Generating customized compound libraries for drug discovery with machine intelligence

Michael Moret[1], Lukas Friedrich[1], Francesca Grisoni[1], Daniel Merk[1,2] & Gisbert Schneider[1]*

[1]ETH Zurich, Department of Chemistry and Applied Biosciences, RETHINK, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland; [2]Goethe University Frankfurt, Institute of Pharmaceutical Chemistry, Max-von-Laue-Str. 9, 60438 Frankfurt, Germany; *Correspondence to Gisbert Schneider, Email: gisbert@ethz.ch

Generative machine learning models sample drug-like molecules from chemical space without the need for explicit design rules. A deep learning framework for customized compound library generation is presented, aiming to enrich and expand the pharmacologically relevant chemical space with new molecular entities 'on demand'. This de novo design approach was used to generate molecules that combine features from 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.

Innovative molecular design methods are needed to support medicinal chemistry by efficient sampling of untapped drug-like chemical space[1,2,3]. Recently, the field of drug design has adopted so-called generative deep learning models to construct new molecules with desired properties[4,5,6,7,8]. Deep learning methods represent a class of machine learning algorithms that learn directly from the input data and do not necessarily depend on rules coded by humans[9,10]. Some of these methods implement a language modeling approach[11], where an artificial neural network aims to learn the probability of a 'token' (*e.g.*, a word or a character) to appear in a sequence based on the distributions all previous tokens in a sequence[12]. Through this process, deep neural networks can learn the features of sequential data. Once trained, these models can generate novel sequences based on the sampled feature distributions.

The language modeling approach for de novo drug design uses the Simplified Molecular Input Line Entry Systems (SMILES) representation of molecules, which encodes molecular structure as a sequence of tokens[13]. Recent prospective applications have experimentally verified the potential of SMILES string-based generative de novo design of small molecules with the desired bioactivity[14,15]. An essential element of these prospective applications is transfer learning[16,17], 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 data with known bioactivities. In the second step, this general model is focused on a certain pharmacological target by performing transfer learning with small sets of molecules possessing the desired bioactivity.

Here, we present an open-access generative deep learning framework for creating virtual libraries of structurally novel and diverse molecules for project-tailored applications in drug discovery and related areas. The computational framework consists of an optimized chemical language model for designing new molecules that populate designated target areas in chemical space. We analyzed the suitability of chemical language models of both synthetic molecules and natural products to enrich libraries with desired characteristics (physicochemical properties, structural diversity and novelty) similar to those of screening compound libraries[18]. The results demonstrate the ability of this computational approach to generate innovative molecules that are focused on a specific target area of the chemical space, *e.g.*, by targeting a specific bioactivity or by enriching sets of structurally diverse de novo generated molecules with natural product characteristics.

Results and Discussion

Generating molecules with a chemical language model

To develop a language model of the chemical constitution of biologically active molecules, a training dataset was compiled from ChEMBL24[19]. Bioactive compounds with annotated bioactivities (IC_{50}, EC_{50}, K_d, K_i) <1 μM were extracted from this chemical database and standardized, resulting in a set of ~365k molecules. Each training molecule was presented to the chemical language model as a one-hot vector encoding, i.e., 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 (equal to 71 in this study). During model training, the chemical language

model learns the conditional probability distribution of a token with respect to all the preceding tokens in the SMILES string (Fig. 1b). To optimize the applicability of the chemical language model to small training data sets, two different strategies were investigated, namely, data augmentation and temperature sampling.

Data augmentation

The amount and quality of the training data are key ingredients to successful language modeling[20]. Using multiple representations of the same entity (data augmentation) is one of the strategies for deep learning to work in a small data regime and obtain generalizing models, i.e., to have a chemically meaningful understanding of the training data[21,22]. To apply data augmentation, we leveraged the nonunivocal property of SMILES string; multiple valid SMILES strings representing the same molecular graph can be obtained by starting the string from any nonhydrogen atom in a molecule[23] (Fig. 2a). We compared the effect of model training with two augmentation levels (10-fold and 20-fold) on the generated SMILES strings in terms of (1) validity, i.e., percentages of SMILES strings that can be translated back to molecular graphs; (2) uniqueness, which is the percentage of nonduplicated SMILES strings; and (3) novelty calculated as 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 the SMILES strings generation is nonredundant. A high degree of *novelty* suggests that the model is suitable for de novo molecule design.

For each augmentation level, the chemical language model was trained for 10 epochs, with an epoch meaning one pass over all of the training data (Fig. 1c). We observed that augmenting the training data was beneficial in terms of all indices compared to the non-augmented scenario (Table 1). However, 20-fold augmentation did not further improve the results obtained with 10-fold augmentation (Table 1).

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 (Eq. 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, so 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 more 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 chemical language model as indicated by the increased levels of validity, uniqueness and novelty of sampled molecules (Table 1). In subsequent experiments, the model trained with 10-fold data augmentation and T = 0.7 was used for generating application-focused libraries.

Generating compound libraries with transfer learning

Building on the general optimization results of the chemical language model, we investigated the potential of transfer learning to create novel and diverse virtual compound libraries for drug discovery. To enrich sets of generated molecules with features relevant for drug discovery[24], 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 ("target space") defined by natural products from the manually curated natural product screening library MEGx (Analyticon Discovery GmbH, Potsdam, Germany).

Generating application-focused compound libraries

As an example of building an application-focused compound library by 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[25] computed on Morgan fingerprints[26] (similarity higher than 0.78). These five natural products were used for transfer learning.

To estimate the coverage of the chemical space during transfer learning, we computed the Fréchet ChemNet Distance (FCD), a distance metric to evaluate the similarity between two populations of molecules based on chemical structure and bioactivity[27]. An FCD value of 0 indicates that the compared molecular spaces are identical, while 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 chemical language models are able to sample

the chemical space between a source and a target space in a continuous fashion although molecules are discrete entities.

During the initial epochs of transfer learning (epochs one to six) the distances of the generated molecules to the target 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 target molecules possibly because ChEMBL24 contains natural products, and many synthetic molecules are natural product-inspired compounds[28]. The increasing distance to the natural product target space during transfer learning might seem initially counterintuitive. A likely explanation for this increasing distance to the natural product target space is the limited size and diversity of the set of five natural products used for transfer learning compared to the whole natural product space.

To highlight the changes of physicochemical properties during transfer learning, we selected the fraction of sp^3-hybridized carbon atoms (Fsp3) as an illustrative example since Fsp3 values typically differ between synthetic and natural compounds[29]. During transfer learning, the Fsp3 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[30], UMAP (Uniform Manifold Approximation and Projection[31]) plots were generated. UMAP creates a two-dimensional representation of high-dimensional data distributions (here: molecules represented as Morgan fingerprints), in which the similarity relations between data points in the original high-dimensional space are largely preserved[31,32]. In this visualization, the molecules sampled from the pretrained chemical language model (light blue) are close to the training data (dark blue), and the molecules are shifted toward 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 target.

We further assessed the coverage of chemical space and the diversity of the generated molecules by analyzing their atom scaffolds (Bemis-Murcko scaffolds)[33]. We examined the five most frequent scaffolds of sampled molecules before, i.e., using the pretrained chemical language model, and during transfer learning (Fig. 4). As a measure of scaffold diversity, we determined the Shannon entropy scaled by the number of investigated scaffolds[34] (scaled Shannon entropy, SSE, Eq. 2). SSE quantitatively reflects the structural diversity of a given set

of scaffolds. Here, SSE = 1 indicates maximum diversity, whereas SSE = 0 indicates full conservation 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, i.e., scaffolds occurring only once in a population, also decreased (Table 2, Supporting 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 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, Supporting Information). Compared to the Enamine compound set (700M drug-like compounds) as an example of a screening compound collection, almost all generated molecules (>99%) were new, and the proportion of new scaffolds among the generated molecules increased from 75% to over 95% during transfer learning (Table 2, Supporting Information).

Generating virtual libraries by expanding chemical space

Having demonstrated the ability of the chemical language model 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 target space. Here, the transfer learning set contained molecule **1** and four dissimilar natural products (**6**–**9,** Fig. 5a) to increase the diversity of the fine-tuning 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). While the distance to the target 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 covering a representative portion of the target space (Fig. 5b,c). In contrast, transfer learning with five similar molecules resulted in the generation of molecules exclusively 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.

