Generative Recurrent Networks for *De Novo* Drug Design – Review

IBN-ML-Study

Week 13 2019/03/31

Eugene Bang

Generative Recurrent Networks for *De Novo* Drug Design

- Introduction
- Methods
  - Datasets
  - Model Structure
  - Model Training and Sampling
  - Fine-tuning for Specific Ligand
  - Fragment Growing Procedure
  - Technical Implementation
- Results and Discussion
  - Molecular Structure Generation
  - Target-specific Fine-tuning
  - Fragment-growing
  - Low-data Drug Design
- Conclusion and Outlook

Introduction

Methods

Datasets

Model Structure

Model Training and Sampling

Fine-tuning for Specific Ligand

Fragment Growing Procedure

Technical Implementation

Results and Discussion

Molecular Structure Generation

Target-specific Fine-tuning

Fragment-growing

Low-data Drug Design

Conclusion and Outlook

- "Exploring this vast chemical space for such compounds which may not have been synthesized before"

- "Here, we present a generative deep learning model based on RNN for de novo drug design"; 3 main use cases :
  - Generating libraries for **high-throughput screening**
  - **Hit-to-lead optimization**
  - Fragment-based hit discovery

- "RNNs based on **LSTM**(long short-term memory) cells have been used to predict protein function from sequence …"

Introduction
Methods
Datasets
Model Structure
Model Training and Sampling
Fine-tuning for Specific Ligand
Fragment Growing Procedure
Technical Implementation
Results and Discussion
Molecular Structure Generation
Target-specific Fine-tuning
Fragment-growing
Low-data Drug Design
Conclusion and Outlook

[Stages of Drug Discovery]

- **Target selection and validation.**
- **Hit discovery.** Screen millions of library compounds to uncover novel activities. *High throughput screening (HTS)* heavily relies on laboratory automation and robotics. In analogy to that, *virtual high throughput screening (VHTS)* aims at a similar screening procedure but using software only
- **Hit to lead.** Limited optimization of promising hits to increase affinity. Because (V)HTS data is huge but of low accuracy, any potential hits has to be confirmed, ideally using multiple independent types of assays.
- **Lead optimization.** Further directed optimization to generate viable drug candidates. Note that target affinity is only one of several factors that decide if a compound can become a practically viable medicine. Other factors are *pharmacodynamics* (biochemical and physical effects of drugs on the living organism) and *pharmacokinetics* (how the living organism acts on the drug). You might also see the acronym *ADMET:* absorption, distribution, metabolism, and excretion, toxicity. Note that toxicity means that the drug binds to targets other than the intended ones, i.e., low *selectivity*.

https://medium.com/@stefan.schroedl/machine-learning-for-drug-discovery-in-a-nutshell-part-i-24ae3f65c135

Introduction

Methods

Datasets

Model Structure

Model Training and Sampling

Fine-tuning for Specific Ligand

Fragment Growing Procedure

Technical Implementation

Results and Discussion

Molecular Structure Generation

Target-specific Fine-tuning

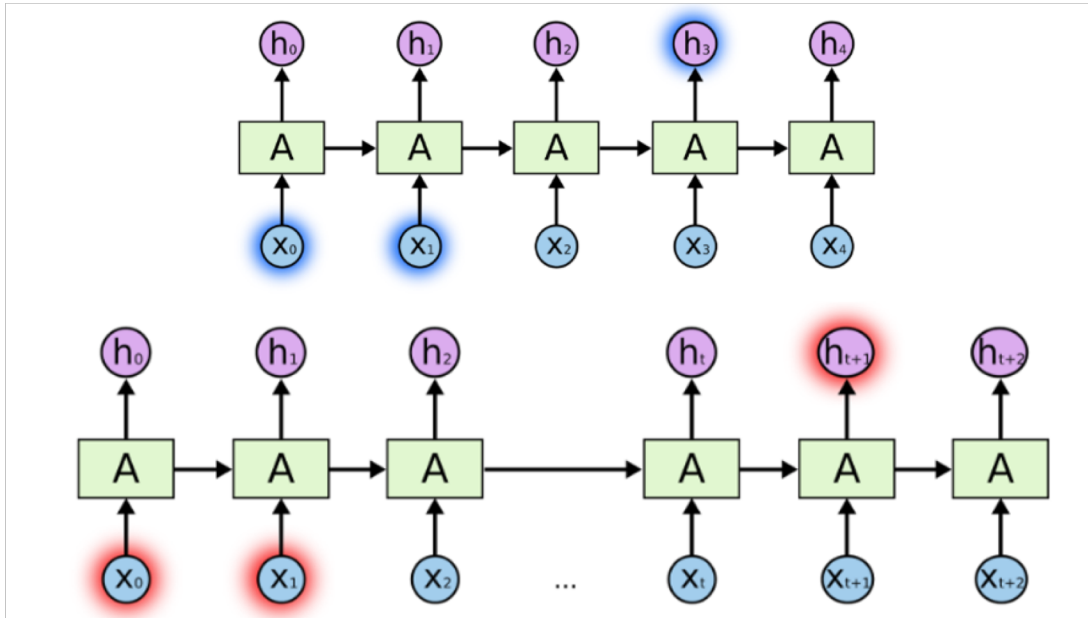
Fragment-growing

Low-data Drug Design

Conclusion and Outlook

LSTM(long short-term memory) cells

- RNN의 Vanishing Gradient Problem
  - RNN은 관련 정보와 그 정보를 사용하는 지점 사이 거리가 멀 경우 역전파시 그래디언트가 점차 줄어 학습능력이 크게 저하됨



