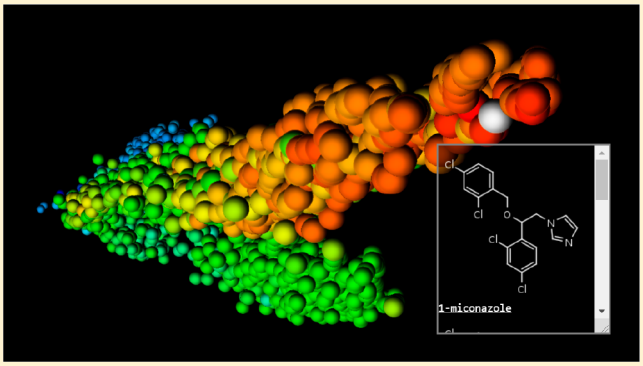
JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING
Cite This: J. Chem. Inf. Model. 2019, 59, 1347−1356
Article
pubs.acs.org/jcim

Drug Analogs from Fragment-Based Long Short-Term Memory Generative Neural Networks

Mahendra Awale,[†] Finton Sirockin,[‡] Nikolaus Stiefl,[‡] and Jean-Louis Reymond*[,†]

[†]Department of Chemistry and Biochemistry, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland
[‡]Novartis Institutes for Biomedical Research, CH-4002 Basel, Switzerland

ABSTRACT: Several recent reports have shown that long short-term memory generative neural networks (LSTM) of the type used for grammar learning efficiently learn to write Simplified Molecular Input Line Entry System (SMILES) of druglike compounds when trained with SMILES from a database of bioactive compounds such as ChEMBL and can later produce focused sets upon transfer learning with compounds of specific bioactivity profiles. Here we trained an LSTM using molecules taken either from ChEMBL, DrugBank, commercially available fragments, or from FDB-17 (a database of fragments up to 17 atoms) and performed transfer learning to a single known drug to obtain new analogs of this drug. We found that this approach readily generates hundreds of relevant and diverse new drug analogs and works best with training sets of around 40,000 compounds as simple as commercial fragments. These data suggest that fragment-based LSTM offer a promising method for new molecule generation.



INTRODUCTION

New druglike small molecules are constantly needed to feed the drug discovery pipeline.[1,2] While the number of possible molecules is extremely large,[3,4] one can narrow the search for such new small molecules by exploiting the accumulated knowledge on drug-target interactions.[5] In one such approach, it was recently discovered that long short-term memory generative neural networks (LSTM) of the type used for grammar learning,[6,7] trained with Simplified Molecular Input Line Entry System (SMILES)[8] representing organic compounds from ChEMBL,[9] a large annotated database of bioactive molecules, can generate new druglike molecular structures, which can even be tailored to specific targets upon transfer learning with focused subsets of bioactive compounds.[10−14] The molecules generated by LSTM retain structural features from the parent molecules, which focuses the generation process on analogs with a higher probability of shared bioactivity, and provide an advantage in terms of synthesis planning because such close analogs may be easier to synthesize using routes known for the parent molecules.

Here we performed transfer learning with a single drug molecule to generate new analogs of this drug, an implementation of LSTM toward analog generation that is simpler than previously reported implementations toward this goal (Figure 1)[13] and is also distinct from analog generation methods based on variational autoencoders.[15,16] We studied the influence of the primary training set of molecules on the outcome of LSTM for 10 drugs covering a broad range of size and complexity from small molecule drugs to macrocyclic natural products (Figure 2). We were specifically interested in comparing the effect of

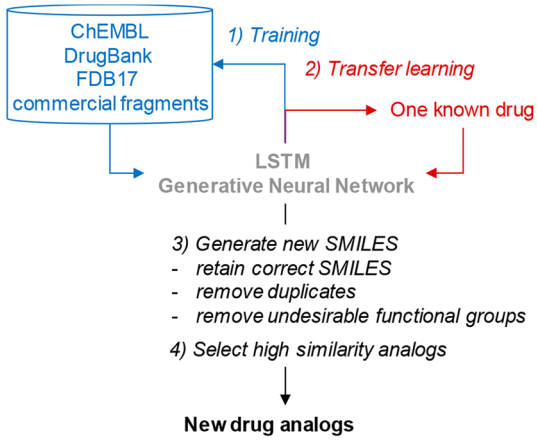


Figure 1. Principle of LSTM neural networks for generating drug analogs.

training with bioactive molecules such as those from ChEMBL[9] or DrugBank[17] with small fragment-sized compounds. We selected fragments either from commercial catalogs or from FDB17,[18] a database of theoretically possible fragments covering the entire chemical space up to 17 atoms.[19] Our data shows that LSTM training with fragment-sized molecules leads to new analogs as efficiently as if training is done with drug type molecules from ChEMBL or DrugBank. Excellent results are obtained by LSTM training with a relatively small set of

Received: December 6, 2018
Published: March 25, 2019

DOI: 10.1021/acs.jcim.8b00902
J. Chem. Inf. Model. 2019, 59, 1347−1356

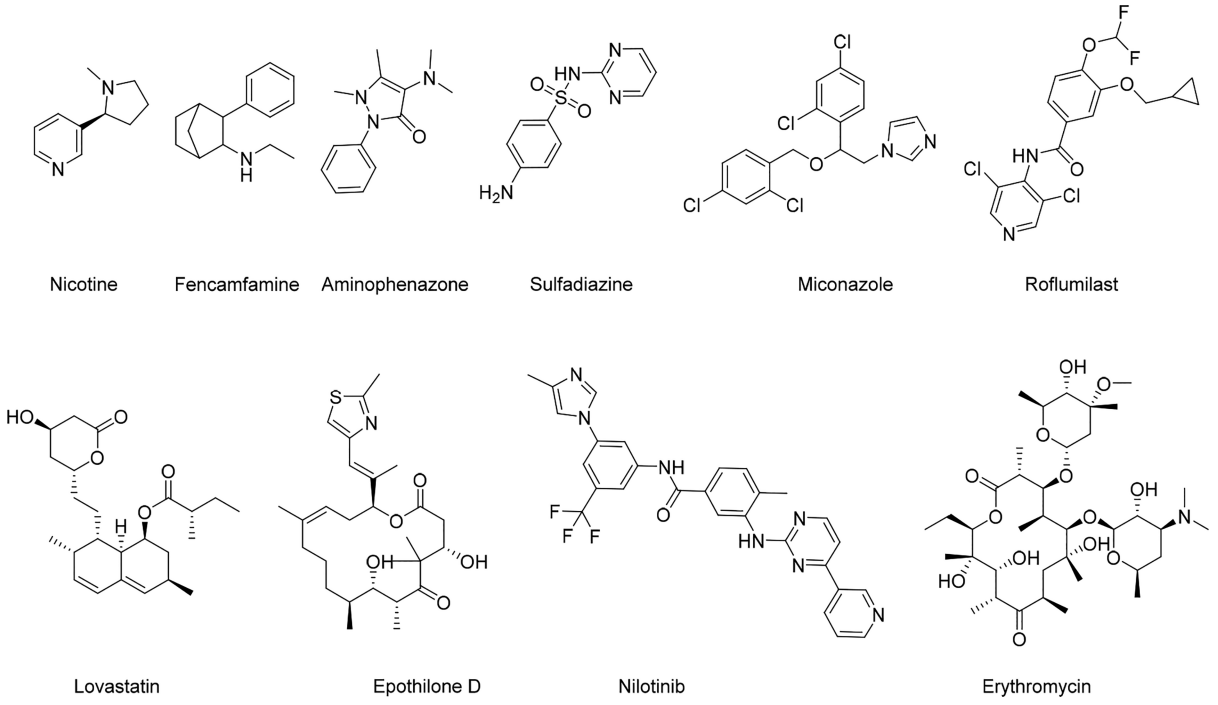


Figure 2. Structure of the 10 drugs used for transfer learning.