The five most frequent scaffolds represented only a small fraction of all generated molecules compared to the analysis with five similar natural products. 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 a novel scaffold compared to the source and target 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 to generate structurally diverse molecules, comprising a broad range of scaffolds and possessing properties of the target space, *e.g.*, an enriched fraction of sp^3 hybridized carbon atoms. This approach could help enrich screening compound collections with innovative compounds and scaffolds for virtual and real high-throughput screening.

Conclusions

Generative deep learning proved applicable to computer-based compound library design for use in medicinal chemistry. The results demonstrate that chemical language models combined with transfer learning support the discovery of new molecular architecture for drug design. Chemical language models proved able to navigate through chemical space using the SMILES molecular representation. By relying on the chemical similarity principle[35,36] and natural products as starting points for drug design, this computational approach successfully generated novel and chemically diverse molecular entities. This pretrained chemical language model is publicly accessible to enable experimentation along with the analysis framework and an interactive map to encourage researchers to apply transfer learning on custom sets of molecules for own 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 for early-stage drug discovery.

Methods

Training compounds and data processing. Compounds with an annotated activity values (IC_{50}, EC_{50}, K_d, K_i) <1 μM (p*Activity* ≥6) were retrieved from ChEMBL24 to cover the chemical space of biologically active compounds. Molecular structures were encoded as canonical SMILES strings[37] with the RDKit package (v2018.03, www.rdkit.org), and only SMILES strings with a length of up to 140 tokens (characters) were retained. SMILES strings of 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 (release date 01.09.2018, Analyticon Discovery GmbH, Potsdam, Germany). To focus on structural features of the central scaffolds of these natural products, all existing carbohydrate moieties were removed by substructure filtering using DataWarrior software (www.openmolecules.org/datawarrior/, 5.0.0)[38]; 2931 molecules were retained. To assess pairwise similarities, all molecules were represented as bit vectors according to the Morgan fingerprint algorithm[26]. Morgan fingerprints numerically encode the presence of radial molecular fragments (length = 2048, radius = 2) as implemented in RDKit (version 2018.03). 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 (TC 0.78 to 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).

Chemical language model implementation. All software programs were implemented in Python (v3.6.5) using Keras (v2.2.0, https://keras.io/) with the TensorFlow GPU backend (v1.9.0, www.tensorflow.org). The chemical language model was implemented as a recurrent neural network with long short-term memory cells (LSTM)[39]. The neural network was composed of four layers having a total of 5,820,515 parameters (layer 1: BatchNormalization, layer 2: LSTM with 1024 units, layer 3: LSTM with 256 units, layer 4: BatchNormalization) and was trained with SMILES strings encoded as one-hot vectors. The Adam optimizer[40] with a learning rate of 0.001 was used for training the chemical language model training (10 epochs, where one

epoch is defined as one pass over all the training data, took ~21 hours on the processed ChEMBL24 data with a 10-fold data augmentation on a single Nvidia GTX 1080 GPU with 256 GB of memory). Each training run was repeated ten times, and 5000 SMILES strings were sampled after each epoch for a total of 10 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 (10^{-4}). Each transfer learning experiment was repeated ten times, and 10000 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 with the softmax function parameterized by the sampling temperature. The probability of the *i*-th token to be sampled from chemical language model predictions was computed as (Eq. 1):

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

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

Data augmentation. Data augmentation was performed with RDKit (v2018.03.3.0) using multiple SMILES string representations of the same molecule. We defined two augmentation levels (10-fold and 20-fold), where a 10-fold augmentation is defined as the SMILES string canonicalized version with 10 alternative representations.

Fréchet ChemNet distance. The Fréchet ChemNet Distance (FCD) was computed by comparing the Fréchet distance[41] 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)[27]. In total, 5000 molecules were randomly selected from each compound set for FCD calculation when possible. A minimum of 2931 molecules was used to compute the FCD to MEGx (i.e., the number of compounds available from MEGx after initial data processing).

Normalized Shannon entropy. We used the following equation to compute the Normalized Shannon entropy (SSE)[34]:

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

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

Code and data availability. The computational framework and the data are available on GitHub at URL: 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

MM and LF contributed equally to this work. MM and LF designed the overall computational workflow. MM implemented the workflow and the open-access software release. LF performed the scaffold and descriptor analysis. All authors contributed to the study design, analyzed 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.

Figure 1 (color)

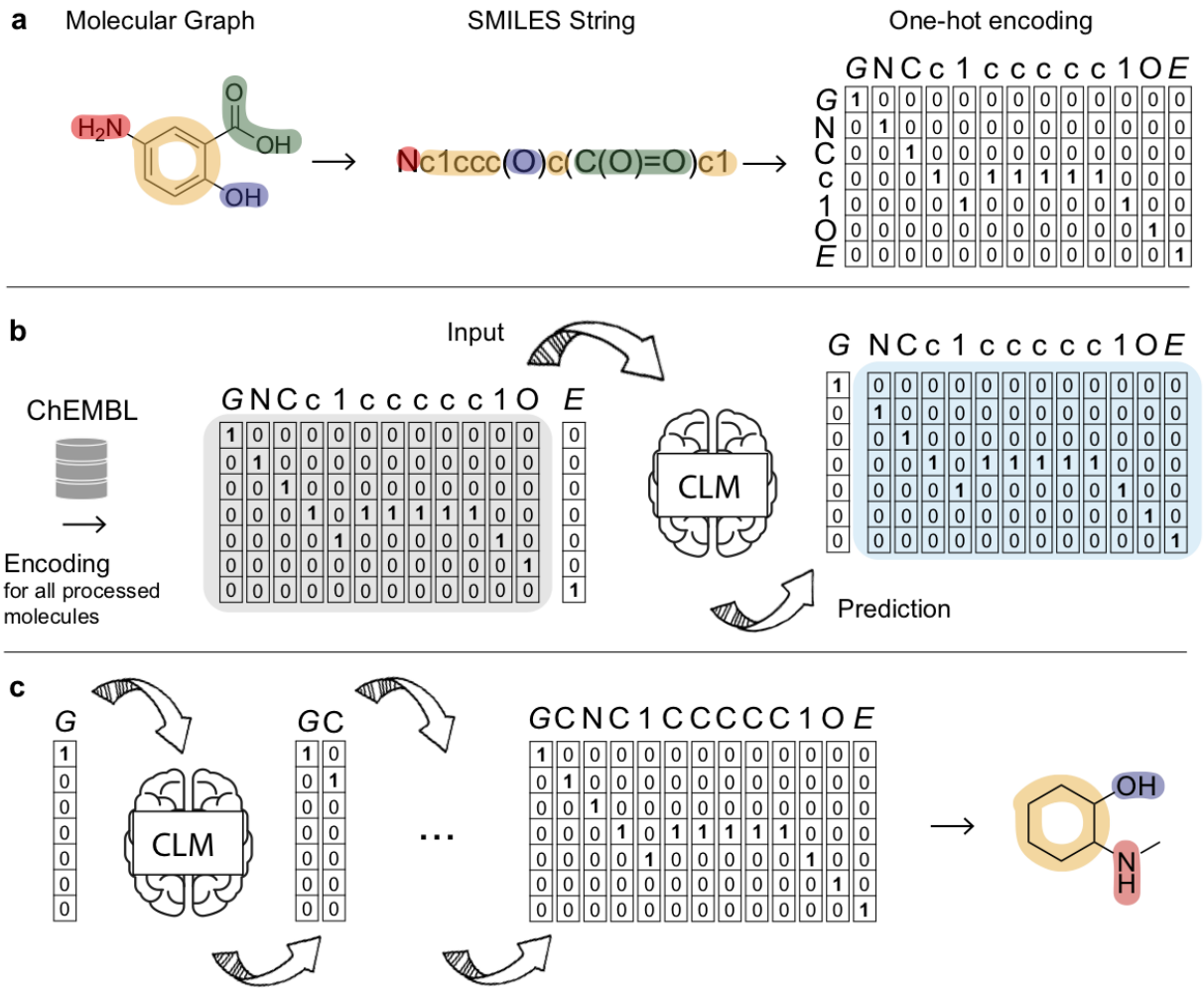

Figure 1. Chemical language model (CLM) training and sampling of new molecules. a, Each molecule is translated to its canonical SMILES string notation from its molecular graph. Combined with a start token ("G") and an end token ("E"), SMILES strings are presented as input to the chemical language model using one-hot encoding. **b,** The chemical language model 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 chemical language model repeatedly samples tokens from the learned distribution until the end token is sampled, indicating the completion of a new SMILES string.

