Article

Perplexity-Based Molecule Ranking and Bias Estimation of Chemical Language Models

Michael Moret,[+] Francesca Grisoni,[*,+] Paul Katzberger, and Gisbert Schneider[*]

Cite This: *J. Chem. Inf. Model.* 2022, 62, 1199−1206

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Chemical language models (CLMs) can be employed to design molecules with desired properties. CLMs generate new chemical structures in the form of textual representations, such as the simplified molecular input line entry system (SMILES) strings. However, the quality of these de novo generated molecules is difficult to assess a priori. In this study, we apply the perplexity metric to determine the degree to which the molecules generated by a CLM match the desired design objectives. This model-intrinsic score allows identifying and ranking the most promising molecular designs based on the probabilities learned by the CLM. Using perplexity to compare "greedy" (beam search) with "explorative" (multinomial sampling) methods for SMILES generation, certain advantages of multinomial sampling become apparent. Additionally, perplexity scoring is performed to identify undesired model biases introduced during model training and allows the development of a new ranking system to remove those undesired biases.
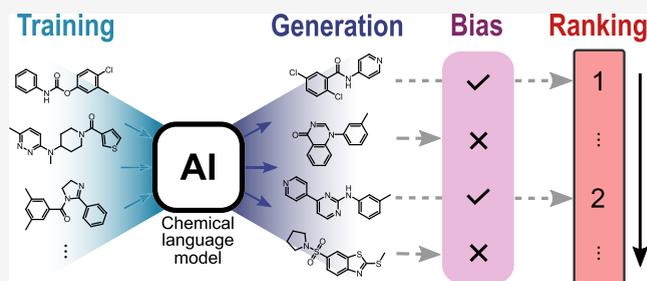
INTRODUCTION

Generative deep learning has become a promising method for chemistry and drug discovery.[1−21] Generative models learn the pattern distribution of the input data and generate new data instances based on learned probabilities.[22] Among the proposed generative frameworks that have been applied to de novo molecular design,[2−19] chemical language models (CLMs) have gained attention because of their ability to generate focused virtual chemical libraries and bioactive compounds.[20,21,23] CLMs are trained on string representations of molecules, *e.g.*, simplified molecular input line entry system (SMILES) strings (Figure 1a),[24] to iteratively predict the next SMILES character using all the preceding portions of the SMILES string (Figure 1b). In this process, CLMs learn the conditional probability of sampling any SMILES character based on the preceding characters in the string. After training, the model can be used for molecular construction. CLMs have been demonstrated to both learn the SMILES syntax and implicitly capture "semantic" features of the training molecules, such as physicochemical properties,[20,21,25,26] bioactivity,[2,21] and chemical synthesizability.[3] Although alternative generative approaches have been proposed for de novo design,[13,27−29] benchmarks have not shown these to outperform CLMs.[30,31] A feature of CLMs is their ability to function in low-data regimes,[25,29] *i.e.*, with limited training data (typically in the range of 5−40 molecules).[2,3,25] One of the most widely employed approaches for low-data model training is transfer learning.[20,32] This method leverages previously acquired information on a related task for which more data are available

("pretraining") before training the CLM on a more specific limited dataset ("fine-tuning").[33]

Several prospective de novo design studies based on CLMs used weighted random sampling (*i.e.*, multinomial sampling, often in the form of temperature sampling) for molecule generation.[2,20,25,34] This method samples the most likely SMILES string characters more frequently than the unlikely characters. This feature enables (i) extensive virtual molecule libraries to be generated and (ii) a certain chemical space to be investigated owing to "fuzzy" (probability-weighted) random sampling. However, such a sampling strategy can result in molecules that do not possess the physicochemical and biological properties of the training data. Furthermore, because the number of molecules that can potentially be sampled from CLMs considerably exceeds synthetic capacities, and a natural ranking of the generated SMILES does not exist, an additional procedure is required for molecule prioritization, *e.g.*, one based on similarity assessment or activity prediction.[2,35] We recently introduced the beam search algorithm as an alternative to multinomial sampling. During beam search, the most likely SMILES strings are generated based on the respective character probabilities, thereby alleviating the strict require-
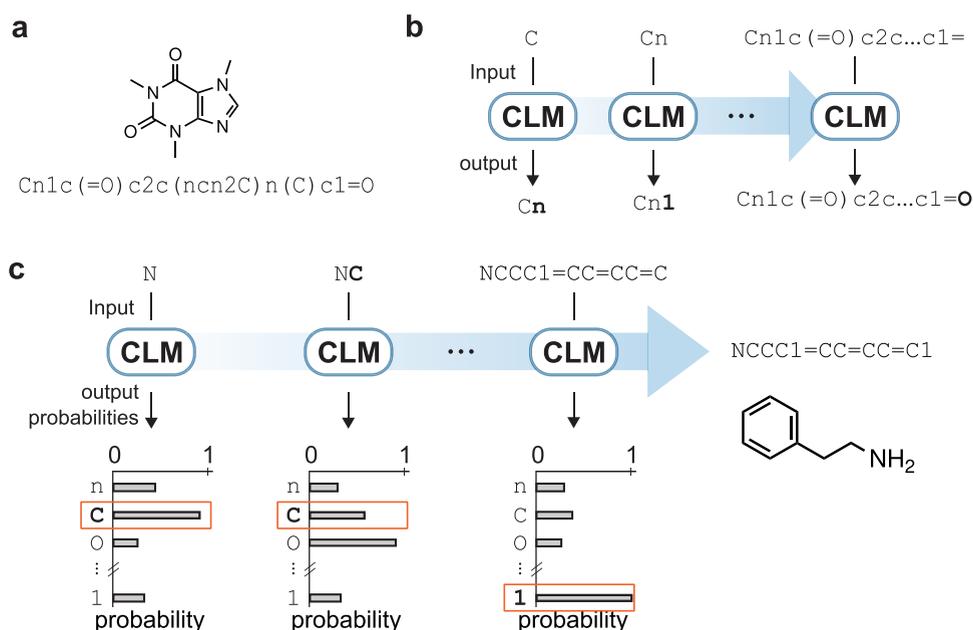
Figure 1. Principles of chemical language models (CLMs). (a) Example of a molecular structure (Kekulé structure) and a corresponding SMILES string. (b) CLMs are trained to iteratively predict the next SMILES character based on the preceding string characters. (c) Multinomial sampling can be used to generate new SMILES strings from trained CLMs, where SMILES characters are sampled with a weighted random sampling of probability distributions learned by the CLM.

