Exploiting Pretrained Biochemical Language Models for Targeted Drug Design

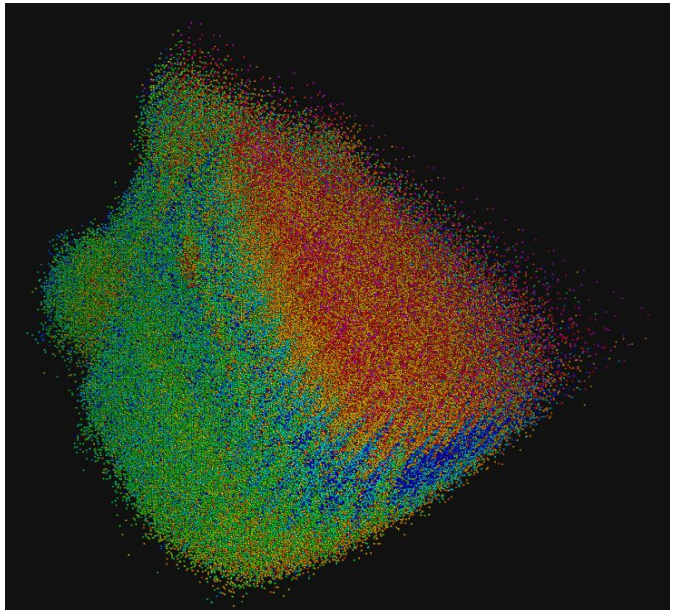
Gökçe Uludoğan[1], Elif Ozkirimli[2], Kutlu O. Ulgen[1], Nilgün Karalı[3], Arzucan Özgür[1]

[1]Bogazici University | [2]F. Hoffmann-La Roche AG | [3] Istanbul University

gokce.uludogan@boun.edu.tr

github.com/gokceuludogan

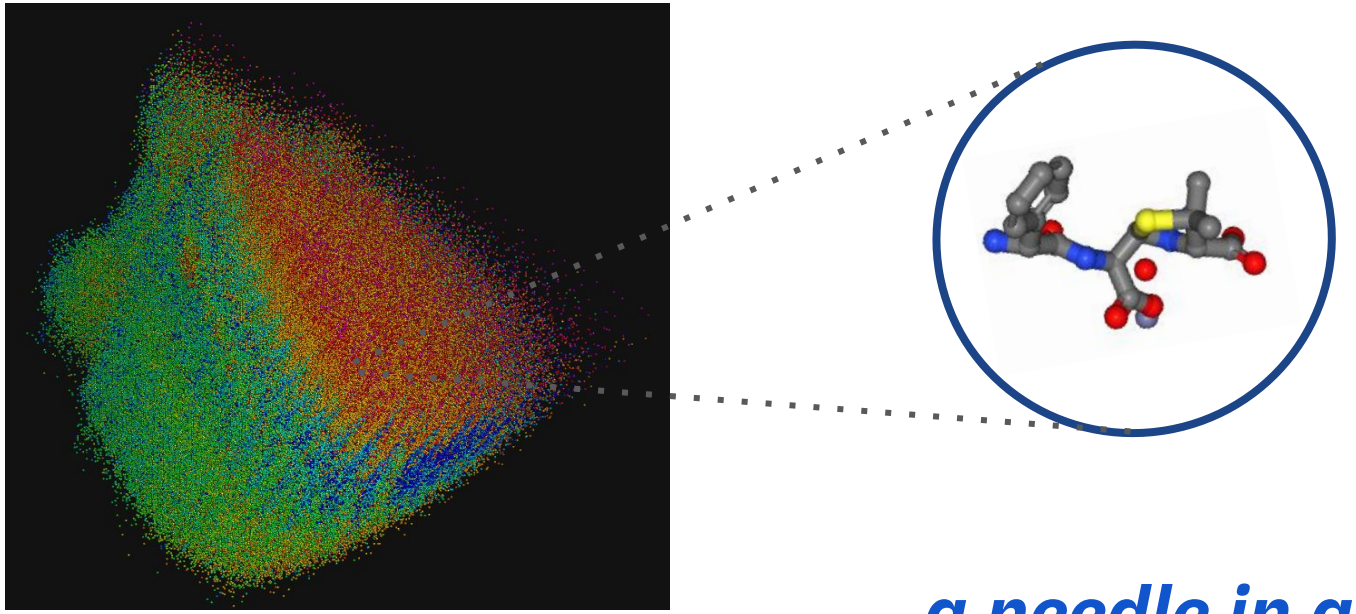
The chemical space is *vast*.