Figure 2 (b/w)

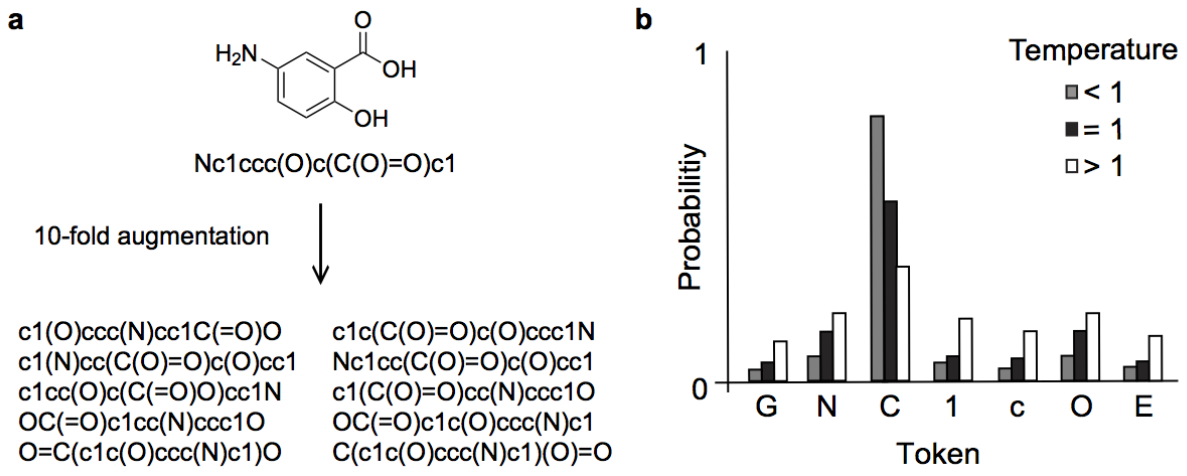

Figure 2. Data augmentation and temperature sampling. a, Example of 10-fold data augmentation. All SMILES strings represent the same molecular graph. **b,** Effect of the sampling temperature *T* on the conditional probability distribution over the SMILES string vocabulary for selected tokens (G, N, C, 1, c, O, E). *T* = 1 represents the probability distribution the chemical language model learned during training. *T* <1 sharpens the distribution. *T* >1 flattens the distribution.

Figure 3 (color)

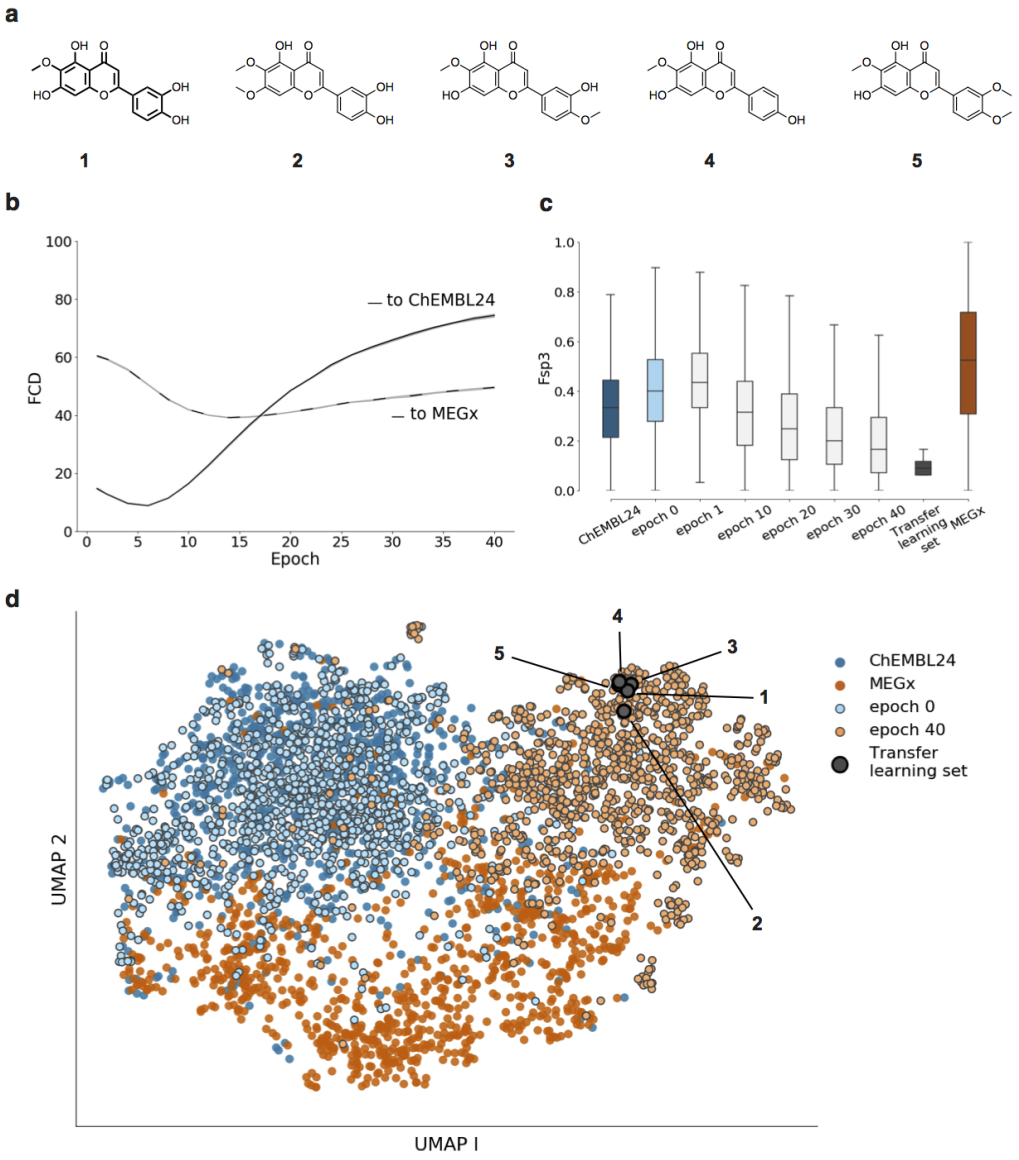

Figure 3. Chemical space navigation by transfer learning with five similar molecules. a, The *Transfer Learning Set* consisted of five structurally similar natural products (**1–5**) from the natural product collection MEGx. **b,** Fréchet ChemNet Distance (FCD) to ChEMBL24 and MEGx, respectively, of generated molecules during chemical space navigation. Mean and 95% confidence interval for ten repeats are shown in shaded area. **c,** Evolution of the fraction of sp^3-hybridized carbon atoms (Fsp3) during chemical space navigation. **d,** UMAP plot of molecules. For each group, 1000 molecules were randomly selected. Dark blue: ChEMBL24. Dark orange: MEGx. Light blue: molecules generated from the pretrained model (training epoch 0). Light orange: molecules generated at epoch 40. Gray circles: transfer learning set (a).