<관련 정보와 그 정보를 사용하는 지점 사이 거리가 멀 경우 RNN 학습능력 저하>

https://ratsgo.github.io/natural%20language%20processing/2017/03/09/rnnlstm/

Introduction

Methods

Datasets

Model Structure

Model Training and Sampling

Fine-tuning for Specific Ligand

Fragment Growing Procedure

Technical Implementation

Results and Discussion

Molecular Structure Generation

Target-specific Fine-tuning

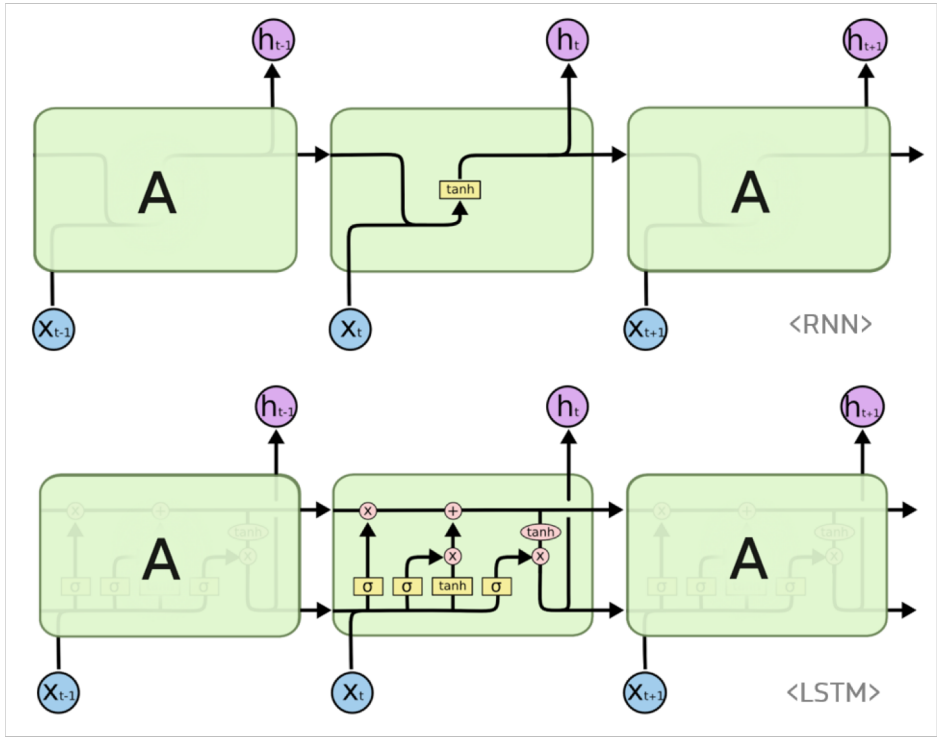
Fragment-growing

Low-data Drug Design

Conclusion and Outlook

LSTM(long short-term memory) cells

- LSTM : Vanishing Gradient Problem 극복을 위해 고안
  - RNN의 hidden state에 cell-state를 추가한 구조



https://ratsgo.github.io/natural%20language%20processing/2017/03/09/rnnlstm/

Introduction

Methods

Datasets

Model Structure

Model Training and Sampling

Fine-tuning for Specific Ligand

Fragment Growing Procedure

Technical Implementation

Results and Discussion

Molecular Structure Generation

Target-specific Fine-tuning

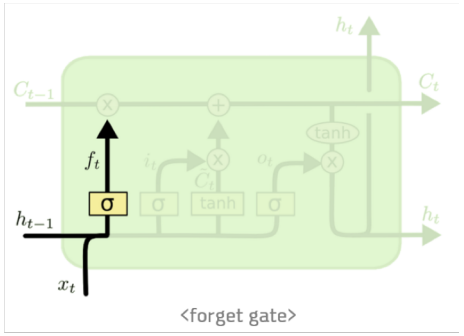
Fragment-growing

Low-data Drug Design

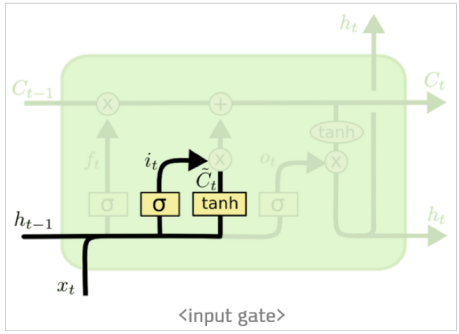
Conclusion and Outlook

LSTM(long short-term memory) cells

- LSTM 의 구조
  - **forget gate** : '과거 정보를 잊기'를 위한 게이트
    - ht−1과 xt를 받아 시그모이드를 취해준 값
    - 출력 값이 0이라면 이전 상태의 정보는 잊고,
    - 1이라면 이전 상태의 정보를 온전히 기억
  - **input gate** it⊙gt : '현재 정보 기억'을 위한 게이트
    - ht−1과 xt를 받아 시그모이드를 취하고,
    - 또 같은 입력으로 하이퍼볼릭탄젠트를 취해준 다음 Hadamard product 연산을 한 값



<forget gate>



<input gate>

https://ratsgo.github.io/natural%20language%20processing/2017/03/09/rnnlstm/

Introduction

Methods

Datasets

Model Structure

Model Training and Sampling

Fine-tuning for Specific Ligand

Fragment Growing Procedure

Technical Implementation

Results and Discussion

Molecular Structure Generation

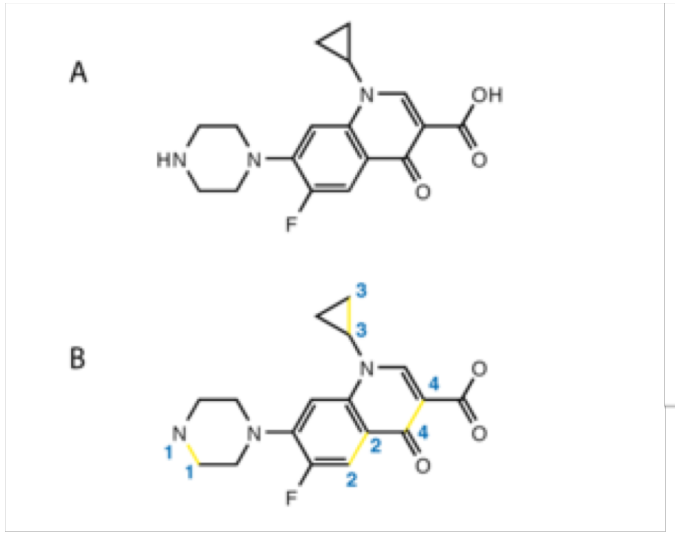
Target-specific Fine-tuning

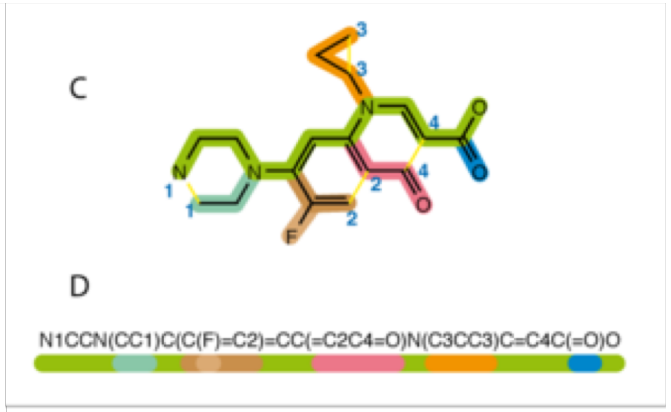
Fragment-growing

Low-data Drug Design

Conclusion and Outlook

- "RNNs can be employed to generate canonical SMILES strings, and can be fine-tuned by transfer learning"
  - **SMILES**(Simplified Molecular-Input Line-Entry System) : specification in the form of a line notation for describing the structure of chemical species using short ASCII strings.