3D visualization of GDBChEMBL
http://faerun.gdb.tools/

Drug-like compounds: ~10^{60}

The chemical space is *vast*.



3D visualization of GDBChEMBL
http://faerun.gdb.tools/

a needle in a haystack

De novo Drug Design

which aims to generate novel molecules from scratch.

Deep generative models in *de novo* drug design

Require either

structural data

or

known active compounds against the target

Targeted drug design as a translation task

eliminates need for structural data and known active compounds

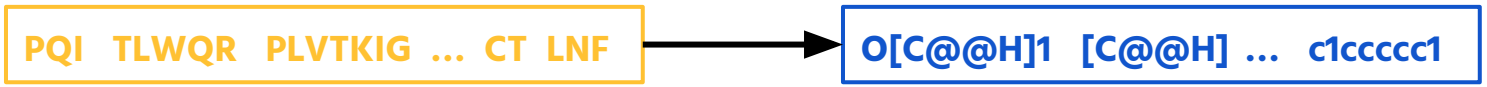
PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLE
EMSLPGRWKPKMIGGIGGFIKVRQYDQILIEICG
HKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF

Amino acid sequence

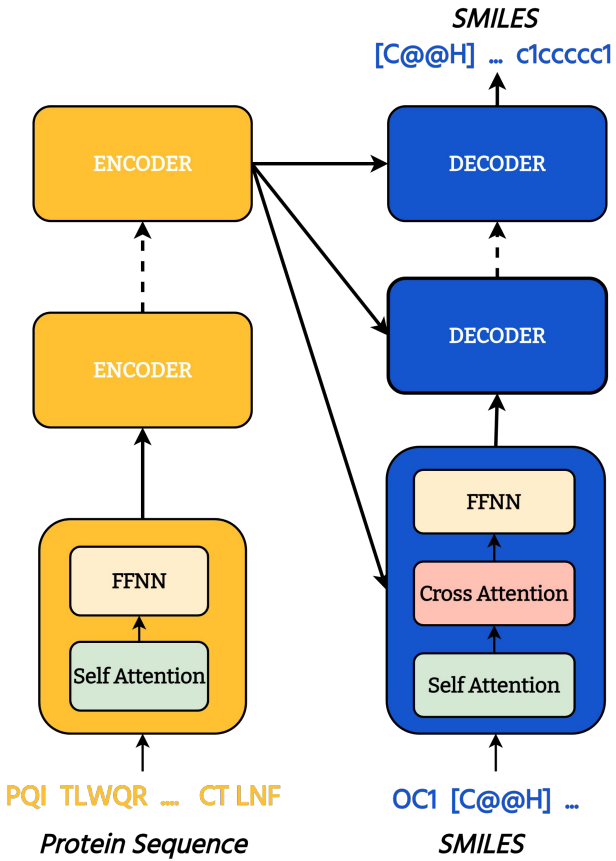
O[C@@H]1[C@@H](O)[C@@H](Cc2ccccc2)N
(CC2CC2)C(=O)N(C\C=C\c2cn[nH]c2)[C@@H
]1Cc1ccccc1