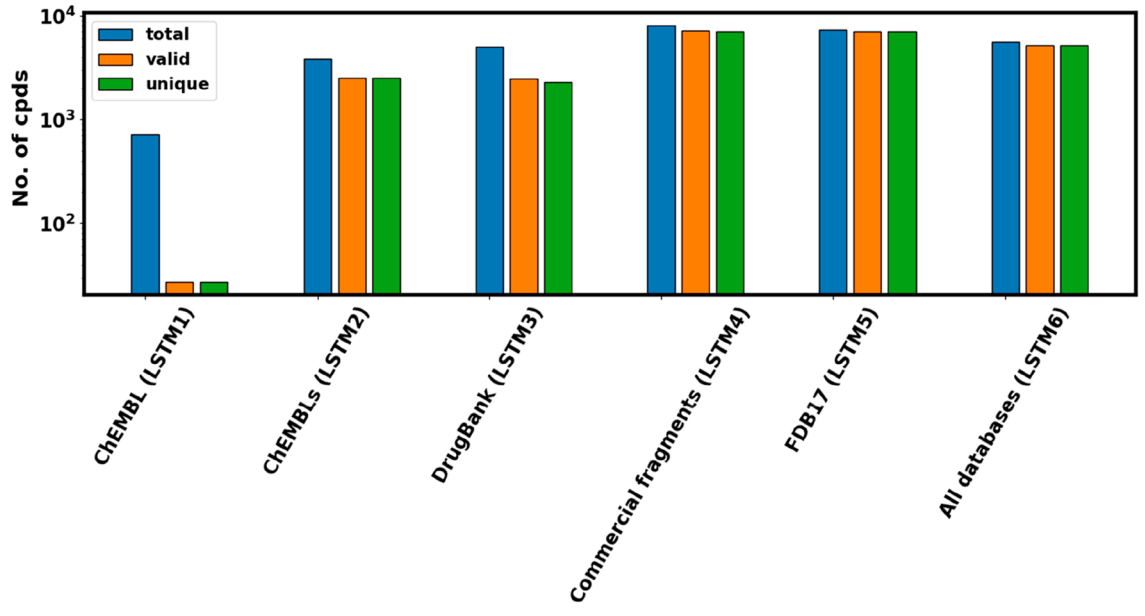


Figure 3. Number of total, valid, and unique compounds generated by different LSTM models after primary training. For each model, 200,000 characters were sampled. Note that the *y*-axis is on log scale.

approximately 40,000 molecules, which covers a relevant portion of chemical diversity at the scale of fragments. Our data suggest that fragment-based LSTM offers a promising method for new molecule generation.

RESULTS AND DISCUSSION

We trained six LSTMs corresponding to six different primary training sets, namely: (1) 344,319 compounds from ChEMBL containing only molecules reported with high quality data points for single protein targets;[20] (2) a random subset of 40,000 molecules from set 1; (3) all compounds up to a size of 50 heavy (non-hydrogen) atoms from DrugBank,[17] which were 5,104 compounds; (4) 40,986 fragments up to 17 atoms collected from various catalogs; (5) 500,000 molecules randomly selected from the fragment database FDB17; and (6) sets 1, 3, 4, and 5 combined, corresponding to 890,409 molecules. As case studies for transfer learning, we selected 10 different drug molecules covering a broad range of size and complexity from very small molecules such as nicotine or aminophenazone, to typical drug molecules such as nilotinib and lovastatin, and up to macrocyclic natural products such as epothilone D and erythromycin (Figure 2).

We performed primary training with the default learning rate of 0.01 for 50 epochs for LSTM1−LSTM5 and for 100 epochs

DOI: 10.1021/acs.jcim.8b00902
J. Chem. Inf. Model. 2019, 59, 1347−1356

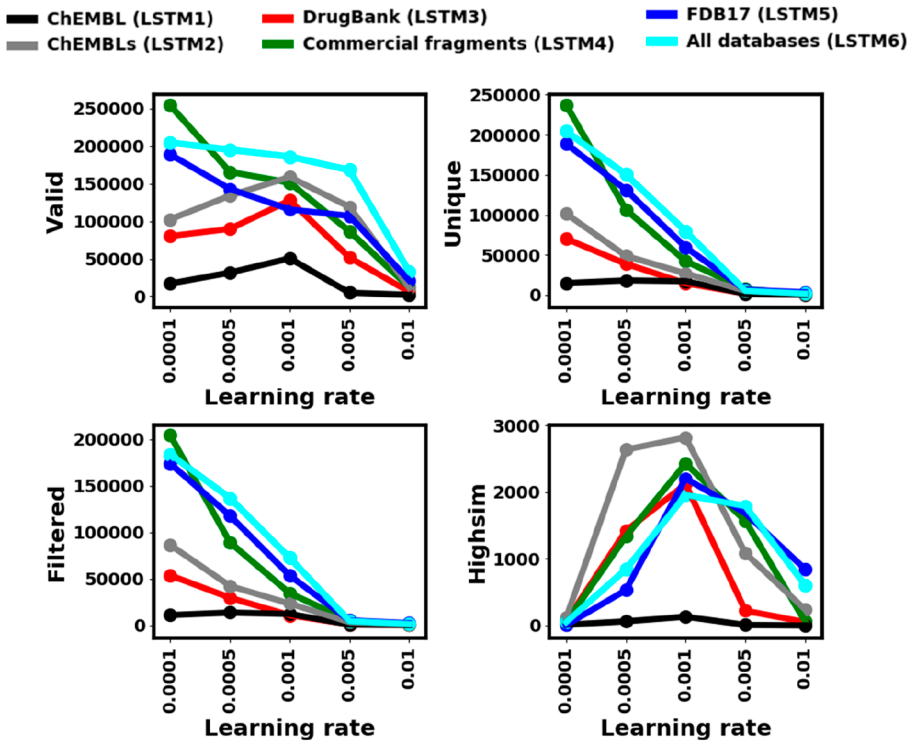


Figure 4. Number of SMILES generated by LSTM1−6 upon transfer learning across all 10 drugs in Figure 2 as a function of learning rates (*x*-axis). For each of the 10 drugs, 200,000 characters were sampled using the respective fine-tuned LSTM model after 5, 10, 15, and 20 epochs. Valid = total number of valid SMILES found across all drugs. Unique = number of valid SMILES remaining after removing duplicates. Filtered: number of SMILES remaining after removing undesirable functional groups. Highsim: number of SMILES for molecules with Avalon Tanimoto similarity >0.7 and Xfp city block distance less than Xfp cutoff distance (Xfp cutoff distance = heavy atom count of a drug ×30).

for the larger and more diverse combined set for LSTM6, which was the smallest amount of training necessary to deliver satisfactory results in terms of the number of valid molecules generated in all cases except for LSTM1 (Figure 3). We did not train further LSTM1, which contains ChEMBL molecules, because LSTM2 corresponding to a smaller subset of the molecules used for LSTM1 gave satisfactory results.