Introduction

Methods

Datasets

Model Structure

Model Training and Sampling

Fine-tuning for Specific Ligand

Fragment Growing Procedure

Technical Implementation

Results and Discussion

Molecular Structure Generation

Target-specific Fine-tuning

Fragment-growing

Low-data Drug Design

Conclusion and Outlook

- "A new approach to de novo drug design using RNN"
  - First part : Model Training
    - Training an LSTM-based RNN model to generate valid SMILES strings
    - Use transfer learning to fine-tune model to generate molecules that are similar to particular class of drugs
      - Especially low-data situations in early-phase drug design
  - Second part : Compound design by sampling
    - Fragment-based drug discovery, 'Fragment-growing'
      - Growing a library of leads starting from a known active fragment

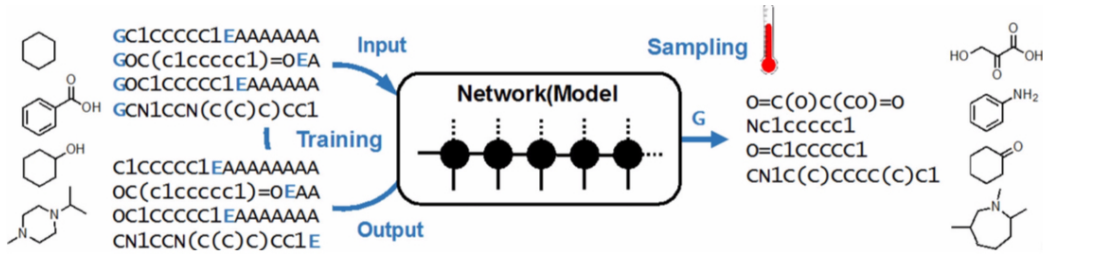


Figure 1. Schematic of model training (left) and compound design by sampling (right).

Introduction

Methods

Datasets

Model Structure

Model Training and Sampling

Fine-tuning for Specific Ligand

Fragment Growing Procedure

Technical Implementation

Results and Discussion

Molecular Structure Generation

Target-specific Fine-tuning

Fragment-growing

Low-data Drug Design

Conclusion and Outlook

- “Our … model provides A fresh concept of
  - generating general compound libraries,
  - target-specific libraries and
  - focus libraries for fragment-based drug discovery”

Introduction

Methods

Datasets

Model Structure

Model Training and Sampling

Fine-tuning for Specific Ligand

Fragment Growing Procedure

Technical Implementation

Results and Discussion

Molecular Structure Generation

Target-specific Fine-tuning

Fragment-growing

Low-data Drug Design

Conclusion and Outlook

- 677,044 SMILES strings from ChEMBL22 (www.ebi.ac.uk/chembl)
  - Annotated nanomolar activities(Kd/i/B, IC/EC50)
  - Then pre-processed to remove
    - Duplicates, salts and stereochemical information
    - Nucleic acids and long peptides
- Finally trained on 541,555 SMILES
  - 34-73 SMILES characters(tokens)

Introduction

Methods

Datasets

Model Structure

Model Training and Sampling

Fine-tuning for Specific Ligand

Fragment Growing Procedure

Technical Implementation

Results and Discussion

Molecular Structure Generation

Target-specific Fine-tuning

Fragment-growing

Low-data Drug Design

Conclusion and Outlook

- "LSTMs possess an input gate, a forget gate, and an output gate. Accordingly, LSTMs are able to specifically control what information passes to the next cell through the hidden state hi."
- "In this way, LSTMs solve the vanishing- or exploding-gradient-problem that RNNs experience due to backpropagation over long sequences"
  - Citation : "Recurrent Neural Networks With Auxiliary Memory Units" J.Wang, L.Zhang, Q.Guo, Z.Yi, IEEE Trans. Neural Netw. Learn. Syst. 2017, doi: 10.1109/TNNLS.2017.2677968
- "Typically, the RNN aims to predict the next roken of a given input"

Introduction

Methods

Datasets

Model Structure

Model Training and Sampling

Fine-tuning for Specific Ligand

Fragment Growing Procedure

Technical Implementation

Results and Discussion

Molecular Structure Generation

Target-specific Fine-tuning

Fragment-growing

Low-data Drug Design

Conclusion and Outlook

- The Structure of the model
  - Two LSTM layers with hidden state vector of size 256, regularized w dropout
  - Followed by dense output layer and softmax activation function
  - Input is one-hot encoded sequence
    - Start : 'G' token / End : 'E' token / Padding : 'A' token

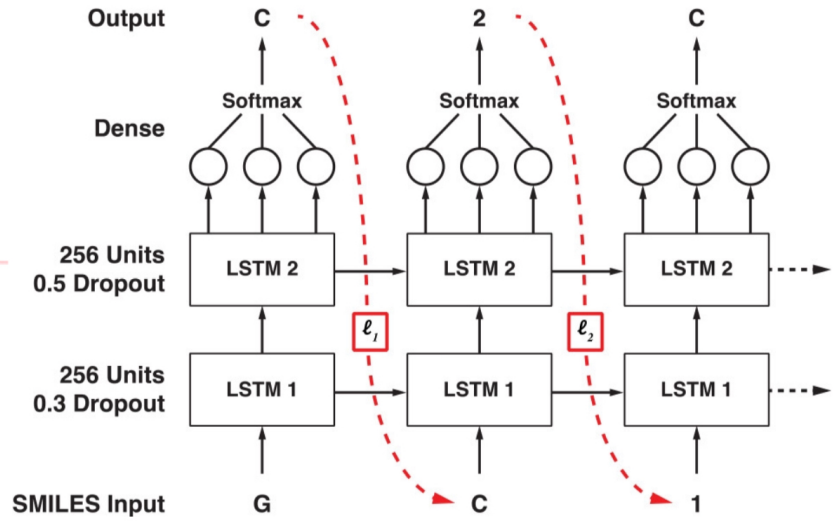


Figure 2. Model of the RNN–LSTM producing SMILES strings, token by token. The token 'G' denotes "GO" at the beginning of the SMILES string. During training, the model predicts the next token for each input token in the sequence. The loss L is calculated at each position as the categorical cross-entropy between the predicted and actual next token.

Introduction

Methods

Datasets

Model Structure

Model Training and Sampling

Fine-tuning for Specific Ligand

Fragment Growing Procedure

Technical Implementation

Results and Discussion

Molecular Structure Generation

Target-specific Fine-tuning

Fragment-growing

Low-data Drug Design

Conclusion and Outlook

- 2 Methods for training the RNN
- 1st method
  - Break each input into some length *l*, and predict the next token
  - Loss was calculated from the likelihood of the predicted token
- 2nd method
  - Pads every input string to *n* tokens(*n* = longest string length)
  - Model predicts the next token in the sequence
  - Loss is averaged over all the target token in all molecules
  - When sampling :
    - Fed the RNN with only token'G' and concatenated the outputs until 'E' was produced
    - Introduced an additional temperature parameter into the softmax function
      - Higher Temp = greater structural diversity/decrease the chemical validity
      - Lower Temp = lower structural diversity/more conservative("safer") predictions

