nature machine intelligence

ARTICLES

https://doi.org/10.1038/s42256-020-0160-y

Generative molecular design in low data regimes

Michael Moret[1], Lukas Friedrich[1], Francesca Grisoni[1], Daniel Merk[1,2] and Gisbert Schneider[1]✉

Generative machine learning models sample molecules from chemical space without the need for explicit design rules. To enable the generative design of innovative molecular entities with limited training data, a deep learning framework for customized compound library generation is presented that aims to enrich and expand the pharmacologically relevant chemical space with drug-like molecular entities on demand. This de novo design approach combines best practices and was used to generate molecules that incorporate features of both bioactive synthetic compounds and natural products, which are a primary source of inspiration for drug discovery. The results show that the data-driven machine intelligence acquires implicit chemical knowledge and generates novel molecules with bespoke properties and structural diversity. The method is available as an open-access tool for medicinal and bioorganic chemistry.

Automated molecular design methods support medicinal chemistry by efficient sampling of untapped drug-like chemical space[1–3]. A variety of so-called generative deep learning models have recently been introduced to the field of drug design to construct new molecules with desired properties from scratch (de novo)[4–11]. Given the rapidly increasing number of theoretical methods published[12,13], we here provide a consolidated deep learning framework for de novo molecular design that combines best practices. This open-access method is based on a validated generative model for creating virtual compound libraries for project-tailored applications in drug discovery and related areas.

Generative deep learning methods represent a class of machine learning algorithms that learn directly from the input data and do not necessarily depend on explicit rules coded by humans[14,15]. Generative models implement probabilistic estimators of data distributions (for example, sets of chemical structures) and specify how to generate new data (for example, new molecules) that fit such distributions[16]. Some of these methods implement a language modelling approach[17], where an artificial neural network aims to learn the probability of a token (for example, a word or a character) appearing in a sequence based on the distributions of all previous tokens in a sequence[18]. Through this process, deep neural networks can learn the features of sequential data. Once trained, these models can generate new sequences based on the sampled feature distributions. The language modelling approach for de novo molecular design relies on string-based molecular representations, such as the simplified molecular input line entry systems (SMILES)[19], which encodes molecular structure as a sequence of tokens. Pioneering prospective applications have experimentally verified the potential of SMILES-generative de novo design of small molecules with the desired bioactivity[4,20,21]. An essential element of these prospective applications is transfer learning[22,23], which is the process of transferring knowledge acquired to solve one task to another related task. In the first step (pretraining), the chemical language of bioactive molecules is learned by training a model on a large set of SMILES strings. In the second step, this general model is focused on a certain pharmacological target by performing transfer learning with small sets of molecules that possess the desired bioactivity.

The computational framework consists of an optimized chemical language model (CLM) for designing new molecules that populate designated areas in chemical space. It implements a recurrent neural network model with long short-term memory (LSTM)[24] for SMILES-based chemical structure generation. In this present study, we show the applicability of the deep learning framework to design molecules that combine features of bioactive synthetic compounds and natural products, which are an important source of inspiration for drug discovery[25,26]. The results demonstrate the ability of this computational approach to generate innovative molecules that are focused on a specific area of chemical space, for example, by enriching sets of structurally diverse de novo-generated molecules with natural product characteristics.

Results and discussion

Generating molecules with a CLM. A training dataset was compiled from ChEMBL24 to develop a language model of the chemical constitution of biologically active molecules[27]. Bioactive compounds with annotated bioactivities (the half maximal inhibitory concentration, IC_{50}; the half maximal effective concentration, EC_{50}; the dissociation constant, K_d; and the inhibition constant, K_i) < 1 μM were extracted from this chemical database and standardized, resulting in a set of ~365 thousand molecules. Each training molecule was presented to the CLM as a one-hot vector; that is, a computer-readable format derived from the respective SMILES string (Fig. 1a). In the one-hot encoding format, each token of the SMILES string vocabulary has a unique mathematical vector representation of a predefined length (71 bit in this study). During model training, the CLM learns the conditional probability distribution of a token with respect to all of the preceding tokens in the SMILES string (Fig. 1b). To further the applicability of the CLM to focused areas of the chemical space in the small-data regime, we combined three established concepts; namely, (1) data augmentation, (2) temperature sampling and (3) transfer learning, and investigated their effects on the quality of the resulting models.

Data augmentation. The amount and quality of the training data are key ingredients for successful language modelling[28]; however, large datasets for deep learning in drug design are scarce. Generative models must be able to handle small datasets to solve project-tailored design tasks in medicinal chemistry. Using multiple representations of the same entity (data augmentation) has been proposed as a strategy to work in small-data regimes and obtain generalizing models[29–31]. Here we employed data augmentation by

[1]Department of Chemistry and Applied Biosciences, RETHINK, ETH Zurich, Zurich, Switzerland. [2]Goethe University Frankfurt, Institute of Pharmaceutical Chemistry, Frankfurt, Germany. ✉e-mail: gisbert@ethz.ch

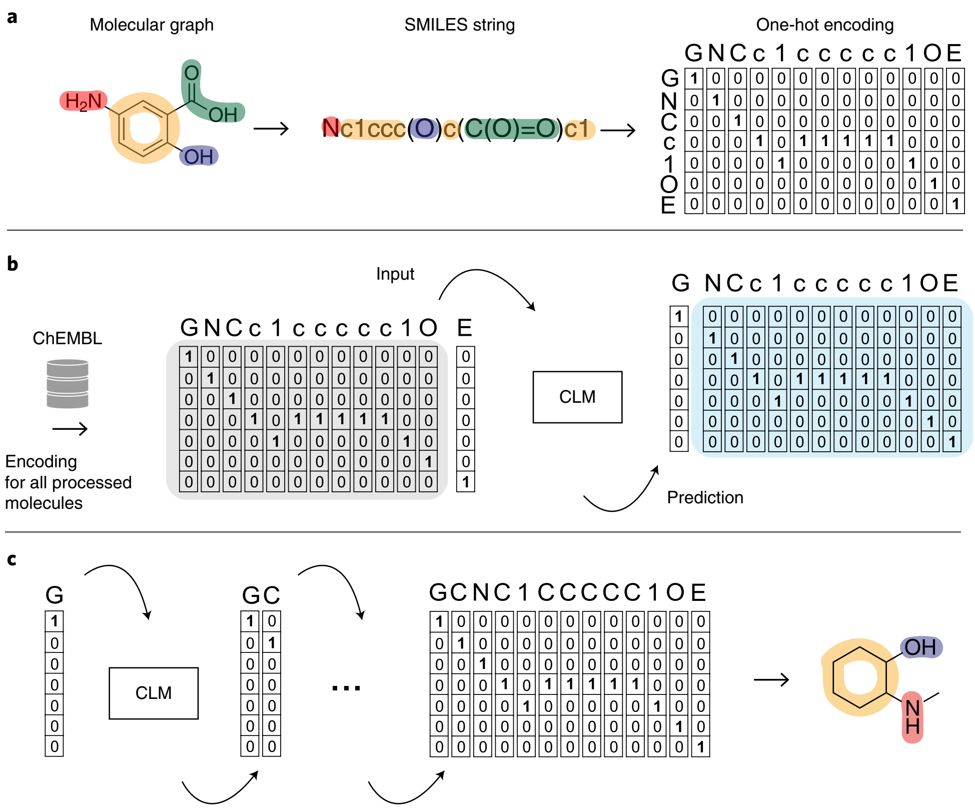
a
Molecular graph
H_2N
O
OH
OH
SMILES string
Nc1ccc(O)c(C(O)=O)c1
One-hot encoding
G N C c 1 c c c c c 1 O E
G 1 0 0 0 0 0 0 0 0 0 0 0 0
N 0 1 0 0 0 0 0 0 0 0 0 0 0
C 0 0 1 0 0 0 0 0 0 0 0 0 0
c 0 0 0 1 0 1 1 1 1 1 0 0 0
1 0 0 0 0 1 0 0 0 0 0 1 0 0
O 0 0 0 0 0 0 0 0 0 0 0 1 0
E 0 0 0 0 0 0 0 0 0 0 0 0 1
b
ChEMBL
Encoding for all processed molecules
G N C c 1 c c c c c 1 O E
Input
CLM
G N C c 1 c c c c c 1 O E
Prediction
c
G
CLM
G C
...
G C N C 1 C C C C C 1 O E
OH
N
H

