Chemical Language Modeling with Structured State Space Sequence Models

Code Paper

Rıza Özçelik, Sarah de Ruiter, Emanuele Criscuolo, Francesca Grisoni

Summary

Motivation

- Chemical space is vast.
- Searched molecules are rare.
- Chemical language models design molecules in no time.
- Capturing and mimicking chemical properties is needed.

Idea

Structured state space sequence models[1] (S4s):

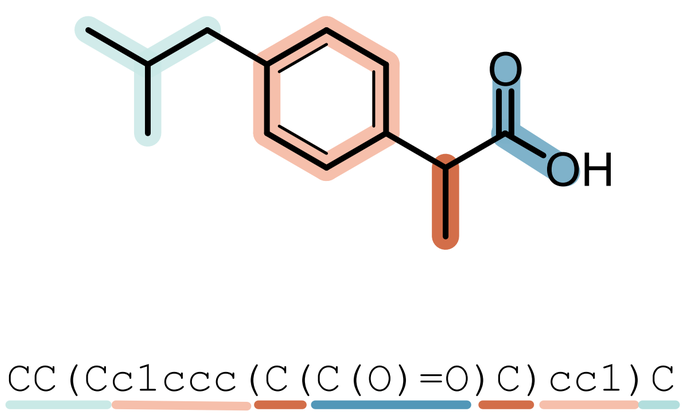
- capture properties with global convolutions;
- design sequences efficiently;
- can mimic syntactic dependencies.

Findings

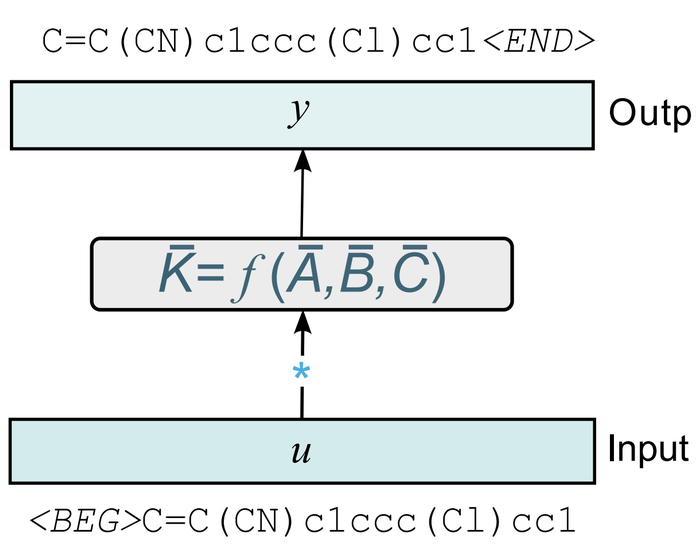
Structured state space sequence models (S4s) can:

- capture global properties better;
- generate complex syntax;
- design MAPK1 inhibitors prospectively.