We then performed transfer learning for each LSTM using each drug for 20 epochs, using learning rates ranging from 0.0001 to 0.01, generating new molecules after 5, 10, 15, and 20 epochs. We collected all generated SMILES, removed invalid SMILES and duplicates, and applied structural filters to eliminate problematic functional groups. Finally, we selected high-similarity analogs to the drug used for transfer learning using a combined filter considering an Avalon fingerprint[21,22] Tanimoto cutoff value to constrain substructures, as well as an Xfp[23] city-*block*-distance cutoff value to constrain overall molecular size, shape, and pharmacophore.

Each LSTM performed differently with each test molecule and learning rate. Analyzing the overall data across all 10 drugs showed that increasing the learning rate of transfer learning led to a strong reduction of the overall number of generated molecules at the level of correct, unique and functionally filtered SMILES. On the other hand, the number of high similarity analogs produced strongly increased with increasing learning rates, such that an optimum was reached at a learning rate of 0.001 (Figure 4). The number of molecules produced increased as a function of the number of epochs of secondary training, however the number of high similarity analogs increased but then remained stable or decreased at high learning rates, in particular for the optimal learning rate of 0.001, indicating that secondary training beyond 20 epochs would lead to overfitting (Figure 5). Training with the large set of ChEMBL molecules produced almost no high similarity analogs (LSTM1, black line), while training all other data sets including the combined set (LSTM2−6) produced comparable numbers of high similarity analogs. It should be noted that the performance of LSTM1 can be tuned by training it for a longer time and tweaking the learning rate.

Analyzing the production of analogs for each drug separately showed that LSTM4 and LSTM5, which were trained with relatively small, fragment sized molecules, produced the largest number of analogs among all six LSTMs tested for the large natural product target molecules lovastatin, epothilone D, and erythromycin (Figure 6). In these cases, we observed a drift toward the size of the target molecule as the learning rate increased, showing that transfer learning partly consisted in learning how to make these large molecules from fragments.

To gain a closer insight into the generated molecules, we grouped all compounds generated for each drug by each LSTM across the different learning rates. All LSTMs produced molecules covering the entire Avalon similarity range, spanning from a vast majority of extremely low similarity compounds, to a small fraction of molecules in the high similarity range (Figure 7). A large fraction of these molecules was unique to each LSTM, indicating that the primary training set strongly influenced the molecule generation process (Table 1). Interestingly LSTM6 trained with all sets combined also produced a majority of molecules not generated by any of the other LSTMs. Note that LSTM5 trained with FDB17 did not produce any analogs with halogen-containing drugs because FDB17 fragments do not contain any halogens.

DOI: 10.1021/acs.jcim.8b00902
J. Chem. Inf. Model. 2019, 59, 1347−1356

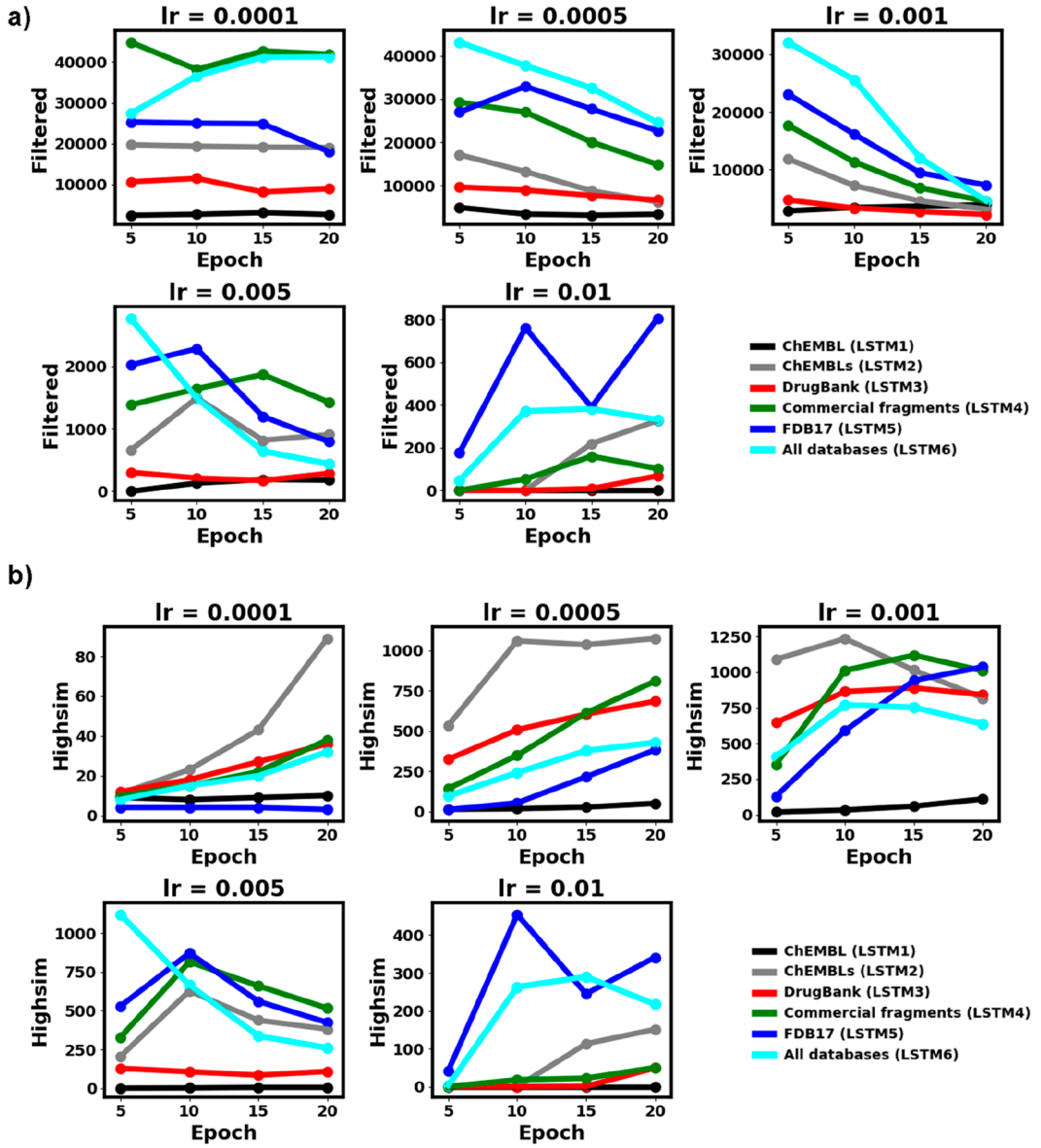


Figure 5. Number of compounds generated by LSTM1−6 upon transfer learning across all 10 drugs in Figure 2 as a function of number of epochs (*x*-axis). For each of the 10 drugs, 200,000 characters were sampled using the respective fine-tuned LSTM model after 5, 10, 15, and 20 epochs. (a) Filtered: number of SMILES remaining after removing undesirable functional groups. (b) Highsim: number of SMILES for molecules with Avalon Tanimoto similarity >0.7 and Xfp city block distance less than Xfp cutoff distance (Xfp cutoff distance = heavy atom count of a drug ×30). (Black line): LSTM1, 344,319 ChEMBL compounds. (Gray line): LSTM2, a random subset of 40,000 of the ChEMBL compounds. (Red line): LSTM3, 5,104 drugs from DrugBank. (Green line): LSTM4, 40,986 commercially available fragments. (Blue line): LSTM5, 500,000 fragments from FDB17. (Cyan line): LSTM6, data sets 1, 3, 4, and 5 combined. lr = learning rate.