Introduction

Methods

Datasets

Model Structure

Model Training and Sampling

Fine-tuning for Specific Ligand

Fragment Growing Procedure

Technical Implementation

Results and Discussion

Molecular Structure Generation

Target-specific Fine-tuning

Fragment-growing

Low-data Drug Design

Conclusion and Outlook

- 2 Methods for training the RNN
- 2nd method

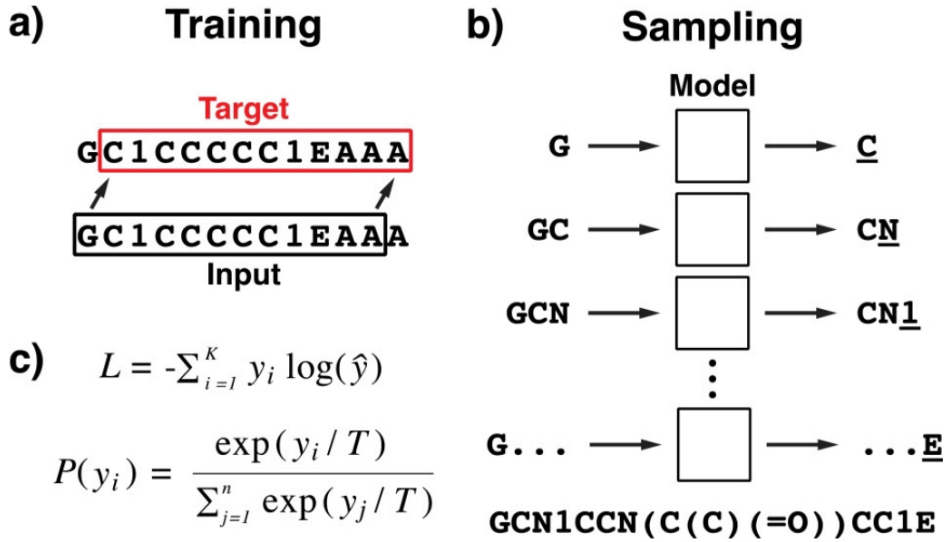


Figure 3. A) The training procedure for the final LSTM model. Each molecule was padded to the length *n* of the longest SMILES string (padding denoted by the token 'A'). The first *n*-1 characters were taken as the input, and the last *n*-1 characters were the target. B) Sampling procedure. The sentinel token 'G' was given to start. At every step of sampling, the last sampled character is taken as the next character in the generated sequence. Sampling continues until the token 'E' denoting "end of sequence" is generated. C) Equations for the calculation of the loss error L, and the *softmax* function $P(y_i)$ with temperature factor *T*.

Introduction

Methods

Datasets

Model Structure

Model Training and Sampling

Fine-tuning for Specific Ligand

Fragment Growing Procedure

Technical Implementation

Results and Discussion

Molecular Structure Generation

Target-specific Fine-tuning

Fragment-growing

Low-data Drug Design

Conclusion and Outlook

- "Fine-tuning the model by further training on smaller subsets of selected compounds"
  - Produce SMILES strings with higher similarity to these target-focused datasets. (T = 0.75)
    - i) 4367 peroxisome proliferator-activated receptor gamma(PPARr) inhibitors – 5 epochs
    - ii) 1490 trypsin inhibitors – 5 epochs
    - iii) five structurally diverse transient receptor potential M8(TRPM8) blockers – 12 epochs
      - Total dataset(448 compounds) clustered by **Tanimoto similarity(MACCS keys)**
      - Yielding 5 clusters with distinct scaffolds; most active compound from each cluster was chosen

- After training, 100 molecules were generated and measured the average Tanimoto similarity between sample and training data.
  - "After every epoch of fine-tuning, the software tool provides the user with the percentage of duplicates in the molecules generated, and how similar the generated molecules are to the provided subset.*"(- Conclusion)*

Introduction

Methods

Datasets

Model Structure

Model Training and Sampling

Fine-tuning for Specific Ligand

Fragment Growing Procedure

Technical Implementation

Results and Discussion

Molecular Structure Generation

Target-specific Fine-tuning

Fragment-growing

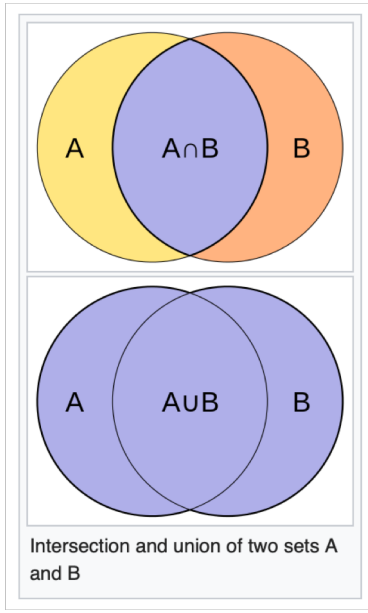
Low-data Drug Design

Conclusion and Outlook

- Tanimoto similarity(Jaccard index)

"The Jaccard index, also known as Intersection over Union and the Jaccard similarity coefficient (originally coined coefficient de communauté by Paul Jaccard), is a statistic used for comparing
the similarityand diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of
the intersection divided by the size of the union of the sample sets"

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$



Intersection and union of two sets A and B

- MACCS keys

The MACCS 166 keys are one of the mainstay fingerprints of cheminformatics, especially regarding molecular similarity. It's rather odd, really, since they were developed for substructure screening and not similarity. I suppose that Jaccard would agree that any relatively diverse feature vector can likely be used to measure similarity, whether it be Alpine biomes or chemical structures.

http://www.dalkescientific.com/writings/diary/archive/2014/10/17/maccs_key_44.html

Introduction

Methods

Datasets

Model Structure

Model Training and Sampling

Fine-tuning for Specific Ligand

Fragment Growing Procedure

Technical Implementation

Results and Discussion

Molecular Structure Generation

Target-specific Fine-tuning

Fragment-growing

Low-data Drug Design

Conclusion and Outlook

- “we allowed the user to enter a fragment which they wish to be present in all SMILES generated”
  - Especially tested the case where the fragment was at one end of the molecule, and provided exit vectors for the model to build upon

Introduction

Methods

Datasets

Model Structure

Model Training and Sampling

Fine-tuning for Specific Ligand

Fragment Growing Procedure

Technical Implementation

Results and Discussion

Molecular Structure Generation

Target-specific Fine-tuning

Fragment-growing

Low-data Drug Design

Conclusion and Outlook

- Tensorflow v1.2
- Keras v2.0
- Python v3.6
- SMILEs string validity and molecular feature calculation in RDkit
- iPython Notebook
- PCA performed using scikit-learn libraries
- MOE(Molecular Operating Environment)

Introduction
Methods
Datasets
Model Structure
Model Training and Sampling
Fine-tuning for Specific Ligand
Fragment Growing Procedure
Technical Implementation
Results and Discussion
Molecular Structure Generation
Target-specific Fine-tuning
Fragment-growing
Low-data Drug Design
Conclusion and Outlook