SMILES

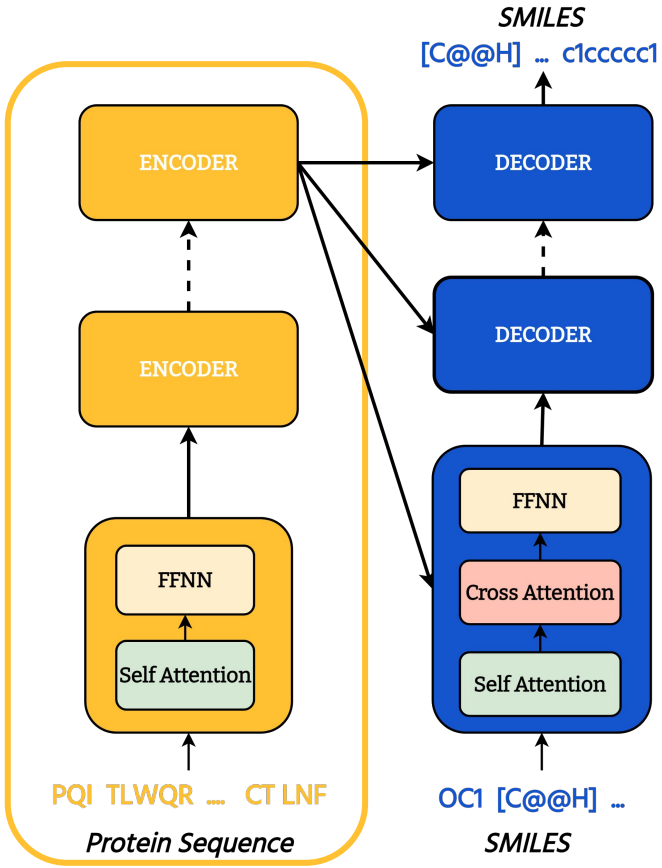
translation from *protein language* to *chemical language*



Model - Transformer-based Encoder Decoder



Model - Transformer-based Encoder Decoder

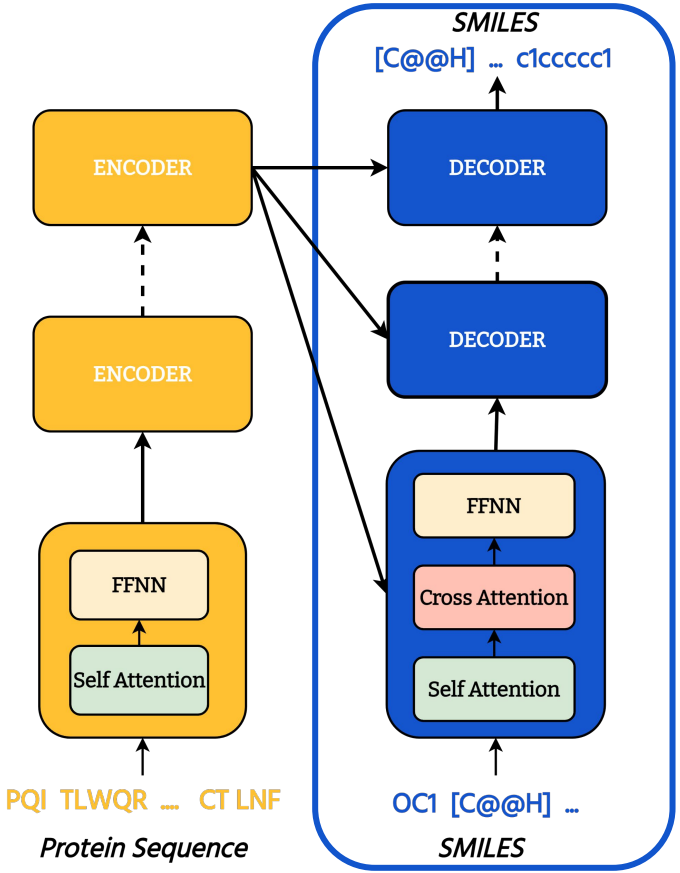


Encoder

Input: **Protein sequence**

Output: **Protein representation**

Model - Transformer-based Encoder Decoder



Encoder

Input: **Protein sequence**
Output: **Protein representation**

Decoder

Input: **SMILES**
Protein representation
Output: **Next chemical unit probabilities**

