Zero-Shot 3D Drug Design by Sketching and Generating

Siyu Long[1], Yi Zhou[2], Xinyu Dai[1*], Hao Zhou[3*]

1 National Key Laboratory for Novel Software Technology, Nanjing University

2 ByteDance AI Lab

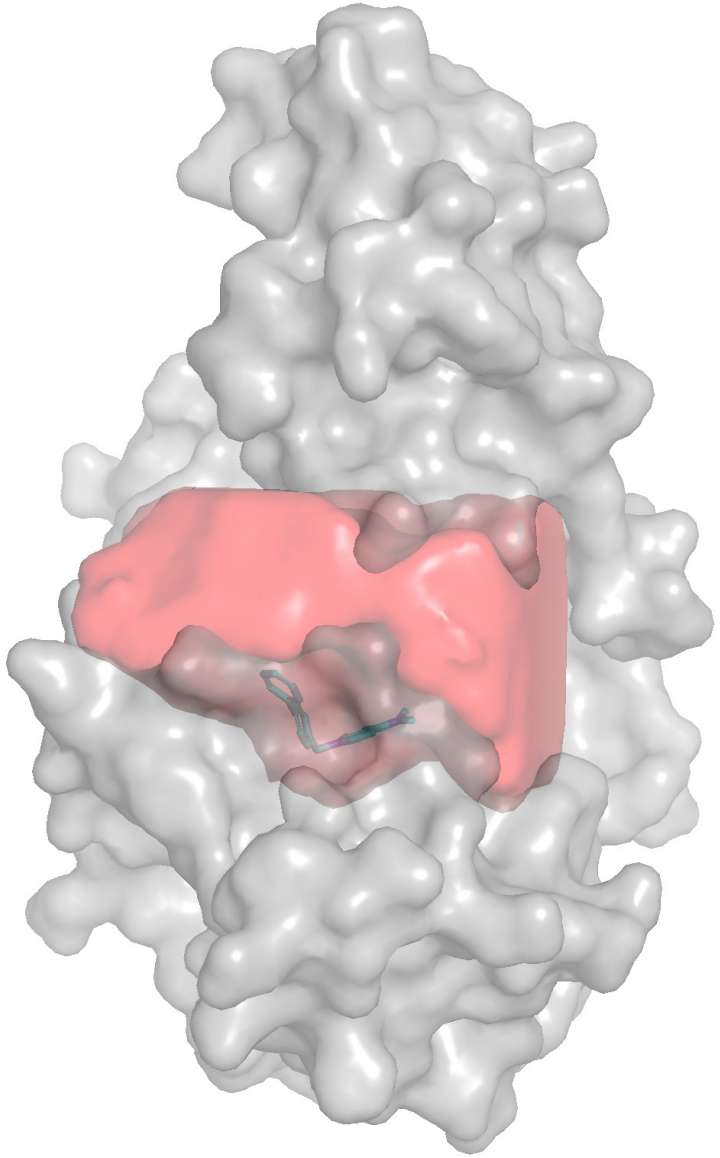
3 Institute for AI Industry Research, Tsinghua University

Outline

- Overview
- Background
- Related Work
- Motivation
- Method (DESERT)
- Experiment
- Conclusion

Overview

- We develop a novel shape-based method for de-novo 3D drug design
- It does not need training data from the wet lab
- It achieves a new SOTA at a fast speed

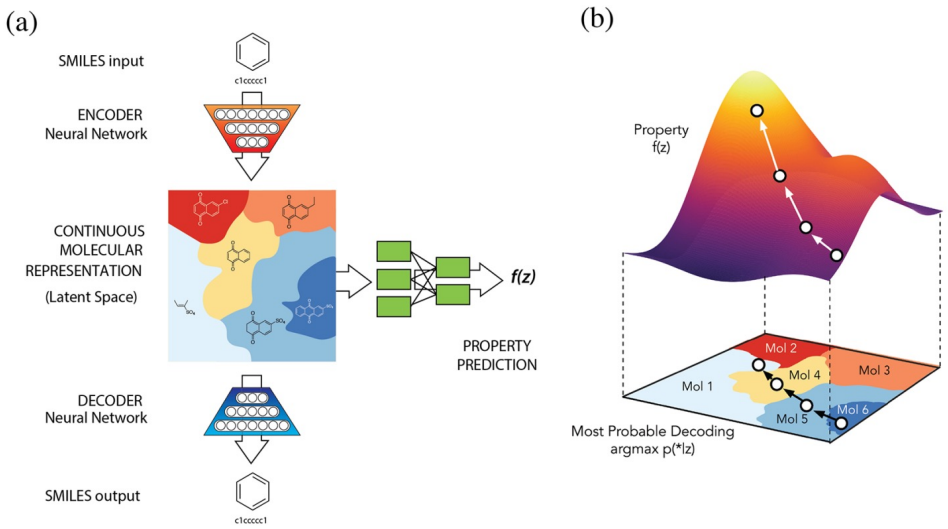


Background

- Drug Design
  - Provide molecules that meet the pharmaceutical requirements for a given protein pocket
- Protein & Protein Pocket
  - Proteins perform a vast array of physiological functions
  - Pockets are the regions where proteins interact with drugs
- Pharmaceutical Requirement
  - Intra-molecule: drug-likeness, synthetic accessibility
  - Inter-molecule: strong binding affinity with proteins