ment for additional molecule prioritization.[36] The beam search performs chemical space "exploitation" as the algorithm searches for the most probable SMILES strings in a greedy manner. This method generates only a few candidate molecules at the expense of chemical space exploration and design diversity.

Herein, we aimed to improve upon an existing CLM that we recently used for prospective application[3,25,36] to increase its potential for automated molecular design and scoring. To this end, we used perplexity to assess the "goodness" of the designs generated by CLMs via multinomial sampling aiming to (i) preserve the advantage of intrinsic molecule ranking[37] as it can be achieved via beam search, (ii) benefit from the chemical space exploration provided by multinomial sampling, and (iii) provide model-based insights into the most promising molecules for follow-up analysis.[36] We systematically trained a CLM on sets of bioactive ligands of 10 different macromolecular targets under four different data-regime scenarios each. On top of its ability to rank the generated SMILES strings, the perplexity metric identified undesired effects of transfer learning using CLMs, thereby qualifying as a criterion for detecting undesired model bias.

■ RESULTS AND DISCUSSION

Perplexity-Based Scoring of Molecular De Novo Designs. This study aimed to leverage the overall likeness of the generated SMILES strings for automated molecule ranking based on the respective character probabilities. Accordingly, the SMILES string(s) with the highest likeness can be considered the best-matching solutions to the CLM sampling problem, reflecting the information learnt by the CLM. We selected the perplexity metric to reflect the probability of sampling a SMILES string as a function of its characters (eq 1).[37] Perplexity has been used to assess the performance of language models in natural language processing.[37−39] For a SMILES string of length N, the

perplexity score can be computed by considering the CLM probability of any ith character (p_i):

$$\text{perplexity} = 2^{-\frac{1}{N}\sum_{i=1}^{N}\log(p_i)} \tag{1}$$

The information on the overall character probabilities is captured into a single metric, which is normalized by the length of the SMILES string (N). Perplexity allows quantifying the CLM confidence that a specific SMILES string could have belonged to the training data. If the assumption that the underlying CLM captured relevant information from the training data is satisfied, then perplexity will be suitable for molecule ranking. Because the training objectives of the CLM are implicitly encoded in the fine-tuning data, the perplexity score allows one to assess whether the generated SMILES strings match the objectives. A SMILES string composed of probable characters (high p_i values) exhibits low perplexity, whereas a string containing many unlikely characters (low p_i values) exhibits high perplexity. Hence, low perplexity scores are desirable.

To analyze the behavior of perplexity, an RNN with long short-term memory (LSTM) cells[40] was pretrained with approximately 1.6 million molecules from ChEMBL (version 28).[41] Ten randomly selected targets were used for fine-tuning (Table 1). For each macromolecular target, ligands that possessed a pChEMBL activity value larger than 6 were selected, where pChEMBL is defined as $-\log_{10}$(molar IC_{50}, XC_{50}, EC_{50}, AC_{50}, K_i, K_d, or potency). To emulate different low-data regimes typical of drug discovery, we prepared fine-tuning sets of different sizes that contained 5, 10, 20, or 40 randomly selected ligands for each target. For each of the 10 targets and each of the four fine-tuning sets, a total of 1000 SMILES strings were sampled after every second epoch during a total of 100 CLM fine-tuning epochs via multinomial sampling. Unlike our previous studies,[2,3,25,35] no temperature parameter was used to modify multinomial sampling to avoid

Table 1. Macromolecular Targets Selected for CLM Fine-Tuning[a]

CHEMBL ID	target	protein classification
CHEMBL1836	prostanoid EP4 receptor	G protein-coupled receptor
CHEMBL1945	melatonin receptor 1A	G protein-coupled receptor
CHEMBL1983	serotonin 1D (5-HT$_{1D}$) receptor	family A G protein-coupled receptor
CHEMBL202	dihydrofolate reductase	oxidoreductase
CHEMBL3522	cytochrome P450 17A1	cytochrome P450
CHEMBL4029	interleukin-8 receptor A	family A G protein-coupled receptor
CHEMBL5073	CaM kinase I delta	kinase
CHEMBL5137	metabotropic glutamate receptor 2	G protein-coupled receptor
CHEMBL5408	serine/threonine-protein kinase TBK1	kinase
CHEMBL5608	NT-3 growth factor receptor	kinase

[a]ChEMBL target identifier, generic target name, and protein classification based on the respective ChEMBL target report card.

the introduction of confounding factors in our analysis. For all fine-tuned models and all the fine-tuning epochs, the mean SMILES string validity consistently exceeded 90% (Figure S1).