S4 for Molecule Design

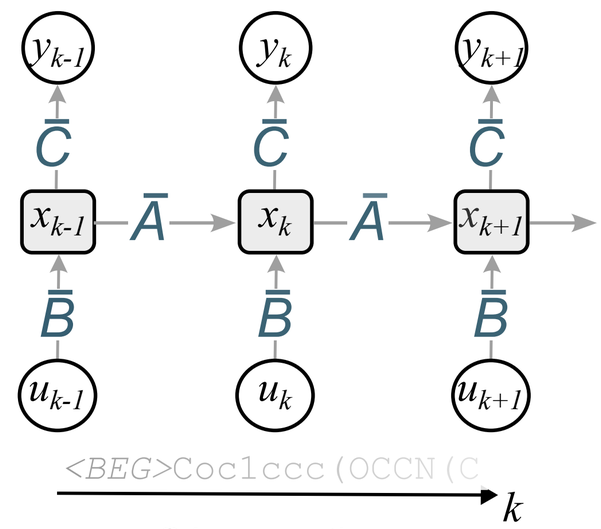


SMILES string



S4 training

Dual formulation

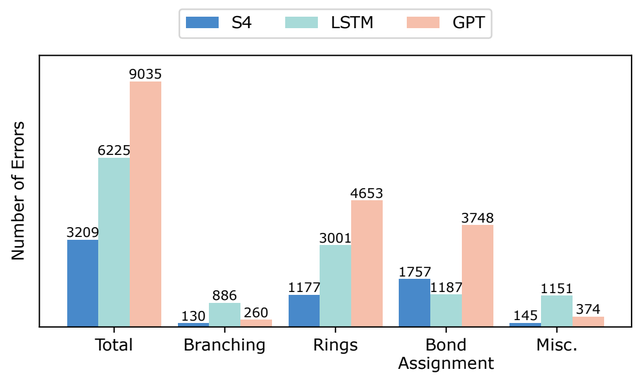


S4 generation

Evaluation

Small molecule design

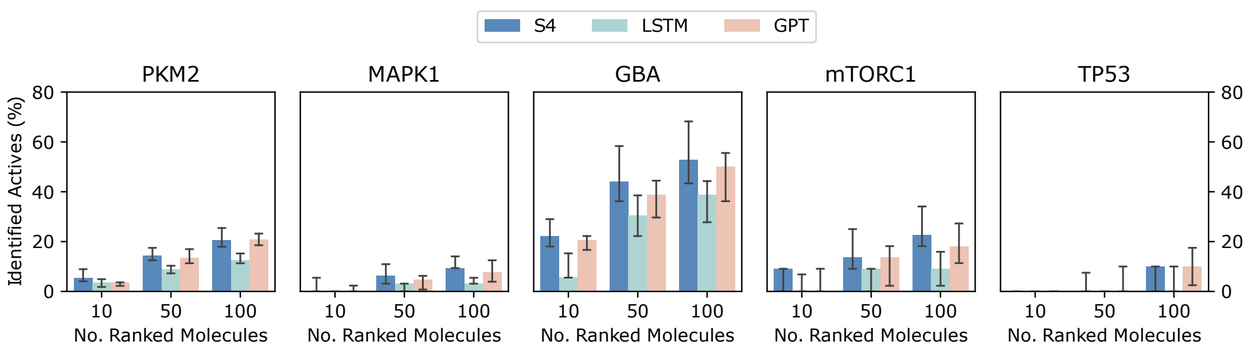
Model	Validity	Uniqueness	Novelty
S4	97%	96%	93%
LSTM	95%	94%	91%
GPT	91%	91%	89%



MOSES benchmark[2]

Model	Valid (%)	Unique@10K (%)	Novelty (%)	IntDiv (↑)	FCD (↓) Test	FCD (↓) TestSF
Train	*n.a*	*n.a*	*n.a*	0.86	0.01	0.48
HMM	7.6	56.7	99.9	0.85	24.5	25.4
NGram	23.8	92.2	96.9	0.87	5.51	6.23
Comb.	100.0	99.1	98.8	0.87	4.24	4.51
CharRNN	97.5	99.9	84.2	0.86	**0.07**	0.52
AAE	93.7	99.7	79.3	0.86	0.56	1.06
VAE	97.7	99.8	69.5	0.86	0.10	0.57
JTN-VAE	**100.0**	100.0	91.4	0.85	0.40	0.94
LatentGAN	89.7	100.0	**95.0**	0.86	0.30	0.83
MD-TF	99.6	99.9	81.6	0.85	0.11	0.51
cMolGPT	98.8	99.9	-	-	-	-
TD-GPT	99.3	99.4	78.1	-	-	-
S4	98.4	100.0	88.1	0.86	0.08	0.43