Related Work

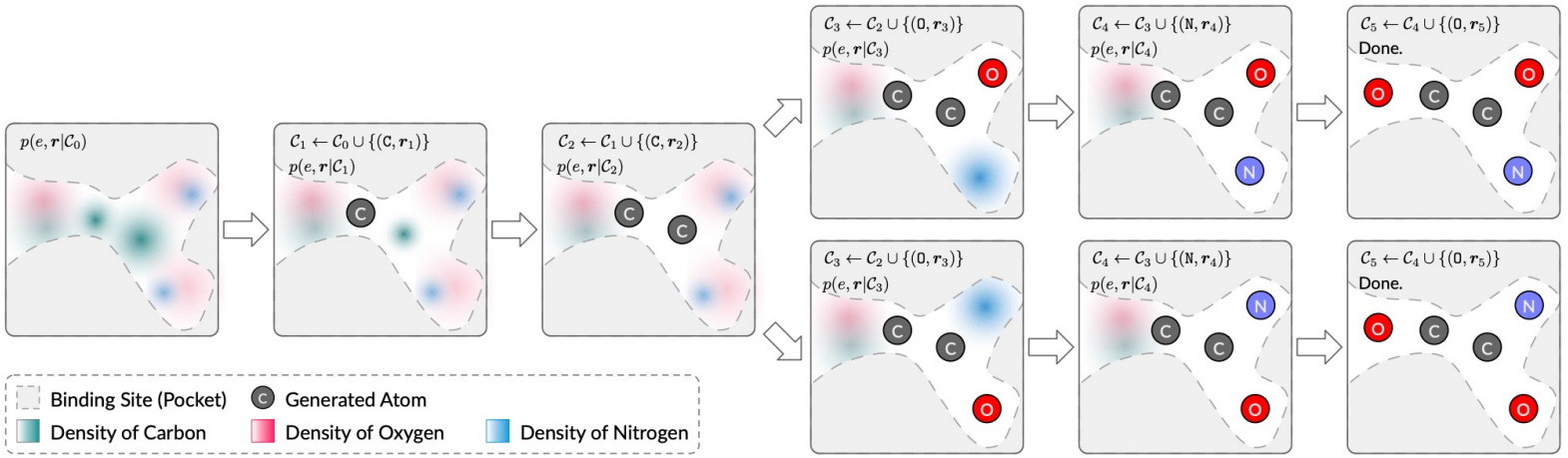
- 1D/2D Drug Design
  - Represent molecules with 1D SMILES or 2D graph
  - Can not design the interaction in 3D space
  - Rely on bioactivity data to optimize molecules



From *Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules*

Related Work

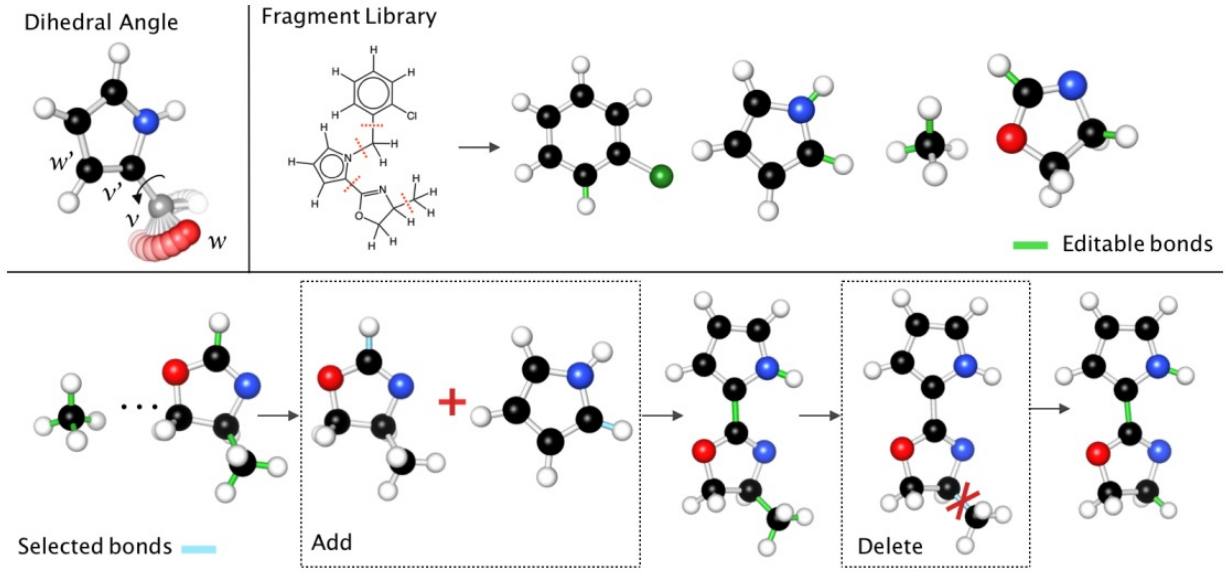
- 3D Drug Design
  - Supervised Method
    - Still need scarce experiment data



From *A 3D Generative Model for Structure-Based Drug Design*

Related Work

- 3D Drug Design
  - Sampling Method
    - Use time-consuming docking simulation to provide supervised signals



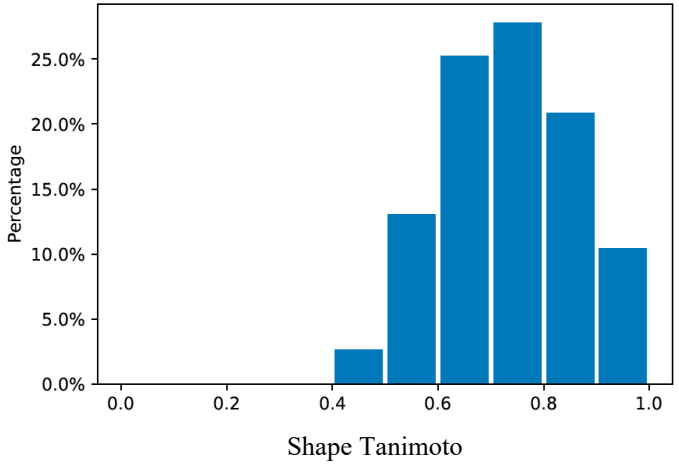
From *Knowledge Guided Geometric Editing for Unsupervised Drug Design*

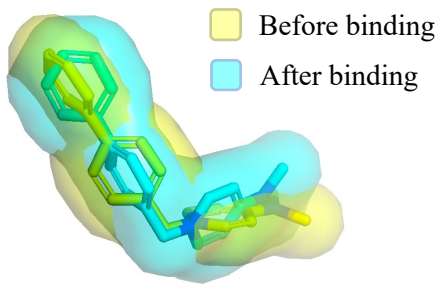
Related Work

- Challenge
  - Reduce the dependence on experiment data
  - Design molecules more efficiently