- Model 1 vs Model 2

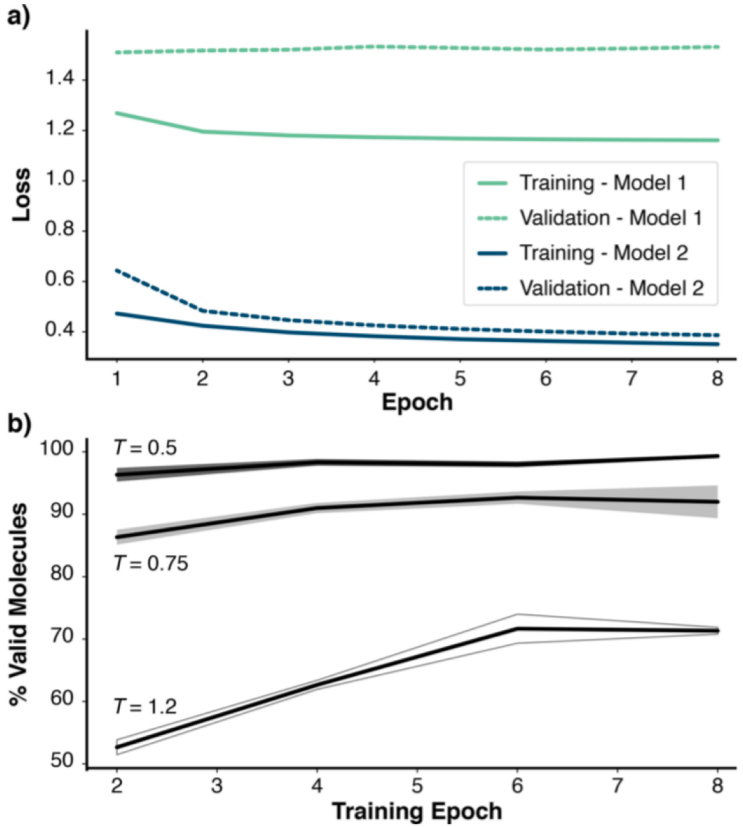


Figure 4. A) Training and Validation loss for Model 1 vs. Model 2. Model 2 was selected as the final model, and its sampling procedure is shown in Figure 2b. B) Percentage of valid molecules generated by Model 2 at different sampling temperatures (*T*). As the temperature increases, the percentage of valid molecules decreases.

Introduction
Methods
Datasets
Model Structure
Model Training and Sampling
Fine-tuning for Specific Ligand
Fragment Growing Procedure
Technical Implementation
Results and Discussion
Molecular Structure Generation
Target-specific Fine-tuning
Fragment-growing
Low-data Drug Design
Conclusion and Outlook

- Sampled SMILES vs original molecules
  - 24 common physiochemical features
  - PCA(Principal component analysis)
    - first two principal componenets(PC1, PC2) were selected
  - Result : generated molecules lie in the same space as the original molecules

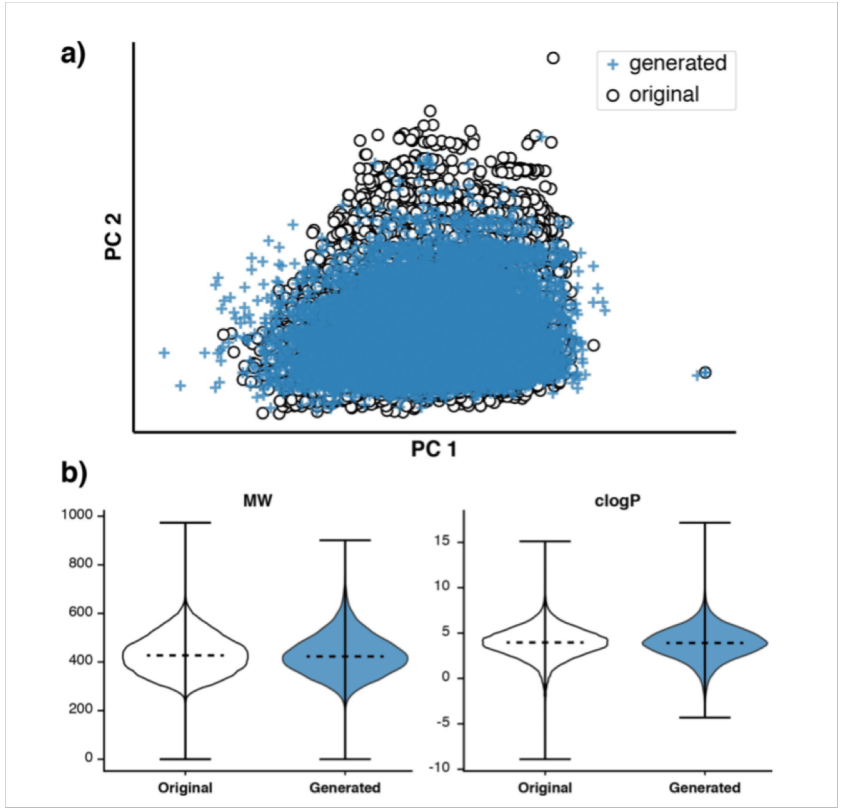


Figure 5. A set of 25,923 valid SMILES strings was generated from the trained Model 2, and 24 physiochemical features were calculated for the generated virtual molecules and the set of 550,000 original training molecules. A) PCA was performed on these 24 generated features from the training molecules, and the first two principal components (PC1, PC2) were selected. The coordinates of the generated molecules were transformed accordingly. We see overlap in the chemical subspace between these two sets of molecules. B) Violin–plots for molecular weight (MW) and clog*P* distributions, with the medians shown as dashed lines. Visual inspection reveals a close match of the generated and original molecules.

* clogP = calculated logP
(P = partition coefficient)

$$\log P_{\mathrm{oct/wat}} = \log\left(\frac{[\mathrm{solute}]^{\text{un-ionized}}_{\text{octanol}}}{[\mathrm{solute}]^{\text{un-ionized}}_{\text{water}}}\right).$$

Introduction
Methods
Datasets
Model Structure
Model Training and Sampling
Fine-tuning for Specific Ligand
Fragment Growing Procedure
Technical Implementation
Results and Discussion
Molecular Structure Generation
Target-specific Fine-tuning
Fragment-growing
Low-data Drug Design
Conclusion and Outlook

- After fine tuning…
  - 1000 sample molecules generated from PPARr
    - 96% valid SMILES / 90% of the valid were unique / 88% were unique from the known
    - Lie in the same physicochemical subspace as the known PPARr ligands
    - The shift due to fine-tuning : as fine-tuning occurs, generated molecules shift towards the part of the spaces that is most densely populated by known PPARr ligands
      - Tanimoto dissimilarity between the known vs
      - Molecules generated w/o fine tuning : 0.425±0.003
      - Molecules generates w/ fine tuning : 0.375±0.003

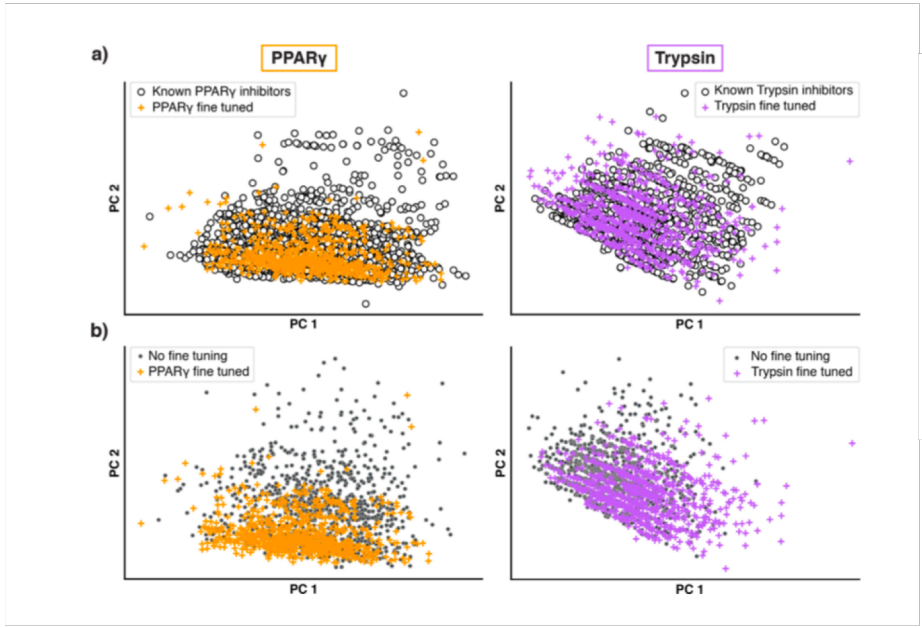


Figure 6. 1000 molecule structures were sampled after fine-tuning on sets of inhibitors of PPARγ (left) and trypsin (right). A) PCA was carried out on 24 physiochemical descriptors and fit to the set of original target inhibitors. The first two principal components (PC1, PC2) were selected for visualization. The molecules generated after RNN model fine-tuning are plotted together with the original ligands from ChEMBL. An analysis of the plots provides an idea of whether the fine-tuned molecules cover the space occupied by the original inhibitors. B) We plot the set of molecules generated without fine-tuning against the set of fine-tuned molecules. The axes are the same principal components as in panel a). We see a clear shift in the distributions of compounds generated with and without fine-tuning.

Introduction
Methods
Datasets
Model Structure
Model Training and Sampling
Fine-tuning for Specific Ligand
Fragment Growing Procedure
Technical Implementation
Results and Discussion
Molecular Structure Generation
Target-specific Fine-tuning
Fragment-growing
Low-data Drug Design
Conclusion and Outlook

- After fine tuning…
  - 1000 sample molecules generated from trypsin inhibitors
    - 93% valid SMILES / 87% of the valid were unique / 93% were unique from the known
    - Lie in the same physicochemical subspace as the known trypsin inhibitors
    - The shift due to fine-tuning : as fine-tuning occurs, generated molecules shift towards the part of the spaces that is most densely populated by known trypsin inhibitoes
      - Tanimoto dissimilarity between the known vs
      - Molecules generated w/o fine tuning : 0.440±0.003
      - Molecules generates w/ fine tuning : 0.409±0.003

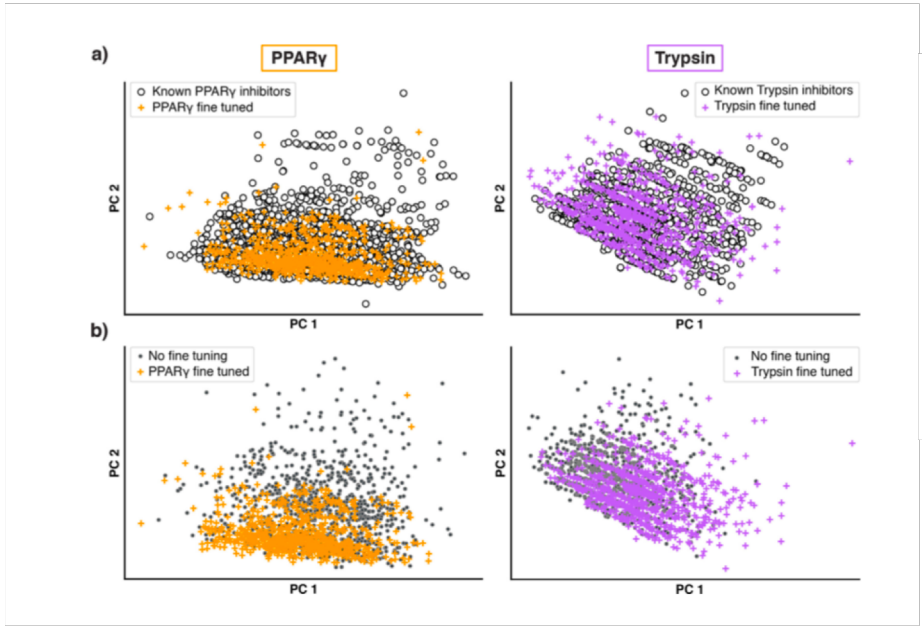


Figure 6. 1000 molecule structures were sampled after fine-tuning on sets of inhibitors of PPARγ (left) and trypsin (right). A) PCA was carried out on 24 physiochemical descriptors and fit to the set of original target inhibitors. The first two principal components (PC1, PC2) were selected for visualization. The molecules generated after RNN model fine-tuning are plotted together with the original ligands from ChEMBL. An analysis of the plots provides an idea of whether the fine-tuned molecules cover the space occupied by the original inhibitors. B) We plot the set of molecules generated without fine-tuning against the set of fine-tuned molecules. The axes are the same principal components as in panel a). We see a clear shift in the distributions of compounds generated with and without fine-tuning.

Introduction
Methods
Datasets
Model Structure
Model Training and Sampling
Fine-tuning for Specific Ligand
Fragment Growing Procedure
Technical Implementation
Results and Discussion
Molecular Structure Generation
Target-specific Fine-tuning
Fragment-growing
Low-data Drug Design
Conclusion and Outlook

- After fine tuning…
  - Selection of the closest/farthest neighbors from one known trypsin inhibitor
  - Limitations
    - Tanimoto similarity gives equal weight to all parts of the molecule
    - For practical applications; must give higher weight to molecules with active fragment

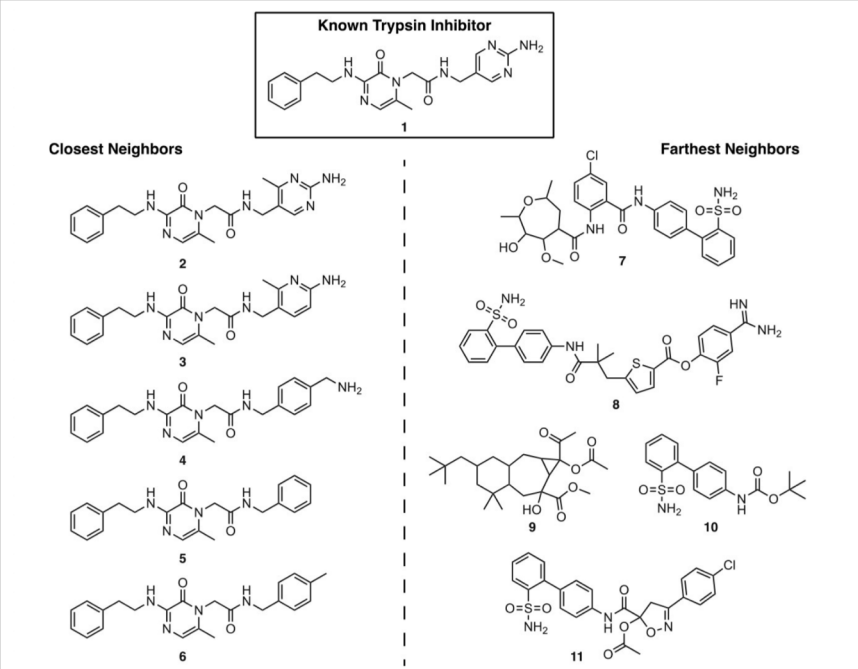


Figure 7. Left: Generated structures (2–6) with highest Tanimoto similarity to the known trypsin inhibitor (1). Right: Generated structures (7–11) with lowest Tanimoto similarity.

Introduction
Methods
Datasets
Model Structure
Model Training and Sampling
Fine-tuning for Specific Ligand
Fragment Growing Procedure
Technical Implementation
Results and Discussion
Molecular Structure Generation
Target-specific Fine-tuning
Fragment-growing
Low-data Drug Design
Conclusion and Outlook

- Conc : useful for hit-to-lead optimization
- “In contrast to other RNN models, ours does not rely on an explicit but limited SMILES vocabulary, which renders this new approach theoretically unlimited with regards to the chemical diversity of the training data.”
- “Model fine-tuning enabled the automated de novo design of target-focused compound sets, without the need of dedicated target prediction tools or other external scoring functions”

Introduction
Methods
Datasets
Model Structure
Model Training and Sampling
Fine-tuning for Specific Ligand
Fragment Growing Procedure
Technical Implementation
Results and Discussion
Molecular Structure Generation
Target-specific Fine-tuning
Fragment-growing
Low-data Drug Design
Conclusion and Outlook

- "One main and novel case of this generative RNN model is in fragment-based drug discovery(FBDD)"
  - "Drug designers might want to start from a fragment known to bind to the target of interest"
  - Thrombin-binding start fragment benzamidine(12)
    - 1000 molecules were generated from the pre-trained RNN model(97% valid)
    - Consequently, our model can be used to generate compound libraries based on a single receptor-binding fragment. Furthermore, this approach can be fine-tuned toward specific scaffolds or proprietary scaffold-centric compound libraries

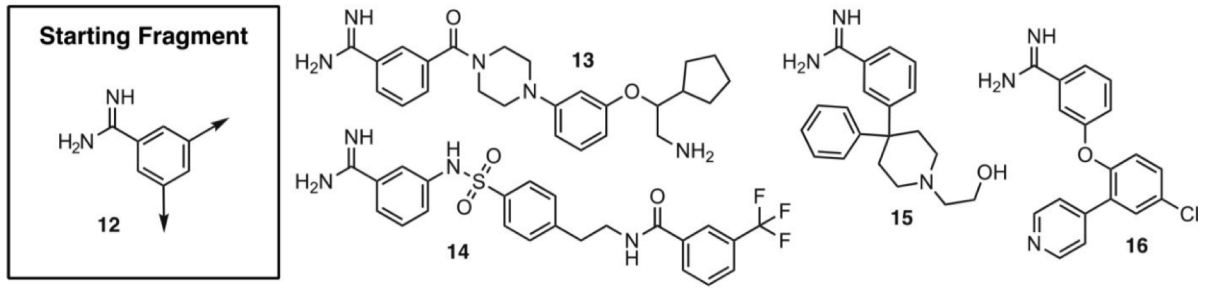


Figure 8. Molecules 13–16 were generated starting from two exit vectors (arrows) of the boxed fragment (benzamidine, 12). 12 has been shown to bind to the thrombin active site and may hence be well-suited for fragment-growing. 1000 molecules were sampled, and a collection of *de novo* generated molecules is shown on the right.

Introduction
Methods
Datasets
Model Structure
Model Training and Sampling
Fine-tuning for Specific Ligand
Fragment Growing Procedure
Technical Implementation
Results and Discussion
Molecular Structure Generation
Target-specific Fine-tuning
Fragment-growing
Low-data Drug Design
Conclusion and Outlook

- "For several targets only a few ligands are known. This situation is characteristic for early-stage drug discovery projects, where de novo design may be especially useful."
  - "For this reason, we extended our generative model to the problem of molecular design in the presence of limited training data availability ("low-data").
    - Dataset was small(5 compunds)
    - Compounds were specifically chosen for the diversity of their scaffolds.

Introduction
Methods
Datasets
Model Structure
Model Training and Sampling
Fine-tuning for Specific Ligand
Fragment Growing Procedure
Technical Implementation
Results and Discussion
Molecular Structure Generation
Target-specific Fine-tuning
Fragment-growing
Low-data Drug Design
Conclusion and Outlook

- 100 de novo generated fine-tuned molecules in PCA projection
  - Closely approximate 4 of 5 reference compounds used for fine-tuning
  - One of the training compounds lie far from the cluster of molecules
    - Because training compounds were selected based on their activity

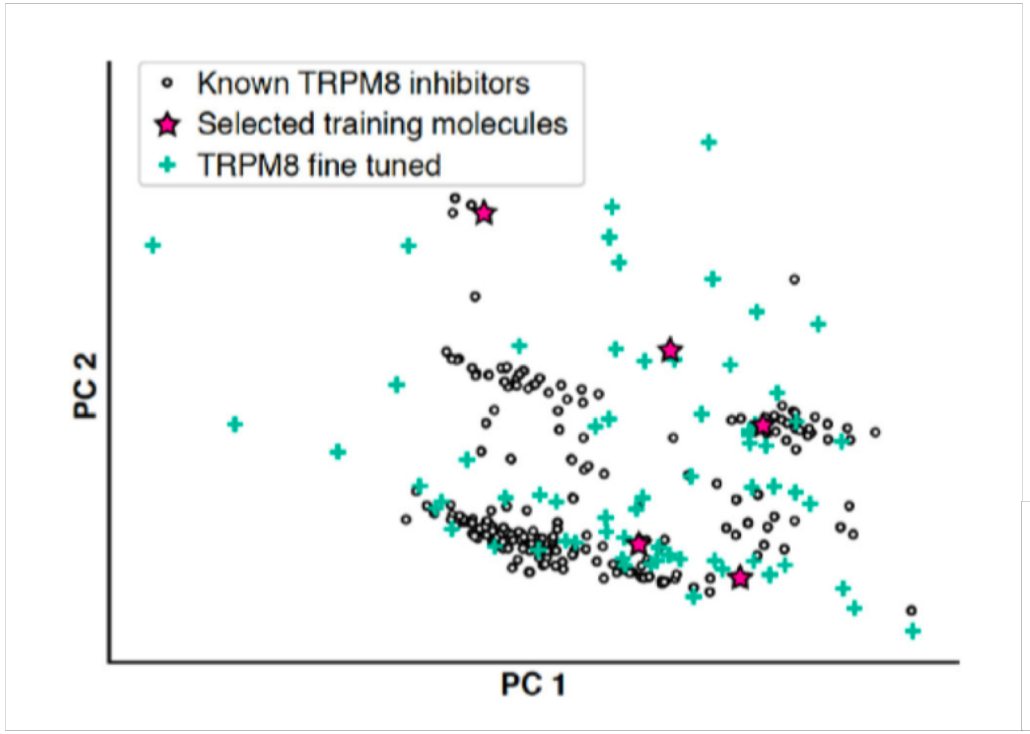


Figure 9. A principal component analysis was conducted on 24 physiochemical features calculated on the full dataset of 448 known TRPM8 inhibitors. The five molecules that were chosen for fine-tuning the model are represented as stars in the coordinate system spanned by the first (PC 1) and second (PC 2) principal components. 100 molecules were sampled, and the positions of the valid molecules are indicated by green crosses.

Introduction
Methods
Datasets
Model Structure
Model Training and Sampling
Fine-tuning for Specific Ligand
Fragment Growing Procedure
Technical Implementation
Results and Discussion
Molecular Structure Generation
Target-specific Fine-tuning
Fragment-growing
Low-data Drug Design
Conclusion and Outlook

- Several generated molecules combine structure motifs from different training ligands

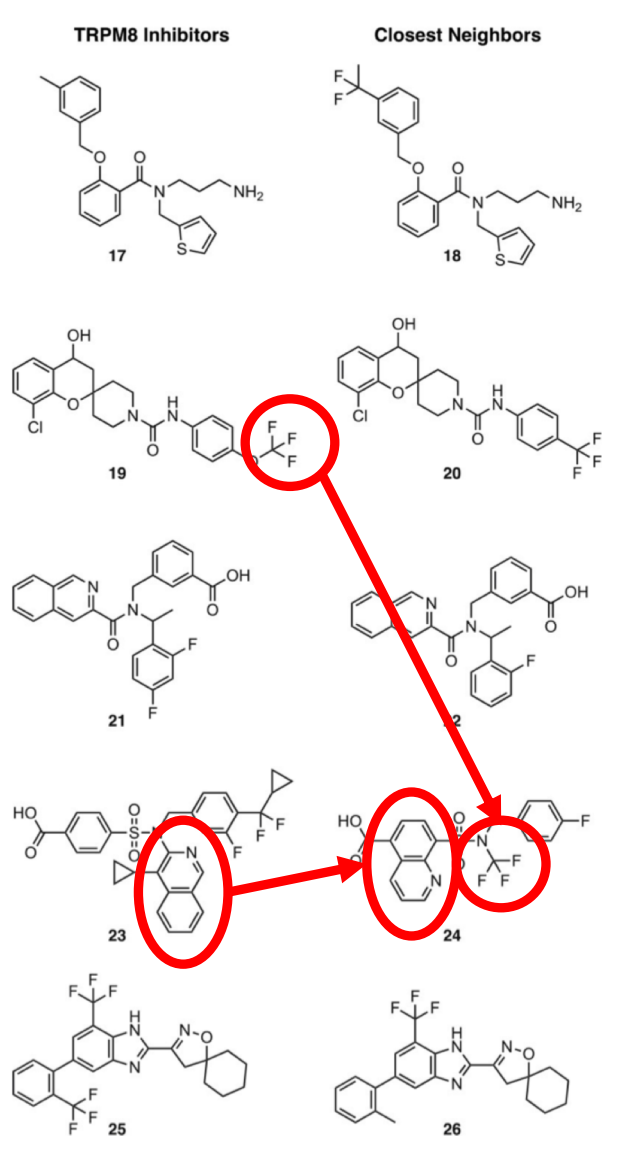


Figure 10. The five TRPM8 inhibitors (17, 19, 21, 23, 25) used to fine-tune the RNN model are shown on the left. The respective generated molecule with highest Tanimoto similarity (nearest neighbor (18, 20, 22, 24, 26)), is shown on the right of every reference compound.

Introduction
Methods
Datasets
Model Structure
Model Training and Sampling
Fine-tuning for Specific Ligand
Fragment Growing Procedure
Technical Implementation
Results and Discussion
Molecular Structure Generation
Target-specific Fine-tuning
Fragment-growing
Low-data Drug Design
Conclusion and Outlook

- Generated a generative RNN-LSTM model that can…
  - (i) generate compound libraries for high-throughput screening,
  - (ii) hit-to-lead optimization for targets, even with a small amount of data
  - (iii) fragment-based drug discovery.
  - The model successfully exhibited transfer learning
    - "warm start" + few epochs of fine tuning for specific subsets
    - fine-tuned samples display significantly higher similarity than the structures generated without fine-tuning
    - Transfer learning was successful, even when only a few ligands were used for fine-tuning.
- "The application of the RNN model to fragment-growing could be useful in several situations. "
  - this approach does not require extensive similarity searching or external scoring. The new molecular structures are generated instantly, which might be attractive for real-time in situ molecular modeling.

Introduction
Methods
Datasets
Model Structure
Model Training and Sampling
Fine-tuning for Specific Ligand
Fragment Growing Procedure
Technical Implementation
Results and Discussion
Molecular Structure Generation
Target-specific Fine-tuning
Fragment-growing
Low-data Drug Design
Conclusion and Outlook

- “Our generative model itself contains fewer parameters than existing models, while achieving the same or improved percentage of valid molecules. “
  - Smaller models are often preferable to larger ones in deep learning because they have a reduced risk of overfitting.

- This present approach does not strictly require an external scoring function for fine-tuning the parameters of the model.
  - optimized the parameters directly from chemical structures with desirable property
  - avoids the risk of potentially error-prone scoring.

Introduction
Methods
Datasets
Model Structure
Model Training and Sampling
Fine-tuning for Specific Ligand
Fragment Growing Procedure
Technical Implementation
Results and Discussion
Molecular Structure Generation
Target-specific Fine-tuning
Fragment-growing
Low-data Drug Design
Conclusion and Outlook

- Downside of this method is :
  - the necessity for available active ligands for parameter optimization.
  - the necessity for model fine-tuning over a particular number of epochs, in order to avoid generating compound duplicates.

- "The necessity for model fine-tuning over a particular number of epochs, in order to avoid generating compound duplicates. "
  - To circumvent this issue; after every epoch of fine-tuning, the software tool provides the user with the percentage of duplicates in the molecules generated, and how similar the generated molecules are to the provided subset.
    - These two quantities are often a trade-off,
    - We currently request the user to make the final decision, rather than applying an arbitrary rule to all fine- tuning sets.

Introduction
Methods
Datasets
Model Structure
Model Training and Sampling
Fine-tuning for Specific Ligand
Fragment Growing Procedure
Technical Implementation
Results and Discussion
Molecular Structure Generation
Target-specific Fine-tuning
Fragment-growing
Low-data Drug Design
Conclusion and Outlook

- "Our approach is useful when combined with some *a priori* knowledge of a specific active fragment that should be kept constant. "
  - However, this current method cannot grow molecules in more than one direction (exit vector) from the start fragment

- "We sought to avoid the need for a **scoring function** based on **synthesizability** that would introduce additional error to the model. "