Figure 4 (color)

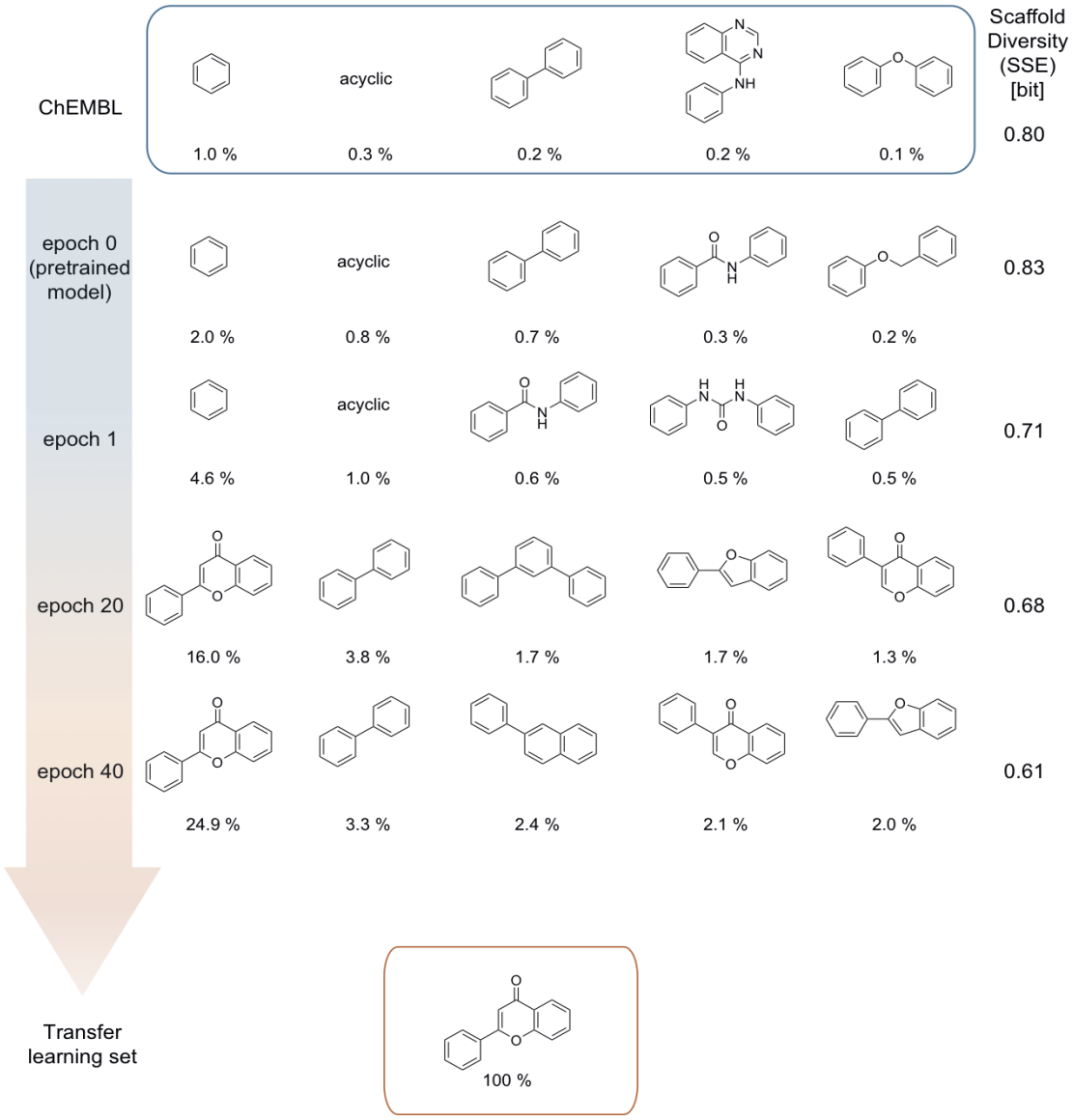

Figure 4. The five most frequent scaffolds from different training epochs during chemical space navigation to de novo generated focused compound libraries. Percentage indicates the fraction of molecules containing the respective scaffold. The scaffold diversity of the five most frequent scaffolds was quantified by Scaled Shannon Entropy (SSE, values from [0,1]) (Eq. 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.

Figure 5 (color)

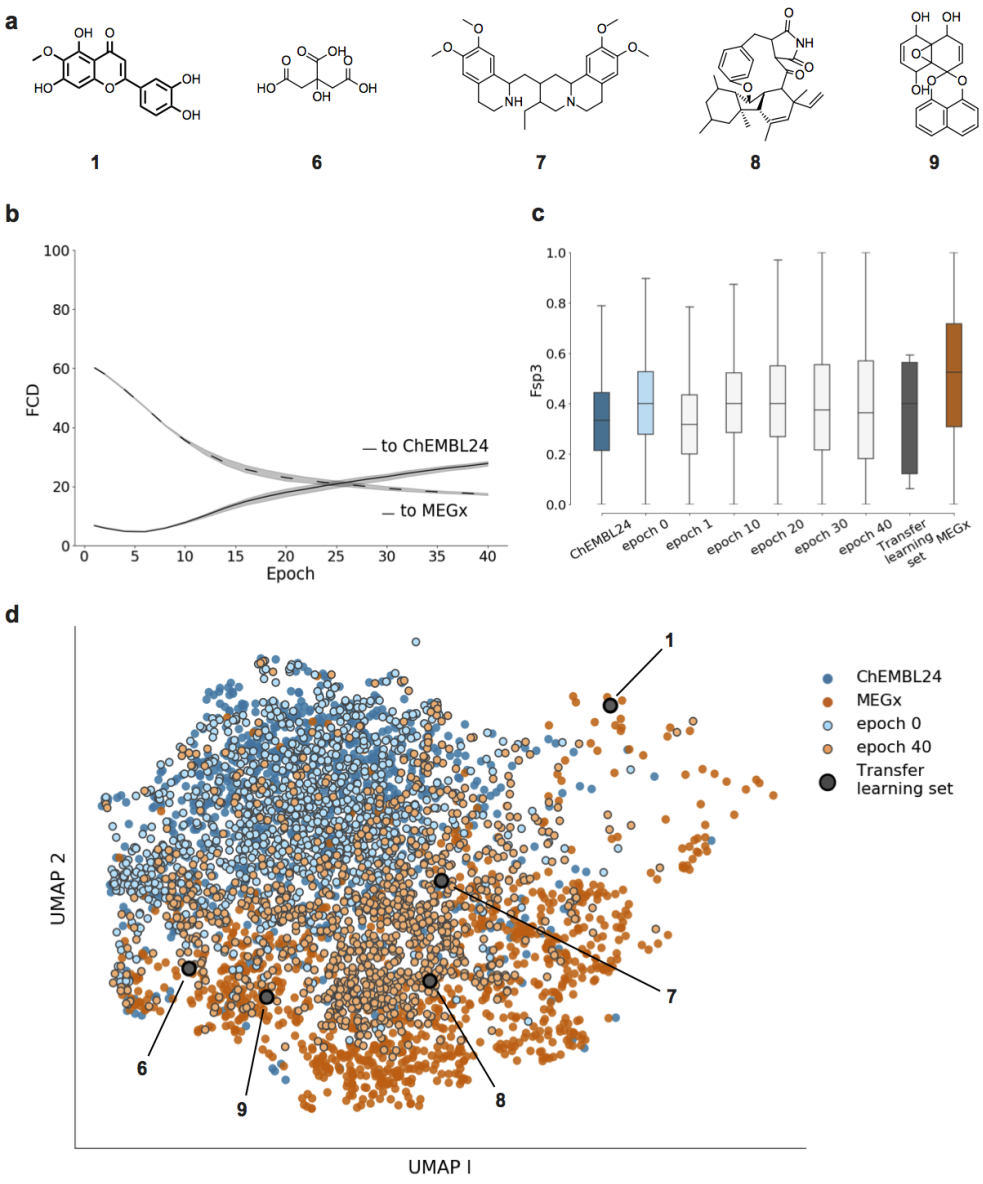

Figure 5. Chemical space navigation by transfer learning with five dissimilar molecules. **a,** Five dissimilar natural products (**1, 6**–**9**) from the MEGx collection. **b,** Fréchet ChemNet Distance (FCD) to ChEMBL24 and MEGx of generated molecules. The mean value and 95% confidence interval (shaded area) for ten repeats are shown. **c,** Evolution of the fraction of sp^3-hybridized carbon atoms (Fsp3) during transfer learning. **d,** UMAP plots of molecule distributions. In total, 1000 molecules were randomly selected from each set. Dark blue: ChEMBL24. Dark orange: MEGx. Light blue: molecules generated with the pretrained model (epoch = 0). Light orange: molecules generated at epoch 40. Gray circles: transfer learning set (a).