Motivation

- Massive data which are not from experiments
  - ZINC database has over 1000M predicted unbound molecules
- Molecular shape is stable when molecules bind to proteins
- Structure determines properties
  - Similar shape, similar properties
  - Complementary shape to the pocket, satisfactory binding affinity





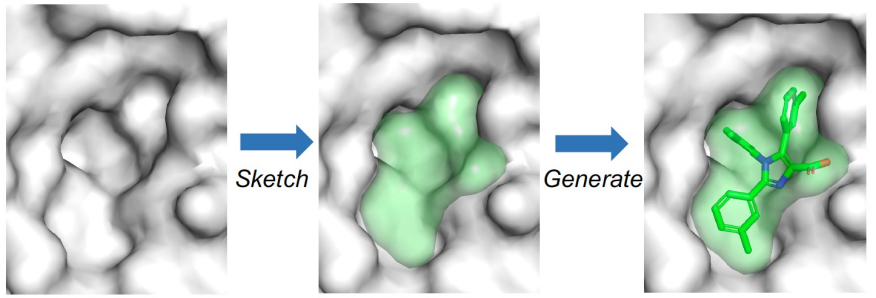
Shape Tanimoto: 0.72

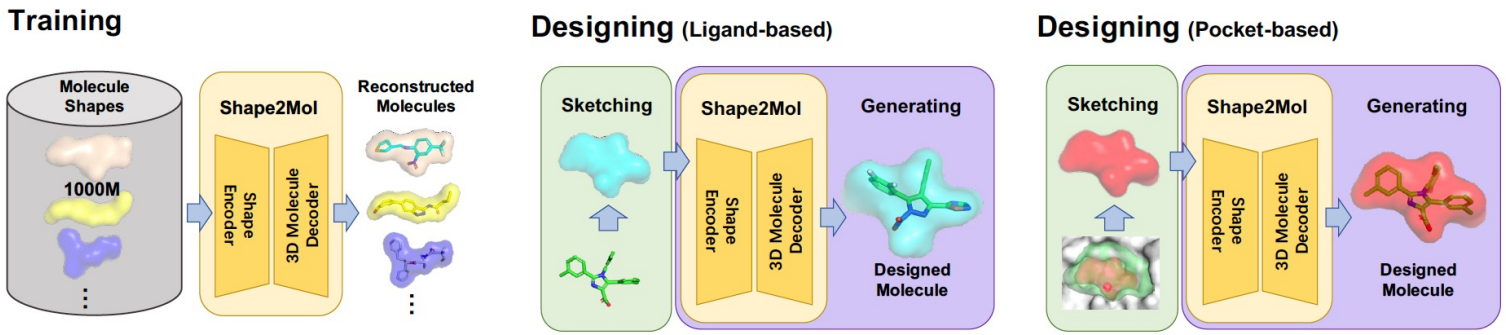
Motivation

- Idea
  - A shape-based molecule design model (pre-trained with massive unbound molecules) has the opportunity to generalize to the situation of designing bound molecules
  - Further, if we can obtain its input shape in a zero-shot way, the model can design drugs without experiment data and does not use the docking process to train itself

DESERT

- Sketch some reasonable shapes complementary to the target pocket
- Generate 3D molecules with a pre-trained model to fill these shapes

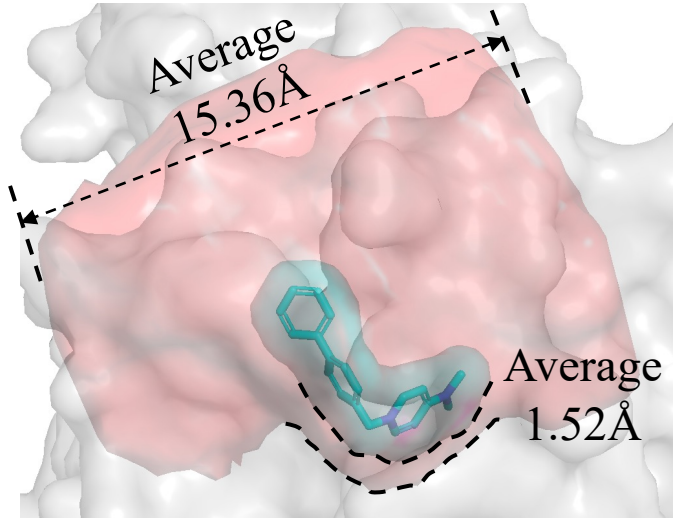




(a) Pre-training with massive unbound molecules (b) DESERT with ligand-based sketching (c) DESERT with pocket-based sketching

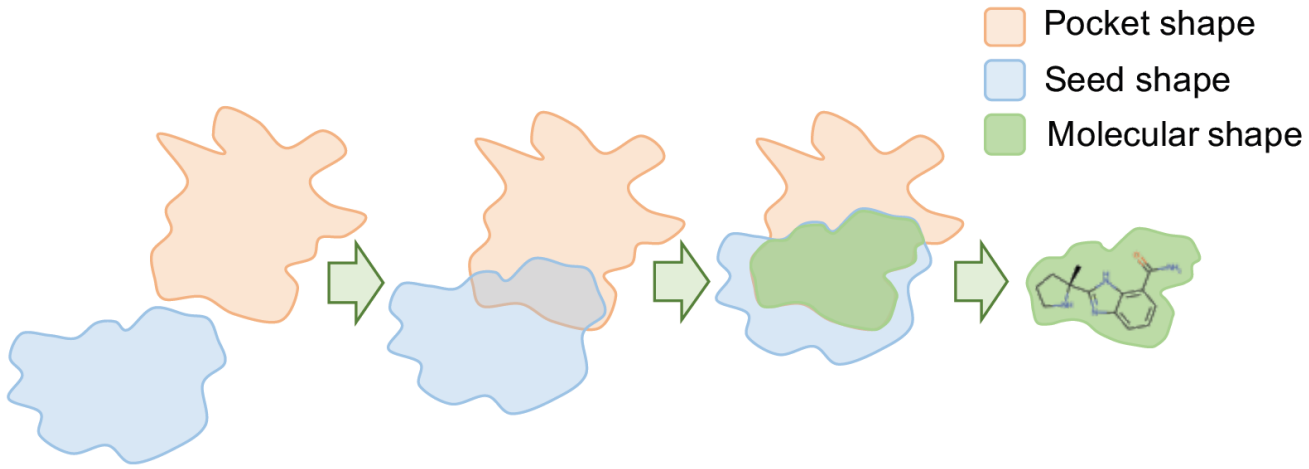
Sketching

- Ligand-based
  - Directly use the ligand's shape as the input
- Pocket-based
  - Pockets are much larger than ligands
  - Ligands lie in the area close to the pocket surface