Chemical Relevance of Perplexity. To investigate the information captured by the perplexity metric, we evaluated its correlation with two measures of molecular similarity computed between the de novo designs and the corresponding fine-tuning sets: (a) Tanimoto similarity on Morgan fingerprints,[42] which captures the presence of common substructures, and (b) Tanimoto similarity on topological pharmaco-phore fingerprints,[43] which captures the presence of shared structural motifs relevant for ligand-target interactions. To allow for an easier comparison with the perplexity metric, the computed Tanimoto similarity values were converted into distances (computed as 1 − similarity).

A Pearson correlation of approximately 0.3 was observed between the perplexity score and the computed distances to the smaller fine-tuning sets (5 and 10 molecules) during the initial fine-tuning epochs before stabilizing at a value of 0.5 (Figure 2). A correlation of 0.5 tends to be reached at earlier epochs with bigger fine-tuning sets compared to using only five molecules (Figure 2). This result suggests that the perplexity score captures common substructure and 2D pharmacophore features while at the same time incorporating additional information modeled by the CLM.

Perplexity as an Indicator for Comparing Molecular Sampling Strategies. In a previous study, new bioactive compounds were successfully identified via beam search sampling,[36] which is a heuristic greedy algorithm. Its search "breadth" is controlled by a width parameter (k), which represents the number of the most probable SMILES strings that the model considers during string extension. Here, beam search was used as a reference method for comparison with multinomial sampling.

We investigated the difference in perplexity scores between molecules generated using a CLM via either beam search or multinomial sampling. To this end, we additionally sampled for each of the 10 targets and each of the four fine-tuning sets, 10 or 50 SMILES strings via beam search sampling (k = 10 or 50).