Figure 6 (color)

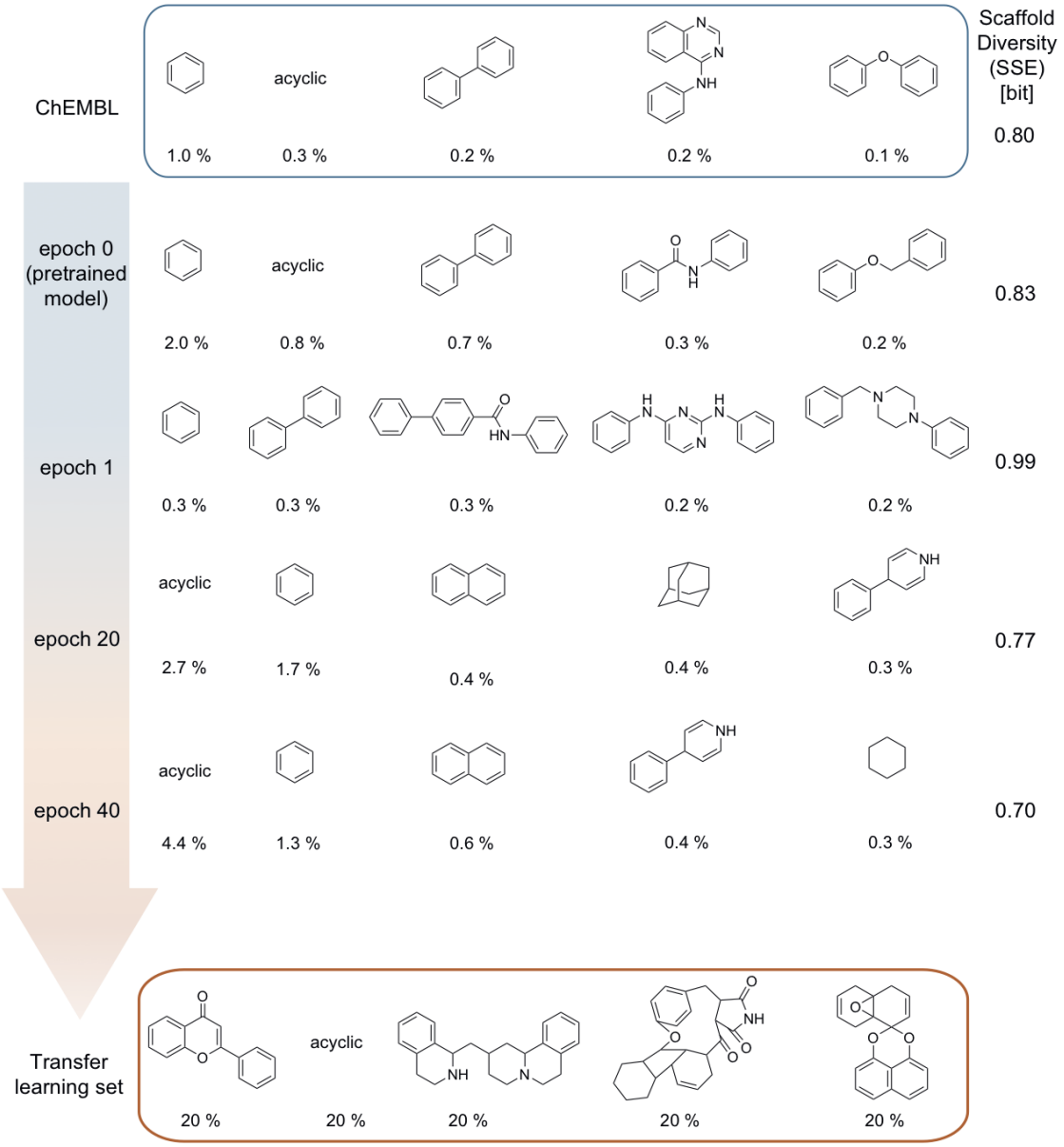

Figure 6. The five most frequent scaffolds from different training epochs during chemical space navigation. Percentage indicates the fraction of molecules containing the scaffold (epoch 0: sampled from the pretrained model.) The diversity of the five most frequent scaffolds is expressed in terms of the Scaled Shannon Entropy (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.

Tables

Table 1. Validity, uniqueness and novelty of molecules depending on data augmentation and sampling temperature. Each experiment was run for 10 epochs and repeated 10 times. After each epoch, 5000 molecules were sampled. The "best epoch" was defined as the one yielding the highest average novelty value. Percentages are reported with respect to the total number of molecules sampled.

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
10-fold	1.0	7	93.5 ± 0.8	93.5 ± 0.8	92.3 ± 0.6
20-fold	1.0	6	94.2 ± 0.4	94.2 ± 0.4	92.6 ± 0.4
10-fold	0.2	7	85.8 ± 6.3	54.9 ± 6.4	52.0 ± 7.2
10-fold	0.7	4	97.4 ± 0.8	97.3 ± 0.8	93.8 ± 1.0
10-fold	1.2	4	84.8 ± 3.4	84.8 ± 3.4	84.4 ± 3.3

Table 2. Scaffold analysis during the transfer learning with five similar and five dissimilar natural products. 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. The fraction of novel scaffolds was calculated by comparing with scaffolds contained in the training data set (ChEMBL24) and the natural product set (MEGx).

Source	Novel Molecules (M)	Scaffolds (N)	Singleton Scaffolds (Ns)	$P_{scaffolds}$ N/M	$P_{singletons}$ Ns/N	Novel % (ChEMBL24 | MEGx)
General base model without transfer learning						
Epoch 0	9567	8296	7844	0.87	0.95	99 | 99
Transfer learning with five similar compounds						
Epoch 1	9035	6622	5983	0.73	0.90	83 | 98
Epoch 20	6543	2522	1998	0.39	0.79	79 | 92
Epoch 40	3373	891	660	0.26	0.74	75 | 85
Transfer learning with five dissimilar compounds						
Epoch 1	9459	8187	7573	0.87	0.93	86 | 99
Epoch 20	8702	7581	7288	0.87	0.96	92 | 97
Epoch 40	8184	7140	6917	0.87	0.97	94 | 97
MEGx	2931	1159	797	0.40	0.69	*n.a.*
ChEMBL24	365,063	135,120	93,174	0.37	0.69	*n.a.*

n.a., not applicable.