Fig. 1 | CLM training and sampling of new molecules. a, Each molecule is translated into a SMILES string from its molecular graph. Combined with a start token (G) and an end token (E), SMILES strings are presented as an input to the CLM using one-hot encoding. **b**, The CLM learns the feature distribution of the dataset by predicting each token from the preceding token(s) in a SMILES string. **c**, For de novo molecule generation (sampling step), the CLM repeatedly samples tokens from the learned distribution until the end token is sampled, indicating the completion of a new SMILES string.

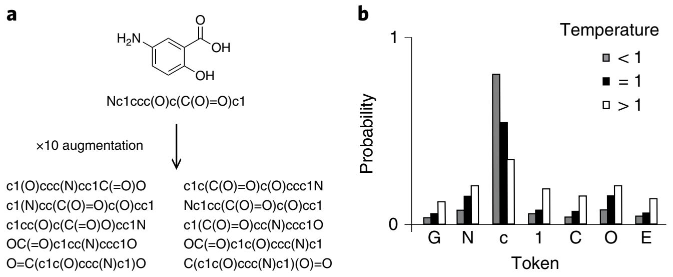

Fig. 2 | Data augmentation and temperature sampling. a, An example of tenfold data augmentation. All SMILES strings represent the same molecular graph. **b**, The effect of the sampling temperature (T) on the conditional probability distribution over the SMILES string vocabulary for selected tokens (G, N, C, 1, c, O, E). $T=1$ represents the probability distribution the CLM learned during training, $T<1$ sharpens the distribution and $T>1$ flattens the distribution.

Table 1 | The validity, uniqueness and novelty of molecules depending on data augmentation and the sampling temperature

Data augmentation	Sampling temperature	Best epoch	Valid (%)	Unique (%)	Novel (%)
None	1.0	10	84.4 ± 1.9	84.3 ± 1.8	82.3 ± 1.3
×1	1.0	10	81.9 ± 1.9	81.9 ± 1.9	81.7 ± 1.8
×3	1.0	9	91.5 ± 0.6	91.5 ± 0.6	90.5 ± 0.6
×10	1.0	7	93.5 ± 0.8	93.5 ± 0.8	92.3 ± 0.6
×20	1.0	6	94.2 ± 0.4	94.2 ± 0.4	92.6 ± 0.4
×10	0.2	7	85.8 ± 6.3	54.9 ± 6.4	52.0 ± 7.2
×10	0.7	4	97.4 ± 0.8	97.3 ± 0.8	93.8 ± 1.0
×10	1.2	4	84.8 ± 3.4	84.8 ± 3.4	84.4 ± 3.3

Each experiment was run for ten epochs and repeated ten times (the mean and standard deviation are reported); 5,000 molecules were sampled after each epoch. The best epoch was defined as the one that yielded the highest average novelty value. Percentages are reported with respect to the total number of molecules sampled.

leveraging the non-univocity of SMILES strings[19]. Multiple valid SMILES strings that represent the same molecular graph were constructed by starting each string from a different non-hydrogen atom in a molecule (Fig. 2a)[32]. We compared the effect of this approach on model training in terms of (1) the validity; that is, the percentage of SMILES strings that can be translated back to molecular graphs; (2) the uniqueness, which is the percentage of non-duplicated SMILES strings; and (3) the novelty of the generated molecules calculated as a percentage of SMILES strings not present in the training set. High validity indicates that the model has learned the necessary features to generate chemically meaningful SMILES strings, high uniqueness indicates that SMILES strings generation is non-redundant, and a high degree of novelty suggests that the model is suitable for generating new molecules from scratch.

Four levels of augmentation were tested (×1, ×3, ×10, ×20) and the CLM was trained for ten epochs at each level, with an epoch defined as one pass over all of the training data (Fig. 1c). We observed that augmenting the training data was beneficial in terms of all indices when compared with the non-augmented scenario,

