Leveraging molecular structure and bioactivity with chemical language models for drug design

Michael Moret[1], Francesca Grisoni[1,2 *], Cyrill Brunner[1] & Gisbert Schneider[1,3 *]

[1]ETH Zurich, Department of Chemistry and Applied Biosciences, RETHINK, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland;
[2]Eindhoven University of Technology, Institute for Complex Molecular Systems, Department of Biomedical Engineering, Groene Loper 7, 5612AZ Eindhoven, Netherlands;
[3]ETH Singapore SEC Ltd, 1 CREATE Way, #06-01 CREATE Tower, Singapore 138602, Singapore;
*Correspondence to Gisbert Schneider (gisbert@ethz.ch) and Francesca Grisoni (f.grisoni@tue.nl)

Abstract

Generative chemical language models (CLMs) can be used for de novo molecular structure generation. These CLMs learn from the structural information of known molecules to generate new ones. In this paper, we show that “hybrid” CLMs can additionally leverage the bioactivity information available for the training compounds. To computationally design ligands of phosphoinositide 3-kinase gamma (PI3Kγ), we created a large collection of virtual molecules with a generative CLM. This primary virtual compound library was further refined using a CLM-based classifier for bioactivity prediction. This second hybrid CLM was pretrained with patented molecular structures and fine-tuned with known PI3Kγ binders and non-binders by transfer learning. Several of the computer-generated molecular designs were commercially available, which allowed for fast prescreening and preliminary experimental validation. A new PI3Kγ ligand with sub-micromolar activity was identified. The results positively advocate hybrid CLMs for virtual compound screening and activity-focused molecular design in low-data situations.

Introduction

Computational methods have become key players in hit and lead discovery in pharmaceutical research, complementing experimental high-throughput screening[1]. Bespoke virtual compound libraries provide access to untapped regions of the chemical space[2], thereby extending the diversity of potential drug candidates. However, owing to the potentially unlimited size of virtual chemical libraries, concerns have been raised over the pragmatism of successfully screening billions of molecules virtually with a potentially high risk of false positives[2,3]. To mitigate some of these challenges, researchers have employed generative deep learning models to construct compounds on demand by de novo design and to obtain small, bespoke virtual compound libraries[4,5]. A variety of data-driven approaches can be used to generate focused virtual chemical libraries and create molecules with the desired properties[5–18]. Chemical language models (CLMs) are based on deep learning networks for processing string representations of molecules (e.g., simplified molecular input line entry system (SMILES) strings; Fig. 1a)[5,7,19]. CLMs have already been successfully employed to generate focused virtual chemical libraries. Examples of de novo designed bioactive molecules include inhibitors of vascular endothelial growth factor receptor 2 kinase and the unfolded protein response pathway[7], and nuclear hormone receptor modulators[20–23].

The creation of a focused virtual chemical library with a CLM generally includes three basic steps: (i) model pretraining with a large set of molecules to learn the SMILES grammar and the feature distribution of the pretraining data, (ii) transfer learning with a smaller set of molecules (fine-tuning set) to bias the molecule generation by the CLM toward the chemical space of interest, and (iii) sampling of new molecules from the data distributions modeled in steps i) and ii)[5,24]. There are alternative approaches for CLM development, e.g., model fine-tuning (step ii) by reinforcement learning[6,25].

In this study, we developed a data-driven molecular design pipeline that leverages both the structural and bioactivity information of known ligands to generate de novo bespoke molecules. We pretrained two CLMs, each with a distinct pretraining strategy, on a large set of patented compound structures (one for molecular generation and one for classification). Both CLMs were fine-tuned on inhibitors of phosphoinositide 3-kinase gamma (PI3Kγ), which is an anticancer, anti-inflammatory, and immunomodulatory drug target[26,27]. For rapid validation, commercially available compounds from the set of de novo generated molecules were tested, as opposed to synthesizing them. A new nanomolar ligand of phosphoinositide 3-kinase gamma (PI3Kγ) was identified.

Results and Discussion

Molecular design and scoring were performed in two steps, each of which was executed by a distinct CLM: (i) molecular de novo design and (ii) refinement of the generated virtual molecule library using the available ligand bioactivity data for the target of interest (PI3Kγ).

Focused library generation

Chemical language model. A CLM based on a long short-term memory (LSTM) model and SMILES strings as input was developed for the de novo generation of a focused virtual chemical library for PI3Kγ[28]. To learn from unlabeled data, CLMs leverage “self-supervised” learning[29]. Specifically, the CLM was trained with an autoregressive approach, i.e., the process of iteratively predicting the next character in a SMILES string given all the previous characters in the string (Fig. 2a)[30]. In previous studies, CLMs were pretrained on molecules with known biological activity (IC_{50}, EC_{50}, K_d, and K_i) <1 μM retrieved from the ChEMBL database[20,23,31–33]. Although the training set can capture the general features of bioactive compounds, it does not necessarily represent the physicochemical properties of approved drugs. Here, to enable the CLM to capture features more related to approved drugs, we used 839,674 molecules from the US patent database for the CLM pretraining[34]. We hypothesized that patented compounds are more likely to become marketed drugs than the molecules deposited in ChEMBL. Transfer learning was performed to properly focus the pretrained CLM toward the target space of PI3Kγ ligands. For transfer learning, 46 PI3Kγ inhibitors with IC_{50} ≤100 nM were selected from the Drug Target Commons (DTC) database[35].

Nucleus sampling for molecule generation. CLMs generate new molecules by extending strings from a "start" character until the "stop" character is sampled or when reaching a preset maximum string length. String characters are iteratively added by weighted random sampling from the probability distribution learned by the CLM during training. The more likely a given character is at a given step according to the probabilities learned by the CLM, the more often it will be sampled, and vice versa. Narrowing the probabilities learned by the CLM with a parameter (the so-called temperature; Fig. 1b) generally improves the SMILES string sampling[31]. This improvement occurs in terms of (i) the quality of the SMILES strings generated, as reflected by their validity (grammatically valid SMILES strings), uniqueness (nonrepetitive molecules), and novelty (molecules not present in the pretraining and fine-tuning data), and (ii) the similarity of the sampled virtual chemical libraries to the reference data in terms of their chemical structures and bioactivities, as measured by the Fréchet ChemNet Distance (FCD)[36]. However, with this "temperature sampling" approach, SMILES characters are unlikely to be sampled, which could result in the construction of molecules that do not match the design objective. To prevent the CLM from picking unlikely SMILES characters by temperature sampling, we employed "nucleus sampling" here[37]. This method reflects the confidence of the model in its predictions by allowing only the most probable character(s) to be sampled using a probability threshold based on the cumulative probabilities of the SMILES characters (Fig. 1c).