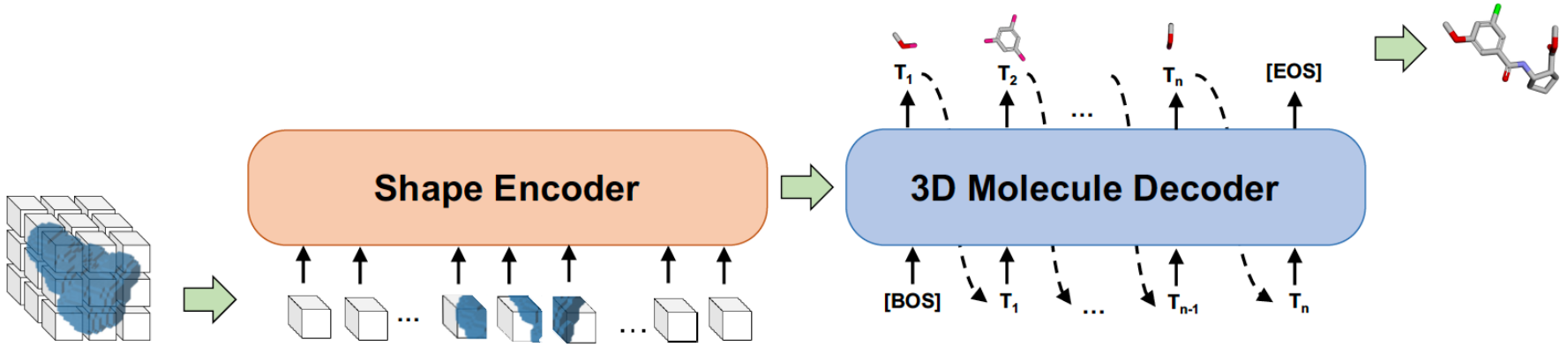
Sketching

- Pocket-based
  - Heuristically, we find a seed shape to intersect the pocket gradually
  - We get the seed shape by overlapping the shapes of several drug-like molecules
  - We stop the sketching process when the intersection has a similar size to a molecule



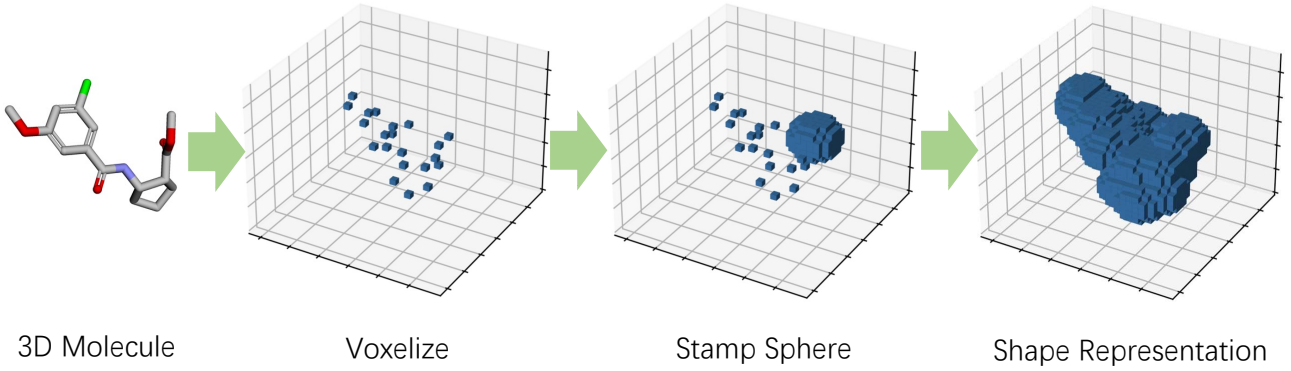
Generating

- Shape2Mol
  - Encoder-Decoder architecture
  - Pre-train with massive unbound molecules



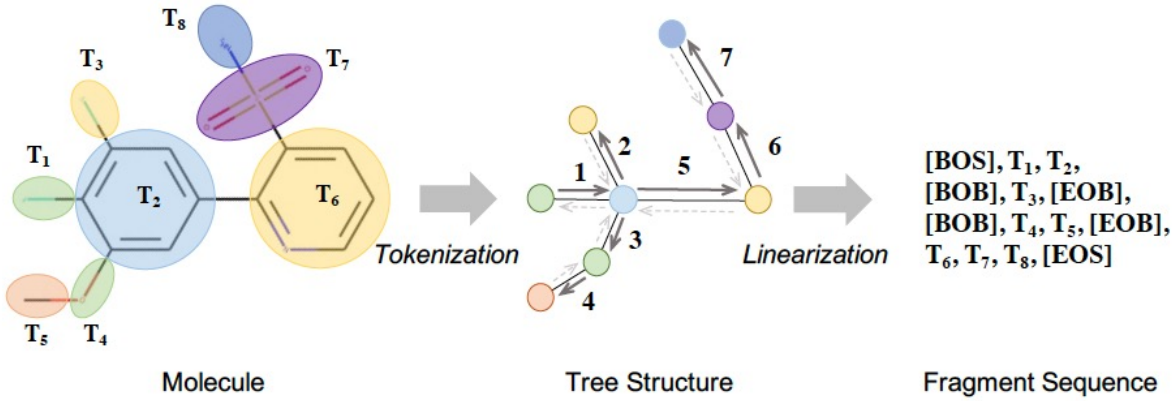
Shape2Mol

- Shape Encoder
  - Input 3D images, i.e., voxelized molecular shapes
  - Extend ViT for 2D images to 3D images