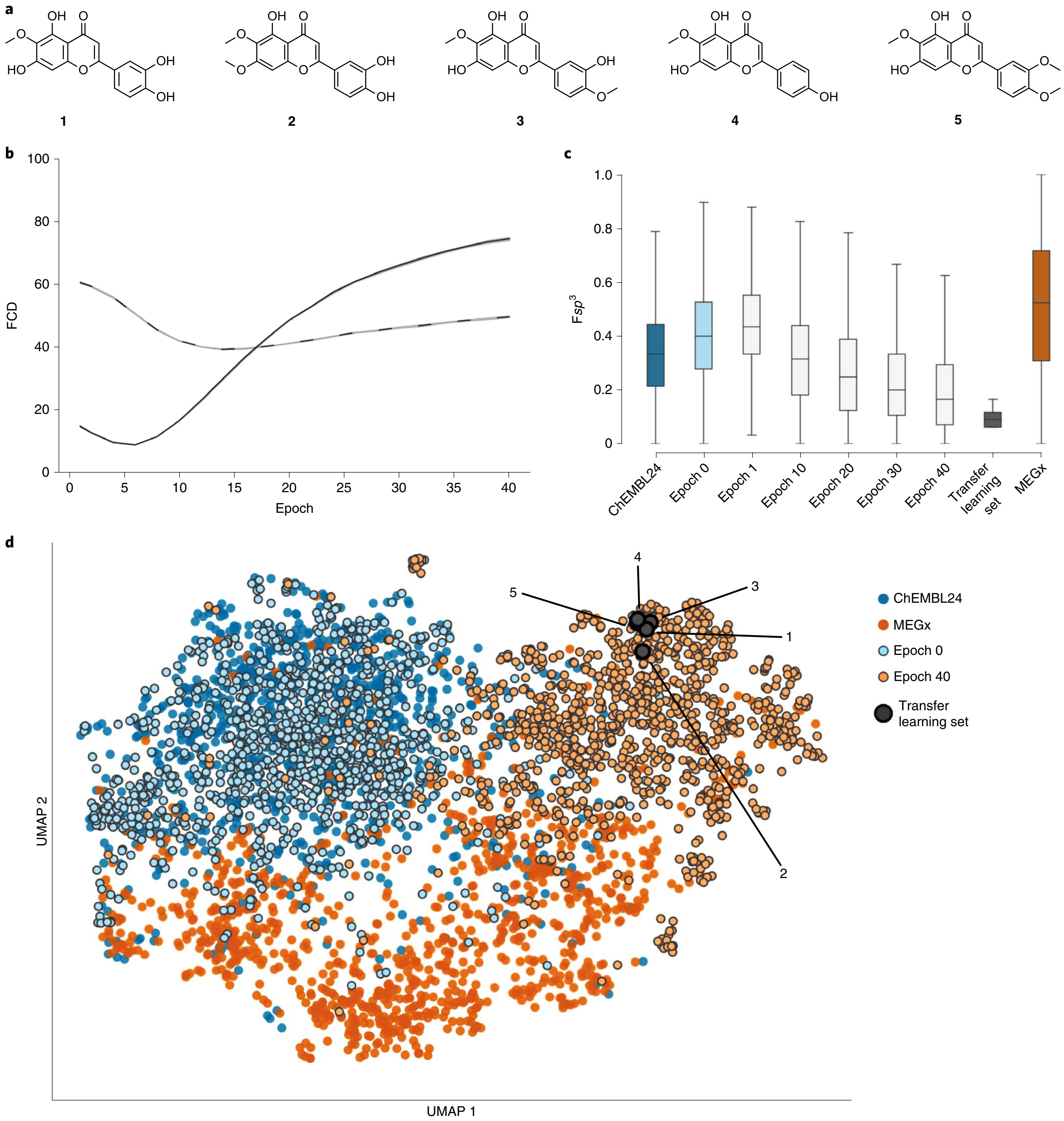

Fig. 3 | Chemical space navigation by transfer learning with five similar molecules. **a**, The transfer learning set comprised five structurally similar natural products (**1–5**) from the natural product collection MEGx. **b**, The FCD of the generated molecules during chemical space navigation to ChEMBL24 (solid line) and MEGx (dashed line), respectively. The mean and 95% confidence interval for ten repeats are shown in the shaded area. **c**, Evolution of the fraction of sp^3-hybridized carbon atoms (Fsp^3) during chemical space navigation. **d**, UMAP plot of molecules (1,000 molecules were randomly selected for each group).

with the exception of onefold augmentation (Table 1). We hypothesize that with only one additional representation of each SMILES string, the underlying statistical patterns of the distribution of tokens are difficult to capture. Moreover, twentyfold augmentation did not further improve on the results that were obtained with tenfold augmentation (Table 1). This result is important because it not only demonstrates that generative models can benefit from SMILES augmentation but also suggests an optimal degree of data augmentation for a given learning task. This result extends previous studies using data augmentation for generative SMILES modelling[33–35].

Temperature sampling. In an attempt to further assess the model's potential to generate valid, unique and novel SMILES strings, we investigated the effect of the so-called sampling temperature, *T*

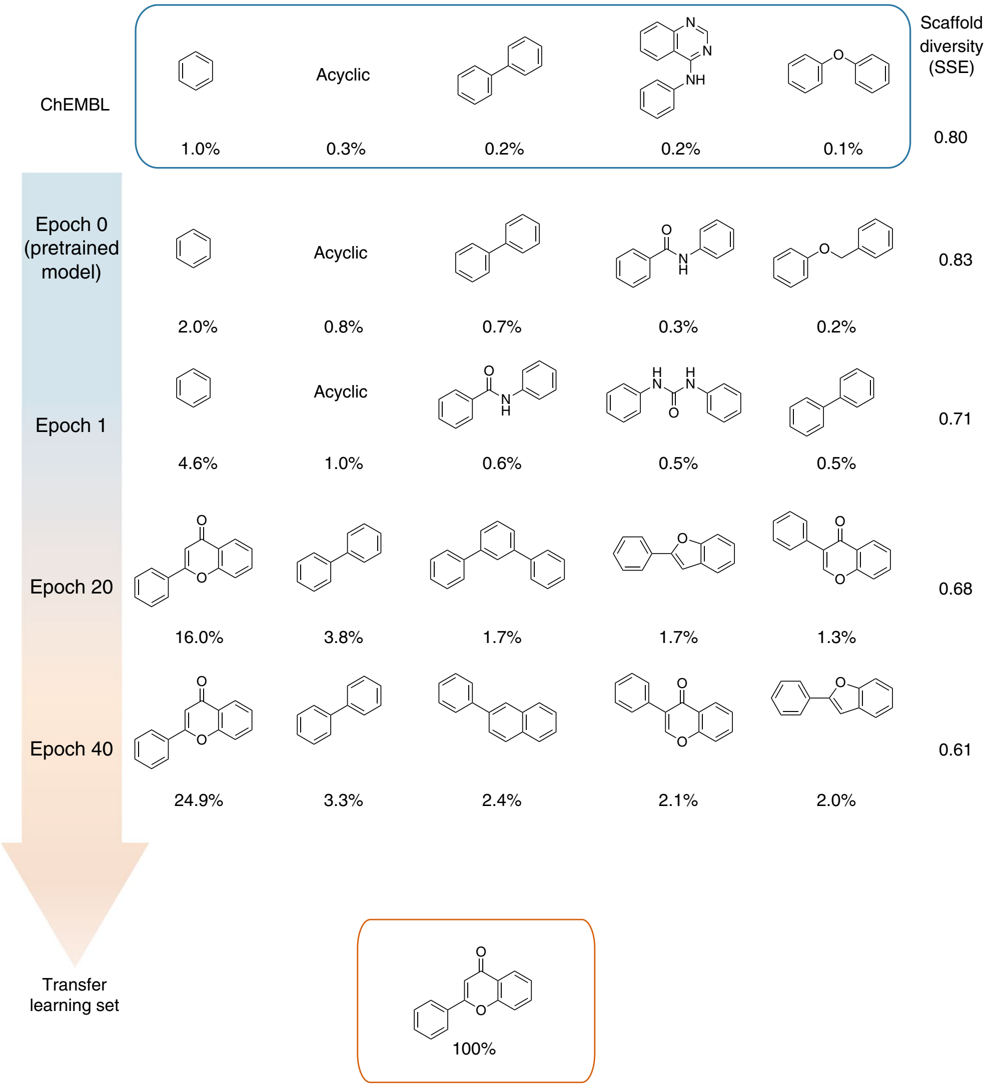

Fig. 4 | The five most frequent scaffolds from different training epochs during chemical space navigation to de novo-generated focused compound libraries. The percentages indicate the fraction of molecules that contain the respective scaffold. The scaffold diversity of the five most frequent scaffolds was quantified by SSE (values from [0,1], see equation (2)). A greater SSE value indicates greater diversity. In total, 7% of all sampled molecules at epoch 1, 25% at epoch 20 and 35% at epoch 40 were represented by the five most frequent scaffolds.