In CLM fine-tuning using the smallest fine-tuning sets (five molecules), multinomial sampling consistently outperformed beam search sampling in terms of the perplexity score as
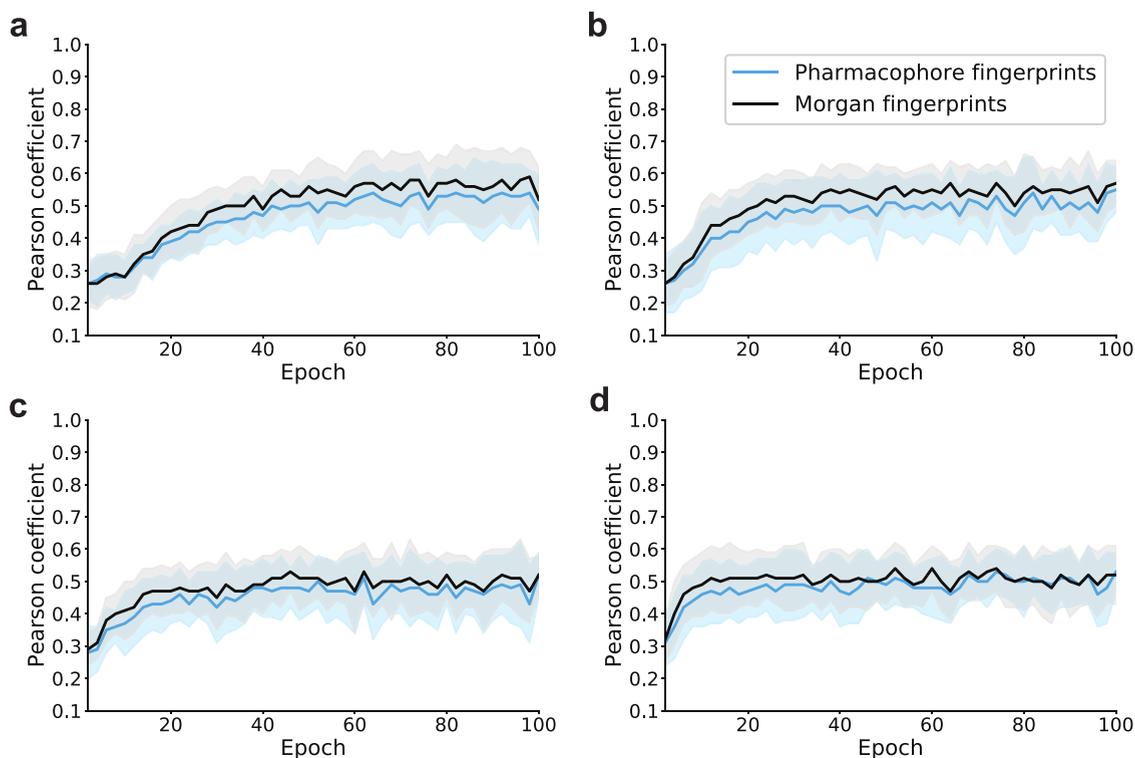
Figure 2. Correlation between the perplexity score and other representations. Pearson correlation coefficient between the perplexity score and the average distance of the generated molecules during fine-tuning (black line, Morgan fingerprints; blue line, 2D pharmacophore fingerprints). Gray shaded areas indicate standard deviations. Fine-tuning sets contained (a) 5, (b) 10, (c) 20, and (d) 40 molecules.
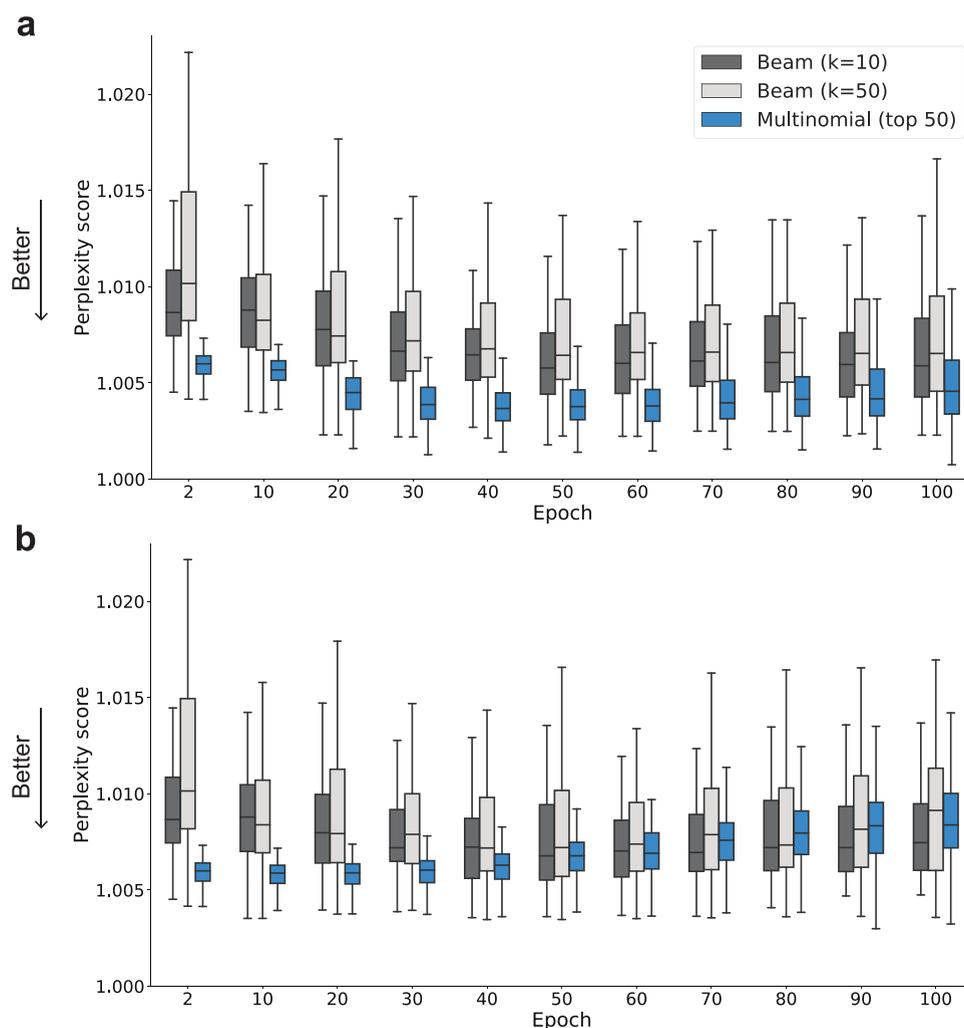
Figure 3. Variation in perplexity during fine-tuning. (a) Distribution of the top-scoring compounds for each method over 100 fine-tuning epochs (only every 10 epochs shown in the graph for clarity). (b) Distribution of the top-scoring compounds by considering only molecules with a similarity below 50% (Tanimoto index computed on Morgan fingerprints) to the closest molecule in their respective fine-tuning set. Median and lower to upper quartile values reported using boxplots for 10 different protein-specific fine-tuning sets, which contain five molecules each. Boxplots for fine-tuning sets with sizes of 10, 20, and 40 molecules are provided in the Supporting Information. Arrows indicate the direction of optimal perplexity values ("Better").

molecules with the best score (lowest perplexity) were obtained (Figure 3a). Increasing the beam search width from $k = 10$ to 50 did not markedly improve the ability of this method in identifying molecules with higher perplexity scores (Figure 3). These observations were confirmed for the larger fine-tuning sets (Figures S2−S4). A potential explanation for this observation is the "greedy" nature of the beam search,[44] which explores only a limited number of possibilities for next-character addition. By contrast, the "fuzzy" nature of multinomial sampling allows the generation of a greater number of molecules and hence a broader exploration of the chemical space of interest.

The 50 top-scoring molecules generated via multinomial sampling not only indicated lower median perplexity values but also spanned a narrower range of values (Figure 3a). This suggests that multinomial sampling yields a greater number of high-scoring designs (low perplexity) for follow-up synthesis and biological testing than the beam search algorithm (Figure 3a and Figures S2−S4).