Exploiting Pretrained Biochemical Language Models for Targeted Drug Design

interacting protein-ligand pairs

vs

Protein sequences and chemical compounds

Exploiting Pretrained Biochemical Language Models for Targeted Drug Design

interacting protein-ligand pairs

vs

Protein sequences and chemical compounds

Protein and Chemical Language Models

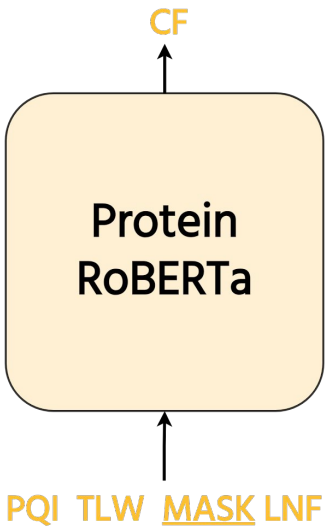
Exploiting Pretrained Biochemical Language Models for Targeted Drug Design

interacting protein-ligand pairs *vs* Protein sequences and chemical compounds

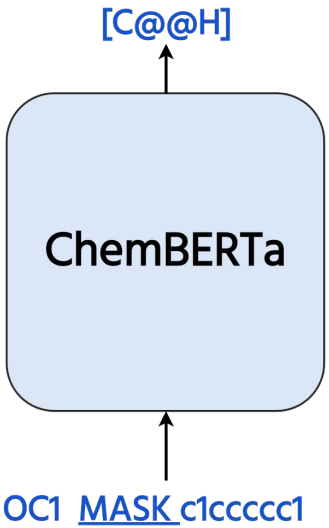
Protein and Chemical Language Models

Warm-start strategies to initialize targeted drug design models with pretrained biochemical language models to enhance generalizability and boost performance

Pretrained Masked Biochemical Language Models



Encoder-only
Transformers



(Filipavicius et. al, 2020)

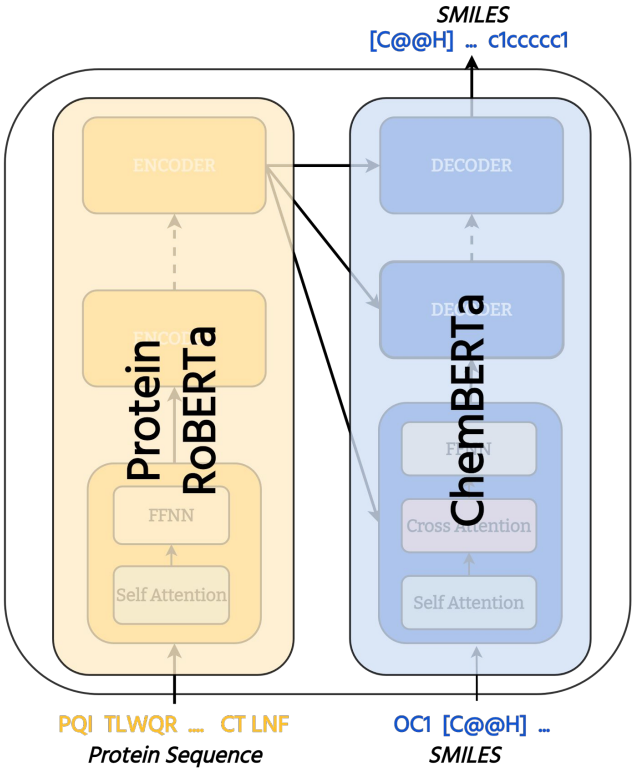
(Chithrananda et. al, 2020)