Shape2Mol

- 3D Molecule Decoder
  - Output the sequence of fragment tuples
    - Tokenization: turn a molecule graph into a tree structure object
    - Linearization: turn the tree into a fragment sequence
  - Modify the vanilla Transformer decoder for fragment sequences



$T_i = (C_i, P_i, R_i)$ where C_i is the fragment category, P_i is the discreterized translation vector, R_i is the discreterized rotation vector

Shape2Mol

- Training Loss

$$\mathcal{L} = -\sum_{i=1}^{n} \left\{ C_i \log \hat{C}_i + P_i^c \log \hat{P}_i^c + R_i^c \log \hat{R}_i^c \right\}$$

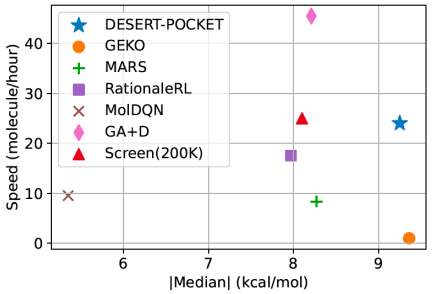
- Decoding Strategy
  - Nucleus sampling decoding strategy
  - Remove duplicate molecules
  - Leverage the docking process to drop results that do not pass the affinity threshold

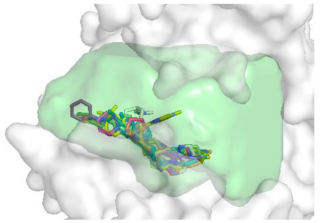
Experiment

- Baseline
  - 1D/2D: JT-VAE, MARS, etc.
  - 3D: liGAN, 3D SBDD, GEKO
  - Virtual Screening: SCREEN
- Dataset
  - 100M drug-like molecules from ZINC as training data
  - Following previous work, we test the method on 12 representative proteins
  - Only 2 of 12 have bioactivity data (Set B) and can be used to test 1D/2D methods
- Metrics
  - Each method provides 100 molecules for evaluation
  - Quality of designed molecular space: Uniq, Succ, Nov, Div, Prod
  - Ability of providing highly active molecules: Median (Median Vina Score)

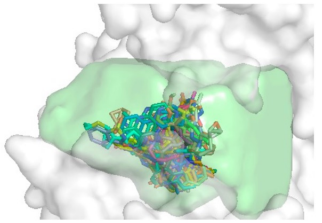
Main Results

Targets		Method	Uniq (%)↑	Succ (%)↑	Nov (%)↑	Div ↑	Prod ↑	Median (kcal/mol)↓
Set A	Guided	GEKO [10]	100.0	55.7	100.0	0.912	0.51	-9.58
	Supervised	liGAN [35]	100.0	0.4	100.0	0.924	0.00	-5.84
		3D SBDD [2]	69.7	13.6	98.9	0.839	0.08	-8.83
	Retrieved	SCREEN (1K)	100.0	25.6	100.0	0.892	0.23	-7.46
		SCREEN (200K)	100.0	64.0	100.0	0.889	**0.57**	-8.66
	Ours	DESERT-LIGAND	100.0	65.3	87.0	0.786	0.41	-8.89
		DESERT-POCKET	100.0	61.1	100.0	0.908	**0.57**	**-9.62**
Set B	Guided	JT-VAE [9]	100.0	13.0	100.0	0.907	0.12	-8.35
		RationaleRL [23]	100.0	27.0	35.0	0.884	0.08	-7.75
		GA + D [30]	39.0	24.0	87.0	0.852	0.06	-7.22
		GraphAF [26]	97.0	0.5	100.0	0.946	0.00	-4.22
		MolDQN [27]	76.5	0.0	100.0	0.742	0.00	-5.52
		MolEvol [69]	99.5	40.5	63.5	0.742	0.17	-8.19
		MARS [33]	86.0	31.5	93.0	0.805	0.22	-7.68
		GEKO [10]	100.0	57.0	100.0	0.910	0.52	-9.19
	Supervised	liGAN [35]	99.8	0.2	100.0	0.923	0.00	-5.34
		3D SBDD [2]	99.9	5.2	100.0	0.853	0.05	-6.39
	Retrieved	SCREEN (1K)	100.0	3.0	100.0	0.891	0.03	-6.94
		SCREEN (200K)	100.0	32.0	100.0	0.882	0.28	-7.95
	Ours	DESERT-LIGAND	100.0	18.0	100.0	0.913	0.17	-7.34
		DESERT-POCKET	100.0	61.0	100.0	0.907	**0.55**	**-9.32**





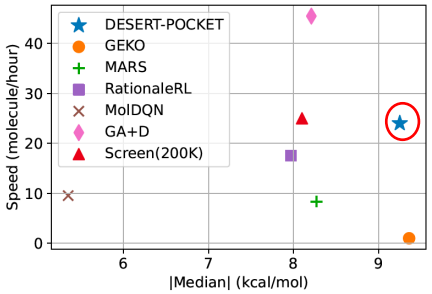
DESERT-LIGAND

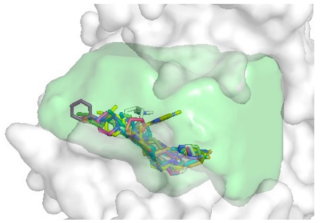


DESERT-POCKET

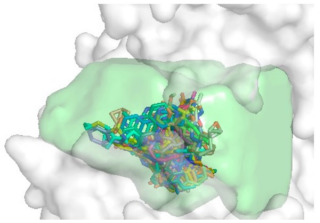
Main Results

Targets		Method	Uniq (%)↑	Succ (%)↑	Nov (%)↑	Div ↑	Prod ↑	Median (kcal/mol)↓
Set A	Guided	GEKO [10]	100.0	55.7	100.0	0.912	0.51	-9.58
	Supervised	liGAN [35]	100.0	0.4	100.0	0.924	0.00	-5.84
		3D SBDD [2]	69.7	13.6	98.9	0.839	0.08	-8.83
	Retrieved	SCREEN (1K)	100.0	25.6	100.0	0.892	0.23	-7.46
		SCREEN (200K)	100.0	64.0	100.0	0.889	**0.57**	-8.66
	Ours	DESERT-LIGAND	100.0	65.3	87.0	0.786	0.41	-8.89
		DESERT-POCKET	100.0	61.1	100.0	0.908	**0.57**	**-9.62**
Set B	Guided	JT-VAE [9]	100.0	13.0	100.0	0.907	0.12	-8.35
		RationaleRL [23]	100.0	27.0	35.0	0.884	0.08	-7.75
		GA + D [30]	39.0	24.0	87.0	0.852	0.06	-7.22
		GraphAF [26]	97.0	0.5	100.0	0.946	0.00	-4.22
		MolDQN [27]	76.5	0.0	100.0	0.742	0.00	-5.52
		MolEvol [69]	99.5	40.5	63.5	0.742	0.17	-8.19
		MARS [33]	86.0	31.5	93.0	0.805	0.22	-7.68
		GEKO [10]	100.0	57.0	100.0	0.910	0.52	-9.19
	Supervised	liGAN [35]	99.8	0.2	100.0	0.923	0.00	-5.34
		3D SBDD [2]	99.9	5.2	100.0	0.853	0.05	-6.39
	Retrieved	SCREEN (1K)	100.0	3.0	100.0	0.891	0.03	-6.94
		SCREEN (200K)	100.0	32.0	100.0	0.882	0.28	-7.95
	Ours	DESERT-LIGAND	100.0	18.0	100.0	0.913	0.17	-7.34
		DESERT-POCKET	100.0	61.0	100.0	0.907	**0.55**	**-9.32**





DESERT-LIGAND

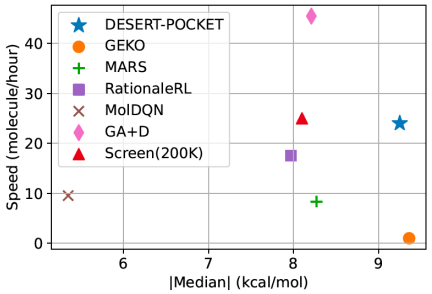


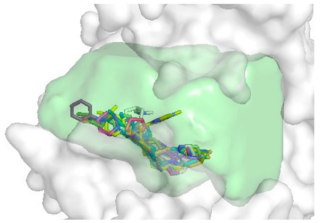
DESERT-POCKET

I. The zero-shot DESERT achieves the SOTA at fast speed

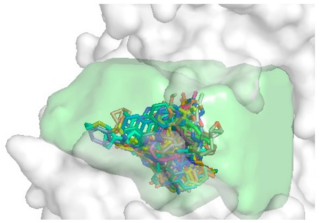
Main Results

Targets		Method	Uniq (%)↑	Succ (%)↑	Nov (%)↑	Div ↑	Prod ↑	Median (kcal/mol)↓
Set A	Guided	GEKO [10]	100.0	55.7	100.0	0.912	0.51	-9.58
	Supervised	liGAN [35]	100.0	0.4	100.0	0.924	0.00	-5.84
		3D SBDD [2]	69.7	13.6	98.9	0.839	0.08	-8.83
	Retrieved	SCREEN (1K)	100.0	25.6	100.0	0.892	0.23	-7.46
		SCREEN (200K)	100.0	64.0	100.0	0.889	**0.57**	-8.66
	Ours	DESERT-LIGAND	100.0	65.3	87.0	0.786	0.41	-8.89
		DESERT-POCKET	100.0	61.1	100.0	0.908	**0.57**	**-9.62**
Set B	Guided	JT-VAE [9]	100.0	13.0	100.0	0.907	0.12	-8.35
		RationaleRL [23]	100.0	27.0	35.0	0.884	0.08	-7.75
		GA + D [30]	39.0	24.0	87.0	0.852	0.06	-7.22
		GraphAF [26]	97.0	0.5	100.0	0.946	0.00	-4.22
		MolDQN [27]	76.5	0.0	100.0	0.742	0.00	-5.52
		MolEvol [69]	99.5	40.5	63.5	0.742	0.17	-8.19
		MARS [33]	86.0	31.5	93.0	0.805	0.22	-7.68
		GEKO [10]	100.0	57.0	100.0	0.910	0.52	-9.19
	Supervised	liGAN [35]	99.8	0.2	100.0	0.923	0.00	-5.34
		3D SBDD [2]	99.9	5.2	100.0	0.853	0.05	-6.39
	Retrieved	SCREEN (1K)	100.0	3.0	100.0	0.891	0.03	-6.94
		SCREEN (200K)	100.0	32.0	100.0	0.882	0.28	-7.95
	Ours	DESERT-LIGAND	100.0	18.0	100.0	0.913	0.17	-7.34
		DESERT-POCKET	100.0	61.0	100.0	0.907	**0.55**	**-9.32**





DESERT-LIGAND



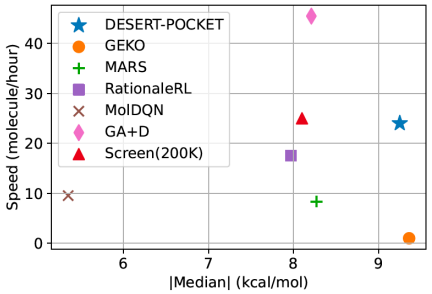
DESERT-POCKET

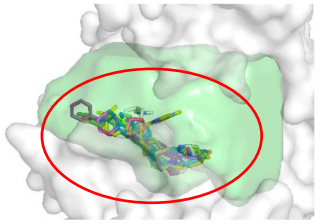
I. The zero-shot DESERT achieves the SOTA at fast speed

II. The shape helps DESERT produce high quality molecules

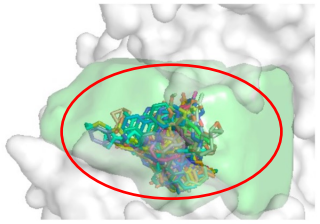
Main Results

Targets		Method	Uniq (%)↑	Succ (%)↑	Nov (%)↑	Div ↑	Prod ↑	Median (kcal/mol)↓
Set A	Guided	GEKO [10]	100.0	55.7	100.0	0.912	0.51	-9.58
	Supervised	liGAN [35]	100.0	0.4	100.0	0.924	0.00	-5.84
		3D SBDD [2]	69.7	13.6	98.9	0.839	0.08	-8.83
	Retrieved	SCREEN (1K)	100.0	25.6	100.0	0.892	0.23	-7.46
		SCREEN (200K)	100.0	64.0	100.0	0.889	**0.57**	-8.66
	Ours	DESERT-LIGAND	100.0	65.3	87.0	0.786	0.41	-8.89
		DESERT-POCKET	100.0	61.1	100.0	0.908	**0.57**	**-9.62**
Set B	Guided	JT-VAE [9]	100.0	13.0	100.0	0.907	0.12	-8.35
		RationaleRL [23]	100.0	27.0	35.0	0.884	0.08	-7.75
		GA + D [30]	39.0	24.0	87.0	0.852	0.06	-7.22
		GraphAF [26]	97.0	0.5	100.0	0.946	0.00	-4.22
		MolDQN [27]	76.5	0.0	100.0	0.742	0.00	-5.52
		MolEvol [69]	99.5	40.5	63.5	0.742	0.17	-8.19
		MARS [33]	86.0	31.5	93.0	0.805	0.22	-7.68
		GEKO [10]	100.0	57.0	100.0	0.910	0.52	-9.19
	Supervised	liGAN [35]	99.8	0.2	100.0	0.923	0.00	-5.34
		3D SBDD [2]	99.9	5.2	100.0	0.853	0.05	-6.39
	Retrieved	SCREEN (1K)	100.0	3.0	100.0	0.891	0.03	-6.94
		SCREEN (200K)	100.0	32.0	100.0	0.882	0.28	-7.95
	Ours	DESERT-LIGAND	100.0	18.0	100.0	0.913	0.17	-7.34
		DESERT-POCKET	100.0	61.0	100.0	0.907	**0.55**	**-9.32**





DESERT-LIGAND



DESERT-POCKET

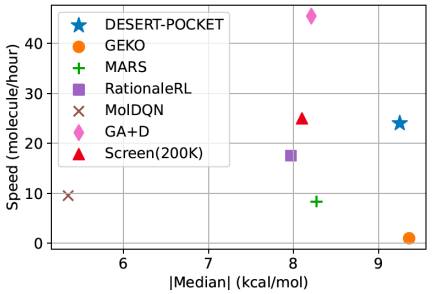
I. The zero-shot DESERT achieves the SOTA at fast speed

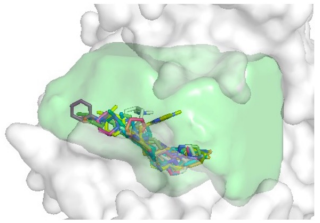
II. The shape helps DESERT produce high quality molecules

III. More comprehensive exploration of protein pockets benefits performance

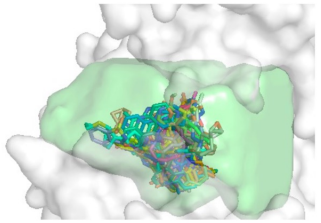
Main Results

Targets		Method	Uniq (%)↑	Succ (%)↑	Nov (%)↑	Div ↑	Prod ↑	Median (kcal/mol)↓
Set A	Guided	GEKO [10]	100.0	55.7	100.0	0.912	0.51	-9.58
	Supervised	liGAN [35]	100.0	0.4	100.0	0.924	0.00	-5.84
		3D SBDD [2]	69.7	13.6	98.9	0.839	0.08	-8.83
	Retrieved	SCREEN (1K)	100.0	25.6	100.0	0.892	0.23	-7.46
		SCREEN (200K)	100.0	64.0	100.0	0.889	**0.57**	-8.66
	Ours	DESERT-LIGAND	100.0	65.3	87.0	0.786	0.41	-8.89
		DESERT-POCKET	100.0	61.1	100.0	0.908	**0.57**	**-9.62**
Set B	Guided	JT-VAE [9]	100.0	13.0	100.0	0.907	0.12	-8.35
		RationaleRL [23]	100.0	27.0	35.0	0.884	0.08	-7.75
		GA + D [30]	39.0	24.0	87.0	0.852	0.06	-7.22
		GraphAF [26]	97.0	0.5	100.0	0.946	0.00	-4.22
		MolDQN [27]	76.5	0.0	100.0	0.742	0.00	-5.52
		MolEvol [69]	99.5	40.5	63.5	0.742	0.17	-8.19
		MARS [33]	86.0	31.5	93.0	0.805	0.22	-7.68
		GEKO [10]	100.0	57.0	100.0	0.910	0.52	-9.19
	Supervised	liGAN [35]	99.8	0.2	100.0	0.923	0.00	-5.34
		3D SBDD [2]	99.9	5.2	100.0	0.853	0.05	-6.39
	Retrieved	SCREEN (1K)	100.0	3.0	100.0	0.891	0.03	-6.94
		SCREEN (200K)	100.0	32.0	100.0	0.882	0.28	-7.95
	Ours	DESERT-LIGAND	100.0	18.0	100.0	0.913	0.17	-7.34
		DESERT-POCKET	100.0	61.0	100.0	0.907	**0.55**	**-9.32**





DESERT-LIGAND



DESERT-POCKET

I. The zero-shot DESERT achieves the SOTA at fast speed

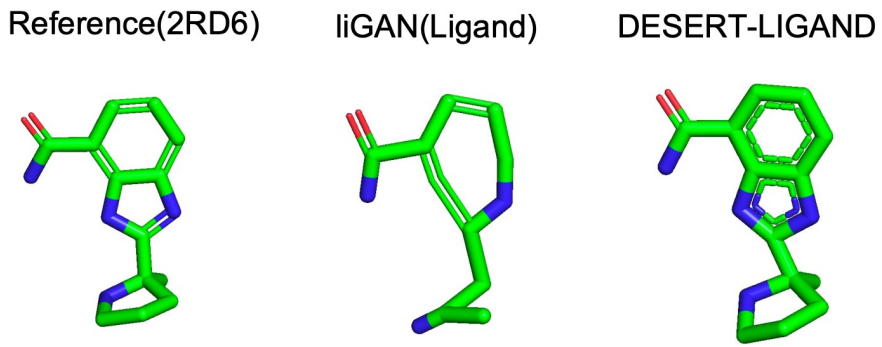
II. The shape helps DESERT produce high quality molecules