(see equation (1)). The sampling temperature ($T>0$) governs the randomness of the chosen token at each step of sequence generation. For $T\rightarrow 0$, the most likely token according to the estimated probability distribution is selected; with increasing values of T, the chances of selecting the most likely token decrease and thus the model generates more diverse sequences (Fig. 2b). In the extreme case of $T\rightarrow\infty$, tokens will be selected with equal probabilities. We investigated the influence of four temperatures with respect to the probability distribution learned by the model: two conservative values ($T=0.2$ and $T=0.7$), one unbiased value ($T=1.0$) and one permissive value ($T=1.2$). The highest levels of valid, unique and novel SMILES strings were obtained at a sampling temperature of $T=0.7$ (Table 1). Combining both data augmentation and temperature sampling led to an optimized CLM, as indicated by the increased levels of validity, uniqueness and novelty of sampled molecules (Table 1). This result is in agreement with a previous study from our group[36] and rectifies a previous recommendation not to use temperature sampling for SMILES generation with recurrent neural networks[37]. The model trained with tenfold data augmentation and $T=0.7$ was used for generating application-focused libraries in all of the subsequent experiments.

Generating compound libraries with transfer learning. Building on the general optimization results of the CLM, we investigated the potential of transfer learning to create novel and diverse virtual compound libraries for drug discovery. Transfer learning has successfully been applied to generative SMILES design before[17,36,38]. Here we investigate the efficiency of this method in a low data regime, which often characterizes early stage drug discovery. To enrich sets of generated molecules with features that are relevant for drug discovery, we applied transfer learning to navigate between two spaces: a synthetic compound space (source space) of bioactive molecules compiled from ChEMBL24, and a chemical space of natural products (objective space) defined by natural products from plants and microorganisms (MEGx collection, Analyticon Discovery GmbH). All of the parameters of the first layer of the neural network were kept constant in an attempt to retain features from the pretraining data. We found that keeping the first layer of the CLM constant helps prevent a performance drop during transfer learning (Supplementary Table 1).

Generating application-focused compound libraries. As an example of building an application-focused compound library by

Table 2 | Scaffold analysis during the transfer learning with five similar and five dissimilar natural products

Source	Novel molecules	Scaffolds	Singleton scaffolds	$P_{scaffolds}$ (N/M)	$P_{singletons}$ (Ns/N)	Novel % (ChEMBL24 | MEGx)
General base model without transfer learning						
Epoch 0	9,567	8,296	7,844	0.87	0.95	99 | 99
Transfer learning with five similar compounds						
Epoch 1	9,035	6,622	5,983	0.73	0.90	83 | 98
Epoch 20	6,543	2,522	1,998	0.39	0.79	79 | 92
Epoch 40	3,373	891	660	0.26	0.74	75 | 85
Transfer learning with five dissimilar compounds						
Epoch 1	9,459	8,187	7,573	0.87	0.93	86 | 99
Epoch 20	8,702	7,581	7,288	0.87	0.96	92 | 97
Epoch 40	8,184	7,140	6,917	0.87	0.97	94 | 97
Datasets						
MEGx	2,931	1,159	797	0.40	0.69	n.a.
ChEMBL24	365,063	135,120	93,174	0.37	0.69	n.a.

n.a., not applicable. Scaffolds (N) were extracted from chemically valid, unique, novel molecules (M). Epoch 0 indicates molecules sampled from the pretrained model (before transfer learning). Singleton scaffolds (Ns) represent scaffolds with a frequency of one; $P_{scaffolds}$ and $P_{singletons}$ represent the fraction of scaffolds and singletons, respectively. The fraction of novel scaffolds was calculated by comparison with scaffolds that are contained in the training dataset (ChEMBL24) and the natural product set (MEGx).

transfer learning, we selected five structurally similar molecules from the MEGx collection of natural product screening compounds (compounds **1**–**5**, Fig. 3a) according to their Jaccard–Tanimoto similarity[39] computed on Morgan fingerprints[40] (with similarity values higher than 0.78). These five natural products were used for transfer learning.

We computed the Fréchet ChemNet Distance (FCD)—a distance metric to evaluate the similarity between two populations of molecules based on their chemical structures and bioactivities[41]—to estimate the coverage of the chemical space during transfer learning. An FCD value of zero indicates that the compared molecular spaces are identical, whereas higher values indicate greater dissimilarity. The FCD curves evolved continuously as a function of the number of training epochs (Fig. 3b). This observation indicates that the CLM is able to sample the chemical space between the source space and the objective space in a continuous fashion despite the molecules being discrete entities.

During the initial epochs of transfer learning (epochs one to six), the distances of the generated molecules to the objective space (MEGx) and the source space (ChEMBL24) decreased before increasing. The lower FCD to the source space during the initial epochs can be explained by the initial effect of transfer learning. The model focused on features that are common between the source space and the molecules of the objective space, possibly because ChEMBL24 contains natural products and many synthetic molecules are natural product-inspired compounds[42]. The increasing distance to the natural product space during transfer learning might seem initially counter intuitive, a likely explanation for this is the limited size and diversity of the set of five natural products used for transfer learning compared with the whole natural product space.

We selected the fraction of sp^3-hybridized carbon atoms (Fsp^3) as an illustrative example to highlight the changes of physicochemical properties during transfer learning; Fsp^3 has been shown to correlate with the odds of a molecule becoming a drug[43] and differs between synthetic compounds and natural products[44]. During transfer learning, the Fsp^3 distribution approximated the transfer learning set distribution (Fig. 3c). This finding confirms that transfer learning from a small set of structurally similar compounds enables the model to implicitly capture relevant physicochemical properties.

In an attempt to visualize the relative location of the computer-generated molecules in chemical space[45], uniform manifold approximation and projection[46] (UMAP) plots were generated. UMAP creates a two-dimensional representation of high-dimensional data distributions, in which the similarity relations between data points in the original high-dimensional space are better preserved than with t-distributed stochastic neighbour embedding[46,47] (here, the molecules were represented as Morgan fingerprints). In this visualization, the molecules sampled from the pretrained CLM (light blue) are close to the training data (dark blue) and the molecules are shifted towards the location of the transfer learning set after transfer learning (epoch 40) (Fig. 3d). This graphical analysis corroborates the effectiveness of transfer learning for navigating in chemical space from the source to the objective. Similar results were observed with 10 and 50 molecules in the transfer learning set (Supplementary Figs. 2 and 4). The same trend is noticeable with only a single molecule used for transfer learning, albeit with a weaker effect (Supplementary Fig. 1).