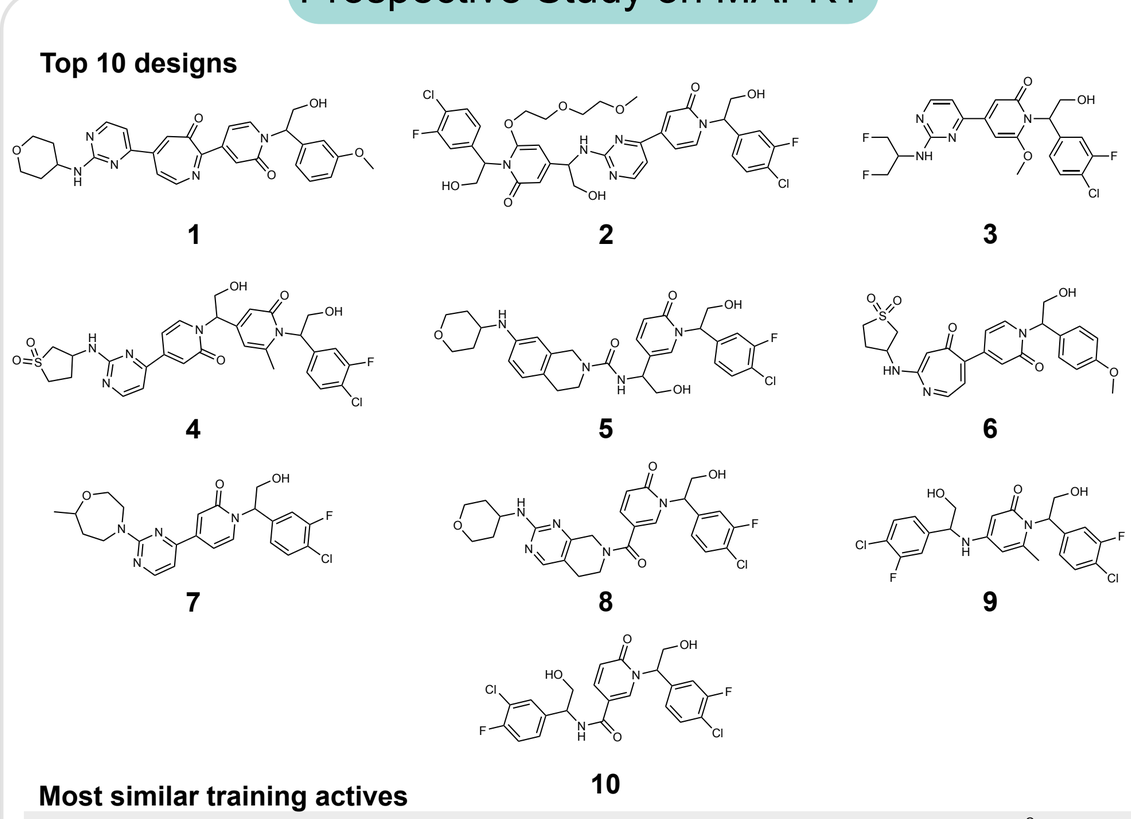
Capturing binding to proteins



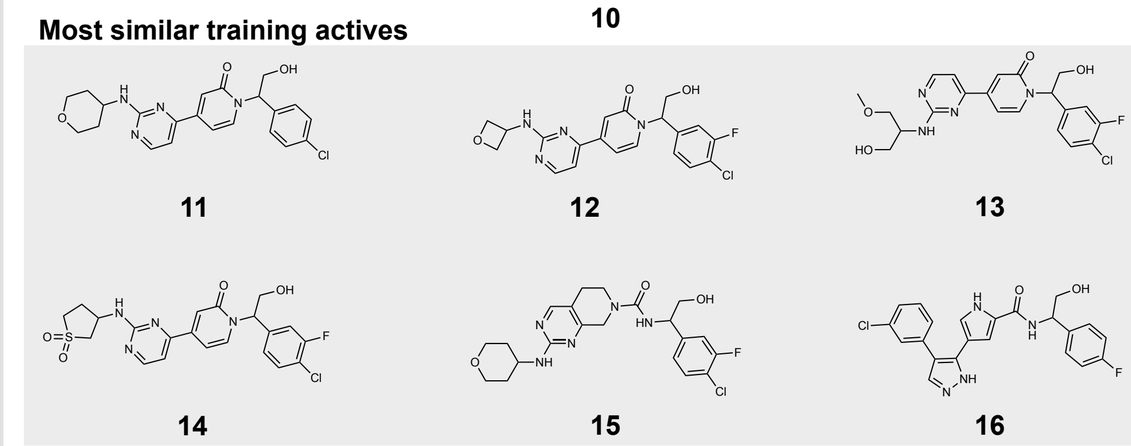
1. Gu, Albert, Karan Goel, and Christopher Re. "Efficiently Modeling Long Sequences with Structured State Spaces." International Conference on Learning Representations.
2. Polykovskiy, Daniil, et al. "Molecular sets (MOSES): a benchmarking platform for molecular generation models." Frontiers in pharmacology 11 (2020): 565644.

Prospective Study on MAPK1

Top 10 designs



Most similar training actives



Molecular dynamics simulations

	S4 design		Most similar training active			Scaffold	Global
ID	ΔG [kcal/mol]	ID	ΔG [kcal/mol]	K_i [nM]	Similarity	Similarity
1	-5.6 ± 0.9	**11**	-9.1 ± 0.8	0.1	79%	65%
2	-23 ± 4	**12**	-12 ± 2	0.4	63%	57%
3	-19.6 ± 0.9	**13**	-10.5 ± 0.7	3.0	100%	65%
4	-13 ± 2	**14**	-11 ± 3	2.5	90%	87%
5	-7 ± 2	**15**	-13 ± 2	0.6	73%	85%
6	-11 ± 3	**14**	-11 ± 3	2.5	56%	56%
7	-10.3 ± 0.6	**11**	-9.1 ± 0.8	0.1	50%	52%
8	-11.2 ± 0.4	**15**	-13 ± 2	0.6	58%	72%
9	-17 ± 2	**13**	-10.5 ± 0.7	3.0	41%	42%
10	-15 ± 2	**16**	-9.1 ± 0.2	63.0	30%	31%

r.ozcelik@tue.nl

MOLECULAR Machine Learning

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY