筑波大学

University of Tsukuba

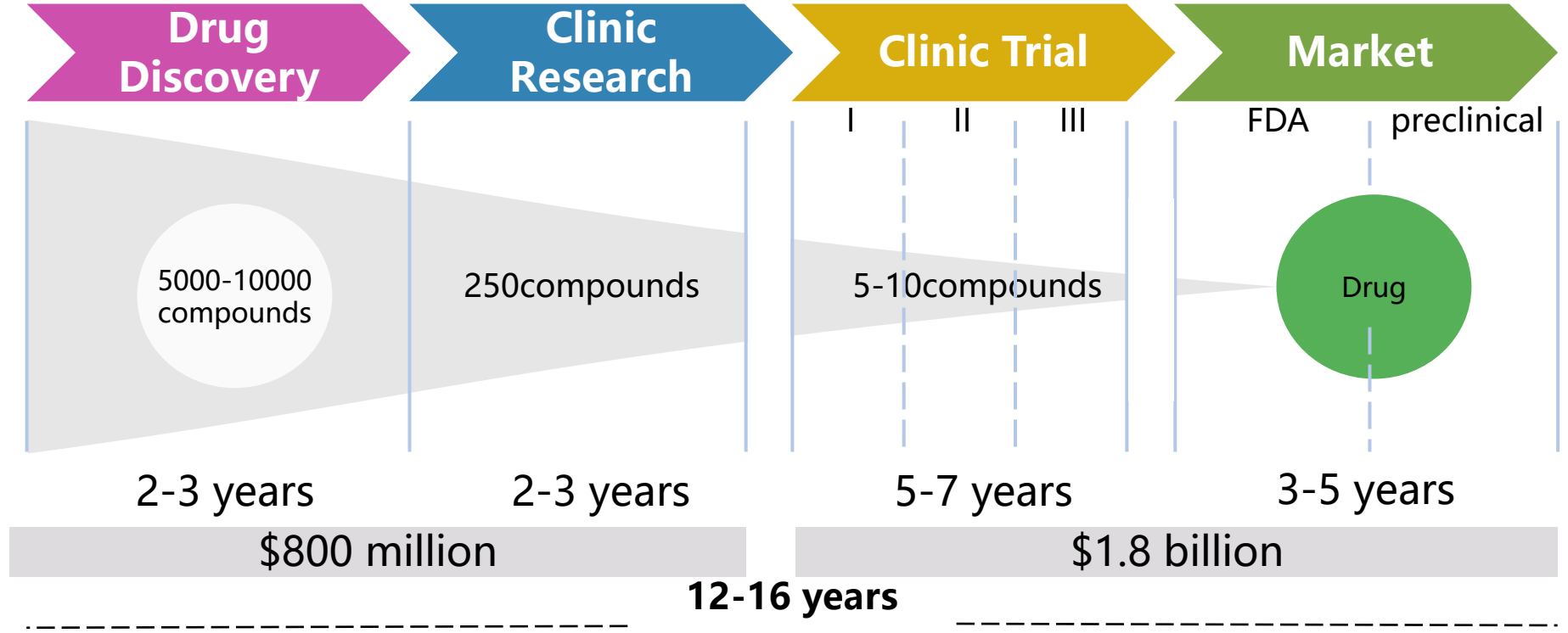
Deep generative model for drug design from protein target sequence

Yangyang Chen

- **Background**
- **Existing research**
- **Our research**
- **Future outlook**

Background

Traditional R&D



Artificial Intelligence (AI)

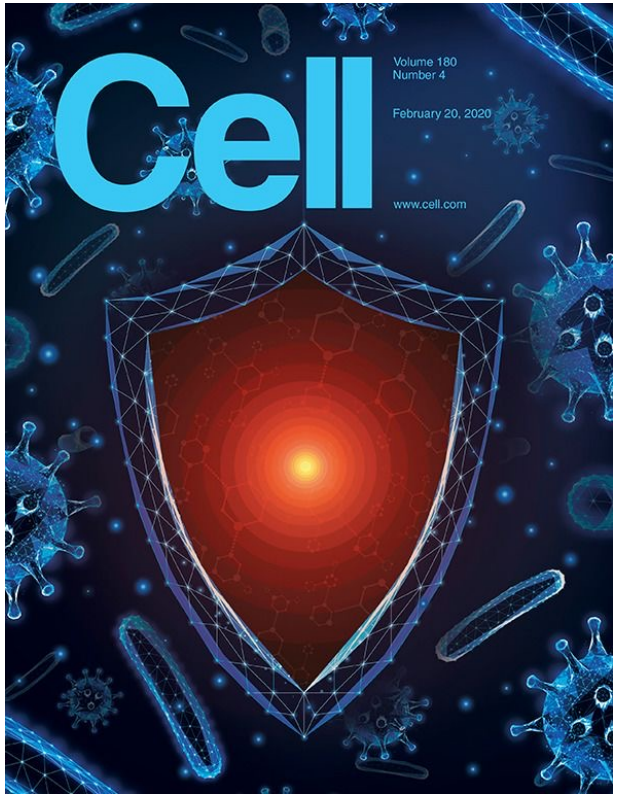
Target Discovery
Lead Compound Design
Lead Compound Screening
Lead Compound Optimization

Drug Repositioning
New indication discovery
Drug Property Prediction
Crystal shape prediction

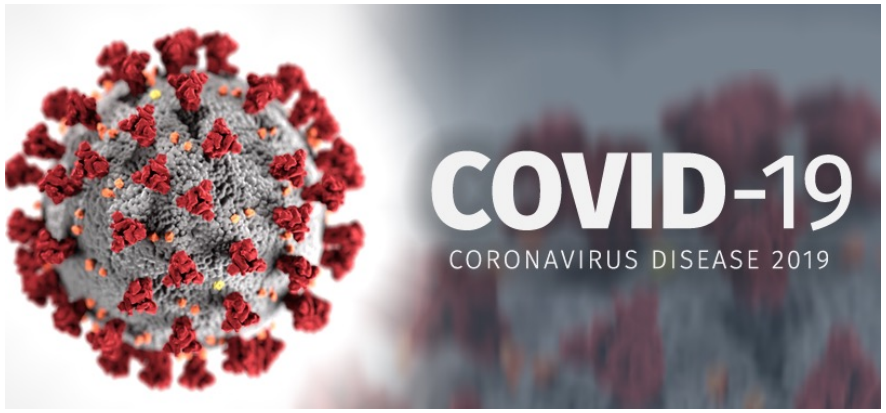
Clinical Trial
Patient Recruitment

Drug Interactions
Product Inspection
Drug Synthesis

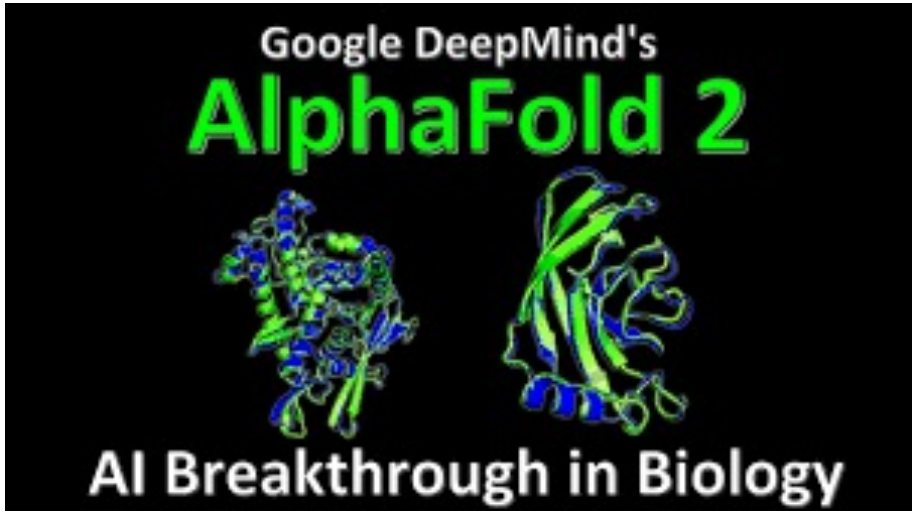
2020-The Year of Artificial Intelligence Drug Development



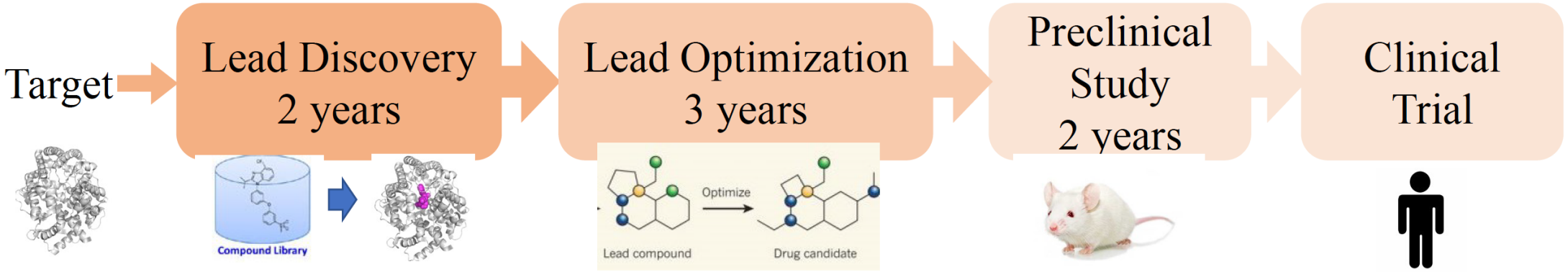
In 2020, Cell cover,
AI discovers the super-powerful antibiotic Halicin



In 2020, the COVID-19 epidemic Outbreak.