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

Supplementary Information

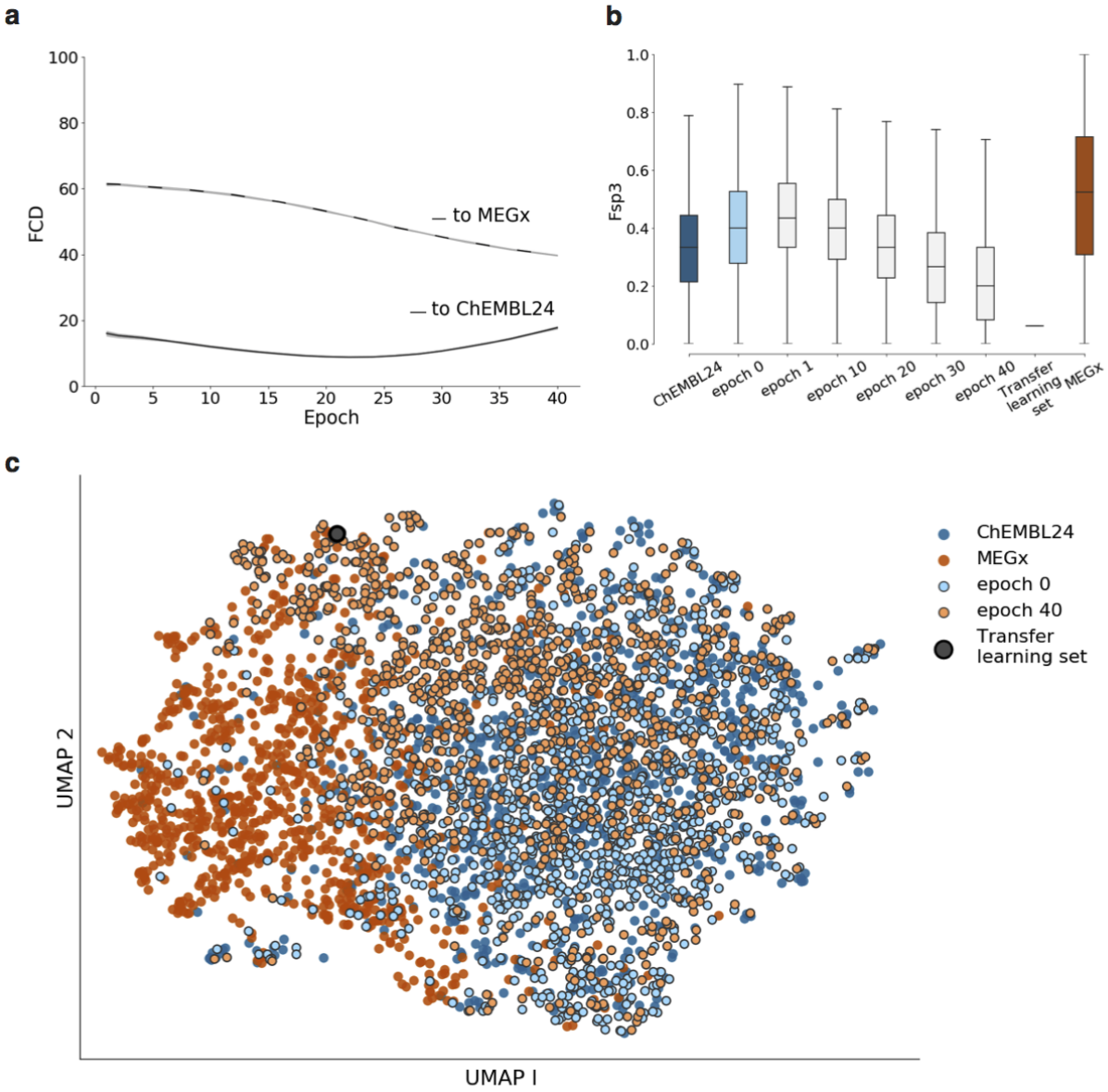

Figure S1. Chemical space navigation by transfer learning with one molecule. **a,** Fréchet ChemNet Distance (FCD) to ChEMBL24 and MEGx of generated molecules during chemical space navigation. Mean and 0.95 confidence interval for ten repeats are shown in shaded area. **b,** Evolution of the fraction of sp^3-hybridized carbon atoms (Fsp3) during the chemical space navigation. **c,** UMAP plot of molecules. For each group, 1k molecules were randomly selected. Dark blue: ChEMBL24. Dark orange: MEGx. Light blue: molecules generated from the pretrained model (i.e., epoch zero). Light orange: molecules generated at epoch 40. Gray circles: transfer learning set.

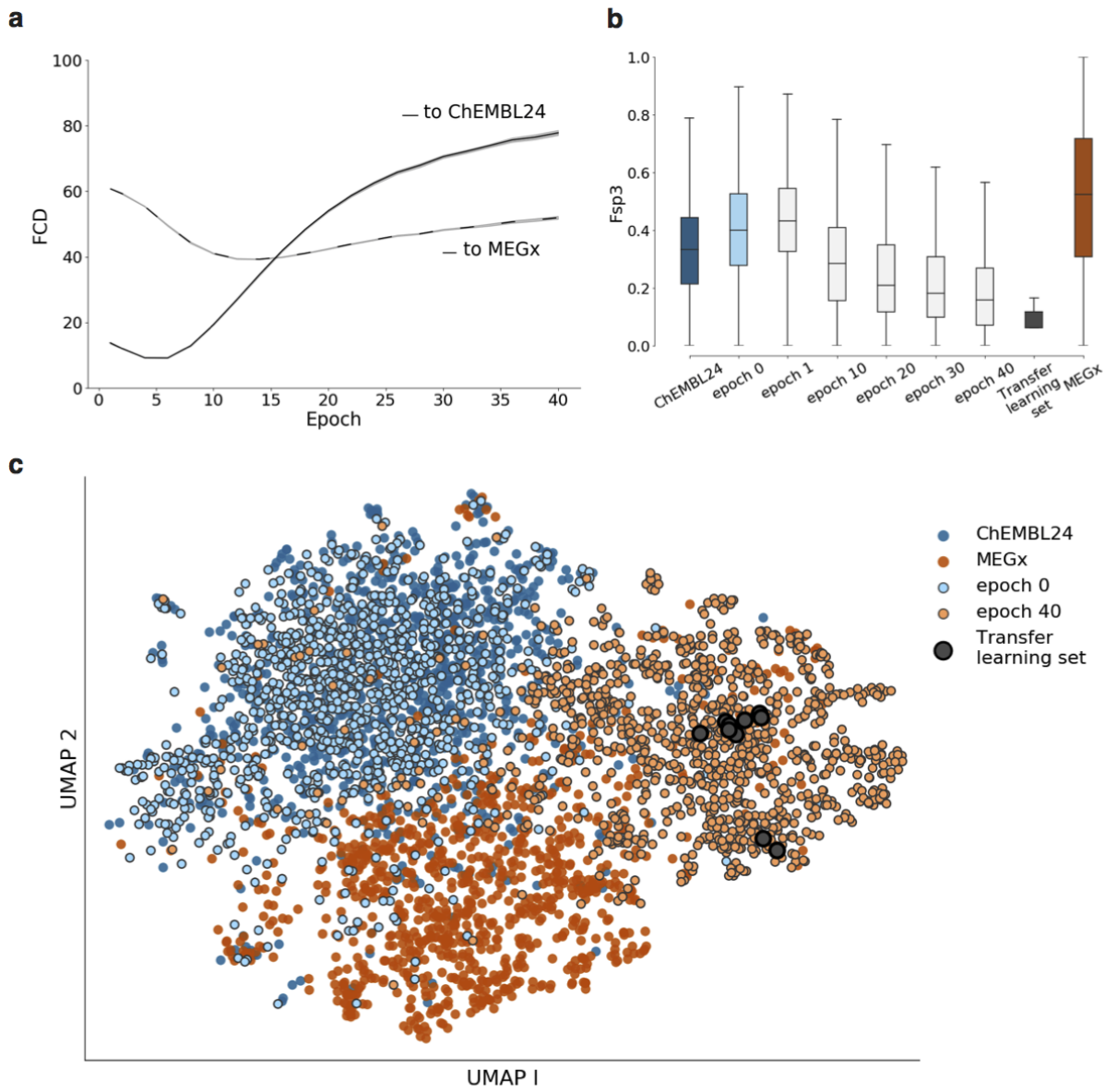

Figure S2. Chemical space navigation by transfer learning with 10 similar molecules. **a,** Fréchet ChemNet Distance (FCD) to ChEMBL24 and MEGx of generated molecules during chemical space navigation. Mean and 0.95 confidence interval for ten repeats are shown in shaded area. **b,** Evolution of the fraction of sp3-hybridized carbon atoms (Fsp3) during the chemical space navigation. **c,** UMAP plot of molecules. For each group, 1k molecules were randomly selected. Dark blue: ChEMBL24. Dark orange: MEGx. Light blue: molecules generated from the pretrained model (i.e., epoch zero). Light orange: molecules generated at epoch 40. Gray circles: transfer learning set.