Nucleus sampling improved upon temperature sampling in terms of lower FCD values (Fig. 1d), indicating a greater overall similarity of the de novo generated molecules to the pretraining set in terms of structural and bioactivity properties. During transfer learning, nucleus sampling generally improved the quality of the sampled molecules in terms of the novelty of the SMILES strings compared to the best temperature sampling data obtained (Fig. 1e)[33]. The results were stable over a range of sampling threshold values (Supplementary Table S1). However, nucleus sampling did not outperform temperature sampling in terms of the uniqueness, validity, and novelty of the SMILES strings generated after the pretraining (Supplementary Table S2). To create a PI3Kγ focused chemical library during transfer learning, we used nucleus sampling with a threshold of 0.85. A total of 5000 SMILES strings were sampled over 50 transfer learning epochs with 10 repetitions (5000 × 50 × 10). A total of 2,500,000 SMILES strings were generated, of which 1,121,735 were valid, unique, and novel compared to both the training and fine-tuning compounds.

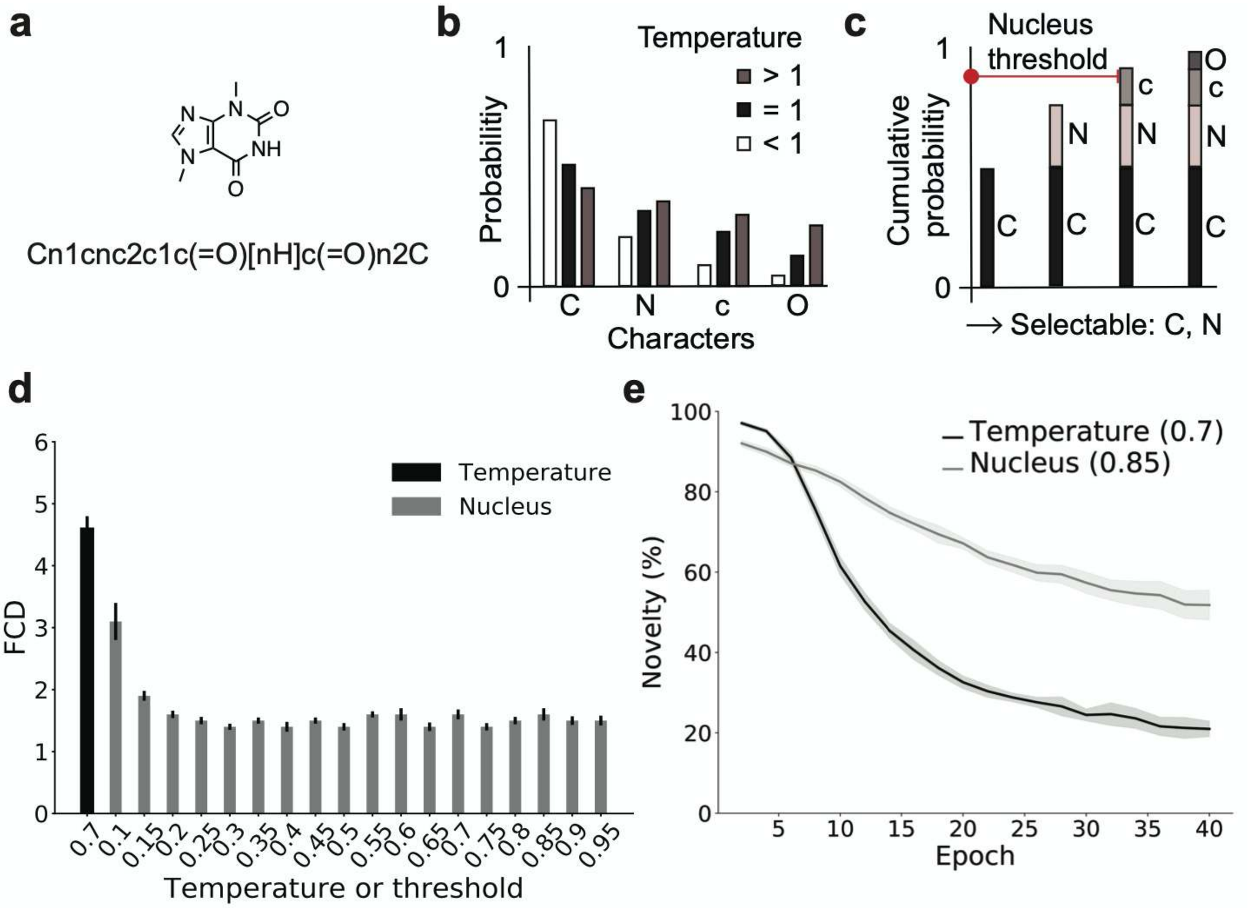

Fig. 1 | De novo molecular generation with the CLM. a, SMILES string representation of a molecule. **b,** Example of the effect of the temperature parameter on the probability distribution learnt by the CLM. **c,** Example of the effect of the nucleus sampling threshold. Only the characters N and C can be sampled here. **d,** Fréchet ChemNet Distance (FCD) comparison between temperature and nucleus sampling after the pretraining (reported as the mean with standard deviation over 10 repeats with 5000 molecules sampled per repeat). **e,** Comparison of the novelty of the generated SMILES strings during the transfer learning between temperature sampling (temperature = 0.7) and nucleus sampling (threshold = 0.85). Mean values (lines) and standard deviations (shaded areas) are shown for 10 repeats (1000 SMILES strings were sampled every second epoch over 40 epochs). Novelty is expressed as the percentage of SMILES strings generated that were valid and not included in either the training or the fine-tuning data.

Bioactivity prediction with a hybrid chemical language model

Leveraging bioactivity data for molecule selection. The availability of bioactivity data for the fine-tuning molecules permitted the training of a bioactivity prediction model to select the most promising de novo designs[38]. Classical chemoinformatics methods often rely on precomputed features (molecular descriptors), combined with a machine learning algorithm for molecular property prediction. In this study, we aimed to explore the potential of a SMILES string-based hybrid CLM to predict the bioactivity. This neural network model combines a generative CLM with a classifier network. Given that (i) inactive molecules were annotated with PI3Kγ pIC_{50} = 4.0 (Fig.