When filtering out designs with a substructure similarity (Tanimoto index on Morgan fingerprints[42]) greater than 50%

of the respective fine-tuning molecules, the difference between multinomial sampling and beam searching was less pronounced (Figure 3 and Figures S5−S7). Multinomial sampling identified molecules with lower perplexity scores than the beam search in 72% of the cases involving the smallest fine-tuning sets (Table S1). The deterioration in performance for highly diverse molecules was less pronounced with the larger fine-tuning sets (Table S1).

The results of this study corroborate the potential of multinomial sampling, not only for chemical space exploration to obtain chemically diverse molecular designs but also for generating high-scoring compounds that are sufficiently diverse from the fine-tuning compounds.

Assessing Pretraining Bias Based on Perplexity. CLM pretraining might impose a greater effect on model performance than CLM fine-tuning as model pretraining is typically performed with data that are at least 2 orders of magnitude higher in amount than fine-tuning.[2,3,25,35,36] If a molecule is generated by a CLM due to pretraining only, then it will not necessarily match the design objectives as represented by the fine-tuning data. We analyzed the degree to which new
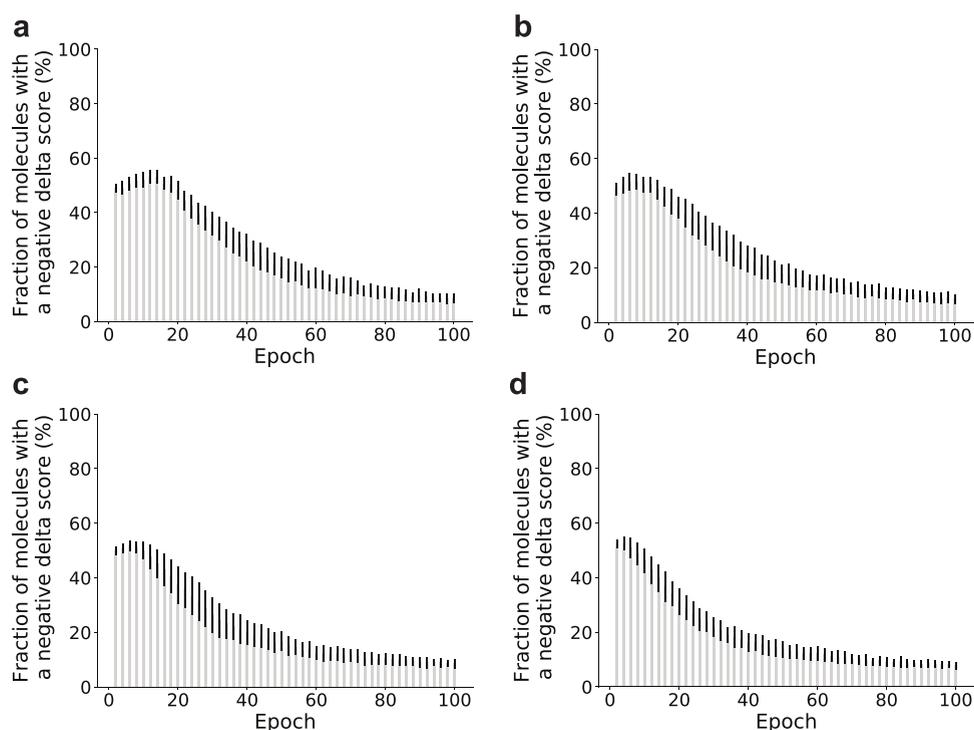
Figure 4. Delta score during fine-tuning experiments in a low-data regime. Percentage of molecules with a negative delta score (1000 sampled SMILES strings; mean ± standard deviation reported for 10 different target proteins). Fine-tuning sets of (a) 5, (b) 10, (c) 20, and (d) 40 molecules.

molecules were generated due to the sole effect of pretraining, *i.e.*, we verified whether "pretraining bias" occurred. In principle, for a CLM, perplexity can be used to score any molecule, including those that are not generated by the model. This can be achieved by computing the conditional probabilities of each SMILES character using the CLM. Therefore, the perplexity score was employed to differentiate between the information learned by the CLM during pretraining and during fine-tuning to score the molecules generated at a specified fine-tuning epoch. First, for each fine-tuning epoch, molecules were scored and ranked by the perplexity of the model used to generate them. Subsequently, each de novo design was scored and ranked based on the perplexity of the CLM after pretraining (*i.e.*, prior to any fine-tuning). We hypothesized that a suitable ranking method should favor molecules that were generated based on information learned by the model during fine-tuning (capturing the final objectives of the experiment) and downrank the molecules generated based solely on pretraining (capturing "generic" information).

To seize this concept quantitatively, we subtracted the rank yielded by the pretrained model ($rank_{pt}$) from that of the fine-tuned CLM ($rank_{ft}$) for each molecule and defined this difference as the "delta" score (eq 2), as follows:

$$delta = rank_{ft} - rank_{pt} \tag{2}$$

Molecules with a positive delta score were considered more likely to be output by the fine-tuned CLM. A negative delta score suggests that the fine-tuning procedure renders a certain molecule less likely to be output than after pretraining, which does not satisfy the design objectives. Thus, when a molecule possesses a good rank based on the perplexity after fine-tuning

but it has a negative delta value, it should not be considered for further experiments.