Warm Start Strategies

One stage strategy

EncDecBase

targeted molecular generation

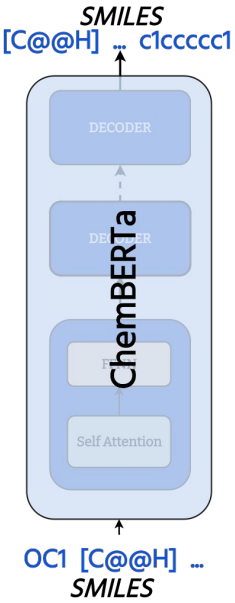


Warm Start Strategies

Two stage strategy

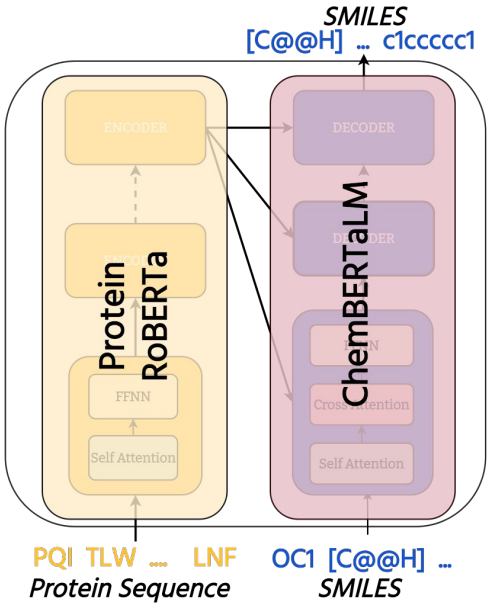
ChemBERTaLM

Stage I:
molecular generation



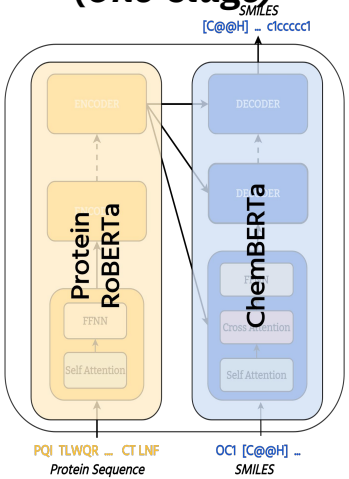
EncDecLM

Stage II:
targeted molecular generation

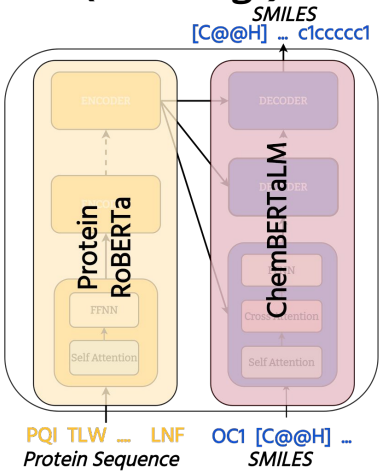


Models

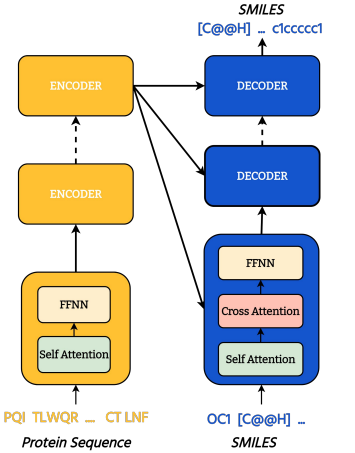
EncDecBase (One-stage)



EncDecLM (Two-stage)



T5 (Raffel et. al, 2019)



Datasets

MOSES

(Polykovskiy et. al, 2020)

BindingDB

(Gilson et. al, 2016)

Decoding

Sampling

Beam Search

Evaluation

MOSES Benchmarking Metrics

Feasibility

- Validity
- Uniqueness
- Novelty

Similarity/Distance

- Frechet ChemNet Distance (FCD) (↓)
- Scaffold Similarity (Scaf) (↑)
- Fragment Similarity (Frag) (↑)
- Nearest Neighbor Tanimoto Similarity (SNN) (↑)

Evaluation

MOSES Benchmarking Metrics

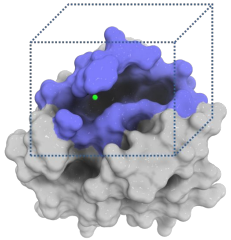
Feasibility

- Validity
- Uniqueness
- Novelty

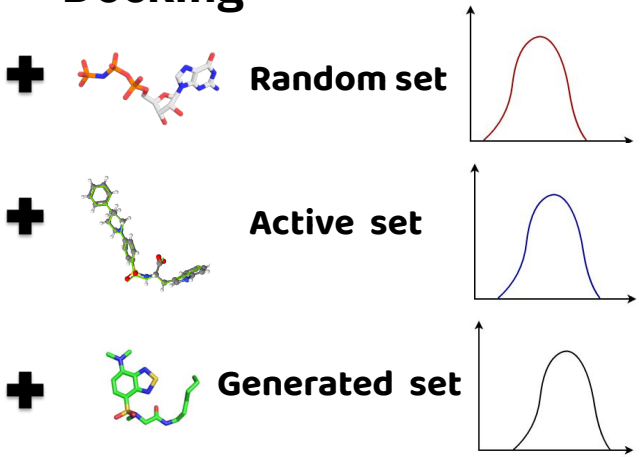
Similarity/Distance

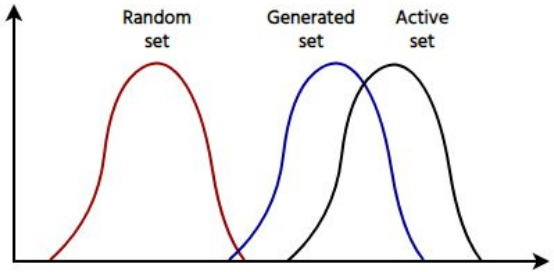
- Frechet ChemNet Distance (FCD) (↓)
- Scaffold Similarity (Scaf) (↑)
- Fragment Similarity (Frag) (↑)
- Nearest Neighbor Tanimoto Similarity (SNN) (↑)

Docking



Target





https://dockthor.lncc.br/v2/

ChemBERTaLM *performs comparably to or better* than the baseline models across metrics

Model	Valid	Filters	Novelty	Test			Scaffold Test (TestSF)		
				FCD (↓)	SNN (↑)	Scaf (↑)	FCD (↓)	SNN (↑)	Scaf (↑)
Train	1	1	1	0.008	0.642	0.991	0.476	0.586	1
AAE	0.937	0.996	0.793	0.556	0.608	0.902	1.057	0.568	0.079
CharRNN	0.975	0.994	0.842	**0.073**	0.601	0.924	0.52	0.565	**0.11**
VAE	0.977	**0.997**	0.695	0.099	**0.626**	**0.939**	0.567	**0.578**	0.059
LatentGAN	0.897	0.973	**0.949**	0.296	0.538	0.886	0.824	0.514	0.100
JTN-VAE	1	0.978	0.914	0.422	0.556	0.892	0.996	0.527	0.100
ChemBERTaLM	**0.991**	**0.997**	0.844	0.090	0.609	0.917	**0.515**	0.572	0.101

Warm started models outperform the baseline T5 model

Model	Beam Search		
	Valid	Unique	Novel
EncDecLM	**0.984**	0.795	0.965
EncDecBase	0.961	0.780	0.978
T5	0.862	**0.909**	**0.999**

in terms of validity

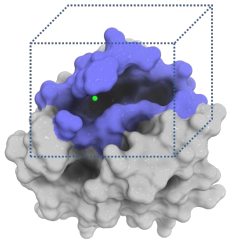
as well as similarity/distance metrics.

Model	Beam Search		
	FCD (↓)	Scaf (↑)	SNN (↑)
EncDecLM	**9.454**	0.090	0.560
EncDecBase	11.652	**0.100**	**0.572**
T5	19.52	0.043	0.506

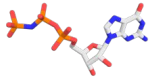
Warm started models can generate molecules *active towards targets of interest.*

Decoding	Model	Generated vs	
		Random	Active
Beam Search	**EncDecBase**	**11**	**8**
	EncDecLM	8	6
	T5	7	5

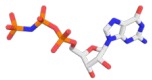
The model warm started with *one-stage strategy* can generate *target specific molecules.*



Target



Molecule set generated for target



Molecule set generated for other targets

*Mann Whitney U test

Target	Generated vs Others	
	p value*	AUC
1ERE	0.418	0.52
2WO6	0.001	0.74
4B6L	1.81E-08	0.9
4I23	2.11E-06	0.83
5K0K	9.28E-05	0.77
5TUY	0.001	0.73
5V1B	4.76E-07	0.85
5XY1	0.001	0.72
6LVL	3.95E-06	0.82
6WJ5	2.69E-06	0.82
6Z1Q	0.175	0.57

Conclusions

We approach targeted drug design problem with ***a language-based view*** and frame it as ***a translation task from protein language to chemical language*** to design molecules ***based only on protein sequence information.***

Conclusions

We approach targeted drug design problem with a language-based view and frame it as a translation task from protein language to chemical language to design molecules based only on protein sequence information.

We investigate *two warm start strategies* to *exploit pretrained biochemical language models* for targeted drug design.

