-Leonard M, Buschmann N, Pissot-Soldermann C, Rynn C, Ruetz S, Stutz S, Chène P, Jeay S, Gessier F (2015) Discovery of a Dihydroisoquinolinone Derivative (NVP-CGM097): A Highly Potent and Selective MDM2 Inhibitor Undergoing Phase 1 Clinical Trials in p53wt Tumors. *J Med Chem* 58 (16):6348-6358. doi:10.1021/acs.jmedchem.5b00810
43. Uesato S, Matsuura Y, Matsue S, Sumiyoshi T, Hirata Y, Takemoto S, Kawaratani Y, Yamai Y, Ishida K, Sasaki T, Enari M (2016) Discovery of new low-molecular-weight p53-Mdmx disruptors and their anti-cancer activities. *Bioorg Med Chem* 24 (8):1919-1926. doi:<https://doi.org/10.1016/j.bmc.2016.03.021>
44. Giustiniano M, Daniele S, Pelliccia S, La Pietra V, Pietrobono D, Brancaccio D, Cosconati S, Messere A, Giuntini S, Cerofolini L, Fragai M, Luchinat C, Taliani S, La Regina G, Da Settimo F, Silvestri R, Martini C, Novellino E, Marinelli L (2017) Computer-Aided Identification and Lead Optimization of Dual Murine Double Minute 2 and 4 Binders: Structure-Activity Relationship Studies and Pharmacological Activity. *J Med Chem* 60 (19):8115-8130. doi:10.1021/acs.jmedchem.7b00912
45. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J (2015) BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res* 44 (D1):D1045-D1053
46. Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. *Neural Computation* 9 (8):1735-1780

47. Zheng S, Yan X, Gu Q, Yang Y, Du Y, Lu Y, Xu J (2019) QBMG: quasi-biogenic molecule generator with deep recurrent neural network. *J Cheminformatics* 11 (1):5. doi:10.1186/s13321-019-0328-9
48. Gupta A, Müller AT, Huisman BJH, Fuchs JA, Schneider P, Schneider G (2018) Generative Recurrent Networks for De Novo Drug Design. *Molecular informatics* 37 (1-2). doi:10.1002/minf.201700111
49. Schwaller P, Gaudin T, Lanyi D, Bekas C, Laino T (2018) "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical science* 9 (28):6091-6098
50. Dieleman S, Schlüter J, Raffel C, Olson E, Sønderby SK, Nouri D, Maturana D, Thoma M, Battenberg E, Kelly J, Fauw JD, Heilman M, de Almeida DM, McFee B, Weideman H, Takács G, de Rivaz P, Crall J, Sanders G, Rasul K, Liu C, French G, Degraeve J (2015) Lasagne: First release. <http://dx.doi.org/10.5281/zenodo.27878>.
51. Al-Rfou R, Alain G, Almahairi A, Angermueller C, Bahdanau D, Ballas N, Bastien F, Bayer J, Belikov A, Belopolsky A (2016) Theano: A Python framework for fast computation of mathematical expressions. arXiv preprint arXiv:160502688
52. Landrum G RDKit: Open-source cheminformatics. <http://www.rdkit.org>.
53. Duchi J, Hazan E, Singer Y (2011) Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *J Mach Learn Res* 12:2121-2159
54. Popowicz G, Czarna A, Holak T (2008) Structure of the human Mdmx protein bound to the p53 tumor suppressor transactivation domain. *Cell Cycle* 7 (15):2441-2443
55. Trott O, Olson AJ (2010) AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 31 (2):455-461
56. Sushko I, Novotarskyi S, Korner R, Pandey AK, Rupp M, Teetz W, Brandmaier S, Abdelaziz A, Prokopenko VV, Tanchuk VY, Todeschini R, Varnek A, Marcou G, Ertl P, Potemkin V, Grishina M, Gasteiger J, Schwab C, Baskin, II, Palyulin VA, Radchenko EV, Welsh WJ, Kholodovych V, Chekmarev D, Cherkasov A, Aires-de-Sousa J, Zhang QY, Bender A, Nigsch F, Patiny L, Williams A, Tkachenko V, Tetko IV (2011) Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J Comput Aided Mol Des* 25 (6):533-554. doi:10.1007/s10822-011-9440-2
57. Sushko I, Novotarskyi S, Körner R, Pandey AK, Kovalishyn VV, Prokopenko VV, Tetko IV (2010) Applicability domain for in silico models to achieve accuracy of experimental measurements. *J Chemom* 24 (3 - 4):202-208. doi:10.1002/cem.1296
58. Karpov P, Godin G, Tetko IV (2019) Transformer-CNN: Fast and Reliable tool for QSAR (subm).
59. Bjerrum EJ (2017) Smiles enumeration as data augmentation for neural network modeling of molecules. arXiv preprint arXiv:170307076