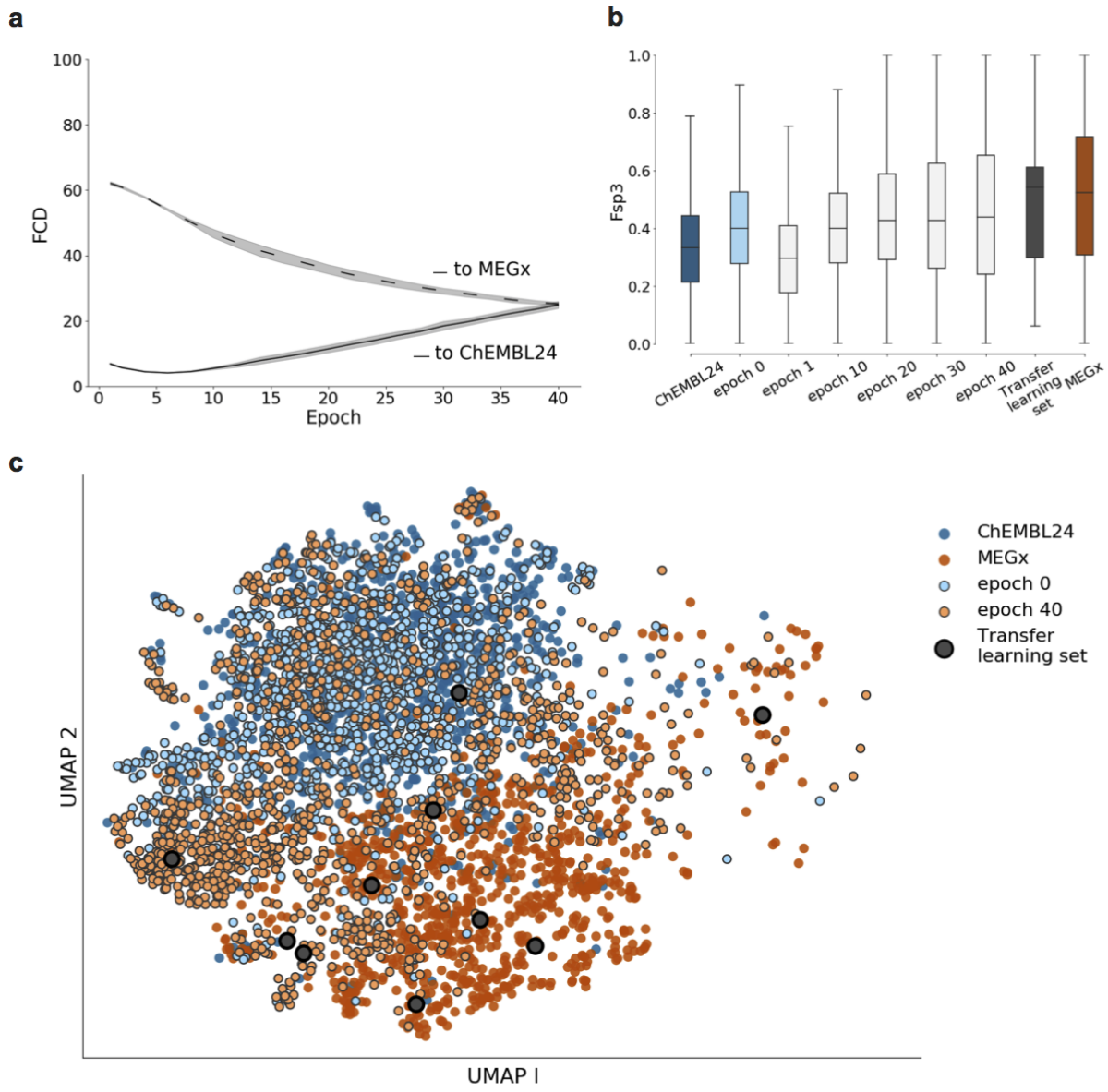

Figure S3. Chemical space navigation by transfer learning with 10 dissimilar molecules. **a,** Fréchet ChemNet Distance (FCD) to ChEMBL24 and MEGx of generated molecules during chemical space navigation. Mean and 0.95 confidence interval for ten repeats are shown in shaded area. **b,** Evolution of the fraction of sp^3-hybridized carbon atoms (Fsp3) during the chemical space navigation. **c,** UMAP plot of molecules. For each group, 1k molecules were randomly selected. Dark blue: ChEMBL24. Dark orange: MEGx. Light blue: molecules generated from the pretrained model (i.e., epoch zero). Light orange: molecules generated at epoch 40. Gray circles: transfer learning set.

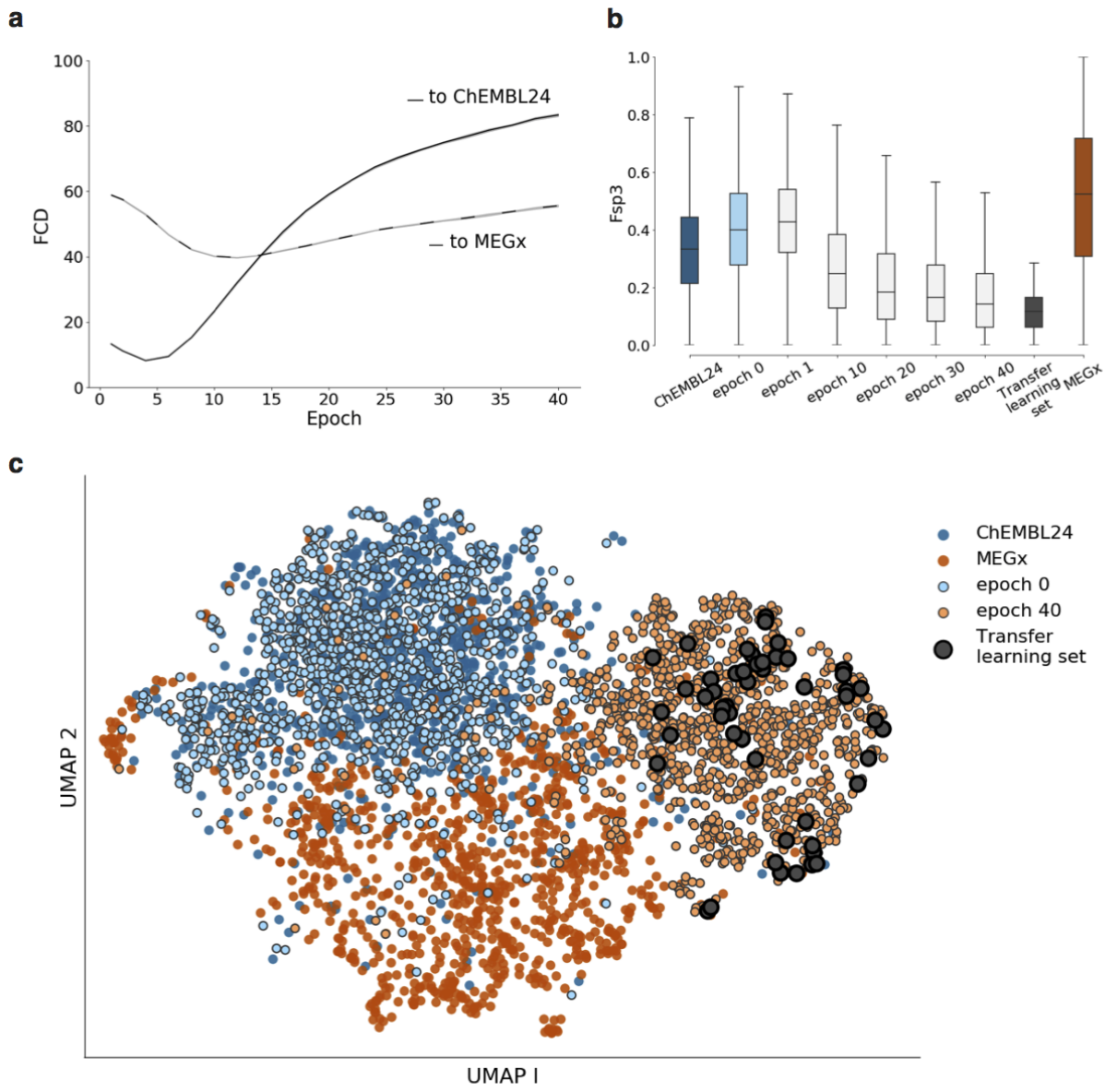

Figure S4. Chemical space navigation by transfer learning with 50 similar molecules. **a,** Fréchet ChemNet Distance (FCD) to ChEMBL24 and MEGx of generated molecules during chemical space navigation. Mean and 0.95 confidence interval for ten repeats are shown in shaded area. **b,** Evolution of the fraction of sp3-hybridized carbon atoms (Fsp3) during the chemical space navigation. **c,** UMAP plot of molecules. For each group, 1k molecules were randomly selected. Dark blue: ChEMBL24. Dark orange: MEGx. Light blue: molecules generated from the pretrained model (i.e., epoch zero). Light orange: molecules generated at epoch 40. Gray circles: transfer learning set.