The delta score was computed for all sampled molecules during fine-tuning experiments (Figure 4). The percentage of molecules with a negative delta score exceeded 40% for the first 20 fine-tuning epochs and remained above 10% until the end of fine-tuning for all fine-tuning set sizes. This observation suggests that 10−40% of the molecules were generated based on "generic" pretraining, instead of "task-focused" fine-tuning. As such, this outcome highlights the practicability of the proposed delta score as an indicator to (i) detect potential pretraining bias, (ii) identify the best-suited epoch for a productive sampling of molecules that fulfill the study goals, and (iii) select the most promising de novo designs.

To expand the analysis, we focused on the 50 top-scoring molecules generated via multinomial sampling. We discovered that, among them, only up to 3% of the molecules received a negative delta score (Figure S8). This shows that using perplexity alone reduces the pretraining bias. However, the pretraining bias was not completely removed, which highlights the benefits of using both the perplexity and delta score for molecule prioritization prior to synthesis and biological testing.

In summary, these results suggest that the potential of generative CLMs in medicinal chemistry can be expanded by employing the SMILES perplexity for molecule prioritization and for detecting potential pretraining bias.

■ CONCLUSIONS

This present study constitutes a step forward toward an automated, self-supervised de novo design. By serving as a model-intrinsic score, perplexity enables the quality assessment of generated molecules. In particular, perplexity might be useful for identifying the most promising molecules, *i.e.*, those

that match the probability distribution of the training data as captured by the CLM. This approach enabled the comparison of two different methods for SMILES sampling from a trained CLM. The results revealed certain advantages of multinomial sampling over the beam search method for molecule generation. Because perplexity can be used to score SMILES strings based on the information learned by a CLM, the pretraining bias can be identified based on the newly introduced delta score. Perplexity combined with the delta score can reveal the most promising molecules, in terms of the fine-tuning objectives, for synthesis and testing. These features can further accelerate drug discovery using CLMs. Future studies will focus on the combination of perplexity with the temperature parameter of multinomial sampling or SMILES augmentation.[34,45] Furthermore, the combination of CLMs and perplexity scoring bears promise for screening large collections of commercially available compounds to accelerate model validation.[46] More experiments should be performed to determine the effect of the new approach on molecular de novo designs involving CLMs.

■ DATA AND SOFTWARE AVAILABILITY

The computational framework presented herein, pretrained neural network weights, and data used for model training are available in a GitHub repository from URL https://github.com/ETHmodlab/CLM_perplexity.

■ METHODS

Data Processing. Molecules were represented as canonical SMILES strings using an RDKit (2019.03.2). SMILES strings were standardized in Python (v3.6.5) by removing salts and duplicates, and only SMILES strings with 20−90 characters were retained.

Pretraining Set. The molecules were retrieved from ChEMBL28.[47] After data processing, the pretraining dataset contained 1,683,181 molecules encoded as SMILES strings. This set was further segregated randomly into a training set (1,599,021 molecules) and a validation set (84,160 molecules).

Fine-Tuning Sets. Target selection was limited to molecules satisfying the following conditions (ChEMBL annotation): (i) organism: *Homo sapiens*; (ii) protein classification (L1): enzymes, membranes receptors, transcription factors, and single proteins; (iii) number of compounds: 962−2057 molecules (range defined by ChEMBL); (iv) number of activities: at least 2000 reported pChEMBL values. Ten target proteins were randomly selected from the list of targets. For each of the 10 selected target proteins, sets of 5, 10, 20, and 40 molecules with pChEMBL >6 were compiled randomly.

CLM Implementation and Training. All computational experiments were implemented in Python (v3.6.5) using Keras (v2.2.0, https://keras.io/) with the Tensorflow GPU backend (v1.9.0, https://www.tensorflow.org/). CLMs were implemented using a recurrent neural network with long short-term memory cells (LSTM).[40] The network, which was composed of four layers comprising 5,820,515 parameters (layer 1: batch normalization; layer 2: LSTM with 1024 units; layer 3: LSTM with 256 units; layer 4: batch normalization), was trained with SMILES strings encoded as one-hot vectors. We used the Adam optimizer with a learning rate of 10^{-4} for the CLM training during 90 epochs,[48] where one epoch was defined as one pass over all the training data. Fine-tuning was performed by further training the CLM on the fine-tuning set for 100 epochs.

Multinomial Sampling. Multinomial sampling was performed based on the CLM output for each SMILES string character. In particular, the probability of each ith character to be sampled (p_i) was computed using eq 3:

$$p_i = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}}$$

(3)

where z_i is the CLM output for the ith character (before applying the softmax function) and j runs over all the characters of the dictionary. T represents the temperature parameter, which in this study was set to $T = 1$. An analysis of the effect of T on the SMILES generation process can be found elsewhere.[25]

Beam Search Sampling. We use the implementation provided in ref[36] (https://github.com/ETHmodlab/molecular_design_with_beam_search) using two different beam widths ($k = 10$ and 50).