Conclusions

We approach targeted drug design problem with a language-based view and frame it as a translation task from protein language to chemical language to design molecules based only on protein sequence information.

We investigate two warm start strategies to exploit pretrained biochemical language models for targeted drug design.

Warm-started models *outperforms* a baseline model trained from scratch.

Conclusions

We approach targeted drug design problem with a language-based view and frame it as a translation task from protein language to chemical language to design molecules based only on protein sequence information.

We investigate two warm start strategies to exploit pretrained biochemical language models for targeted drug design.

Warm-started models outperforms a baseline model trained from scratch.

Warm started models can generate *compounds with affinity towards target of interest.*

Conclusions

We approach targeted drug design problem with a language-based view and frame it as a translation task from protein language to chemical language to design molecules based only on protein sequence information.

We investigate two warm start strategies to exploit pretrained biochemical language models for targeted drug design.

Warm-started models outperforms a baseline model trained from scratch.

Warm started models can generate compounds with affinity towards target of interest.

Warm started model with the one-stage strategy can design *target specific molecules.*

Acknowledgements

People ♥


Arzucan Özgür


Elif Ozkirimli


Kutlu O. Ulgen


Nilgün Karalı

TUBITAK-119E133 Project Group

- Asu Büşra Temizer
- Taha Khoulani
- Rıza Özçelik
- Berk Atıl
- Cansu Yılmaz
- Can Koban
- Nural Özel
- Selen Parlar

Funding

TUBITAK Grant Numbers:
119E133
2210-A

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Travel Fellowship Award

TUBA GEBİP Award

References

Chithrananda, S., Grand, G., & Ramsundar, B. (2020). ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885.

Filipavicius, M., Manica, M., Cadow, J., & Martinez, M. R. (2020). Pre-training Protein Language Models with Label-Agnostic Binding Pairs Enhances Performance in Downstream Tasks. Retrieved from http://arxiv.org/abs/2012.03084.

Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., & Chong, J. (2016). BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Research, 44(D1), D1045–D1053. https://doi.org/10.1093/nar/gkv1072

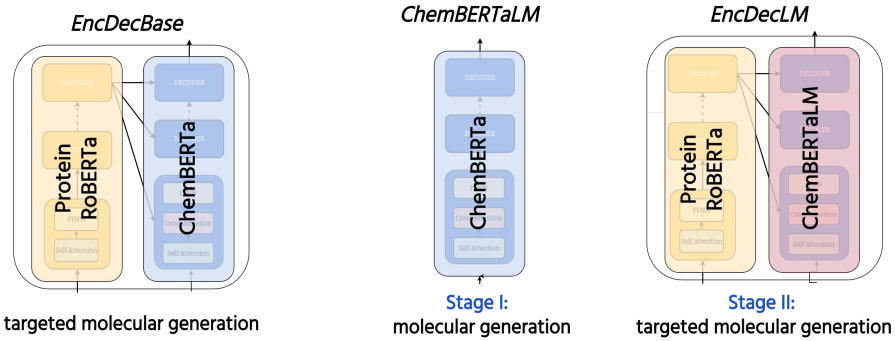
Grechishnikova, D. (2021). Transformer neural network for protein-specific de novo drug generation as a machine translation problem. Scientific Reports, 11(1), 1–13. https://doi.org/10.1038/s41598-020-79682-4

Polykovskiy, D., Zhebrak, A., Sanchez-Lengeling, B., Golovanov, S., Tatanov, O., Belyaev, S., … Zhavoronkov, A. (2020). Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. Frontiers in Pharmacology, 11, 1–19. https://doi.org/10.3389/fphar.2020.565644

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., … Liu, P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research, 21, 1–67. Retrieved from http://arxiv.org/abs/1910.10683

boun-tabi	biochemical-lms-for-drug-design
spaces	gokceuludogan/WarmMolGen
arXiv	2209.00981
doi zenodo	10.5281/zenodo.6832145

Warm Start Strategies



Findings

Warm-started models outperforms a baseline model trained from scratch.

Warm started models can generate compounds with affinity towards targets of interest.

Warm started model with the one-stage strategy can design target specific molecules.