We further assessed the coverage of chemical space and the diversity of the generated molecules by analysing their Bemis–Murcko scaffolds[48]. We examined the five most frequent scaffolds of sampled molecules before (using the pretrained CLM) and during transfer learning (Fig. 4). As a measure of scaffold diversity, we determined the Shannon entropy scaled by the number of investigated scaffolds (that is, the scaled Shannon entropy or SSE, equation (2))[49]. The SSE quantifies the structural diversity of a given set of scaffolds: SSE = 1 indicates maximum diversity, whereas SSE = 0 indicates the presence of a single molecular scaffold. During the transfer learning process, the number of molecules containing one of the five most frequent scaffolds increased, whereas their diversity decreased in terms of SSE. When assessing the whole population, the number of unique scaffolds decreased by approximately 50% during transfer learning. The fraction of singletons (scaffolds occurring only once in a population) also decreased (Table 2 and Supplementary Information). This result shows that transfer learning with the structurally conserved natural products **1**–**5** (Fig. 3a) led to the de novo design of a structurally focused compound collection that predominantly contains the chemical scaffold of the transfer learning set.

We then examined the novelty of the generated molecules and their corresponding scaffolds. The total number of novel molecules

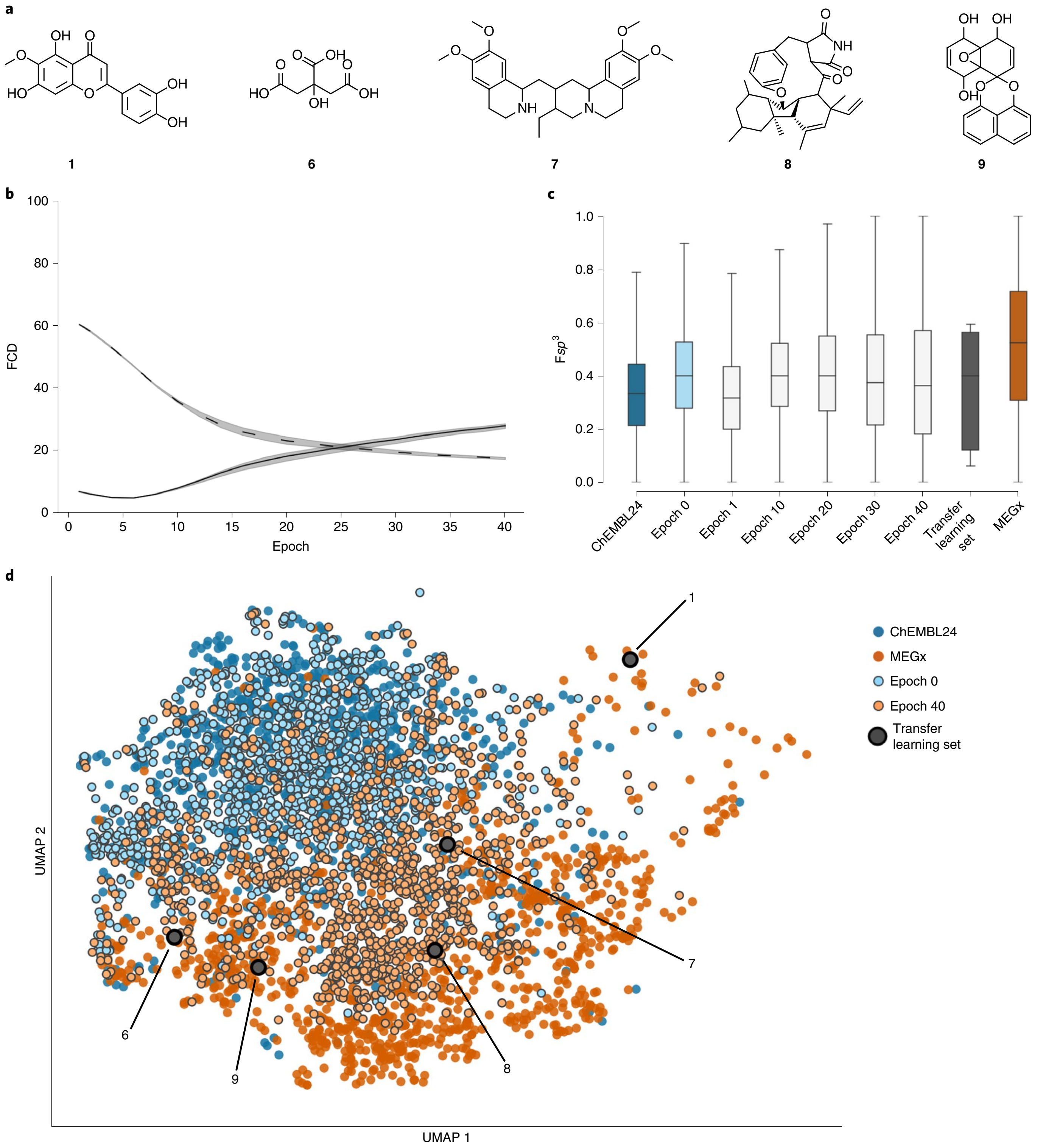

Fig. 5 | Chemical space navigation by transfer learning with five dissimilar molecules. a, Five dissimilar natural products (**1, 6–9**) from the MEGx collection are shown. **b**, The FCDs of the generated molecule to ChEMBL24 (solid line) and MEGx (dashed line), respectively. The mean value and 95% confidence interval (shaded area) for ten repeats are shown. **c**, The evolution of the fraction of sp^3-hybridized carbon atoms (Fsp^3) during transfer learning. **d**, UMAP plot of molecule distributions is shown. In total, 1,000 molecules were randomly selected from each set.

with respect to the training and transfer learning set was reduced by 60% at the end of the transfer learning process, whereas the number of novel scaffolds only decreased marginally (Table 2 and Supplementary Information). We compared the de novo designs with the Enamine compound set (700 M drug-like compounds)—which is among the largest available sets of screening compounds—to further assess the novelty of the generated compounds. More than 99% of the molecules generated by the model were new and the proportion of new scaffolds compared with Enamine increased from 75% to over 95% during transfer learning (Table 2 and

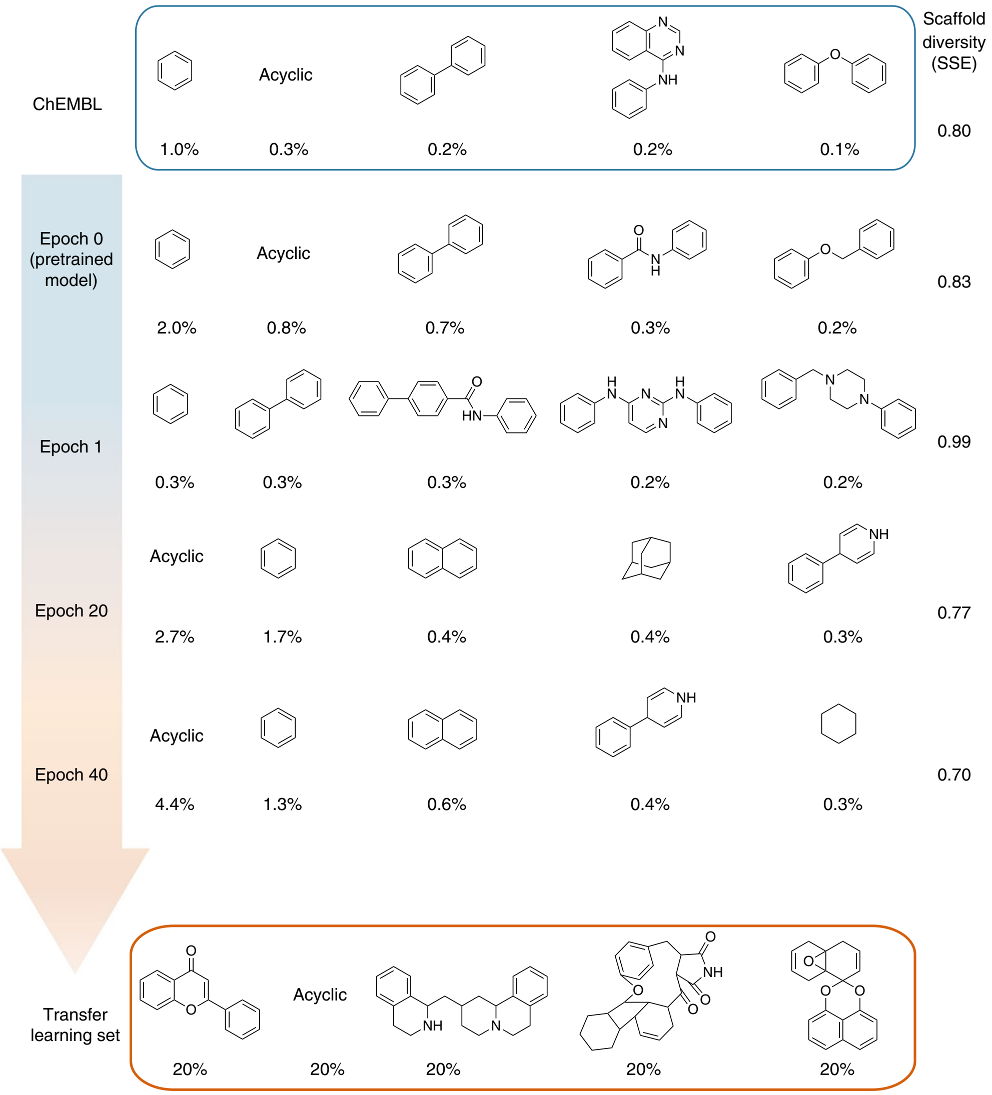

Fig. 6 | The five most frequent scaffolds from different training epochs during chemical space navigation. The percentages indicate the fraction of molecules containing the scaffold (epoch 0: sampled from the pretrained model). The diversity of the five most frequent scaffolds is expressed in terms of the SSE. Here, 1% of all sampled molecules at epoch 1, 6% at epoch 20 and 7% at epoch 40, contained one of the five most frequent scaffolds.

Supplementary Information). The FCD to Enamine also increased during transfer learning (Supplementary Fig. 6). Overall, the molecular design process benefitted from transfer learning, in terms of the FCD and increased scaffold diversity compared with the Enamine screening compound collection. These new scaffolds might provide innovative starting points for drug design.

Generating virtual libraries by expanding the chemical space. Having demonstrated the ability of the CLM to generate scaffold-focused de novo sets, we explored the application of transfer learning to expand the sampled chemical space from the training space to the objective space. Here the transfer learning set contained molecule **1** as well as four dissimilar natural products (**6–9**, Fig. 5a) to increase the diversity of the transfer learning set and observe its effect on the structure of the generated molecules. We observed that both FCD curves evolved continuously as a function of the number of epochs (Fig. 5b). Although the distance to the objective space (MEGx) continuously decreased with the number of epochs, the distance to the source space (ChEMBL24) remained initially stable but increased after the fifth epoch. The Fsp^3 distribution of the sampled molecules (Fig. 5c) after the last transfer learning epoch (epoch 40) reflects the distributions of both the transfer learning set and the whole MEGx collection. This result suggests that pronounced structural diversity of the transfer learning set permits sampling of molecules with structural characteristics that cover a representative portion of the objective space (Fig. 5b,c). By contrast, transfer learning with five similar molecules resulted in the generation of molecules predominantly with characteristics of the transfer learning set (Fig. 3b,c). UMAP visualization indicates that many molecules were sampled from areas in the vicinity of the natural products **6–9**. Overall, the compound distribution at epoch 40 corroborates extended coverage of chemical space with de novo-generated molecules. Similar results were observed with 10 and 50 molecules in the transfer learning set (Supplementary Figs. 3 and 5).

The five most frequent scaffolds represented only a small fraction of all generated molecules compared to the analysis with five similar natural products (Fig. 6). The diversity (SSE) of the five most frequent scaffolds decreased during transfer learning. The fractions of scaffolds and singletons were high and slightly increased throughout the transfer learning process (Table 2). The generated sets comprised a large fraction of molecules with

Table 3 | Novelty comparison of de novo-generated molecules with the Enamine REAL database (720 million compounds)

Transfer learning set	Novel molecules (%)	Novel Scaffolds (%)
Five similar compounds		
Epoch 1	99.75 ± 0.05	79.15 ± 0.61
Epoch 20	99.93 ± 0.04	92.23 ± 0.45
Epoch 40	99.97 ± 0.03	95.05 ± 1.01
Five dissimilar compounds		
Epoch 1	99.73 ± 0.04	82.68 ± 0.20
Epoch 20	99.94 ± 0.02	97.71 ± 0.06
Epoch 40	99.99 ± 0.01	99.68 ± 0.02

The mean and standard deviation are reported for five sets independently and randomly sampled from Enamine, each containing at least 3,373 molecules (see Methods).

a novel scaffold compared to the source and objective spaces (Table 2). After transfer learning, the majority of the generated molecules and scaffolds (>99%) were not contained in the Enamine collection (Table 3).

We conclude that transfer learning with a structurally diverse transfer learning set allows the generation of structurally diverse molecules that comprise a broad range of scaffolds and possess properties of the objective space; for example, an enriched fraction of sp^3-hybridized carbon atoms. This approach could help enrich screening compound collections with innovative, natural product-inspired compounds and scaffolds for virtual screening and high-throughput screening.

Conclusions

Generative deep learning extends the medicinal chemistry toolbox by providing a complementary approach to exhaustive or chemical-transformation-based structure enumeration and combinatorial sampling. The results of this study demonstrate that CLMs combined with data augmentation, transfer learning and temperature sampling enable the discovery of new molecular entities in a low data regime, which is often encountered in early stage drug-discovery projects.

The SMILES-based model proved able to generate new molecules with bespoke properties at the interface between synthetic compounds and natural products. By relying on the chemical similarity principle[50] and natural products as starting points for drug design, this computational approach successfully generated novel and chemically diverse molecular entities. This pretrained CLM and the analysis framework are publicly accessible to encourage researchers to apply transfer learning on custom sets of molecules for chemical space exploration. It should be noted that this computational framework does not explicitly assess the synthesizability of molecules, and further compound ranking and prioritization may be required. Keeping these constraints in mind, only broad prospective application of this machine learning model will reveal if the underlying data-driven approach has the potential to accelerate the identification of novel bioactive compounds. We envision these de novo structure generators an integral part of future drug discovery teams for decision making with collaborative intelligence.