60. Tetko IV, Karpov P, Bruno E, Kimber TB, Godin G Augmentation Is What You Need! In: Tetko IV, Kůrková V, Karpov P, Theis F (eds) 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings. Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions. Cham: Springer International Publishing, pp 831–835
61. Bruns RF, Watson IA (2012) Rules for Identifying Potentially Reactive or Promiscuous Compounds. *J Med Chem* 55 (22):9763-9772. doi:10.1021/jm301008n
62. Dixon SL, Smondyrev AM, Knoll EH, Rao SN, Shaw DE, Friesner RA (2006) PHASE: a new engine for pharmacophore perception, 3D QSAR model development, and 3D database screening: 1. Methodology and preliminary results. *J Comput Aided Mol Des* 20 (10-11):647-671. doi:10.1007/s10822-006-9087-6
63. Dixon SL, Smondyrev AM, Rao SN (2006) PHASE: a novel approach to pharmacophore modeling and 3D database searching. *Chemical biology & drug design* 67 (5):370-372. doi:10.1111/j.1747-0285.2006.00384.x
64. Boettcher A, Buschmann N, FuretJean-Marc P, KALLEN G, Lisztwan JH, Masuya K, Mayr L, Vaupel A (2008) 3-imidazolyl-indoles for the treatment of proliferative diseases. WO2008119741A2, Oct. 9, 2008
65. Popowicz GM, Czarna A, Wolf S, Wang K, Wang W, Dömling A, Holak TA (2010) Structures of low molecular weight inhibitors bound to MDMX and MDM2 reveal new approaches for p53-MDMX/MDM2 antagonist drug discovery. *Cell Cycle* 9 (6):1104-1111
66. Case DA, Cheatham TE, 3rd, Darden T, Gohlke H, Luo R, Merz KM, Jr, Onufriev A, Simmerling C, Wang B, Woods RJ (2005) The Amber biomolecular simulation programs. *J Comput Chem* 26 (16):1668-1688
67. Miller BR, McGee TD, Swails JM, Homeyer N, Gohlke H, Roitberg AE (2012) MMPBSA.py: An Efficient Program for End-State Free Energy Calculations. *Journal of Chemical Theory and Computation* 8 (9):3314-3321
68. Hou T, Wang J, Li Y, Wang W (2011) Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J Chem Inf Model* 51 (1):69-82. doi:10.1021/ci100275a
69. Huuskonen JJ, Livingstone DJ, Tetko IV (2000) Neural Network Modeling for Estimation of Partition Coefficient Based on Atom-Type Electrotopological State Indices. *Journal of Chemical Information and Computer Sciences* 40 (4):947-955. doi:10.1021/ci9904261
70. Karlov DS, Sosnin S, Tetko IV, Fedorov MV (2019) Chemical space exploration guided by deep neural networks. *RSC Adv* 9:5151-5157
71. Chen S-K, Zhao P, Shao Y-X, Li Z, Zhang C, Liu P, He X, Luo H-B, Hu X (2012) Moracin M from *Morus alba* L. is a natural phosphodiesterase-4 inhibitor. *Bioorg Med Chem Lett* 22 (9):3261-3264. doi:10.1016/j.bmcl.2012.03.026

72. Kallen J, Izaac A, Chau S, Wirth E, Schoepfer J, Mah R, Schlapbach A, Stutz S, Vaupel A, Guagnano V, Masuya K, Stachyra TM, Salem B, Chene P, Gessier F, Holzer P, Furet P (2019) Structural States of Hdm2 and HdmX: X-ray Elucidation of Adaptations and Binding Interactions for Different Chemical Compound Classes. *ChemMedChem*. doi:10.1002/cmdc.201900201
73. Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. *J Med Chem* 39 (15):2887-2893. doi:10.1021/jm9602928
74. Popova M, Isayev O, Tropsha A (2018) Deep reinforcement learning for de novo drug design. *Sci Adv* 4 (7):eaap7885. doi:10.1126/sciadv.aap7885
75. Stahl N, Falkman G, Karlsson A, Mathiason G, Bostrom J (2019) Deep Reinforcement Learning for Multiparameter Optimization in de novo Drug Design. *Journal of chemical information and modeling* 59 (7):3166-3176. doi:10.1021/acs.jcim.9b00325
76. van Deursen R, Ertl P, Tetko I, Godin G (2019) GEN: Highly Efficient SMILES Explorer Using Autodidactic Generative Examination Networks. doi:10.26434/chemrxiv.9796874.v1