2c) and (ii) there is a natural ordering of the PI3Kγ ligands according to their pIC_{50} values, the bioactivity prediction task was framed as an ordinal classification task, i.e., classification with a class order[39]. Such a model considers both the active and inactive compounds for training and preserves both the class labels and the class order. For model training, we defined three class labels: "inactive" ($pIC_{50} \leq 4.0$, 34 molecules), "moderately active" ($4.0 < pIC_{50} \leq 6.5$, 121 molecules), and "highly active" ($pIC_{50} > 6.5$, 43 molecules). The CLM generated a focused virtual chemical library by leveraging the structural information of the molecules used for fine-tuning, while the classifier layer factored their activity labels into the model (Fig. 2d).

We explored two different pretraining strategies for feature learning with a large amount of unlabeled data.

1. *Autoregressive pretraining* (Fig. 2a). This strategy is analogous to the one performed for the generative CLM.
2. *ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) pretraining* (Fig. 2b)[40]. The ELECTRA approach is based on training a model to distinguish between "real" input characters and "corrupt" ones, which was previously shown to be useful for contextual representation of natural language[40]. We adapted ELECTRA for the CLM training with an LSTM model and SMILES strings as input[28]. The training data contained corrupted input SMILES strings generated by randomly substituting multiple characters with other characters of the SMILES language. The CLM was trained to spot the corrupted characters.

We hypothesized that, compared to autoregressive pretraining, ELECTRA pretraining has a more appropriate inductive bias (i.e., the set of algorithmic assumptions to solve a given task) to extract useful features for ordinal classification. The inductive bias of autoregressive pretraining is particularly suited for generating SMILES strings because the training and generative tasks are the same, namely, adding characters iteratively. However, ligands of the same macromolecular target tend to have similar chemical substructures, and, therefore, the ability of a model to distinguish small structural changes was deemed relevant. At the same time, small structural changes might lead to drastic variation of the biological activity (the so-called activity cliffs)[41]. Hereinafter, the model that was pretrained with the ELECTRA method is referred to as "E-CLM."

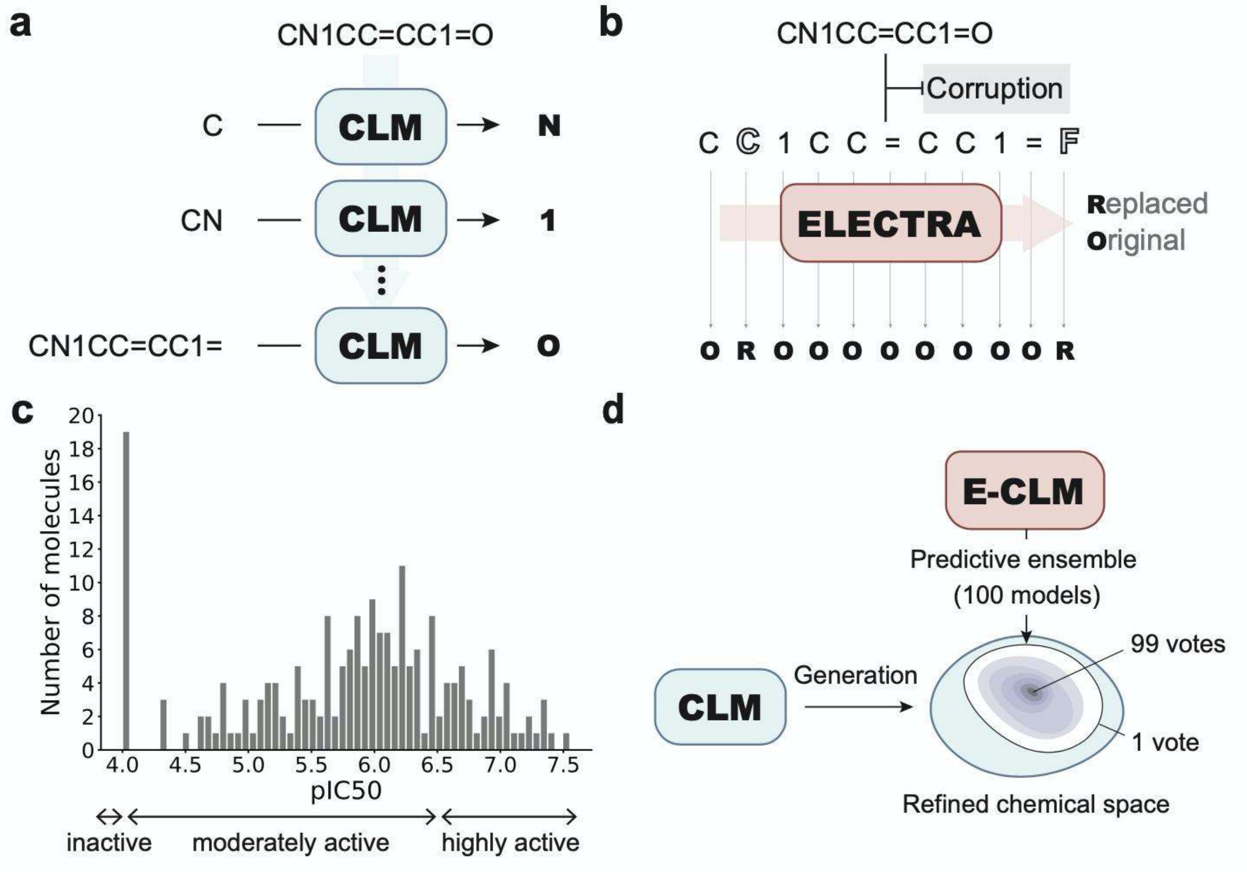

Fig. 2 | Bioactivity prediction. a, A CLM for molecule generation iteratively predicts the next character in a SMILES string given the preceding characters ("autoregressive" approach). **b,** An E-CLM (a CLM pretrained with the ELECTRA method) is trained on corrupted SMILES strings aiming to predict, for each string character, whether it is the original (correct) or a corrupted (substituted) character. **c,** Activity distribution of the PI3Kγ ligands. Compounds with annotated $pIC_{50} \leq 4.0$ were considered "inactive", and a pIC_{50} value of 6.5 was used to separate the "moderately active" from the "highly active" compounds. **d,** The molecular structures (in the form of a SMILES string) of the fine-tuning set were used to focus the CLM (pretrained on the US patent database) on the chemical space of the target of interest (PI3Kγ). To account for the uncertainty in the predictions, we employed an ensemble of 100 models to rank the generated molecules by the number of "votes".