The pooled set of all filtered molecules for each LSTM had a size distribution close to that of the primary training set, while high similarity analogs covered the range of target molecules (Figure 8a,c). The synthetic accessibility scores[24] of the generated molecules was as expected from the training set, with LSTMs trained with experimental molecules (ChEMBL, DrugBank, commercial fragments) producing molecules with favorable low-value scores, while FDB17 consisting of possible but not synthesized molecules gave less favorable, high-value score, for both filtered and high-similarity analogs (Figure 8b,d).

Most of these high similarity analogs were produced by more than one LSTM, with often less than half of the generated molecules being unique to the LSTM (Table 2). Some of the analogs shared by more than one LSTM were in fact already documented in ChEMBL (Table 3). These known analogs often featured one atom insertions, deletions, substitutions, or inversions (Figure 9). Despite their high similarity to the targets, these analogs were structurally diverse, as shown by the large number of different Bemis−Murcko scaffolds[25] present among these high similarity analogs (Table 4). Note that performing additional molecule generation runs on a single LSTM produced new analogs at each run throughout the Avalon similarity range except for values above 0.9 where only a limited number of compounds are possible. This is illustrated here for the case of nicotine, miconazole, and erythromycin analogs produced by LSTM4 trained with commercial fragments, considering only compounds passing the shape and pharmacophore similarity (Xfp) cutoff (Figure 10a−c). The range of

DOI: 10.1021/acs.jcim.8b00902
J. Chem. Inf. Model. 2019, 59, 1347−1356

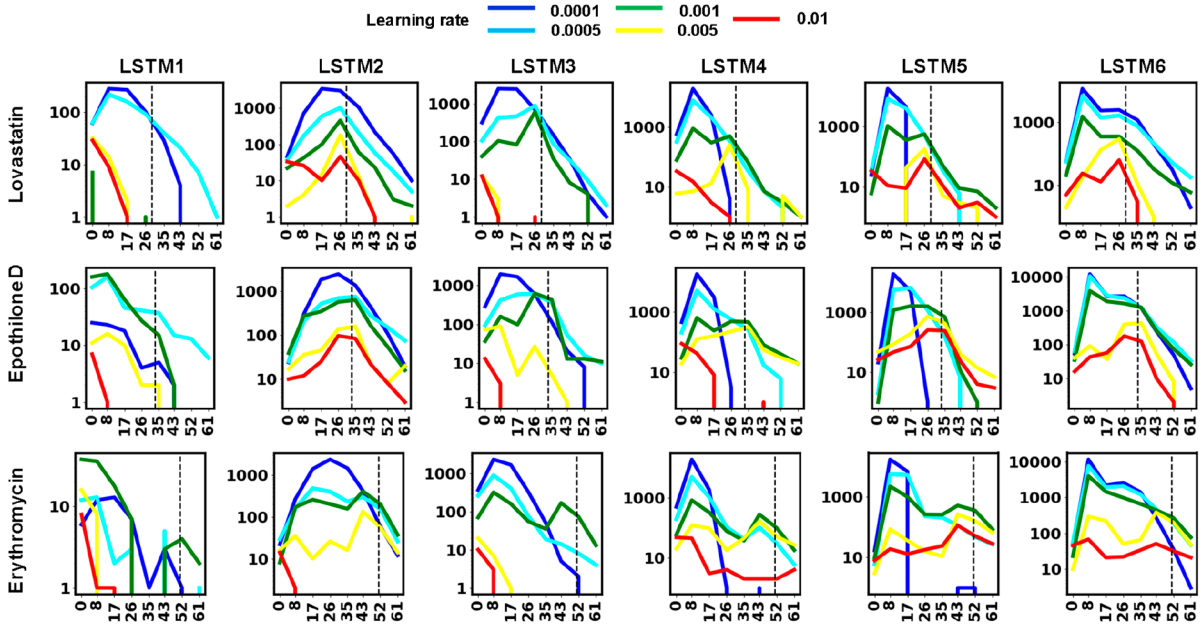


Figure 6. Molecular size (heavy atom count) histogram of generated molecules as a function of learning rate for lovastatin, epothilone D, and erythromycin. The vertical dashed line indicates the size of the drug.

molecules produced is illustrated here for miconazole by an interactive substructure fingerprint similarity map (Figure 10d).[26] This suggests that many additional such analogs of each drug can potentially be produced by running LSTMs for longer periods.

CONCLUSION

Here we trained LSTM generative neural networks with SMILES of druglike molecules from ChEMBL and DrugBank, or of fragments from commercial catalogs or from FDB17, and performed transfer learning with single drug compounds to generate new analogs of these drugs. We found that LSTMs trained with fragments produced drug analogs as efficiently as LSTMs trained with full size druglike molecules from ChEMBL or DrugBank. In the case of large natural products such as lovastatin, epothilone D, or erythromycin, LSTMs trained with fragments readily learned to assemble large molecules and produced more high similarity analogs of these drugs than LSTMs trained with full-sized molecules, showing that transfer learning informs rules to assemble small fragments into druglike molecules. Several of the high similarity analogs produced by LSTMs were already documented in ChEMBL and featured one atom changes, however the overall structural diversity of these analogs was high, as attested by a large number of scaffolds. Neural networks trained with approximately 40,000 molecules as simple as a set of commercially available fragments performed excellently in this application, suggesting that fragment-based LSTM neural networks offer a promising method for new molecule generation.

METHODS

Compound Databases for LSTM Training. (1) ChEMBL: The ChEMBL22 database was downloaded from http://www.ebi.ac.uk/chembl. Thereafter, the database was filtered to retain only the compounds reported against a "single protein" target where the source organism was either human or rat, having an activity value (IC_{50}, EC_{50}, EC_{50}, K_i or K_D) of ≤10 μM and heavy atom count of ≤50.[20] (2) ChEMBLs: this set was created by randomly selecting 40,000 compounds from the ChEMBL set mentioned above. (3) DrugBank: DrugBank database version 5.0.11 was downloaded from http://www.drugbank.ca and filtered to retain only the compounds having ≤50 heavy atoms. (4) Commercial fragments: fragment-like molecules were collected from various suppliers after which molecules obeying Congreve's rule of three criteria and having a heavy atom count of ≤17 were retained in the set. (5) FDB17: this set was created by randomly sampling 500,000 compounds from the entire FDB17 database. (6) All databases: this set was created by combining the databases 1, 3, 4, and 5. All molecules were process using the JChem Chemaxon package. Molecules were parsed in nonisomeric unique SMILES format, counterions were removed from molecules, valence errors were checked, molecules were protonated at pH 7.4, and duplicate molecules were removed in the context of each database. For each database, the plain text file containing the unique SMILES notation of molecules was used as input to train the LSTM model.