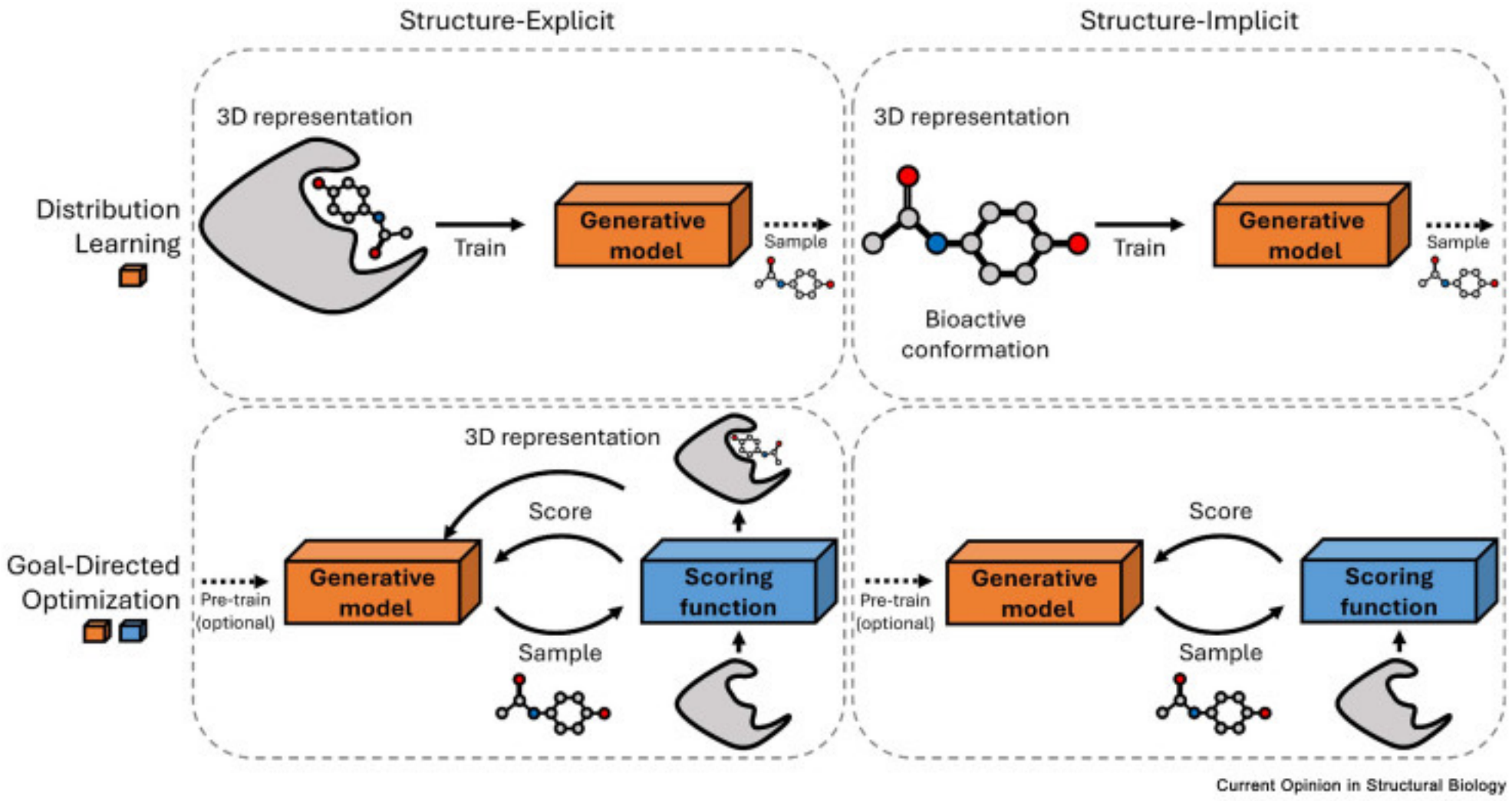
In 2020, the AlphaFold2 comes out.



- Drug discovery: the process of finding new medicines to treat diseases.
- Finding the right drug ligand for a protein is a complex task.
- Traditional methods: biological experiments.
- Intrinsic: Finding the most suitable ligand from a large range of drug molecules.

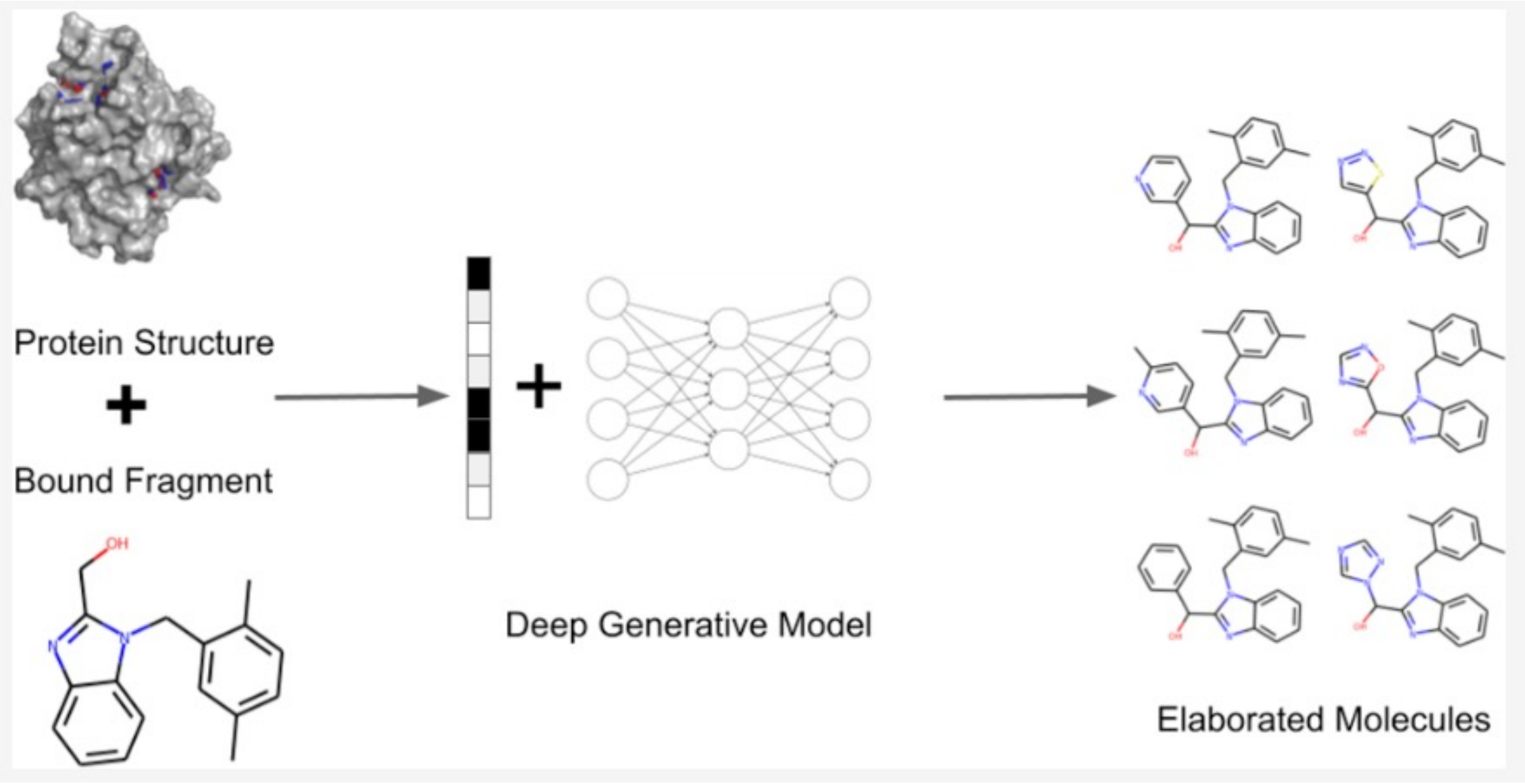
Existing research

distribution learning or goal-directed optimization and structure-explicit/implicit