To probe the effect of the pretraining scheme on the predictions, we added only a single feedforward layer to the pretrained CLM and E-CLM for bioactivity prediction. This additional network layer consisted of three neurons, one for each of the three bioactivity classes. It was added to fine-tune the entire network for bioactivity prediction[42,43]. To mitigate the class data imbalance, we applied oversampling to the classes with fewer data (i.e., the "inactive" and "highly active" classes)[44].

Overall, we found that the E-CLM performed better than the standard CLM for the task of identifying the most active molecules, while minimizing the number of inactive molecules misclassified as "highly active". For the chosen threshold (0.4), the E-CLM had a false positive rate of 10.0% compared to 46.7% for the CLM for the same true positive rate (71.3%) (Supplementary Figs. S2a and S3a). Fine-tuning of all neural network weights performed better than keeping the weights of one of the two layers constant (Supplementary Figs. S2a, S2c, and S2d). These results highlight the importance of choosing an appropriate pretraining method depending on the downstream task, e.g., data generation or classification.

Increasing the prediction confidence using deep ensemble learning. Deep learning models suffer from a decrease in performance when applied to out-of-domain data[45], a well-known issue in quantitative structure-activity relationship modeling[46,47]. To increase the confidence in the bioactivity predictions, we used a deep ensemble model by combining the predictions of multiple models with a majority voting approach[48,49]. Owing to the nondeterministic optimization process, repeats of the same CLM training procedure will lead to different models. Deep ensemble learning has been shown to perform well across different domains to account for the predictive uncertainty of the models, while having the benefit of being straightforward to implement[50]. Accordingly, 100 different E-CLM classifiers were trained on the bioactivity prediction task. The level of confidence in a prediction was defined as the number of models that classified a given input molecule as "highly active".

With increasing confidence levels, the number of molecules predicted as "highly active" decreased (Fig. 3a), a documented effect of ensemble voting[51]. None of the molecules from the focused virtual library was predicted as "highly active" with all 100 votes. Forty-seven de novo designs were predicted as highly active, with 99 votes. Among these top-ranked molecules, 64% featured a new atom scaffold and 62% featured a new graph scaffold with respect to the fine-tuning set[52,53]. Higher confidence was reflected in the increased substructure similarity of the predicted actives to the molecules of the fine-tuning set, as captured by the Tanimoto index computed on Morgan fingerprints (Fig. 3b)[54]. In line with the chemical similarity principle[55], this observation suggests that there is a greater chance of identifying active molecules when the number of votes is high. The five most dissimilar molecules among the top-ranked molecules had a similarity to their respective nearest neighbors of the fine-tuning set, ranging from 53% to 58% (Fig. 3c). The closest molecules of the fine-tuning set have a similarity ranging between 77% and 100%, meaning that one molecule of the fine-tuning set was re-created by the CLM, although with a different stereochemistry (Fig. 3d), a structural feature that is not captured by Morgan fingerprints. This result highlights the potential of the approach to explore both closely related molecules to known bioactives, e.g., for structure-activity relationship studies or hit-to-lead expansion, as well as more structurally innovative compounds for "scaffold hopping".