LSTM model and Primary Training. All the LSTM models reported herein were constructed using Keras version 2.0.9, a Python-based high-level neural network learning library with a TensorFlow-gpu backend. For primary training, we trained six different LSTM models using six different databases namely ChEMBL, ChEMBLs, DrugBank, Commercial fragments, FDB17, and All databases. The architecture of each of these LSTM models contains three LSTM layers (each of size 512), each followed by a Dropout layer of size 512, with dropout-rate of 0.2 to avoid model-overfitting. The output of hidden layer is then processed through the TimeDistributed Layer and Output layer with the softmax as an activation function. All LSTM models were trained using "adagrad" as an optimizer with a learning rate of 0.01 and categorical cross entropy as the loss function. The number of epochs, batch size, sequence length, and vocabulary size for each of these models were: ChEMBL (50, 64, 64, 35), ChEMBLs (50, 32, 64, 35), DrugBank (50, 16,

DOI: 10.1021/acs.jcim.8b00902
J. Chem. Inf. Model. 2019, 59, 1347−1356

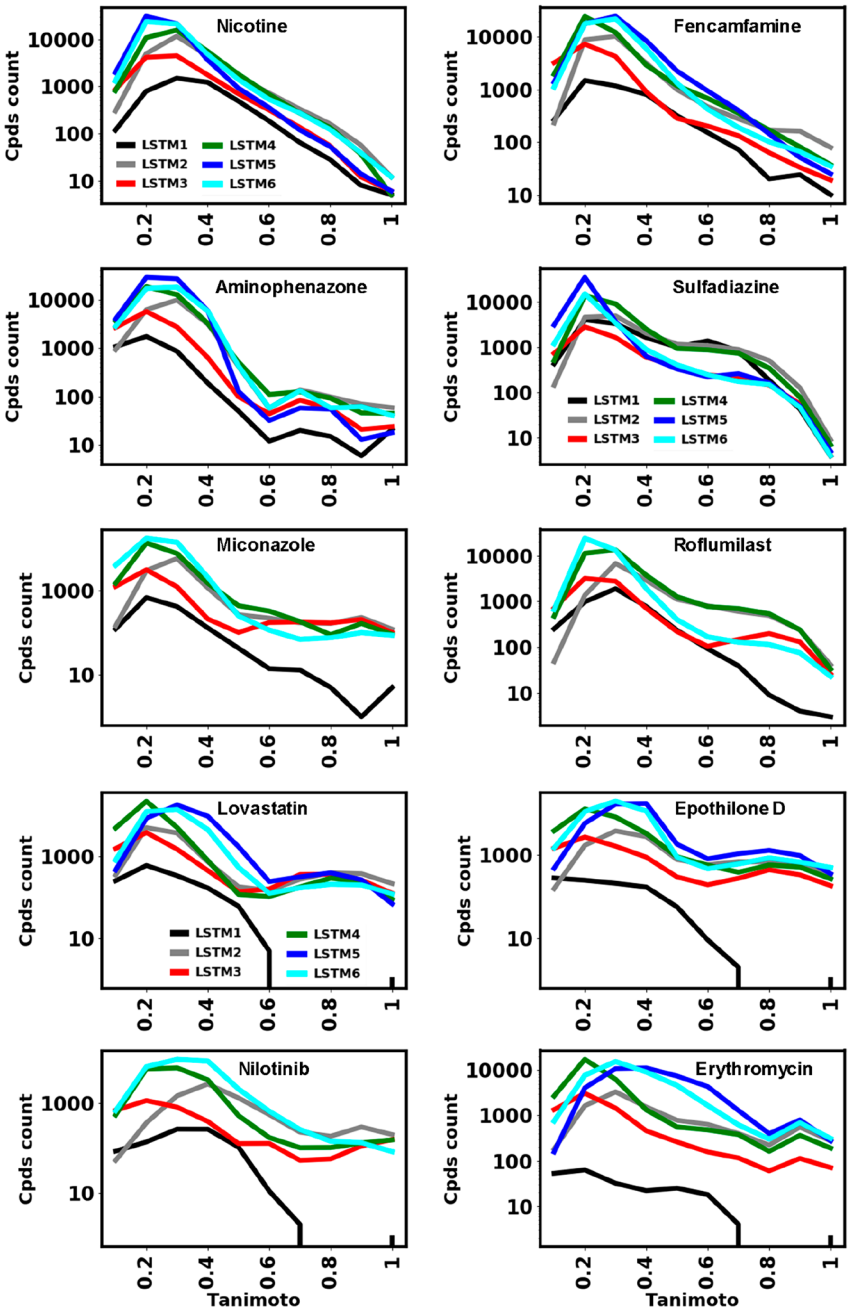


Figure 7. Avalon fingerprint similarity histogram (logarithmmic scale) for all molecules produced by the LSTMs upon transfer learning with the indicated drugs and passing functional group filters.

64, 27), commercial fragments (50, 32, 54, 34), FDB17 (50, 64, 64, 21) and all databases (100, 64, 64, 40) respectively.

Transfer Learning. Each of the six primary LSTM models were fine-tuned (transfer learning) with respect to each of the 10 drugs mentioned in this paper. For each drug, five independent fine-tuned LSTM models were generated using five different learning rates: 0.0001, 0.0005, 0.001, 0.005, and 0.01. For each model, the plain text file containing the unique SMILES notation of the drug repeated for 20 times was used as the input. Each model was fine-tuned for 20 epochs using a batch size of 4 and a sequence length of 64. Each batch was different since the number of characters in each drug SMILES was different from 64 (aminophenazone, 3; epothilone, 62; erythromycin, 97; fencamfamin, 24; lovastatin, 53; miconazole, 43; nicotine, 16; nilotinib, 70; roflumilast, 43; sulfphadiazine, 29). We used a dropout rate of 20% to avoid instant overfitting.

Sampling and Processing of SMILES. For each drug, 200,000 characters were sampled using the respective fine-tuned LSTM model after 5, 10, 15, and 20 epochs. The Numpy *random.choice* method was used to select the character given the predicted probabilities over the vocabulary. After sampling, the generated SMILES (newline character was used as delimiter to separate the SMILES of different molecules) were processed using the RDkit library. Molecules which were successfully processed by RDkit were considered as valid molecules. Thereafter, molecules were protonated at pH 7.4 using the JChem Chemaxon library; duplicate molecules and molecules

DOI: 10.1021/acs.jcim.8b00902
J. Chem. Inf. Model. 2019, 59, 1347−1356

Table 1. Number of Unique/Overall Filtered Compounds Produced by the Different LSTM Neural Networks

neural network	LSTM1	LSTM2	LSTM3	LSTM4	LSTM5	LSTM6	
source database	ChEMBL	ChEMBLs	DrugBank	commercial fragments	FDB17	all databases	unique across LSTMs
training cpds.	344,319	40,000	5,104	40,986	500,000	890,409	
nicotine	3008/4389	21K/24K	10K/12K	31K/35K	56K/59K	50K/53K	179K
fencamfamine	3377/4302	22K/24K	15K/16K	40K/43K	54K/55K	47K/48K	187K
aminophenazone	2953/3968	19K/20K	10K/11K	35K/38K	64K/65K	42K/44K	179K
sulfadiazine	11K/12K	14K/15K	5614/6791	26K/28K	42K/43K	20K/21K	123K
miconazole	1292/1449	10K/11K	6320/6851	25K/25K	0/0	38K/39K	83K
roflumilast	4036/4268	13K/14K	7550/8183	31K/32K	0/0	39K/40K	97K
lovastatin	1304/1422	10K/11K	8032/8631	32K/33K	38K/38K	31K/32K	124K
epothilone D	868/962	11K/11K	7768/8264	29K/30K	45K/45K	46K/46K	142K
nilotinib	806/862	6941/7216	3384/3636	16K/16K	0/0	27K/28K	55K
erythromycin	180/219	9235/9535	6764/7116	29K/29K	40K/40K	40K/41K	127K

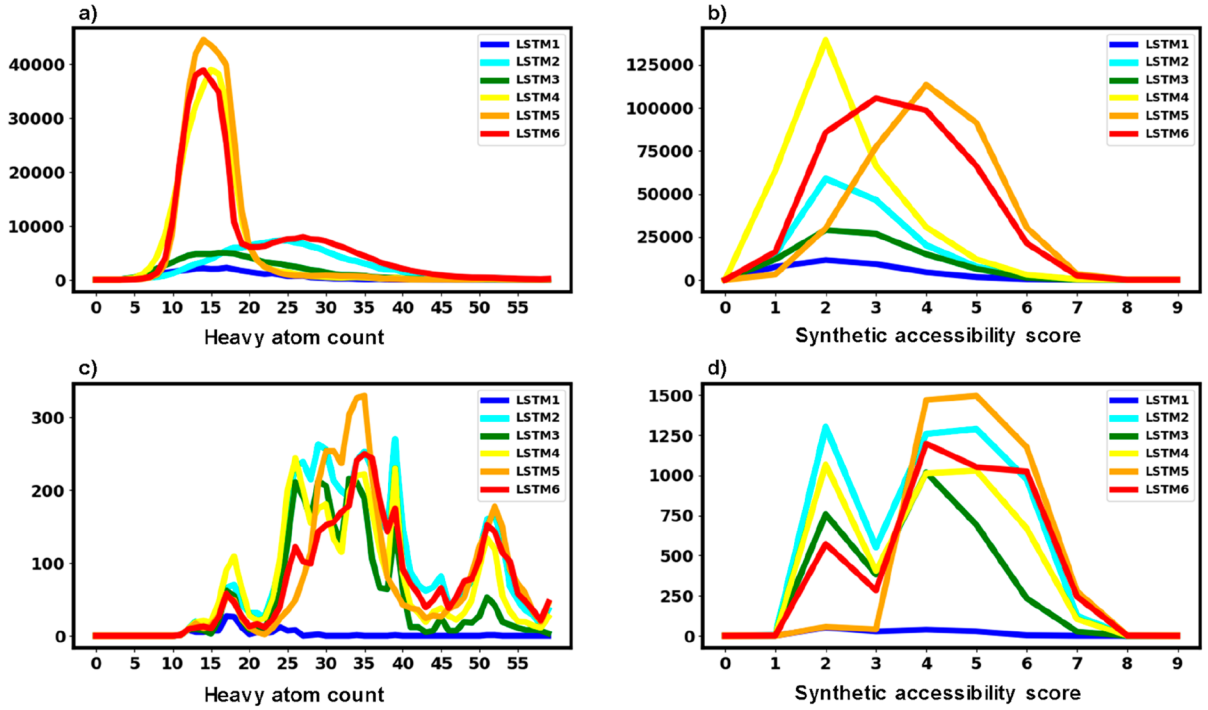


Figure 8. Molecular size histograms and synthetic accessibility score for all filtered (a,b) and Highsim (c,d) molecules produced by LSTMs summed over the 10 different drugs.

Table 2. Number of Unique/Total High Similarity Drug Analogs Produced by the Different LSTM Neural Networks

neural network	LSTM1	LSTM2	LSTM3	LSTM4	LSTM5	LSTM6	
source database	ChEMBL	ChEMBLs	DrugBank	commercial fragments	FDB17	all databases	unique across LSTMs
training cpds.	344,319	40,000	5,104	40,986	500,000	890,409	
nicotine	0/23	32/82	1/32	32/93	9/47	16/67	166
fencamfamine	15/42	126/218	40/96	130/231	92/164	41/114	580
aminophenazone	5/26	34/96	23/71	38/99	22/66	19/65	223
sulfadiazine	6/27	19/59	11/37	28/74	8/30	2/25	124
miconazole	2/10	301/500	268/438	174/336	0/0	153/256	1134
roflumilast	8/15	319/557	117/283	351/585	0/0	45/166	1126
lovastatin	0/1	631/986	460/757	352/625	487/728	289/530	2729
epothilone D	0/1	911/1301	561/831	807/1160	1595/2039	1163/1511	5707
nilotinib	0/1	506/666	180/321	218/381	0/0	243/355	1362
erythromycin	0/2	832/1042	174/243	524/709	1243/1444	1105/1288	4190

containing unstable functional groups were also removed from the list.

Avalon and Xfp Fingerprints. An Avalon substructure fingerprint containing 1,024 bits was computed using RDkit and the Avalon toolkit. The Xfp topological pharmacophore and

DOI: 10.1021/acs.jcim.8b00902
J. Chem. Inf. Model. 2019, 59, 1347−1356

Table 3. Known Bioactive from ChEMBL Produced by LSTMs[a]

	known	LSTM1	LSTM2	LSTM3	LSTM4	LSTM5	LSTM6	unique
nicotine	41	4	10	5	7	5	6	12
fencamfamine	12	0	0	0	0	0	0	0
aminophenazone	134	2	4	3	4	3	4	5
sulfadiazine	24	0	0	0	2	0	0	2
miconazole	150	1	17	12	8	0	9	19
roflumilast	19	1	1	0	0	0	2	2
lovastatin	49	0	11	12	10	6	11	16
epothilone D	75	0	5	4	6	4	7	9
nilotinib	41	0	3	2	1	0	3	4
erythromycin	201	0	4	1	2	2	1	5

[a]Number of known bioactive analogs of each drug found by similarity search in ChEMBL and numbers found in high similarity analogs for each LSTM. Nearest neighbors of each drug were extracted by performing similarity search for a given drug on the ChEMBL Web site.

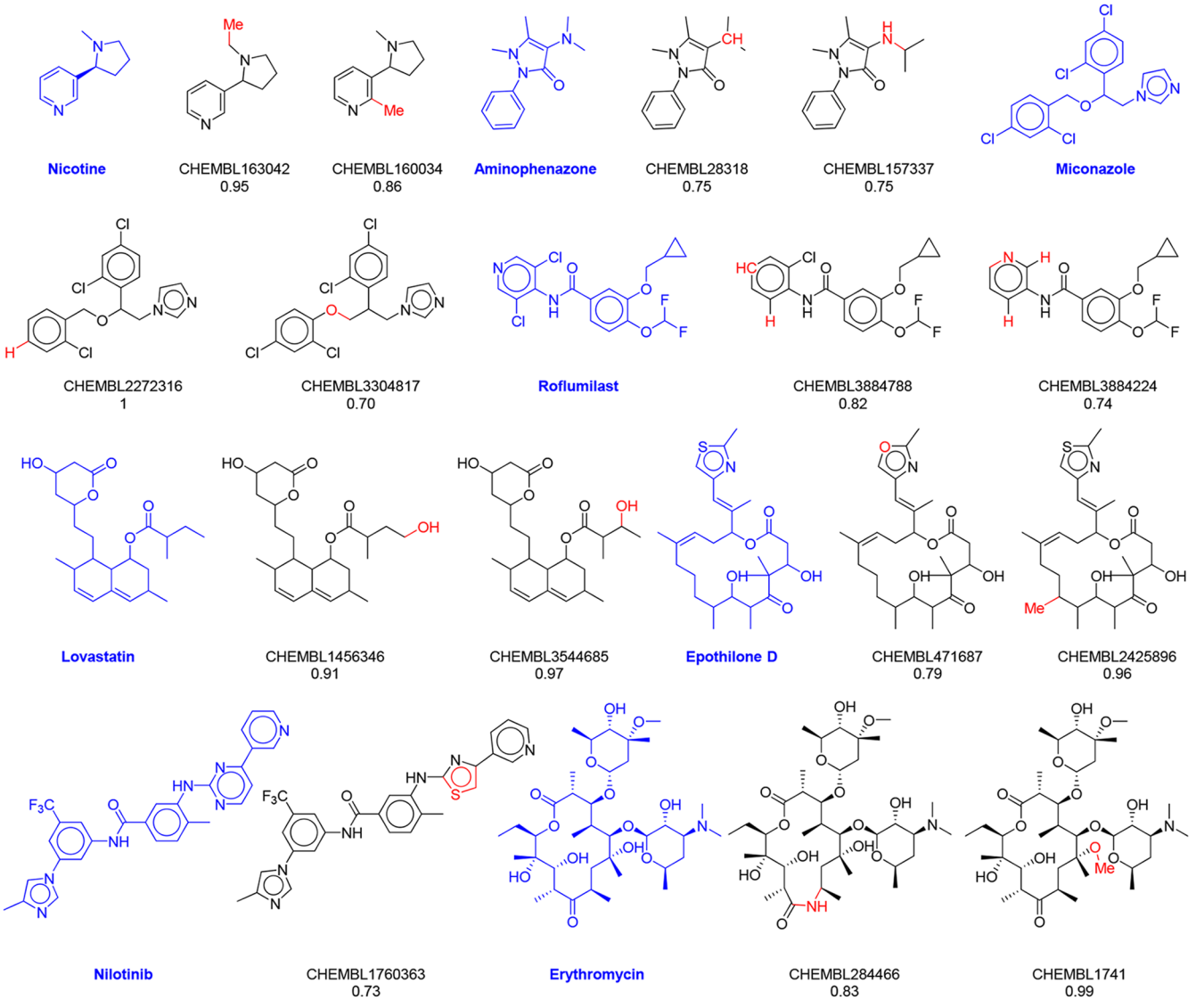


Figure 9. Examples of high similarity drug analogs produced by LSTM and documented in ChEMBL. For each drug (blue) analogs are shown (black) with structural changes highlighted (red), ChEMBL ID code, and Avalon Tanimoto similarity to the parent drug. None of the analogs shown were included in the training sets. The chirality is shown in part of the drugs for clarity of the drawing, however LSTM is run on achiral SMILES and produces analogs without defined stereochemistry.

shape fingerprint was computed using a Java program written in-house using the Jchem Chemaxon library as a starting point. Similarities between molecules were quantified using the Tanimoto coefficient and the City block distance, respectively, for Avalon and Xfp fingerprints.

DOI: 10.1021/acs.jcim.8b00902
J. Chem. Inf. Model. 2019, 59, 1347−1356

Table 4. Number of Unique/Total Bemis−Murcko Scaffolds among High Similarity Drug Analogs Produced by the Different LSTM Neural Networks

neural network	LSTM1	LSTM2	LSTM3	LSTM4	LSTM5	LSTM6	
source database	ChEMBL	ChEMBLs	DrugBank	commercial fragments	FDB17	all databases	total unique scaffolds
training cpds.	344,319	40,000	5,104	40,986	500,000	890,409	
nicotine	0/11	4/20	0/16	4/29	0/16	1/23	35
fencamfamine	7/14	55/92	17/39	41/82	54/90	18/44	247
aminophenazone	0/7	3/19	3/18	12/30	5/20	3/15	51
sulfadiazine	1/13	6/21	1/12	5/20	3/15	1/12	32
miconazole	0/4	65/115	34/78	23/68	0/0	40/76	225
roflumilast	0/6	75/152	12/75	101/186	0/0	9/49	288
lovastatin	0/1	143/276	66/175	71/189	113/223	38/108	613
epothilone D	0/1	491/807	280/538	410/675	781/1145	481/756	2960
nilotinib	0/1	270/387	109/215	146/278	0/0	91/161	773
erythromycin	0/1	372/546	87/145	263/413	611/806	510/705	2123