Molecular Fingerprints. All fingerprints were computed with the RDKit (2019.03.2) with default settings: (a) Morgan fingerprints were used with a radius of 2 and length of 1024; (b) 2D pharmacophore fingerprints were used with the pre-configured signature factory as published by Gobbi and Poppinger.[43]

■ ASSOCIATED CONTENT

ⓈI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c00079.

> Table S1. Comparison of multinomial sampling and beam search for identifying high-scoring molecules, Figure S1. Validity of the sampled SMILES strings during fine-tuning, Figure S2. Variation of the perplexity score during fine-tuning (10 fine-tuning molecules), Figure S3. Variation of the perplexity score during fine-tuning (20 fine-tuning molecules), Figure S4. Variation of the perplexity score during fine-tuning (40 fine-tuning molecules), Figure S5. Variation of the perplexity score during fine-tuning (10 fine-tuning molecules; threshold on similarity), Figure S6. Variation of the perplexity score during fine-tuning (20 fine-tuning molecules; threshold on similarity), Figure S7. Variation of the perplexity score during fine-tuning (40 fine-tuning molecules; threshold on similarity), and Figure S8. Variation of the delta score during fine-tuning (50 top-ranked molecules) (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Francesca Grisoni − *Institute for Complex Molecular Systems, Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven 5612AZ, Netherlands; Center for Living Technologies, Alliance TU/e, WUR, UU, UMC Utrecht, Utrecht 3584 CB, The Netherlands;* ⓞ orcid.org/0000-0001-8552-6615; Email: f.grisoni@tue.nl

Gisbert Schneider − *Department of Chemistry and Applied Biosciences, ETH Zurich, Zurich 8093, Switzerland; ETH Singapore SEC Ltd., Singapore 138602, Singapore;* ⓞ orcid.org/0000-0001-6706-1084; Email: gisbert@ethz.ch

Authors

Michael Moret − *Department of Chemistry and Applied Biosciences, ETH Zurich, Zurich 8093, Switzerland*

Paul Katzberger − *Department of Chemistry and Applied Biosciences, ETH Zurich, Zurich 8093, Switzerland*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jcim.2c00079

Author Contributions

[+]M.M. and F.G. contributed equally to this work. F.G., M.M., and G.S. conceived the study. M.M. and F.G. designed the methodology, and M.M. implemented the software. All authors analyzed the results and contributed to the writing of the manuscript.

Notes

The authors declare the following competing financial interest(s): G.S. declares a potential financial conflict of interest as he is a consultant to the pharmaceutical industry and a co-founder of inSili.com GmbH, Zurich, Switzerland. No other potential conflicts of interest are declared.

■ REFERENCES

(1) Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Netw.* **2015**, *61*, 85−117.

(2) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inf.* **2018**, *37*, 1700153.

(3) Grisoni, F.; Huisman, B. J. H.; Button, A. L.; Moret, M.; Atz, K.; Merk, D.; Schneider, G. Combining Generative Artificial Intelligence and on-Chip Synthesis for de Novo Drug Design. *Sci. Adv.* **2021**, *7*, eabg3338.

(4) Tang, B.; Ewalt, J.; Ng, H.-L. Generative AI Models for Drug Discovery. In *Topics in Medicinal Chemistry*; Springer: Berlin, Heidelberg, 2021; pp. 1−23.

(5) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1241−1250.

(6) Putin, E.; Asadulaev, A.; Ivanenkov, Y.; Aladinskiy, V.; Sanchez-Lengeling, B.; Aspuru-Guzik, A.; Zhavoronkov, A. Reinforced Adversarial Neural Computer for de Novo Molecular Design. *J. Chem. Inf. Model.* **2018**, *58*, 1194−1204.

(7) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-Novo Design through Deep Reinforcement Learning. *Aust. J. Chem.* **2017**, *9*, 48.

(8) Born, J.; Manica, M.; Oskooei, A.; Cadow, J.; Markert, G.; Martínez, M. R. PaccMannRL: De Novo Generation of Hit-like Anticancer Molecules from Transcriptomic Data via Reinforcement Learning. *iScience* **2021**, *24*, 102269.

(9) De Cao, N.; Kipf, T. MolGAN: An Implicit Generative Model for Small Molecular Graphs. *Preprint at* http://arxiv.org/abs/1805.11973, 2018.

(10) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A. Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. *Nat. Biotechnol.* **2019**, *37*, 1038−1040.

(11) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361*, 360−365.

(12) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.

(13) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. *Preprint at* http://arxiv.org/abs/1802.04364, 2018.

(14) 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.

(15) Skalic, M.; Jiménez, J.; Sabbadin, D.; De Fabritiis, G. Shape-Based Generative Modeling for de Novo Drug Design. *J. Chem. Inf. Model.* **2019**, *59*, 1205−1214.

(16) Kang, S.; Cho, K. Conditional Molecular Design with Deep Generative Models. *J. Chem. Inf. Model.* **2019**, *59*, 43−52.

(17) Winter, R.; Montanari, F.; Steffen, A.; Briem, H.; Noé, F.; Clevert, D.-A. Efficient Multi-Objective Molecular Optimization in a Continuous Latent Space. *Chem. Sci.* **2019**, *10*, 8016−8024.