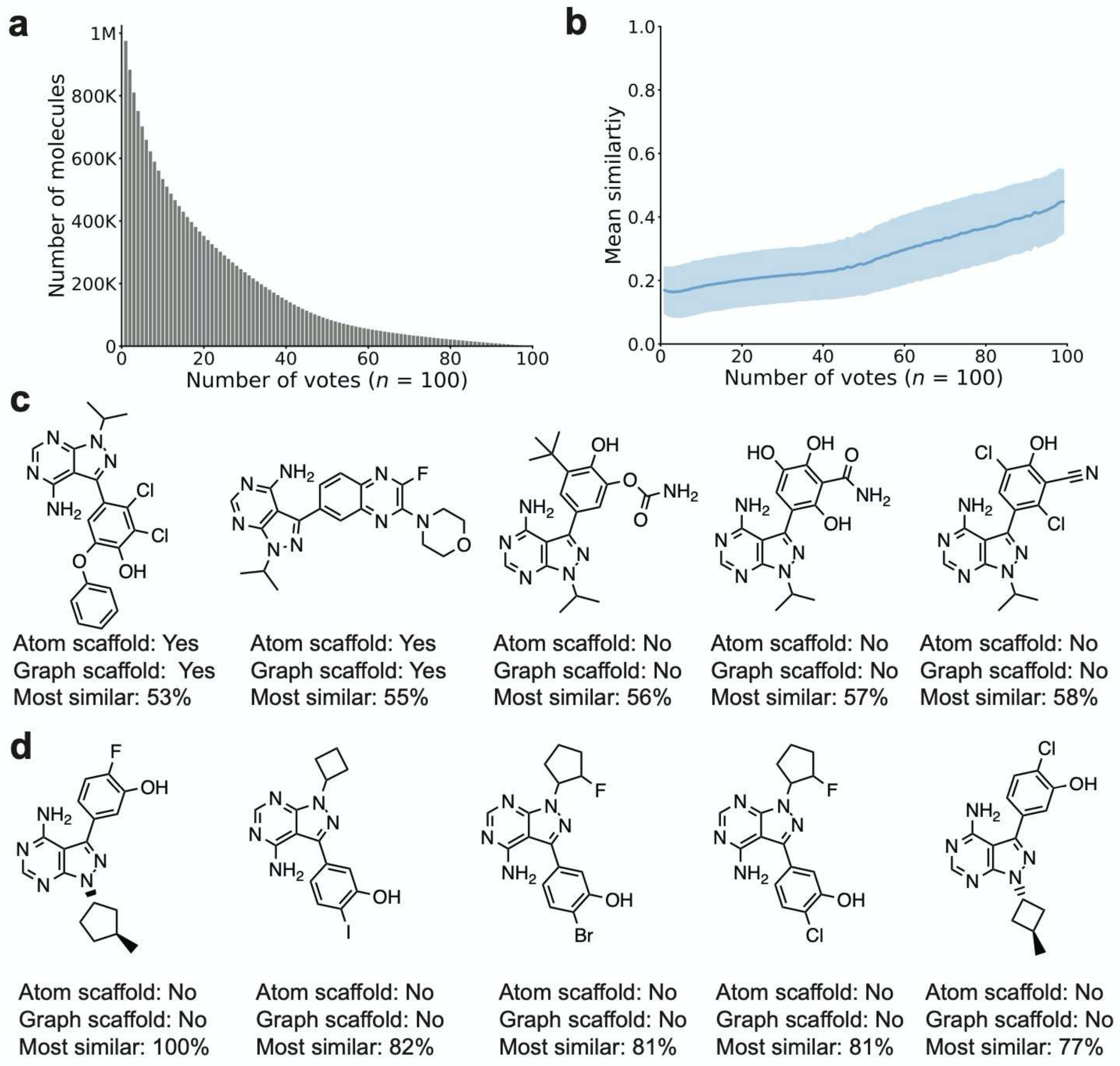

Fig. 3 | Molecule ranking with a deep ensemble model. a, Number of molecules in the refined virtual chemical library that were predicted as "highly active" as a function of the number of votes (confidence level). **b,** Average structural similarity (Tanimoto similarity index computed on Morgan fingerprints) of each de novo design to the fine-tuning set as a function of the number of votes. The solid line represents the mean value, with the shaded area representing the standard deviation. **c,** Top-ranked designs (99/100 votes) selected with the most distant nearest neighbor, whose similarity is indicated below the structure ("Most similar") in the fine-tuning set. The atom ("Atom scaffold") and graph ("Graph scaffold") scaffold novelty of the structure with respect to the fine-tuning set is indicated below each structure ("Yes": new, "No": not new). **d,** Top-ranked designs (99/100 votes) selected with the closest nearest neighbor in the fine-tuning set.

In vitro bioactivity testing

For a proof of concept, some of the molecules generated by the CLM were tested for PI3Kγ binding in vitro. To optimize the efficiency in terms of both time and resources, we selected the test molecules from the refined virtual chemical libraries that could be purchased from commercial suppliers, as opposed to synthesizing the de novo designs. In total, 16 computer-generated molecules were commercially available. Their predictive confidence ranged from 80/100 votes for compound **1** to 24/100 votes for compound **16** (Fig. 4).

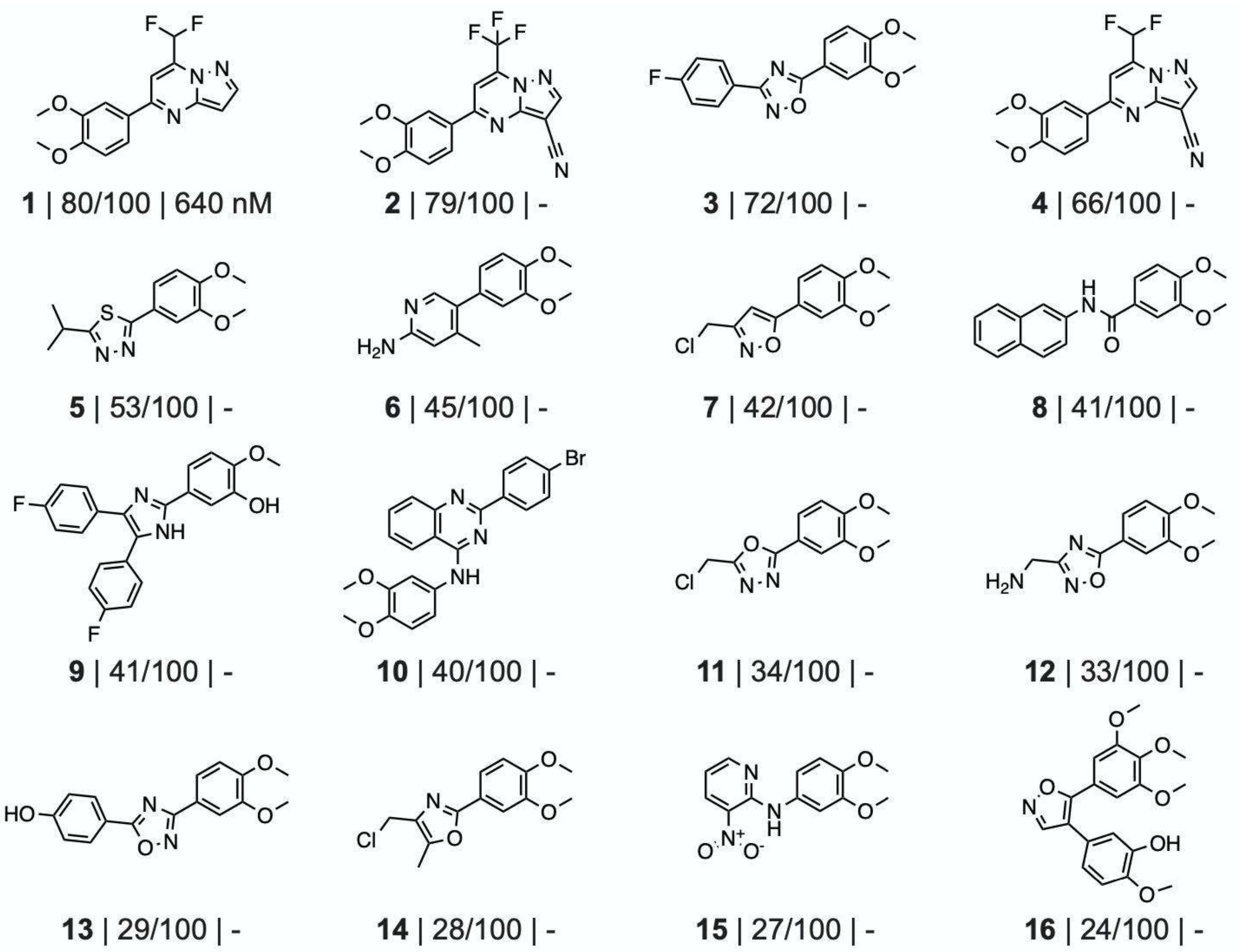

Fig. 4 | Compounds tested for PI3Kγ inhibition. Compounds **1–16** are shown, together with the number of votes from the ensemble of the maximum number of 100 possible votes and the experimentally determined binding constant K_d. Absence of a value (-) indicates no observed binding of the compound to the target.

Although none of the ordered molecules was part of the top-ranked set (i.e., receiving 99/100 votes), compound **1**, the molecule with the highest number of votes (80/100), was a hit, with K_d ranging between 0.6 and 0.7 μM (n = 2; 670 nM and 620 nM) (Figs. 4 and 5, and Supplementary Table S3). None of the lower-ranking compounds inhibited PI3Kγ in the biochemical assay (Fig. 4). The confidence level of our ensemble correctly prioritized compound **1** (active in vitro) over compounds **2** and **4** (inactive in vitro), despite all of them having the same scaffold but with different substituents (Fig. 4). We hypothesize that this might be due to the positive effect of the ELECTRA pretraining, which was aimed at recognizing the effect of small structural changes.

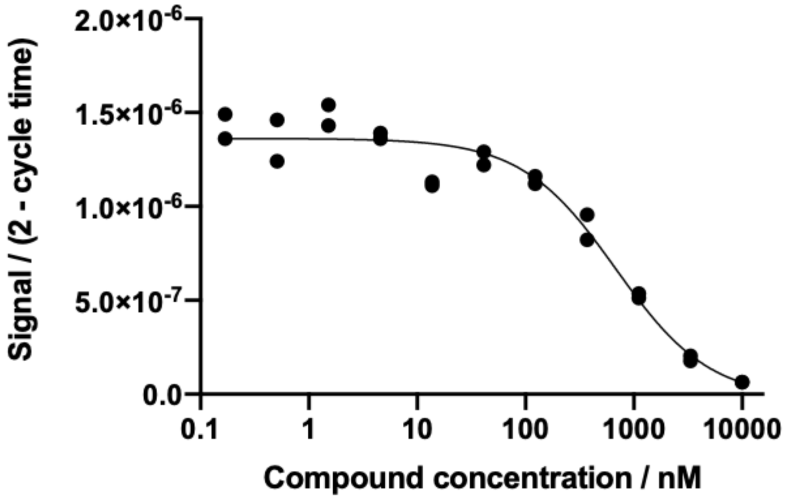

Fig. 5 | In vitro characterization of compound 1. Kinase-ligand binding was determined in a competition assay (*n* = 2), using an immobilized ligand of PI3Kγ and quantitative polymerase chain reaction (qPCR) measuring the competing DNA-tagged PI3Kγ protein. The signal is expressed as a transformation of the qPCR cycle time (2 - cycle time).

Hit compound **1** has a new atom scaffold compared to all molecules in the ChEMBL database (version 28) annotated with "pActivity" ≥ 5.0 on PI3Kγ ("pActivity": -log(molar IC_{50}, XC_{50}, EC_{50}, AC_{50}, K_i, K_d, or "potency")). The most similar molecule among these has a Tanimoto similarity of 34% to that of compound **1** (Supplementary Fig. S4). This preliminary in vitro validation advocates an ensemble prediction approach for virtual compound screening and ranking of the computer-generated molecular designs.

Conclusion

Methodological improvements in CLM training advanced the sampling of target-focused virtual molecule libraries. The hybrid CLM classifier successfully included both structural and bioactivity information of the fine-tuning molecules for the design of a virtual chemical library, thereby complementing the available methodological repertoire for virtual screening. It remains to be determined in more detail to what extent the CLM pretraining method affects model performance in the downstream task, i.e., molecular generation or ordinal classification. Importantly, CLM training was performed without data augmentation to study the positive effect of nucleus sampling on the generation of a SMILES string. Future improvement might be possible by combining nucleus sampling with data augmentation for CLM transfer learning[33,56–58]. Given the setup of the present study, it was not possible to determine whether our hypothesis regarding the beneficial effect of model pretraining on patented chemical structures holds true. The long time required for hit-to-lead expansion and for preclinical and clinical drug development until a marketed drug is obtained will likely preclude any such analysis. Nonetheless, a PI3Kγ ligand with a new scaffold was computationally generated. Hit compound **1** was correctly predicted to be active with 80% confidence (80/100 votes). This result now motivates the synthesis of the top-ranking compounds with novel scaffolds. Obtaining rapid experimental validation of a set of readily available de novo designed molecules prior to embarking on de novo synthesis might help assess the value of computationally generated activity-focused chemical libraries. Future prospective studies will also have to assess the general applicability of this approach to other targets from different target families.

This study highlights the versatility of generative deep learning for hit and lead finding in drug discovery, where the same computational pipeline can be used to both create new molecules and screen libraries of existing compounds. We envision future projects in which de novo design methods are first validated for physically available molecules from a compound repository or commercial suppliers before investing in potentially more expensive and time-consuming syntheses. This approach could help accelerate the design-make-test,-analyze cycles, and reduce the risk of failure[59].

Code and data availability

The computational framework presented in this study, along with the pretrained neural network weights and the data used for training, is available as a GitHub repository at https://github.com/ETHmodlab/hybridCLMs.

Methods

Target selection. The protein target PI3Kγ[26] was selected on the basis of the data available in the DTC[35] database. We selected one of the targets with the most annotated data. Molecules with activity entries satisfying all of the following conditions were kept: standard relation: "=", standard unit: "nM", substrate value: "10", substrate unit: "µM", test inhibitor type: "competitive inhibitor", compound concentration value: "0.001–50", test assay format: "biochemical", test assay type: "functional", test assay subtype: "enzyme activity". This filtering step resulted in a dataset containing 198 molecules (Supplementary Fig. S1).

Training data. The training molecules were represented as canonical SMILES strings using the RDKit package (v. 2019.03.2, https://www.rdkit.org). SMILES strings with a length of up to 90 characters were retained and standardized in Python (v. 3.6.5) by removing salts and duplicates. The CLM was pretrained on the pharmaceutical subset of the US patent database[60,61]. After the processing, 839,674 unique molecules encoded as canonical SMILES strings constituted the pretraining data. PI3Kγ inhibitors with a reported bioactivity ≤100 nM in the DTC database were used for the CLM transfer learning ("fine-tuning set"). This criterion resulted in a fine-tuning set containing 43 molecules.

CLM pretraining and fine-tuning for the generation of SMILES strings. The CLM model was implemented in Python (v. 3.6.5) using Keras (v. 2.2.0, https://keras.io/) with the TensorFlow GPU backend (v. 1.9.0, https://www.tensorflow.org). The model was implemented as a recurrent neural network with LSTM cells. The neural network was composed of four layers with a total of 5,820,515 parameters (layer 1: BatchNormalization, layer 2: 1024 LSTM cells, layer 3: 256 LSTM cells, and layer 4: BatchNormalization) and trained with SMILES data encoded as one-hot vectors. The CLM was trained using the Adam optimizer (learning rate = 10^{-3}) and the categorical cross-entropy loss function. Training was performed over 40 epochs, where one epoch was defined as one pass over all the training data. Transfer learning was performed by keeping the parameters of the first network layer constant and training the second layer with a learning rate of 10^{-4}.

ELECTRA pretraining. The E-CLM model was implemented in Python (v. 3.6.5) using Keras (v. 2.2.0, https://keras.io/) with the TensorFlow GPU backend (v. 1.9.0, https://www.tensorflow.org). The ELECTRA model was implemented with the same architecture as that of the generative CLM, i.e., as a recurrent neural network with LSTM cells. Model training was performed with the Adam optimizer (learning rate = 10^{-3}, 50 epochs) and the binary cross-entropy loss function.

Ordinal classifier training. The hybrid CLM network contained the weights of the pretrained E-CLM plus an additional feedforward layer with three sigmoidal neurons. The model was trained to solve an ordinal classification task, where each of the three output neurons corresponded to one class. *k*-Means clustering (*k* = 5, Scikit-learn; https://scikit-learn.org/stable/) was performed to group the fine-tuning molecules according to their similarity based on Morgan fingerprints. Four groups were used for cross-validation and one for classifier testing. The output threshold values, the number of transfer learning epochs, and the oversampling values of the less represented classes were defined by cross-validation. The best settings were selected on the basis of the performance on the test set, which was used once (oversampling: +40 molecules for the two less represented classes, sigmoid threshold: 0.4, number of transfer learning epochs: 200). Each of the 100 CLM models of the final ensemble was trained with the best settings on all available data. The neural network architecture was composed of six layers with a total of 5,646,982 parameters (layer 1: BatchNormalization, layer 2: 1024 LSTM cells, layer 3: 256 LSTM cells, layer 4: BatchNormalization, layer 5: Dropout, and layer 6: Dense, with three units, each with a sigmoid activation function) and was trained with SMILES encoded as one-hot vectors. The models were trained with the Adam optimizer and the binary cross-entropy loss function (learning rate = 10^{-4}, 200 epochs).

Temperature sampling. SMILES characters were sampled using the softmax function parameterized by the sampling temperature. The probability of the *i*-th character being sampled from the CLM predictions was computed as (Eq. 1)

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

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

Nucleus sampling. SMILES characters were sampled with a temperature value equal to 1 (Eq. 2), considering only characters whose cumulative probability was greater than the nucleus parameter (“top vocabulary”):

$$\sum_{x \in V^{(p)}} P(x \mid x_{1:i-1}) > p\,, \qquad (2)$$

where $V^{(p)}$ is the top vocabulary, *x* is an element of the vocabulary, and *p* is the nucleus parameter.

Biochemical kinase binding assay. PI3Kγ binding assays were performed by Eurofins Discovery (https://www.eurofinsdiscoveryservices.com) on a fee-for-service basis. KINOMEscan™ was used to determine the dissociation constant K_d of compounds **1–16**. The assay was based on the ability of a test compound to compete with an immobilized active site-directed ligand. Competition of the test compound with the immobilized ligand was measured via quantitative PCR (qPCR) of the DNA tag of DNA-tagged kinase[62]. An 11-point three-fold serial dilution of each test compound was prepared in 100% DMSO at 100× final test concentration and subsequently diluted to 1× in the assay (final DMSO concentration = 1%). Dissociation constants were estimated with a standard dose-response curve using the Hill equation (Eq. 3)[63]:

$$Response\ =\ Background\ +\ \frac{Signal - Background}{1 + (K_d^{Hill\,slop}\ /\ Dose^{\ Hill\,Slope})}\,, \qquad (3)$$

where the Hill slope was set to -1. Curves were fitted using a nonlinear least square fit with the Levenberg-Marquardt algorithm[64].

Acknowledgments

This work was financially supported by the Swiss National Science Foundation (grant no. 205321_182176 to G.S.) and by the RETHINK initiative at ETH Zurich. The authors thank Alexander Button, Damian Gautschi, Jonas Bossart, and Kenneth Atz for their helpful discussions. Daniel Lowe and Lukas Friedrich are thanked for their technical assistance with the US patent database.

Author contributions

M.M., F.G., and G.S. conceived the study. M.M. implemented the software. C.B. curated the data. All authors analyzed the results and contributed to the writing of 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, Switzerland. No other potential conflicts of interest are declared.

References

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

1. Neves, B. J. *et al.* QSAR-based virtual screening: advances and applications in drug discovery. *Front. Pharmacol.* **9**, 1275 (2018).
2. Walters, W. P. Virtual chemical libraries. *J. Med. Chem.* **62**, 1116–1124 (2019).
3. Jain, A. N. Scoring noncovalent protein-ligand interactions: A continuous differentiable function tuned to compute binding affinities. *J. Comput. Aided Mol. Des.* **10**, 427–440 (1996).
4. Schneider, G. & Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discov.* **4**, 649–663 (2005).
5. 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).
6. Popova, M., Isayev, O. & Tropsha, A. Deep reinforcement learning for de novo drug design. *Sci. Adv.* **4**, eaap7885 (2018).
7. Yuan, W. *et al.* Chemical space mimicry for drug discovery. *J. Chem. Inf. Model.* **57**, 875–882 (2017).
8. 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.* **52**, 2864–2875 (2012).
9. Liu, Q., Allamanis, M., Brockschmidt, M. & Gaunt, A. Constrained graph variational autoencoders for molecule design. in *The Thirty-second Conference on Neural Information Processing Systems,* pp. 7795-7804 (2018).
10. You, J., Liu, B., Ying, R., Pande, V. & Leskovec, J. Graph convolutional policy network for goal-directed molecular graph generation. *Preprint at http://arxiv.org/abs/1806.02473* (2018).
11. Ikebata, H., Hongo, K., Isomura, T., Maezono, R. & Yoshida, R. Bayesian molecular design with a chemical language model. *J. Comput. Aided Mol. Des.* **31**, 379–391 (2017).
12. Nigam, A., Friederich, P., Krenn, M. & Aspuru-Guzik, A. Augmenting genetic algorithms with deep neural networks for exploring the chemical space. *Preprint at http://arxiv.org/abs/1909.11655* (2019).
13. Skalic, M., Jiménez, J., Sabbadin, D. & De Fabritiis, G. Shape-based generative modeling for de novo drug design. *J. Chem. Inf. Model.* **59**, 1205–1214 (2019).
14. Wallach, I., Dzamba, M. & Heifets, A. AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *Preprint at http://arxiv.org/abs/1510.02855* (2015).
15. Gómez-Bombarelli, R. *et al.* Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
16. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **361**, 360–365 (2018).
17. Soleimany, A. P. *et al.* Evidential deep learning for guided molecular property prediction and discovery. *ACS Cent. Sci.* **7**, 1356–1367 (2021).