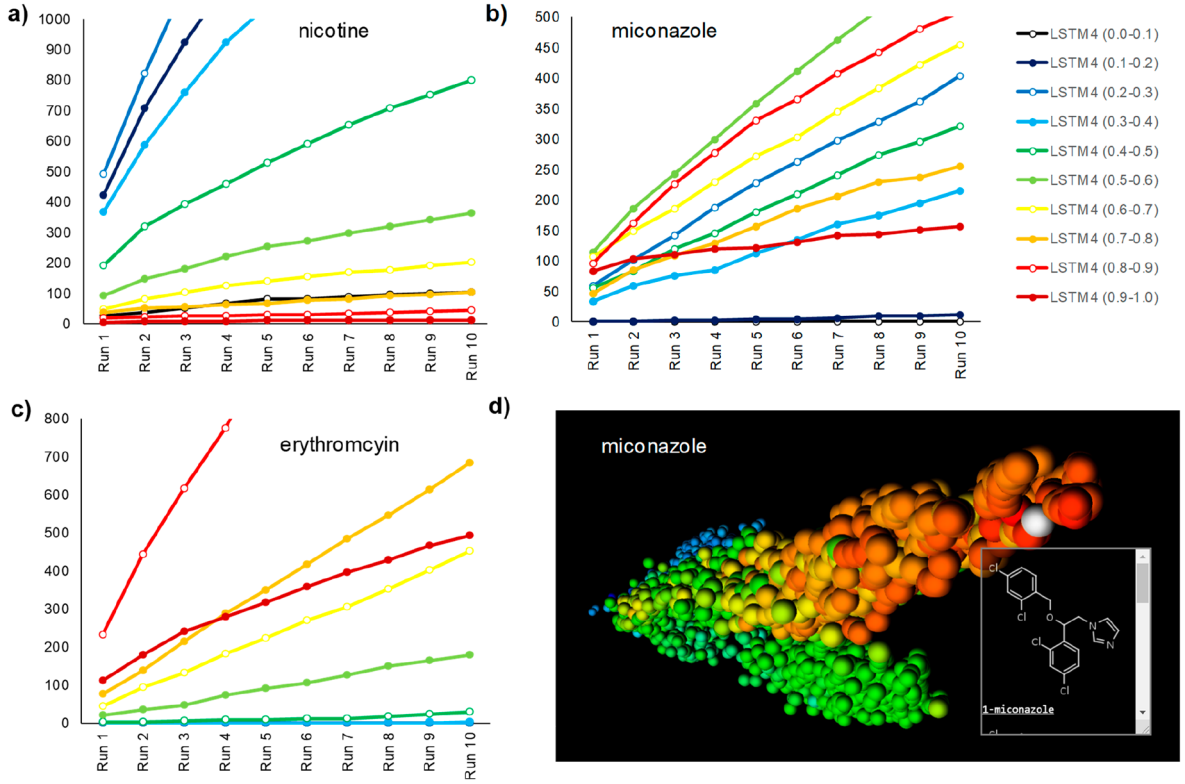


Figure 10. Production of analogs with LSTM4 at 0.001 learning rate upon additional LSTM runs, as function of Avalon Tanimoto similarity to the drug. Only analogs passing the Xfp similarity cutoff were retained. (a) Cumulative number of unique nicotine analogs upon additional runs as function of Avalon similarity, (b) same as (a) for miconazole, (c) same as (a) for erythromycin. (d) Substructure fingerprint similarity map of miconazole analog produced in (b), color-coded by the Avalon Tanimoto similarity from highest (red) to lowest (blue). The interactive 3D-map is available at the following link: http://gdbtools.unibe.ch:8080/webMolCS/yourSIM.html?jobID=1537890147494&fp=Sfp.

AUTHOR INFORMATION

Corresponding Author

*E-mail: jean-louis.reymond@dcb.unibe.ch.

ORCID

Finton Sirockin: 0000-0003-2536-7485

Jean-Louis Reymond: 0000-0003-2724-2942

Author Contributions

M.A. designed and realized the study and wrote the paper. F.S., N.S., and J.L.R. codesigned and supervised the study and wrote the paper.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported financially by a grant of NIBR to M.A.

REFERENCES

(1) Bleicher, K. H.; Bohm, H. J.; Muller, K.; Alanine, A. I. Hit and Lead Generation: Beyond High-Throughput Screening. *Nat. Rev. Drug Discovery* **2003**, *2*, 369−378.

(2) Renner, S.; Popov, M.; Schuffenhauer, A.; Roth, H. J.; Breitenstein, W.; Marzinzik, A.; Lewis, I.; Krastel, P.; Nigsch, F.; Jenkins, J.; Jacoby, E.

DOI: 10.1021/acs.jcim.8b00902
J. Chem. Inf. Model. 2019, 59, 1347−1356

Recent Trends and Observations in the Design of High-Quality Screening Collections. *Future Med. Chem.* **2011**, *3*, 751−766.
(3) Reymond, J. L.; Ruddigkeit, L.; Blum, L. C.; Van Deursen, R. The Enumeration of Chemical Space. *WIREs comput. Mol. Sci.* **2012**, *2*, 717.
(4) Awale, M.; Visini, R.; Probst, D.; Arus-Pous, J.; Reymond, J. L. Chemical Space: Big Data Challenge for Molecular Diversity. *Chimia* **2017**, *71*, 661−666.
(5) Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.; Martinez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis, D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review. *J. Chem. Inf. Model.* **2012**, *52*, 867−881.
(6) Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Networks* **2015**, *61*, 85−117.
(7) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. arXiv:1603.04467 **2016** (accessed August 20, 2018).
(8) Weininger, D. Smiles, a Chemical Language and Information-System 0.1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.
(9) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The Chembl Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083−D1090.
(10) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120−131.
(11) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inf.* **2018**, *37*, 1700153.
(12) Ertl, P.; Lewis, R.; Martin, E.; Polyakov, V.. In Silico Generation of Novel, Drug-Like Chemical Matter Using the Lstm Neural Network. arXiv:1712.07449v2. **2017** (accessed August 20, 2018).
(13) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De-Novo Design through Deep Reinforcement Learning. *J. Cheminform.* **2017**, *9*, 48.
(14) Li, Y.; Zhang, L.; Liu, Z. Multi-Objective De Novo Drug Design with Conditional Graph Generative Model. *J. Cheminform.* **2018**, *10*, 33.
(15) Harel, S.; Radinsky, K. Prototype-Based Compound Discovery Using Deep Generative Models. *Mol. Pharmaceutics* **2018**, *15*, 4406−4416.
(16) Polykovskiy, D.; Zhebrak, A.; Vetrov, D.; Ivanenkov, Y.; Aladinskiy, V.; Mamoshina, P.; Bozdaganyan, M.; Aliper, A.; Zhavoronkov, A.; Kadurin, A. Entangled Conditional Adversarial Autoencoder for De Novo Drug Discovery. *Mol. Pharmaceutics* **2018**, *15*, 4398−4405.
(17) Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; Tang, A.; Gabriel, G.; Ly, C.; Adamjee, S.; Dame, Z. T.; Han, B.; Zhou, Y.; Wishart, D. S. Drugbank 4.0: Shedding New Light on Drug Metabolism. *Nucleic Acids Res.* **2014**, *42*, D1091−D1097.
(18) Visini, R.; Awale, M.; Reymond, J.-L. Fragment Database Fdb-17. *J. Chem. Inf. Model.* **2017**, *57*, 700−709.
(19) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J. L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database Gdb-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.
(20) Awale, M.; Reymond, J. L. The Polypharmacology Browser Ppb2: Target Prediction Combining Nearest Neighbors with Machine Learning. *J. Chem. Inf. Model.* **2019**, *59*, 10−17.
(21) Gedeck, P.; Rohde, B.; Bartels, C. Qsar–How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *J. Chem. Inf. Model.* **2006**, *46*, 1924−1936.
(22) Rhode, B.. Avalon Toolkit. Https://Sourceforge.Net/P/Avalontoolkit/Wiki/Home/ (accessed February 21, 2019).
(23) Awale, M.; Reymond, J. L. Atom Pair 2d-Fingerprints Perceive 3d-Molecular Shape and Pharmacophores for Very Fast Virtual Screening of Zinc and Gdb-17. *J. Chem. Inf. Model.* **2014**, *54*, 1892−1897.
(24) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-Like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminform.* **2009**, *1*, 8.
(25) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(26) Awale, M.; Probst, D.; Reymond, J. L. Webmolcs: A Web-Based Interface for Visualizing Molecules in Three-Dimensional Chemical Spaces. *J. Chem. Inf. Model.* **2017**, *57*, 643−649.

DOI: 10.1021/acs.jcim.8b00902
J. Chem. Inf. Model. 2019, 59, 1347−1356