Methods

Training compounds and data processing. Compounds with annotated activity values (IC_{50}, EC_{50}, K_d, K_i) < 1 μM were retrieved from ChEMBL24 to cover the chemical space of biologically active compounds. Molecular structures were encoded as canonical SMILES strings[51] using the RDKit package (v.2018.03, www.rdkit.org) and only SMILES strings with a length of up to 140 tokens (characters) were retained. SMILES strings were standardized in Python (v3.6.5, www.python.org) by removing stereochemical information, salts and duplicates. This data preparation resulted in a set of 365,063 bioactive molecules encoded as unique SMILES strings (referred to as ChEMBL24).

Transfer learning sets. Molecules for transfer learning were retrieved from the natural product collection MEGx (released 01 September 2018, Analyticon Discovery GmbH). All existing sugar moieties were removed by substructure filtering using DataWarrior software (www.openmolecules.org/datawarrior/, v.5.0.0) to focus on structural features of the central scaffolds of these natural products[52]; 2,931 molecules were retained. To assess pairwise similarities, all molecules were represented as bit vectors according to the Morgan fingerprint algorithm[20] (length = 2,048 bits, radius = 2 bonds) as implemented in RDKit (version 2018.03). Morgan fingerprints numerically encode the presence of radial molecular fragments. Molecule **1** was randomly selected from the dataset. The four most similar molecules according to their Tanimoto similarity (Tanimoto coefficient, T_C) to molecule **1** were chosen from MEGx (T_C = 0.78 to T_C = 0.82). Based on molecule **1**, the MaxMinPick algorithm, as implemented in RDKit (LazyBitVectorPick), was used to select a subset of four dissimilar natural products (T_C = 0.04 to T_C = 0.10).

CLM implementation. All software programs were implemented in Python (v3.6.5) using Keras (https://keras.io/, v2.2.0) with the TensorFlow GPU backend (www.tensorflow.org, v1.9.0). The CLM was implemented as a recurrent neural network with LSTM cells[24]. The neural network was composed of four layers that have a total of 5,820,515 parameters: layer 1, BatchNormalization; layer 2, LSTM with 1,024 units; layer 3, LSTM with 256 units; layer 4, BatchNormalization) and was trained with SMILES strings encoded as one-hot vectors. We used the categorical cross-entropy loss and the Adam optimizer[53] with a learning rate of 0.001 for training the CLM (ten epochs, where one epoch is defined as one pass over all the training data, took approximately 21 h on the processed ChEMBL24 data with tenfold data augmentation on a single Nvidia GTX 1080 GPU with 256 GB of memory). Additional model details are provided in the Supplementary Information.

Each training run was repeated ten times and 5,000 SMILES strings were sampled after each epoch for a total of ten epochs. Transfer learning was performed by keeping the parameters of the first model layer constant and training the second layer with a smaller learning rate (equal to 10^{-4}). Each transfer learning experiment was repeated ten times and 10,000 molecules were sampled after each second epoch for a total of 40 epochs.

Sampling of new SMILES strings. Sampling of SMILES string characters was performed using the softmax function parameterized by the sampling temperature. The probability of the ith token to be sampled from CLM predictions was computed as (equation (1)):

$$q_i = \frac{e^{(z_i/T)}}{\sum_j e^{(z_j/T)}} \quad (1)$$

where z_i is the CLM prediction for token i, T is the temperature and q_i is the sampling probability of token i given by the CLM.

Data augmentation. Data augmentation was performed with RDKit (v2018.03.3.0) using multiple SMILES string representations of the same molecule. We implemented four levels of data augmentation (×1, ×3, ×10, ×20) for pretraining, where a tenfold augmentation is defined as the SMILES string canonicalized version plus ten alternative representations. Tenfold augmentation was also applied to the transfer learning sets.

FCD. The FCD was computed by comparing the Fréchet distance[54] between molecules from the training set and generated molecules. The FCD was calculated following the implementation provided by Preuer et al. (https://github.com/bioinf-jku/FCD)[41]. In total, 5,000 molecules were randomly selected from each compound set for FCD calculation when possible. A minimum of 2,931 molecules was used to compute the FCD to MEGx (that is, the number of compounds available from MEGx after initial data processing). To compute the FCD to the Enamine compound collection, five sets of up to 5,000 molecules each were independently and randomly sampled from Enamine, containing at least 3373 molecules (that is, the number of compounds available from epoch 40 of transfer learning with five similar molecules).

Normalized SSE. We used the following equation to compute the normalized SSE[49]:

$$\mathrm{SSE} = \frac{-\sum_{i=1}^{n} p_i \ln\left(\frac{c_i}{P}\right)}{\log_2(n)} \quad (2)$$

where the numerator is the Shannon entropy, n is the number of unique scaffolds considered, c_i is the number of compounds containing the ith scaffold and P is the total number of compounds among the considered *n* scaffolds. The denominator bounds the values to the interval [0,1].

Data availability

The training data used for model building are available from the supplementary Code Ocean capsule[55] at https://doi.org/10.24433/CO.0753661.v1, or as GitHub repository[56] at https://github.com/ETHmodlab/virtual_libraries.

Code availability

The computational framework for generative molecular design, along with the pretrained neural network weights, is available from the supplementary Code Ocean capsule[55] at https://doi.org/10.24433/CO.0753661.v1, or as a GitHub repository[56] at https://github.com/ETHmodlab/virtual_libraries.

Received: 18 August 2019; Accepted: 10 February 2020;
Published online: 16 March 2020