18. Born, J. *et al.* PaccMannRL: De novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning. *iScience* **24**, 102269 (2021).
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., Friedrich, L., Grisoni, F. & Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Mol. Inf.* **37**, 1700153 (2018).
21. Grisoni, F. *et al.* Combining generative artificial intelligence and on-chip synthesis for de novo drug design. *Sci. Adv.* **7**, eabg3338 (2021).
22. Moret, M., Helmstädter, M., Grisoni, F., Schneider, G. & Merk, D. Beam search for automated design and scoring of novel ROR ligands with machine intelligence. *Angew. Chem. Int. Ed.* **60**, 19477–19482 (2021).
23. 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).
24. 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).
25. Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminform.* **9**, 48 (2017).
26. Yang, J. *et al.* Targeting PI3K in cancer: Mechanisms and advances in clinical trials. *Mol. Cancer* **18**, 26 (2019).
27. Kaneda, M. M. *et al.* PI3Kγ is a molecular switch that controls immune suppression. *Nature* **539**, 437–442 (2016).
28. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
29. Wani, M. A., Bhat, F. A., Afzal, S. & Khan, A. I. *Advances in deep learning*. (Springer, Singapore, 2020).
30. Bengio, Y., Ducharme, R., Vincent, P. & Jauvin, C. A neural probabilistic language model. *J. Mach. Learn. Res.* **3**, 1137–1155 (2003).
31. Gupta, A. *et al.* Generative recurrent networks for de novo drug design. *Mol. Inf.* **37**, 1700111 (2018).
32. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2017).
33. Moret, M., Friedrich, L., Grisoni, F., Merk, D. & Schneider, G. Generative molecular design in low data regimes. *Nat. Mach. Intell.* **2**, 171–180 (2020).
34. Lowe, D. Chemical reactions from US patents (1976-Sep2016). (2017).
35. Tanoli, Z. *et al.* Drug Target Commons 2.0: A community platform for systematic analysis of drug-target interaction profiles. *Database* **2018**, 1–13 (2018).
36. 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).
37. Holtzman, A., Buys, J., Forbes, M. & Choi, Y. The curious case of neural text degeneration. Preprint at *http://arxiv.org/abs/1904.09751* (2019).
38. David, L. *et al.* Applications of deep-learning in exploiting large-scale and heterogeneous compound data in industrial pharmaceutical research. *Front. Pharmacol.* **10**, 1303 (2019).

39. Li, L. & Lin, H. T. Ordinal regression by extended binary classification. *Advances in Neural Information Processing Systems* **19**, 856–872 (2007).
40. Clark, K., Luong, M.-T., Le, Q. V. & Manning, C. D. ELECTRA: Pre-training text encoders as discriminators rather than generators. Preprint at *http://arxiv.org/abs/2003.10555* (2020).
41. Dimova, D., Stumpfe, D. & Bajorath, J. Systematic assessment of coordinated activity cliffs formed by kinase inhibitors and detailed characterization of activity cliff clusters and associated SAR information. *Eur. J. Med. Chem.* **90**, 414–427 (2015).
42. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. Preprint at *http://arxiv.org/abs/1810.04805* (2018).
43. Howard, J. & Ruder, S. Universal language model fine-tuning for text classification. Preprint at *http://arxiv.org/abs/1801.06146* (2018).
44. Japkowicz, N. Learning from imbalanced data sets: A comparison of various strategies. in *AAAI Workshop on Learning From Imbalanced Data Sets,* pp. 10–15 (2000).
45. Hirschfeld, L., Swanson, K., Yang, K., Barzilay, R. & Coley, C. W. Uncertainty quantification using neural networks for molecular property prediction. *J. Chem. Inf. Model.* **60**, 3770–3780 (2020).
46. Sahigara, F. *et al.* Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* **17**, 4791–4810 (2012).
47. Gadaleta, D., Mangiatordi, G. F., Catto, M., Carotti, A. & Nicolotti, O. Applicability domain for QSAR models: where theory meets reality. *Int. J. Quant. Struct. Prop. Relat.* **1**, 45–63 (2016).
48. Lam, L. & Suen, S. Y. Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Trans. Syst, Man Cybern. - Part A: Systems and Humans* **27**, pp. 553–568 (1997).
49. Koch, C. P. *et al.* Exhaustive proteome mining for functional MHC-I ligands. *ACS Chem. Biol.* **8**, 1876–1881 (2013).
50. Lakshminarayanan, B., Pritzel, A., & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. in *Advances in Neural Information Processing Systems* **30**, 6 (2017).
51. Valsecchi, C., Grisoni, F., Consonni, V. & Ballabio, D. Consensus versus individual QSARs in classification: comparison on a large-scale case study. *J. Chem. Inf. Model.* **60**, 1215–1223 (2020).
52. Xu, Y. & Johnson, M. Algorithm for naming molecular equivalence classes represented by labeled pseudographs. *J. Chem. Inf. Comput. Sci.* **41**, 181–185 (2001).
53. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).
54. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
55. Johnson, M. A. & Maggiora, G. M. Concepts and applications of molecular similarity. (Wiley, New York, 1990).
56. Skinnider, M. A., Greg Stacey, R., Wishart, D. S. & Foster, L. J. Chemical language models enable navigation in sparsely populated chemical space. *Nat. Mach. Intell.* **3**, 759–770 (2021).

57. Bjerrum, E. J. SMILES enumeration as data augmentation for neural network modeling of molecules. Preprint at *https://arxiv.org/abs/1703.07076* (2017).
58. Dao, T., Gu, A., Ratner, A. J. Smith, V., De Sa, C. & Ré, C. A kernel theory of modern data augmentation. *Proceedings of the 36th International Conference on Machine Learning,* pp. 1528–1537 (2019).
59. Schneider, G. Automating drug discovery. *Nat. Rev. Drug Discov.* **17**, 97–113 (2018).
60. Lowe, D. M. *Extraction of chemical structures and reactions from the literature*. (Cambridge, University of Cambridge, 2012).
61. Schneider, N., Lowe, D. M., Sayle, R. A., Tarselli, M. A. & Landrum, G. A. Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter. *J. Med. Chem.* **59**, 4385–4402 (2016).
62. Fabian, M. A. *et al.* A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **23**, 329–336 (2005).
63. HILL & A. V. The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *J. Physiol.* **40**, 4–7 (1910).
64. Levenberg, K. A method for the solution of certain non-linear problems in least squares. *Quart. Appl. Math.* **2**, 164–168 (1944).