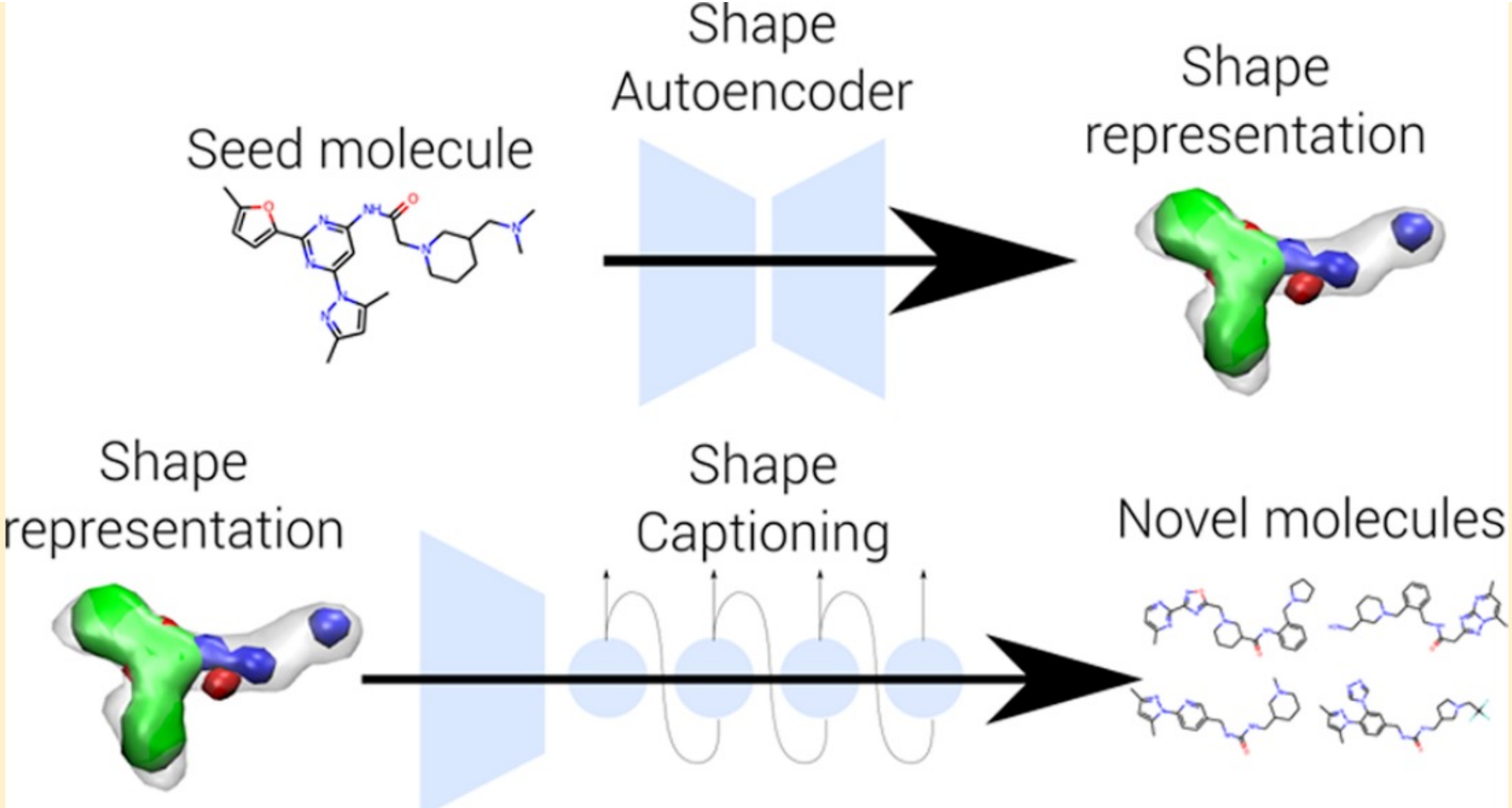
Thomas, M., Bender, A., & de Graaf, C. (2023). Integrating structure-based approaches in generative molecular design. *Current Opinion in Structural Biology*, *79*, 102559.

Distribution Learning & Structure-Explicit



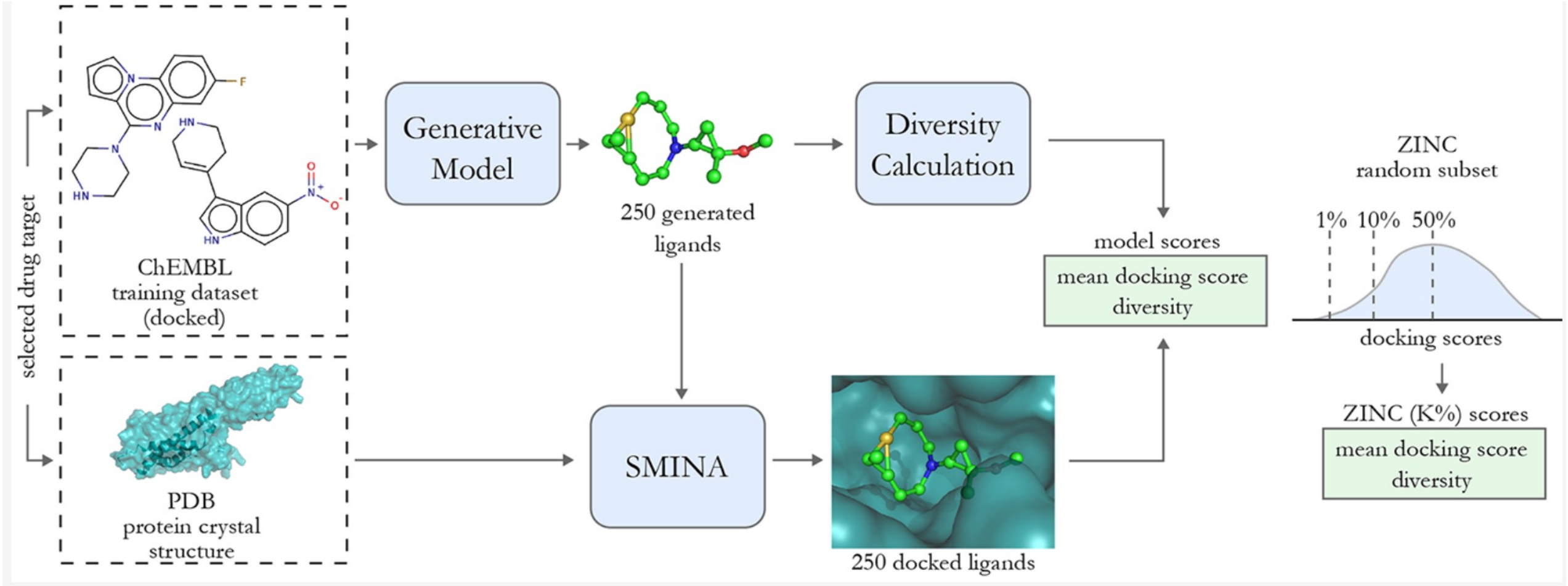
Hadfield, T. E., Imrie, F., Merritt, A., Birchall, K., & Deane, C. M. (2022). Incorporating target-specific pharmacophoric information into deep generative models for fragment elaboration. *Journal of Chemical Information and Modeling*, *62*(10), 2280-2292.

Distribution Learning & Structure-Implicit



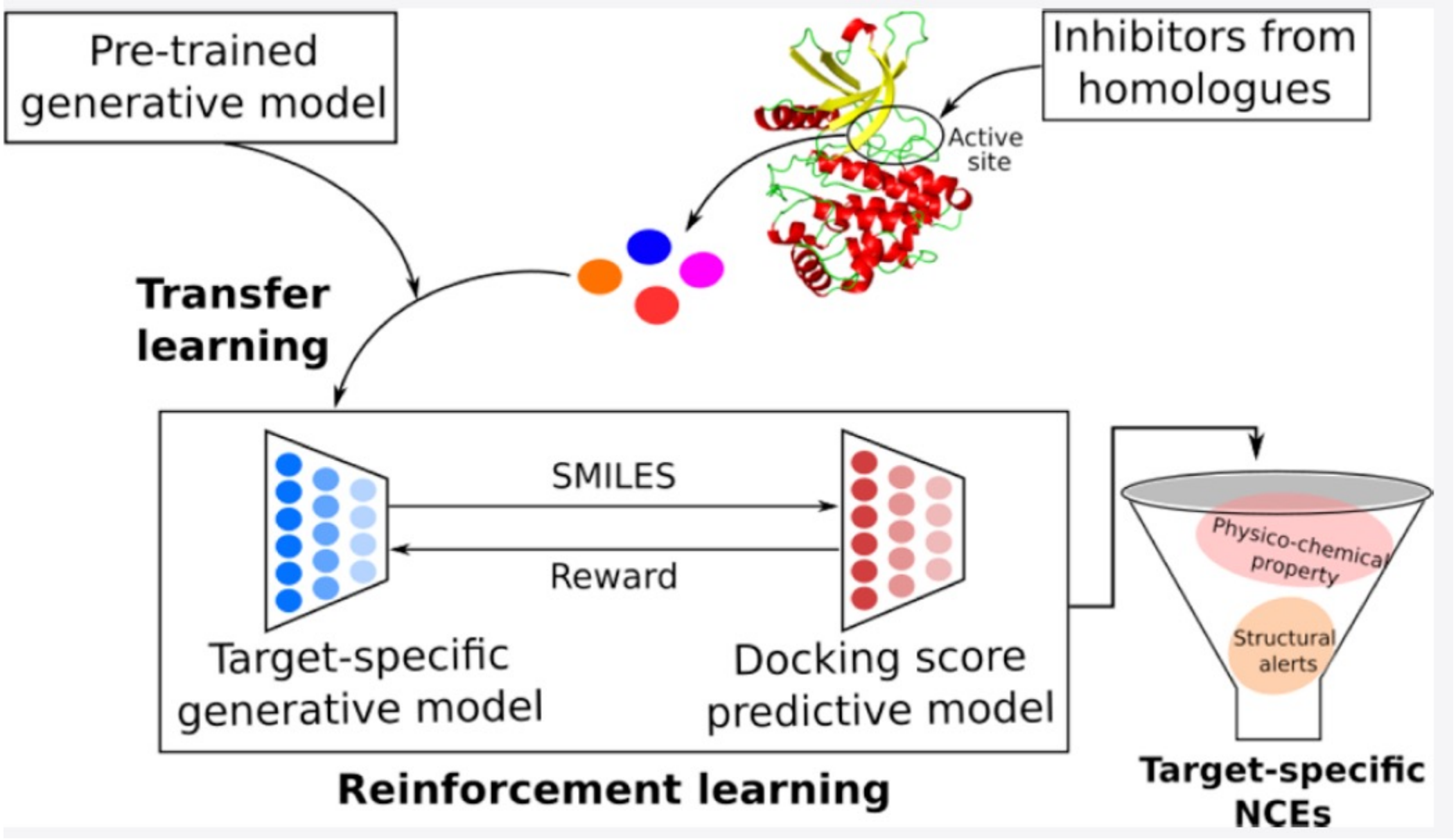
Skalic, Miha, et al. "Shape-based generative modeling for de novo drug design." *Journal of chemical information and modeling* 59.3 (2019): 1205-1214.

Goal-directed Optimization & Structure-Explicit



Ciepliński, Tobiasz, et al. "Generative Models Should at Least Be Able to Design Molecules That Dock Well: A New Benchmark." *Journal of Chemical Information and Modeling* (2023).

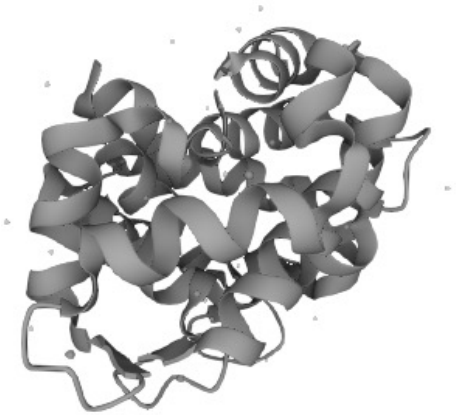
Goal-directed Optimization & Structure-Implicit

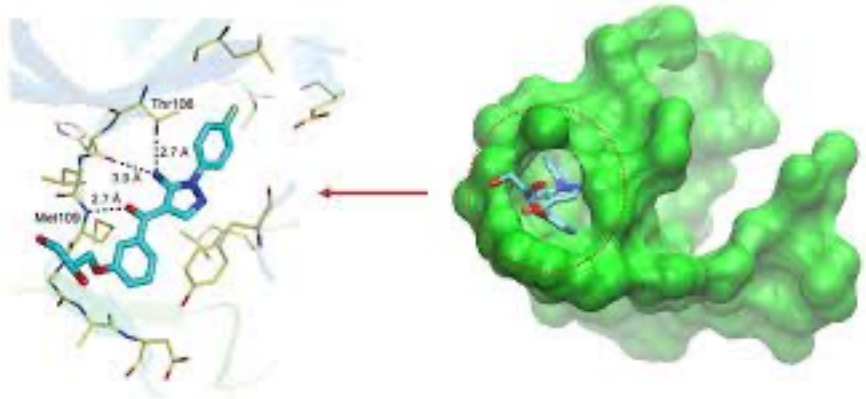


Krishnan, Sowmya Ramaswamy, et al. "Accelerating de novo drug design against novel proteins using deep learning." *Journal of Chemical Information and Modeling* 61.2 (2021): 621-630.

Problems and limitations

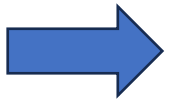
- Complexity of protein structure
- Structure of some proteins unknown (pocket)





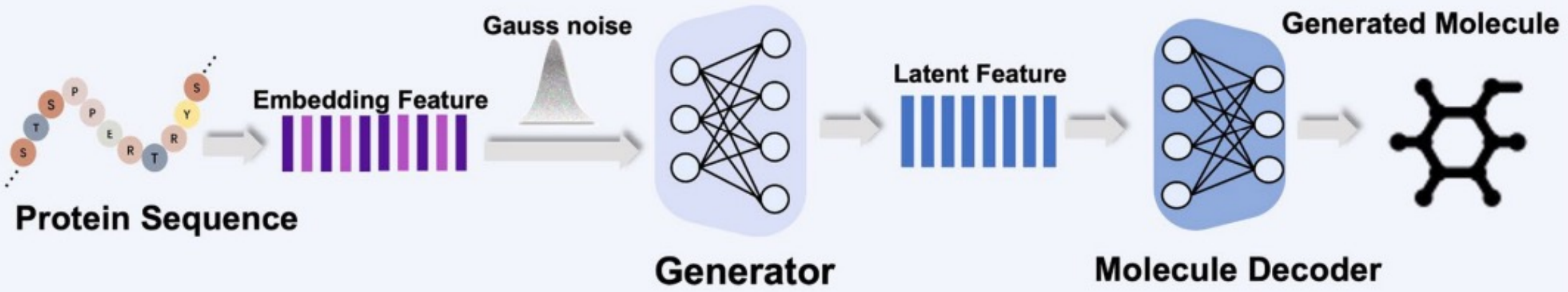
- Simple form:

  protein sequence



```
>sp|P14416|DRD2_HUMAN D(2) dopamine receptor OS=Homo sapiens OX=9606 GN=DRD2 PE=1 SV=2
MDPLNLSWYD DDLERQNWSR PFNGSDGKAD RPHYNYYATL LTLLIAVIVF GNVLVCMAVS
REKALQTTTN YLIVSLAVAD LLVATLVMPW VVYLEVVGEW KFSRIHCDIF VTLDVMMCTA
SILNLCAISI DRYTAVAMPM LYNTRYSSKR RVTVMISIVW VLSFTISCPL LFGLNNADQN
ECIIANPAFV VYSSIVSFYV PFIVTLLVYI KIYIVLRRRR KRVNTKRSSR AFRAHLRAPL
KGNCTHPEDM KLCTVIMKSN GSFPVNRRRV EAARRAQELE MEMLSSTSPP ERTRYSPIPP
SHHQLTLPDP SHHGLHSTPD SPAKPEKNGH AKDHPKIAKI FEIQTMPNGK TRTSLKTMSR
RKLSQQKEKK ATQMLAIVLG VFIICWLPFF ITHILNIHCD CNIPPVLYSA FTWLGYVNSA
VNPIIYTTFN IEFRKAFLKI LHC
```

Our research



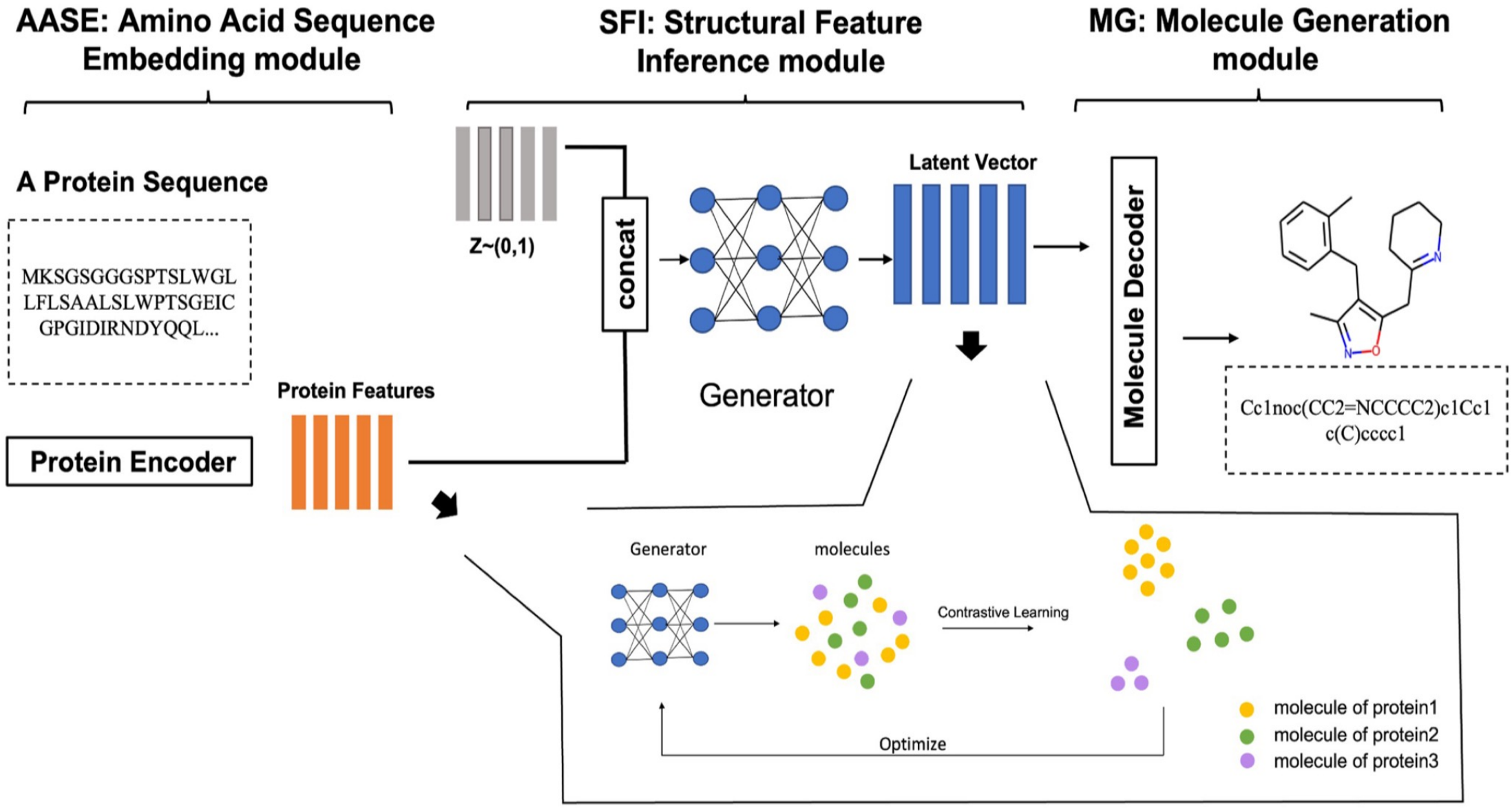
Model overflow

- Input: protein sequence Output: molecule SMILES
- Three modules:

  AASE: Amino Acid Sequence Embedding module

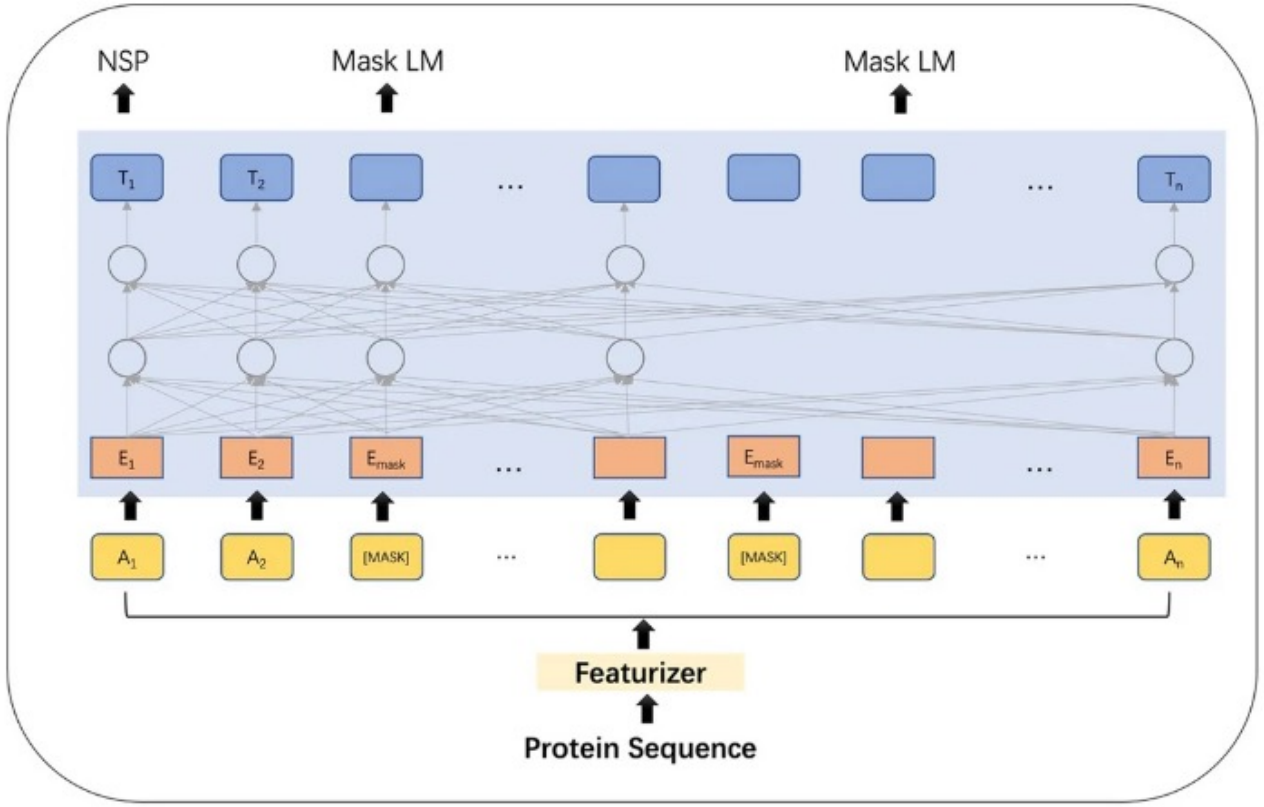
  SFI: Structural Feature Inference module

  MG: Molecule Generation module

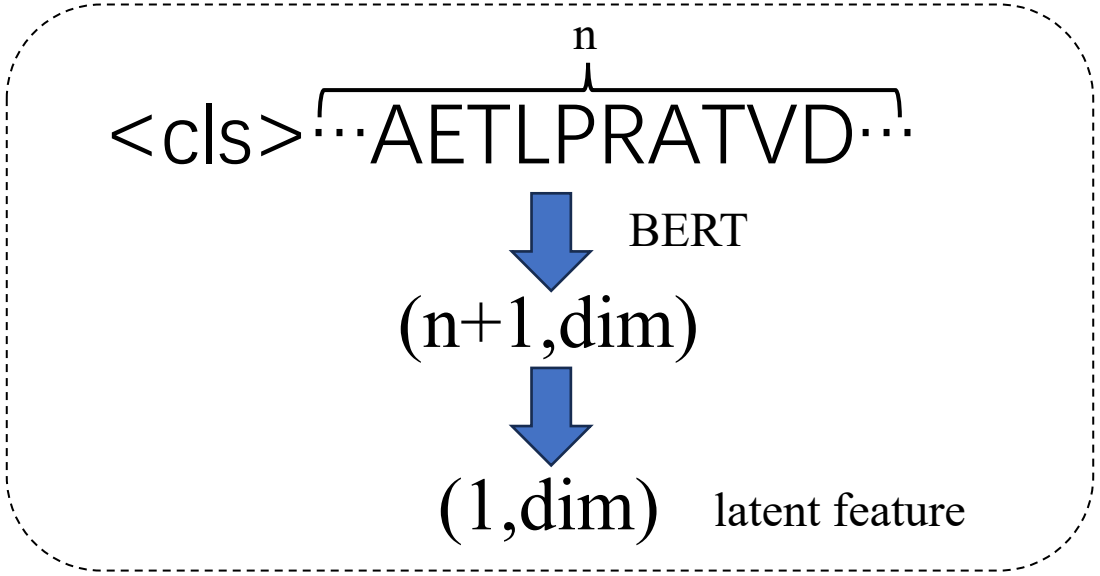


Chen, Y., Wang, Z., Wang, L. *et al.* Deep generative model for drug design from protein target sequence. *J Cheminform* **15**, 38 (2023). https://doi.org/10.1186/s13321-023-00702-2

Amino Acid Sequence Embedding module



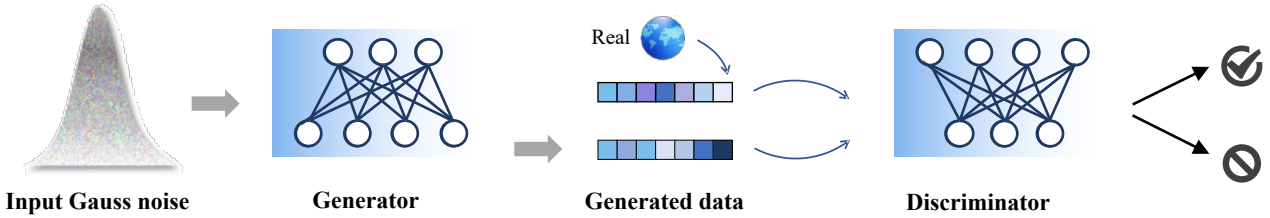
pre-trained protein encoder



- Protein sequence is too long.
- BERT: Pre-training of deep bidirectional transformers for language understanding.
- Latent feature to represent the protein.

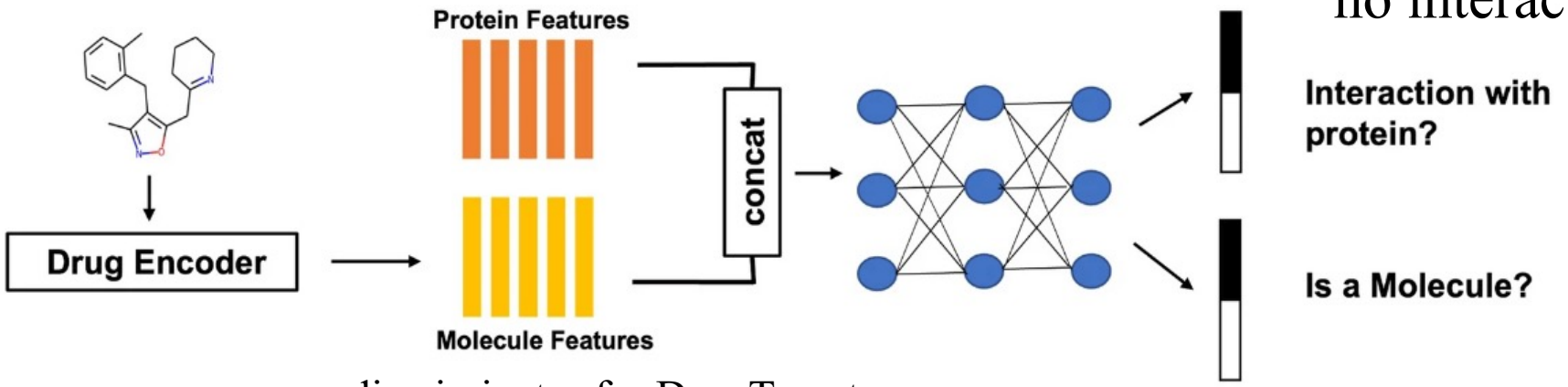
Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Structural Feature Inference module



GANs: generative adversarial networks

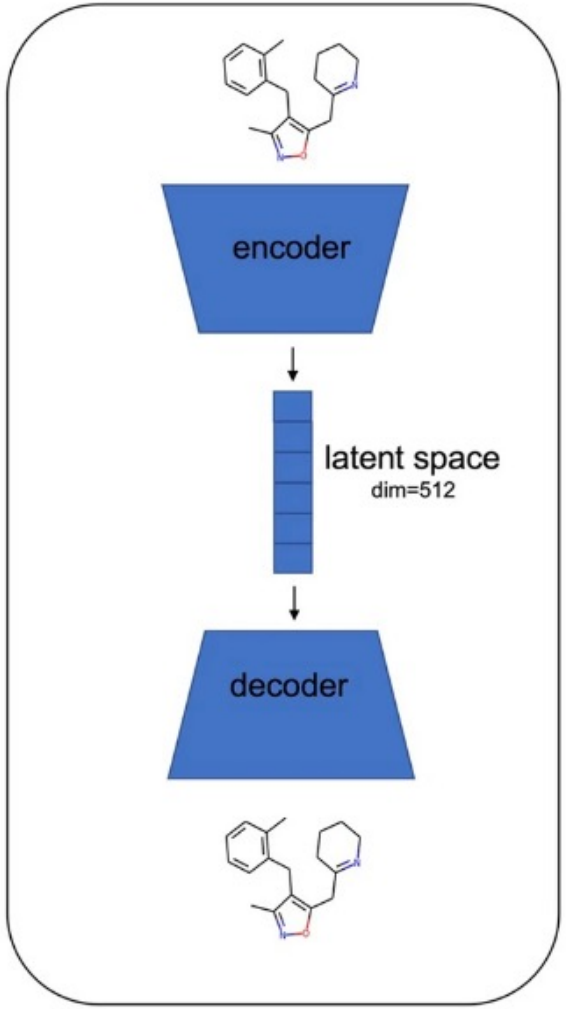
- Generative Adversarial Networks
- Generator:
  protein feature as input
- Discriminator:
  one output->two outputs
- Real ligands: interact and valid
- Generated ligands:
  no interaction and invalid

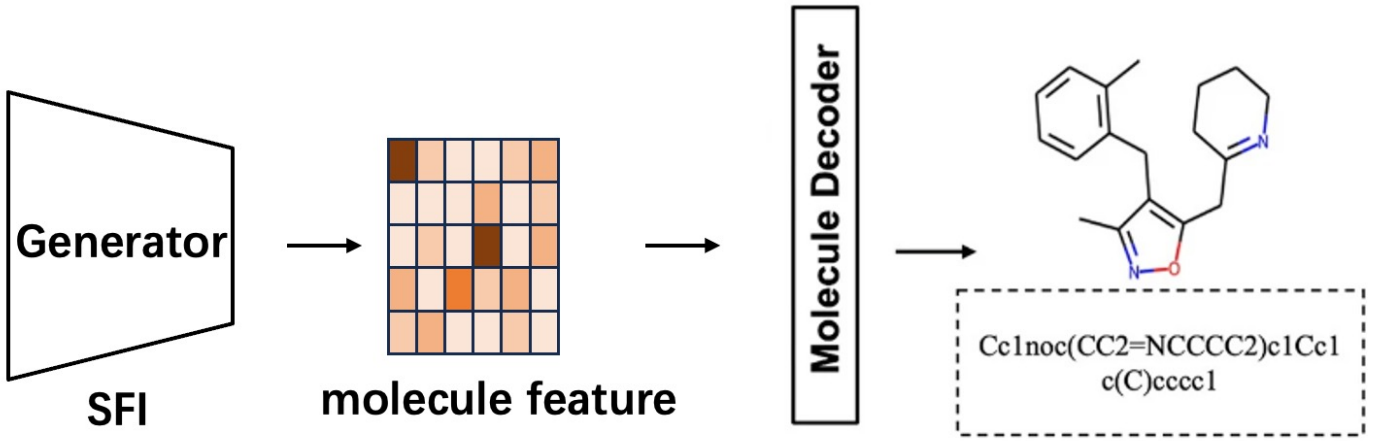


discriminator for DeepTarget

Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems* 27 (2014)

Molecule Generation module





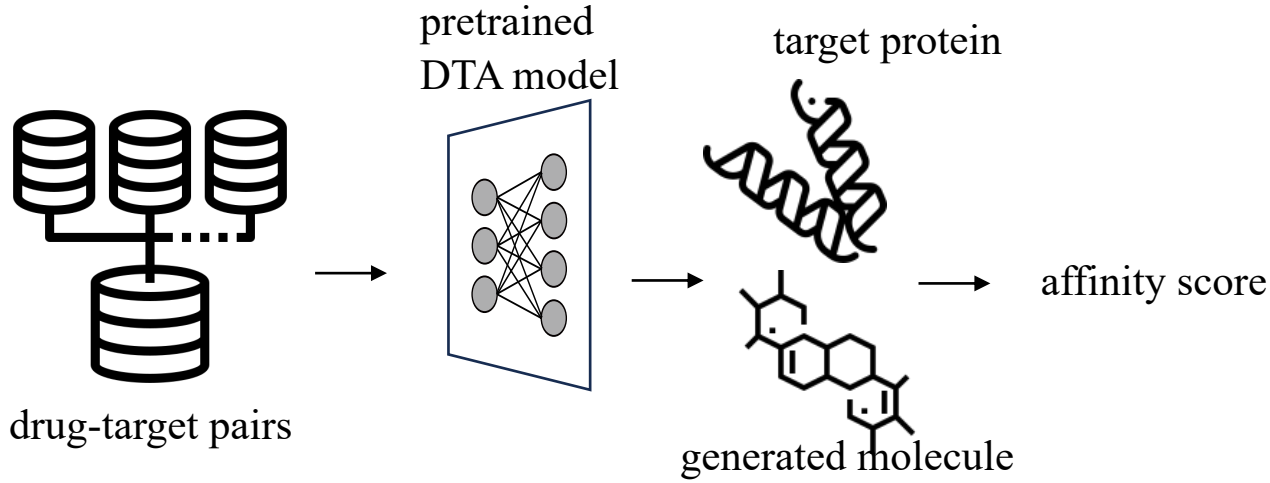
- Autoencoder: pre-trained in ChEMBL
- Encoder:
  embedding the molecules to latent feature
- Latent Feature:
  represent the generated molecules
- Decoder
  decoding feature to molecules

Tschannen, Michael, Olivier Bachem, and Mario Lucic. "Recent advances in autoencoder-based representation learning." *arXiv preprint arXiv:1812.05069* (2018).

Evaluation Metrics

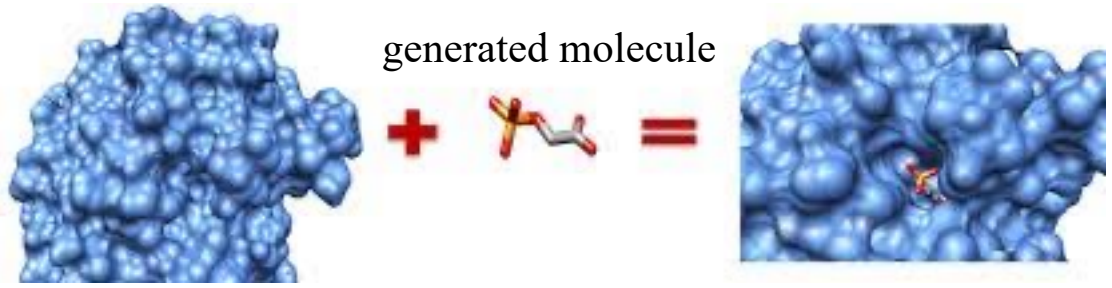
Basic Prosperities:

- The generated molecules should satisfy the basic property distribution
- Molecular weight (MW)
- Lipophilicity (LogP)
- Natural product-likeness (NP-likeness)
- Synthetic accessibility score (SAscore)
- Quantitative estimation of drug-likeness (QED)



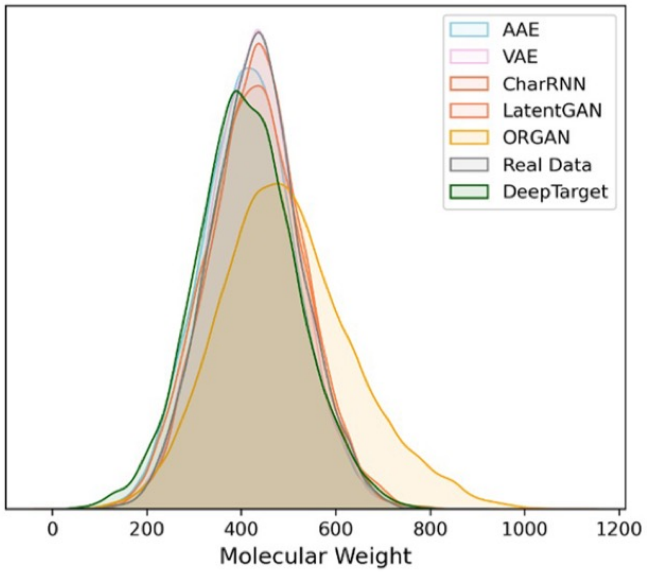
Interaction with protein target:

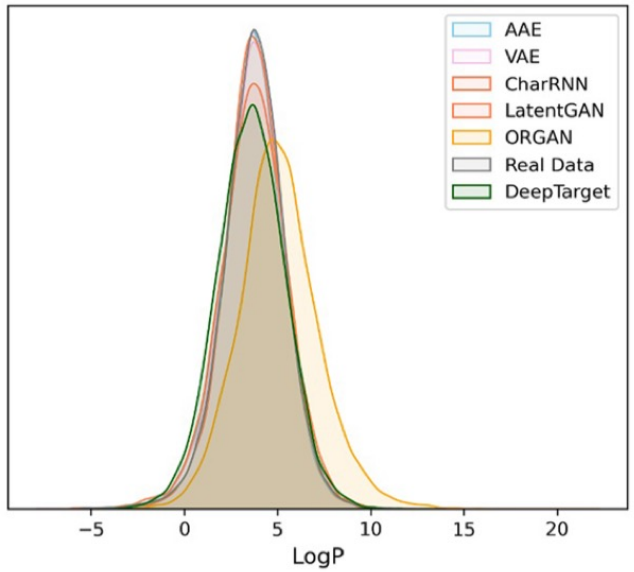
- Affinity score
- Docking score

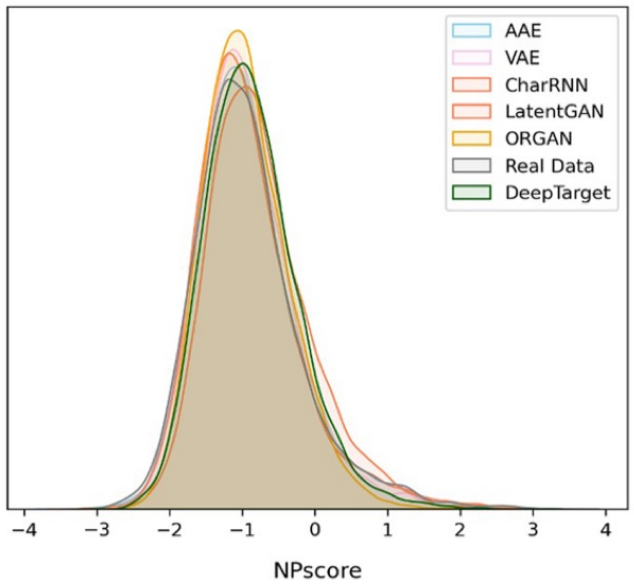


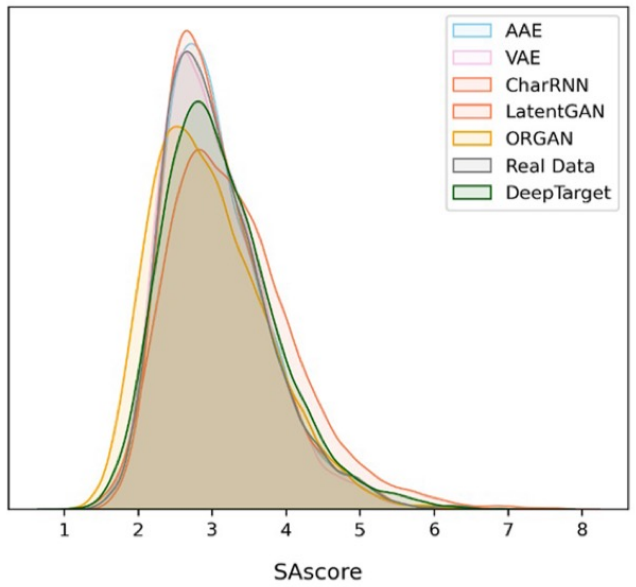
Results

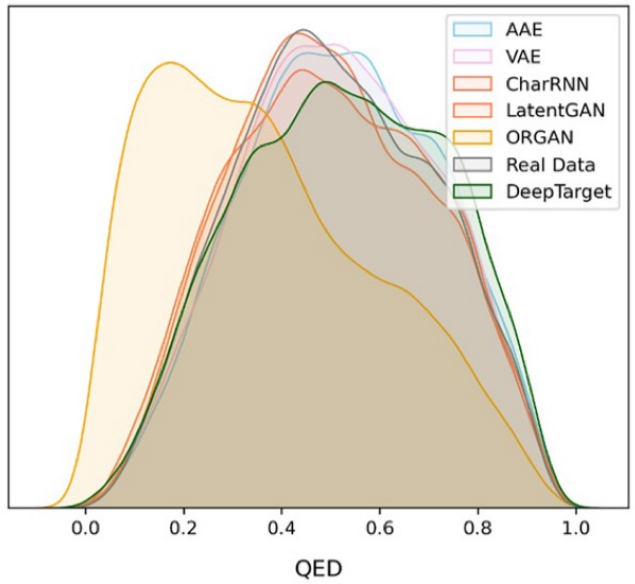
Polykovskiy, Daniil, et al. "Molecular sets (MOSES): a benchmarking platform for molecular generation models." *Frontiers in pharmacology* 11 (2020): 565644.











Results

Two target examples:

P14416 · DRD2_HUMAN

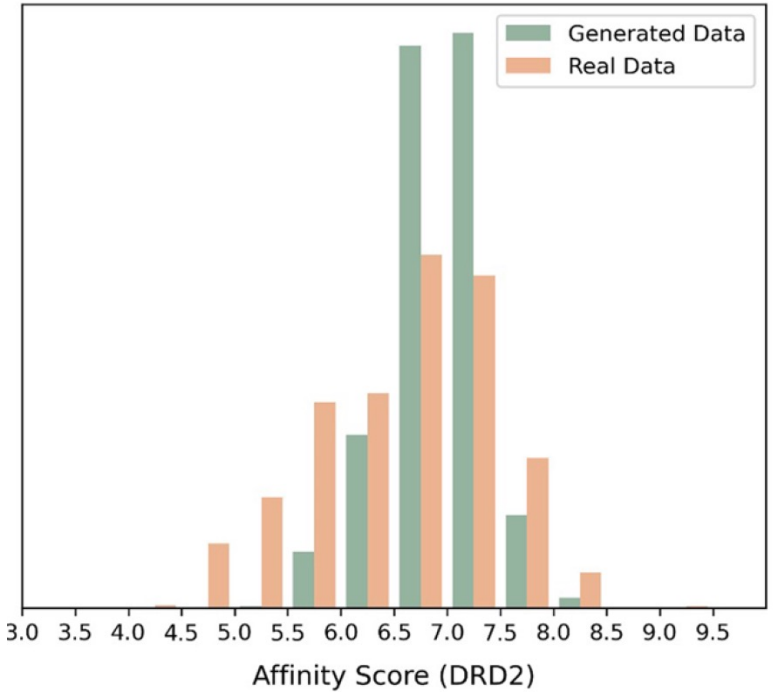
Protein	D(2) dopamine receptor
Gene	DRD2
Status	UniProtKB reviewed (Swiss-Prot)
Organism	Homo sapiens (Human)

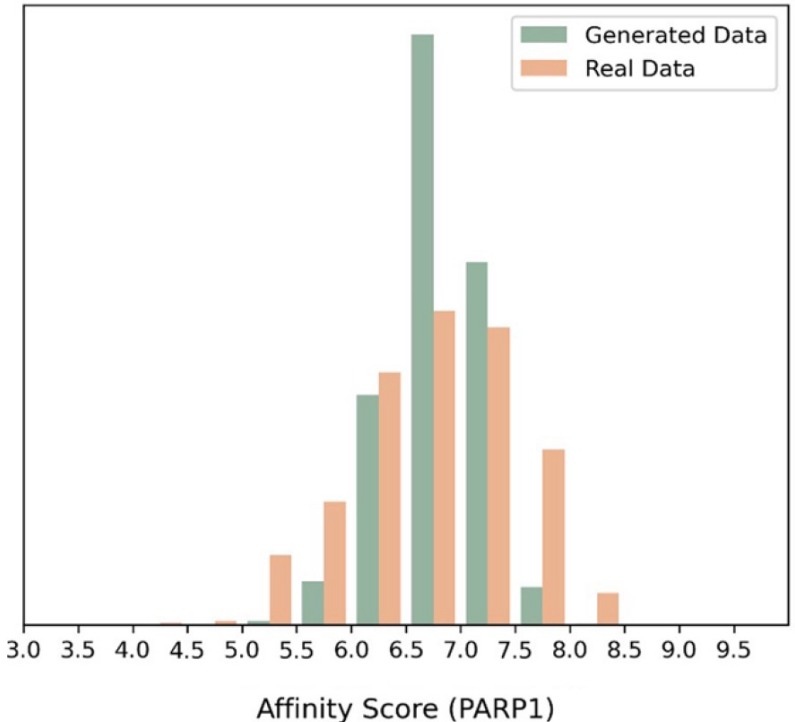
https://www.uniprot.org/uniprotkb/P14416/entry

P09874 · PARP1_HUMAN

Protein	Poly [ADP-ribose] polymerase 1
Gene	PARP1
Status	UniProtKB reviewed (Swiss-Prot)
Organism	Homo sapiens (Human)

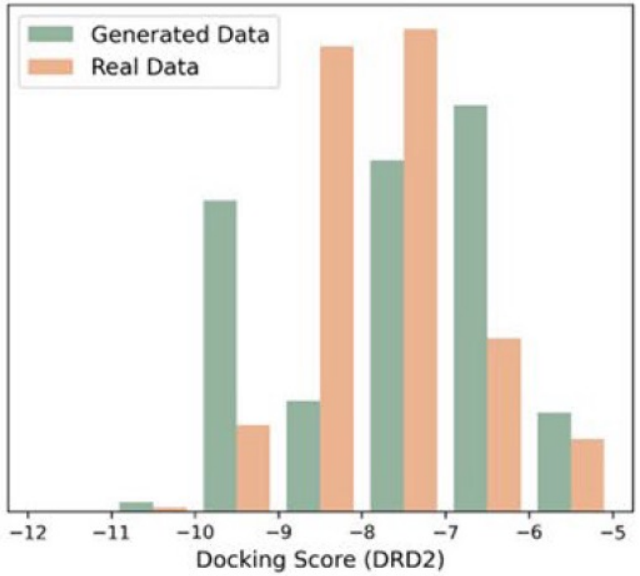
https://www.uniprot.org/uniprotkb/P09874/entry

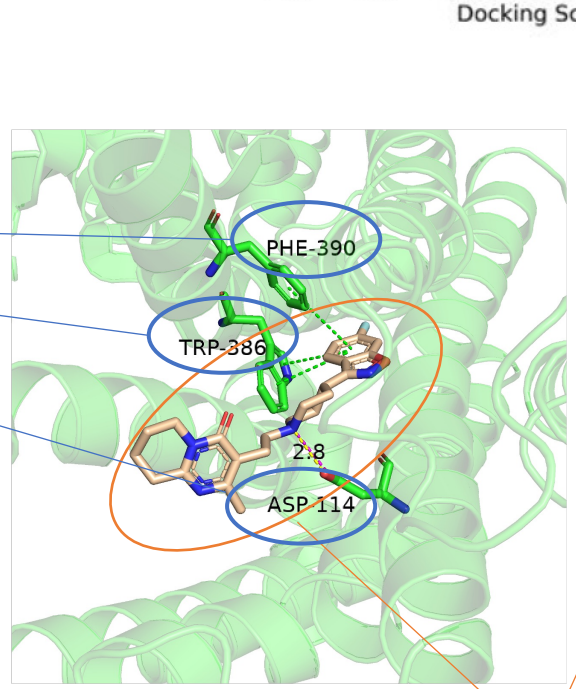




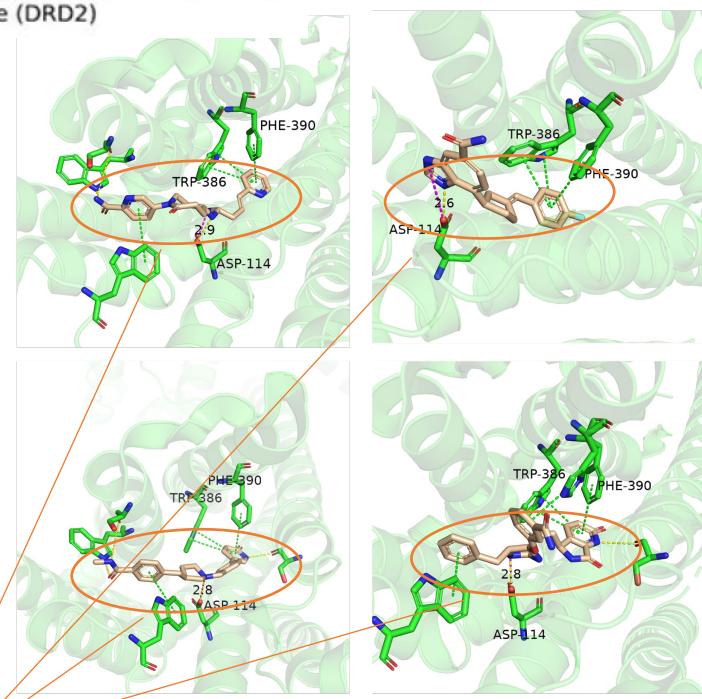
Result: Affinity

- Molecular docking is a computer technique that predicts how drugs and proteins interact to help discover new drugs.
- Three key residues interact with ligand.
- The generated molecules could interact with same key residues of real ligand.





Real data



Generated data

ligand molecules

Future outlook

Future outlook

- How to better represent features from protein and small molecules?
- How to further improve the affinity towards the target?
- ……

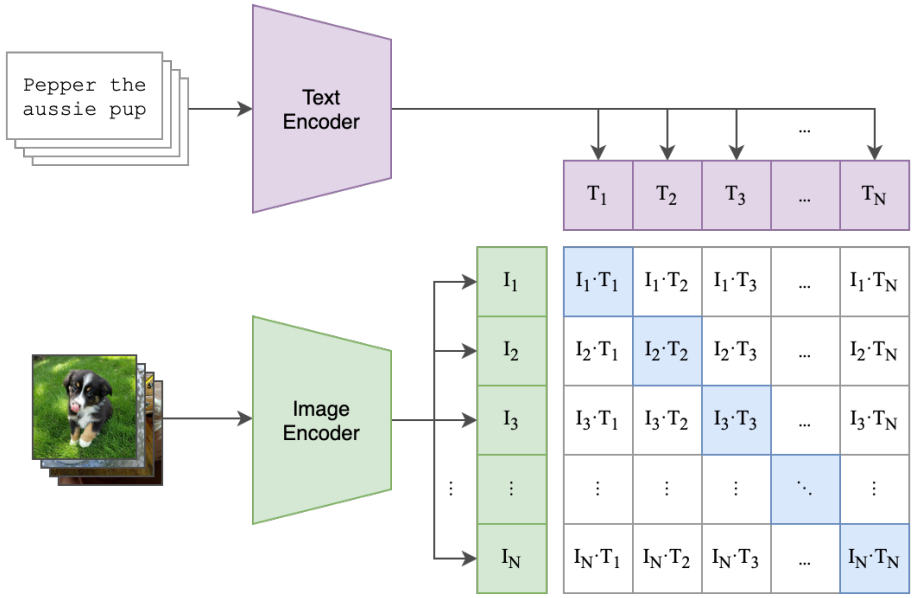
Future outlook

- How to better represent features from protein and small molecules?

Contrastive Learning

- multi-view
- close the feature distance of similar data

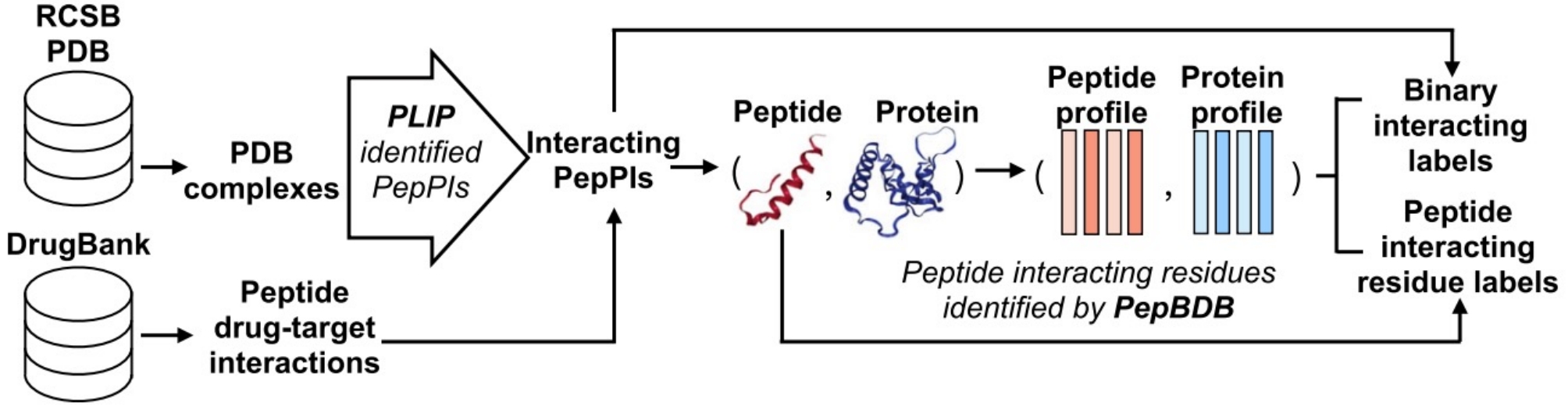
$$score\left(f(\mathrm{I}), f(\mathrm{T}^{+})\right) \gg score\left(f(\mathrm{I}), f(\mathrm{T}^{-})\right)$$



Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PMLR, 2021.

Future outlook

- How to further improve the affinity towards the target?



multiple loss function constraints

Lei, Yipin, et al. "A deep-learning framework for multi-level peptide–protein interaction prediction." *Nature communications* 12.1 (2021): 5465.

Thanks