(18) Gao, K.; Nguyen, D. D.; Tu, M.; Wei, G.-W. Generative Network Complex for the Automated Generation of Drug-like Molecules. *J. Chem. Inf. Model.* **2020**, *60*, 5682−5698.

(19) Sattarov, B.; Baskin, I. I.; Horvath, D.; Marcou, G.; Bjerrum, E. J.; Varnek, A. De Novo Molecular Design by Combining Deep Autoencoder Recurrent Neural Networks with Generative Topographic Mapping. *J. Chem. Inf. Model.* **2019**, *59*, 1182−1196.

(20) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120−131.

(21) Yuan, W.; Jiang, D.; Nambiar, D. K.; Liew, L. P.; Hay, M. P.; Bloomstein, J.; Lu, P.; Turner, B.; Le, Q.-T.; Tibshirani, R.; Khatri, P.; Moloney, M. G.; Koong, A. C. Chemical Space Mimicry for Drug Discovery. *J. Chem. Inf. Model.* **2017**, *57*, 875−882.

(22) Salakhutdinov, R. Learning Deep Generative Models. *Annu. Rev. Stat. Appl.* **2015**, *2*, 361−385.

(23) Walters, W. P. Virtual Chemical Libraries. *J. Med. Chem.* **2019**, *62*, 1116−1124.

(24) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(25) Moret, M.; Friedrich, L.; Grisoni, F.; Merk, D.; Schneider, G. Generative Molecular Design in Low Data Regimes. *Nat. Mach. Intell.* **2020**, *2*, 171−180.

(26) Skinnider, M.; Wang, F.; Pasin, D.; Greiner, R.; Foster, L.; Dalsgaard, P.; Wishart, D. S. A Deep Generative Model Enables Automated Structure Elucidation of Novel Psychoactive Substances. *Preprint at* https://chemrxiv.org/articles/preprint/A_Deep_Generative_Model_Enables_Automated_Structure_Elucidation_of_Novel_Psychoactive_Substances/14644854/1, 2021.

(27) Li, Y.; Zhang, L.; Liu, Z. Multi-Objective de Novo Drug Design with Conditional Graph Generative Model. *Aust. J. Chem.* **2018**, *10*, 33.

(28) Zhou, Z.; Kearnes, S.; Li, L.; Zare, R. N.; Riley, P. Optimization of Molecules via Deep Reinforcement Learning. *Sci. Rep.* **2019**, *9*, 10752.

(29) Skinnider, M. A.; Stacey, R. G.; Wishart, D. S.; Foster, L. J. Chemical Language Models Enable Navigation in Sparsely Populated Chemical Space. *Nat. Mach. Intell.* **2021**, *3*, 759−770.

(30) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front. Pharmacol.* **2020**, *11*, 565644.

(31) Flam-Shepherd, D.; Zhu, K.; Aspuru-Guzik, A. Keeping It Simple: Language Models Can Learn Complex Molecular Distributions. *Preprint at* http://arxiv.org/abs/2112.03041, 2021.

(32) 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.

(33) 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.

(34) Gupta, A.; Müller, A. T.; Huisman, B. J. H.; Fuchs, J. A.; Schneider, P.; Schneider, G. Generative Recurrent Networks for de Novo Drug Design. *Mol. Inf.* **2018**, *37*, 1700111.

(35) 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.* **2018**, *1*, 68.

(36) 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.* **2021**, *60*, 19477−19482.

(37) Manning, C.; Schutze, H. *Foundations of Statistical Natural Language Processing*; MIT Press: Cambridge, MA, 1999.

(38) Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models Are Unsupervised Multitask Learners. *OpenAI blog* **2019**, *1*, 9.

(39) Hu, J.; Gauthier, J.; Qian, P.; Wilcox, E.; Levy, R. P. A Systematic Assessment of Syntactic Generalization in Neural Language Models. *Preprint at* http://arxiv.org/abs/2005.03692, 2020.

(40) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735−1780.

(41) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Others. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945−D954.

(42) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(43) Gobbi, A.; Poppinger, D. Genetic Optimization of Combinatorial Libraries. *Biotechnol. Bioeng.* **1998**, *61*, 47−54.

(44) Wilt, C. M.; Thayer, J. T.; Ruml, W. A Comparison of Greedy Search Algorithms. In *Third Annual Symposium on Combinatorial Search*; 2010; pp. 129−136.

(45) Arús-Pous, J.; Johansson, S. V.; Prykhodko, O.; Bjerrum, E. J.; Tyrchan, C.; Reymond, J.-L.; Chen, H.; Engkvist, O. Randomized SMILES Strings Improve the Quality of Molecular Generative Models. *Aust. J. Chem.* **2019**, *11*, 71.

(46) Moret, M.; Grisoni, F.; Brunner, C.; Schneider, G. Leveraging Molecular Structure and Bioactivity with Chemical Language Models for Drug Design. *Preprint at* https://chemrxiv.org/engage/api-gateway/chemrxiv/assets/orp/resource/item/615580ced1fc334326f9356e/original/leveraging-molecular-structure-and-bioactivity-with-chemical-language-models-for-drug-design.pdf.

(47) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Marañón, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47*, D930−D940.

(48) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *Preprint at* http://arxiv.org/abs/1412.6980, 2014.