References

1. Walters, W. P. Virtual chemical libraries. *J. Med. Chem.* **62**, 1116–1124 (2019).
2. Mullard, A. 2018 FDA drug approvals. *Nat. Rev. Drug Discov.* **18**, 85–89 (2019).
3. Dowden, H. & Munro, J. Trends in clinical success rates and therapeutic focus. *Nat. Rev. Drug Discov.* **18**, 495 (2019).
4. Yuan, W. et al. Chemical space mimicry for drug discovery. *J. Chem. Inf. Model.* **57**, 875–882 (2017).
5. Guimaraes, G. L., Sanchez-Lengeling, B., Outeiral, C., Farias, P. L. C. & Aspuru-Guzik, A. Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. Preprint at http://arxiv.org/abs/1705.10843 (2017).
6. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **9**, 48 (2017).
7. Putin, E. et al. Reinforced adversarial neural computer for de novo molecular design. *J. Chem. Inf. Model.* **58**, 1194–1204 (2018).
8. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
9. Popova, M., Shvets, M., Oliva, J. & Isayev, O. MolecularRNN: generating realistic molecular graphs with optimized properties. Preprint at https://arxiv.org/abs/1905.13372 (2019).
10. Jin, W., Barzilay, R. & Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. Preprint at https://arxiv.org/abs/1802.04364 (2018).
11. You, J., Liu, B., Ying, Z., Pande, V. & Leskovec, J. Graph convolutional policy network for goal-directed molecular graph generation. *Adv. NIPS* **32**, 6410–6421 (2018).
12. Vamathevan, J. et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
13. Yang, X., Wang, Y., Byrne, R., Schneider, G. & Yang, S. Concepts of artificial intelligence for computer-assisted drug discovery. *Chem. Rev.* **119**, 10520–10594 (2019).
14. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
15. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **61**, 85–117 (2015).
16. Jebara, T. *Machine Learning: Discriminative and Generative* (Kluwer Academic, Springer, 2004).
17. 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.* **4**, 120–131 (2018).
18. Bengio, Y., Ducharme, R., Vincent, P. & Jauvin, C. A neural probabilistic language model. *J. Mach. Learn. Res.* **3**, 1137–1155 (2003).
19. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
20. Merk, D., Grisoni, F., Friedrich, L. & Schneider, G. Tuning artificial intelligence on the *de novo* design of natural-product-inspired retinoid X receptor modulators. *Commun. Chem.* **1**, 68 (2018).
21. Merk, D., Friedrich, L., Grisoni, F. & Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Mol. Inf.* **37**, 1700153 (2018).
22. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? *Adv. NIPS* **27**, 3320–3328 (2014).
23. Peters, M. E., Ruder, S. & Smith, N. A. To tune or not to tune? Adapting pretrained representations to diverse tasks. In *Proc. 4th Workshop on Representation Learning for NLP* 7–14 (RepL4NLP, 2019).
24. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
25. Rodrigues, T., Reker, D., Schneider, P. & Schneider, G. Counting on natural products for drug design. *Nat. Chem.* **8**, 531–541 (2016).
26. Follmann, M. et al. An approach towards enhancement of a screening library: the next generation library initiative (NGLI) at Bayer—against all odds? *Drug Discov. Today* **24**, 668–672 (2019).
27. Gaulton, A. et al. The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2016).
28. Radford, A. et al. Language models are unsupervised multitask learners. Preprint at https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (2019).
29. Simard, P., Victorri, B., LeCun, Y. & Denker, J. Tangent prop—a formalism for specifying selected invariances in an adaptive network. *Adv. NIPS* **4**, 895–903 (1991).
30. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Adv. NIPS* **25**, 1097–1105 (2012).
31. Dao, T. et al. A kernel theory of modern data augmentation. *Proc. Mach. Lern. Res.* **97**, 1528–1537 (2019).
32. Bjerrum, E. J. SMILES enumeration as data augmentation for neural network modeling of molecules. Preprint at https://arxiv.org/abs/1703.07076 (2017).
33. Bjerrum, E. & Sattarov, B. Improving chemical autoencoder latent space and molecular de novo generation diversity with heteroencoders. *Biomolecules* **8**, 131 (2018).
34. Arús-Pous, J. et al. Randomized SMILES strings improve the quality of molecular generative models. *J. Cheminformatics* **11**, 71 (2019).
35. Prykhodko, O. et al. A de novo molecular generation method using latent vector based generative adversarial network. *J. Cheminformatics* **11**, 74 (2019).
36. Gupta, A. et al. Generative recurrent networks for de novo drug design. *Mol. Inf.* **37**, 1700111 (2018).
37. Neil, D. et al. Exploring deep recurrent models with reinforcement learning for molecule design. In *The Sixth International Conference on Learning Representations. Vancouver Convention Center* Workshop paper (ICLR, 2018); https://iclr.cc/Conferences/2018
38. Awale, M., Sirockin, F., Stiefl, N. & Reymond, J. L. Drug analogs from fragment-based long short-term memory generative neural networks. *J. Chem. Inf. Model.* **59**, 1347–1356 (2019).
39. Tanimoto, T. T. *An Elementary Mathematical Theory of Classification and Prediction* (International Business Machines Corporation, 1958).
40. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
41. Preuer, K., Renz, P., Unterthiner, T., Hochreiter, S. & Klambauer, G. Fréchet ChemNet distance: a metric for generative models for molecules in drug discovery. *J. Chem. Inf. Model.* **58**, 1736–1741 (2018).
42. Boufridi, A. & Quinn, R. J. Harnessing the properties of natural products. *Annu. Rev. Pharmacol. Toxicol.* **58**, 451–470 (2018).
43. Lovering, F., Bikker, J. & Humblet, C. Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **52**, 6752–6756 (2009).
44. Stratton, C. F., Newman, D. J. & Tan, D. S. Cheminformatic comparison of approved drugs from natural product versus synthetic origins. *Bioorg. Med. Chem. Lett.* **25**, 4802–4807 (2015).
45. Reutlinger, M. & Schneider, G. Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *J. Mol. Graph. Model.* **34**, 108–117 (2012).
46. McInnes, L. & Healy, J. UMAP: Uniform manifold approximation and projection for dimension reduction. Preprint at http://arxiv.org/abs/1802.03426v1 (2018).
47. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2018).
48. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).
49. Medina-Franco, J. L. & Martínez-Mayorga, K. Scaffold diversity analysis of compound data sets using an entropy-based measure. *QSAR Comb. Sci.* **28**, 1551–1560 (2009).
50. Johnson, M. A. & Maggiora, G. M. *Concepts and Applications of Molecular Similarity* (John Wiley & Sons, 1990).
51. O'Boyle, N. M. Towards a universal SMILES representation—a standard method to generate canonical SMILES based on the InChI. *J. Cheminform.* **4**, 22 (2012).
52. Sander, T., Freyss, J., von Korff, M. & Rufener, C. DataWarrior: an open-source program for chemistry aware data visualization and analysis. *J. Chem. Inf. Model.* **55**, 460–473 (2015).
53. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at http://arxiv.org/abs/1412.6980 (2014).
54. Fréchet, M. Sur la distance de deux lois de probabilité. *Comp. Rend. Hebdom. Séances l'Acad. Sci.* **244**, 689–692 (1957).
55. Moret M., Friedrich L., Grisoni F., Merk D. & Schneider G. *Generative Molecular Design in Low Data Regimes* (CodeOcean, 2020); https://doi.org/10.24433/CO.0753661.v1

56. Moret M., Friedrich L., Grisoni F., Merk D. & Schneider G. *Generative Molecular Design in Low Data Regimes* (GitHub, ETH Zurich, 2020); https://github.com/ETHmodlab/virtual_libraries

Acknowledgements

This work was financially supported by the Novartis Forschungsstiftung (FreeNovation: AI in Drug Discovery), the Swiss National Science Foundation (grant no. 205321_182176 to G.S.) and the RETHINK initiative at ETH Zurich.

Author contributions

M.M. and L.F. contributed equally to this work. M.M. and L.F. designed the overall computational workflow. M.M. implemented the workflow and the open-access software release. L.F. performed the scaffold and descriptor analysis. All authors contributed to the study design, analysed the data and jointly wrote the manuscript.

Competing interests

G.S. declares a potential financial conflict of interest as a consultant to the pharmaceutical industry and co-founder of inSili.com GmbH, Zurich. No other potential conflicts of interest are declared.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-020-0160-y.

Correspondence and requests for materials should be addressed to G.S.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.