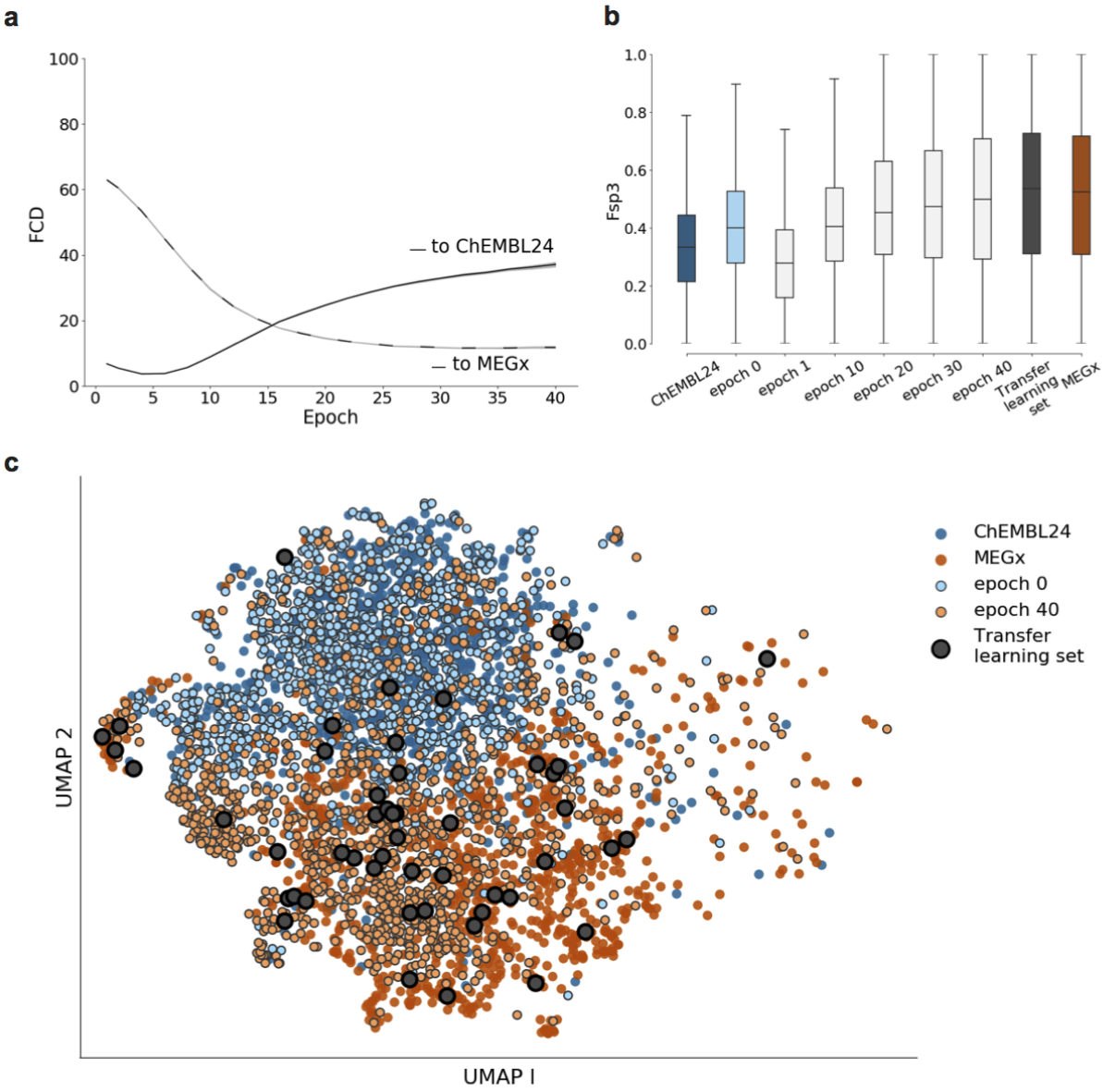

Figure S5. Chemical space navigation by transfer learning with 50 dissimilar molecules. **a,** Fréchet ChemNet Distance (FCD) to ChEMBL24 and MEGx of generated molecules during chemical space navigation. Mean and 0.95 confidence interval for ten repeats are shown in shaded area. **b,** Evolution of the fraction of sp^3-hybridized carbon atoms (Fsp3) during the chemical space navigation. **c,** UMAP plot of molecules. For each group, 1k molecules were randomly selected. Dark blue: ChEMBL24. Dark orange: MEGx. Light blue: molecules generated from the pretrained model (i.e., epoch zero). Light orange: molecules generated at epoch 40. Gray circles: transfer learning set.

References

Note: Several references point to non–peer–reviewed texts and preprints. These are cited to account for the actuality of the topic of this article.

[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] 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).
[5] Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **9**, 48 (2017).
[6] Putin, E. *et al.* Reinforced adversarial neural computer for de novo molecular design. *J. Chem. Inf. Model.* **58**, 1194–1204 (2018).
[7] Gómez-Bombarelli, R. *et al*. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
[8] 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).
[9] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
[10] Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **61**, 85–117 (2015).
[11] 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).
[12] Bengio, Y., Ducharme, R., Vincent, P. & Jauvin, C. A neural probabilistic language model. *J. Mach. Learn. Res.* **3**, 1137–1155 (2003).
[13] 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).
[14] 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).
[15] Merk, D., Friedrich, L., Grisoni, F. & Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Mol. Inf.* **37**, 1700153 (2018).
[16] Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? *Preprint at http://arxiv.org/abs/1411.1792* (2014).
[17] Peters, M., Ruder, S. & Smith, N. A. To tune or not to tune? Adapting pretrained representations to diverse tasks. *Preprint at http://arxiv.org/abs/1903.05987* (2019).
[18] 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).
[19] Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2016).

[20] 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).
[21] Simard, P., Victorri, B., LeCun, Y. & Denker, J. Tangent Prop - A formalism for specifying selected invariances in an adaptive network. In *Advances in Neural Information Processing Systems* **4**, pp. 895–903 (1992).
[22] Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *in Advances in Neural Information Processing Systems* 1097–1105 (2012).
[23] Bjerrum, E. J. SMILES enumeration as data augmentation for neural network modeling of molecules. *Preprint at https://arxiv.org/abs/1703.07076* (2017).
[24] Rodrigues, T., Reker, D., Schneider, P. & Schneider, G. Counting on natural products for drug design. *Nat. Chem.* **8**, 531–541 (2016).
[25] Tanimoto, T. T. *An Elementary Mathematical Theory of Classification and Prediction.* (International Business Machines Corporation, New York, NY, USA, 1958).
[26] Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
[27] 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).
[28] Boufridi, A. & Quinn, R. J. Harnessing the properties of natural products. *Annu. Rev. Pharmacol. Toxicol.* **58**, 451–470 (2018).
[29] 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).
[30] Reutlinger, M. & Schneider, G. Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *J. Mol. Graph. Model.* **34**, 108–117 (2012).
[31] McInnes, L. & Healy, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *Preprint at http://arxiv.org/abs/1802.03426v1* (2018).
[32] Becht, E. *et al*. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol. 37*, **38** (2018).
[33] Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).
[34] 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).
[35] Johnson, M. A. & Maggiora, G. M. *Concepts and Applications of Molecular Similarity.* (John Wiley & Sons, New York, NY, USA, 1990).
[36] Maggiora, G. M. & Bajorath, J. Chemical space networks: a powerful new paradigm for the description of chemical space. *J. Comput. Aided Mol. Des.* **28**, 795–802 (2014).
[37] 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).
[38] 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).
[39] Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).

[40] Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. *Preprint at http://arxiv.org/abs/1412.6980* (2014).
[41] Fréchet, M. Sur la distance de deux lois de probabilité. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences* **244**, 689–692 (1957).