III. More comprehensive exploration of protein pockets benefits performance

IV. Unsupervised methods have larger potential than the supervised counterparts

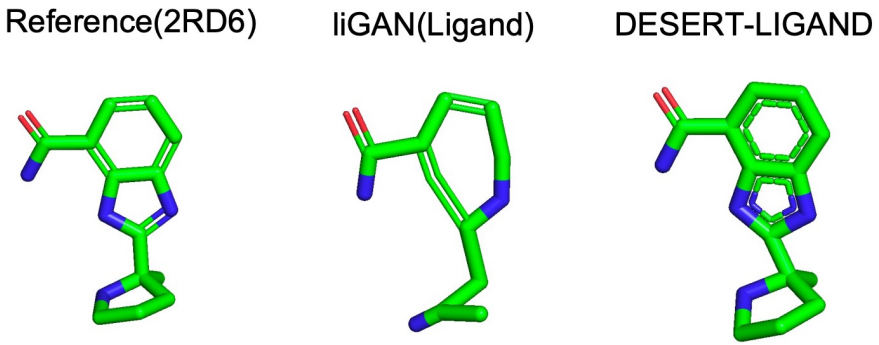
Shape Faithfulness & Structure Rationality

Method	Shape Tanimoto	Free Energy (kcal/mol)
Random	0.325	/
Real	/	167.28
liGAN (Ligand)	0.869	289.55
DESERT-LIGAND	0.875	188.54



Shape Faithfulness & Structure Rationality

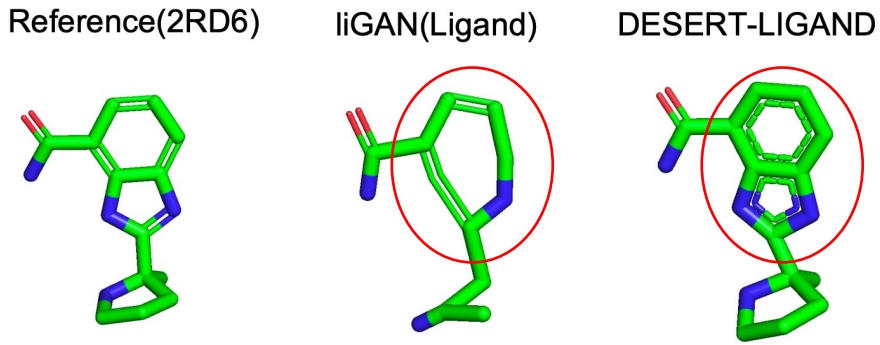
Method	Shape Tanimoto	Free Energy (kcal/mol)
Random	0.325	/
Real	/	167.28
liGAN (Ligand)	0.869	289.55
DESERT-LIGAND	0.875	188.54



I. DESERT can design molecules that fit shapes

Shape Faithfulness & Structure Rationality

Method	Shape Tanimoto	Free Energy (kcal/mol)
Random	0.325	/
Real	/	167.28
liGAN (Ligand)	0.869	289.55
DESERT-LIGAND	0.875	188.54

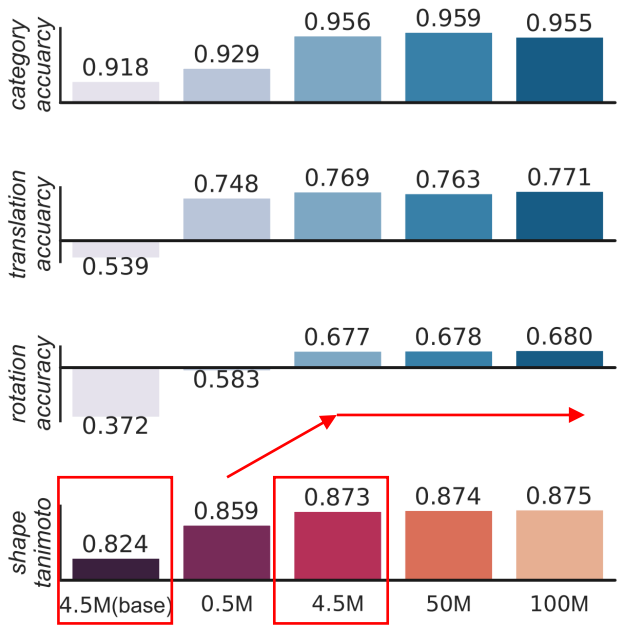


I. DESERT can design molecules that fit shapes

II. Fragments make DESERT's results have more correct molecular structures

Ablation Study (Generating)

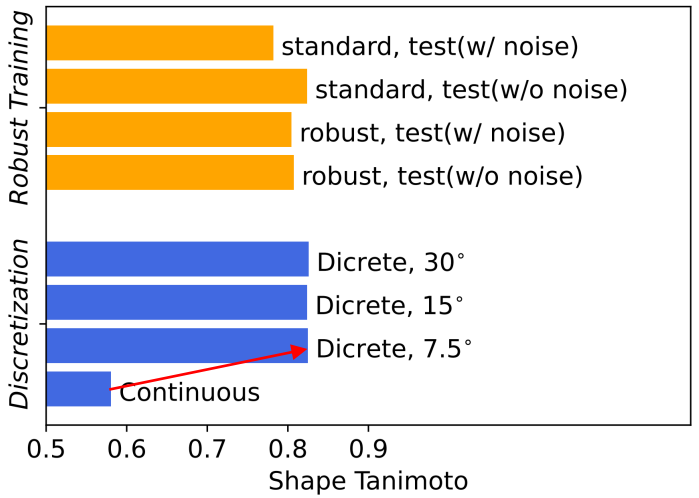
Pre-training Configuration



I. Larger model achieves better performance

II. Performance saturation occurs when the dataset is of moderate size

Robust Training & Discretization



I. Model trained with shape noise shows better performance when test data are noisy

II. Discretization consistently improves the result

Decoding Strategy

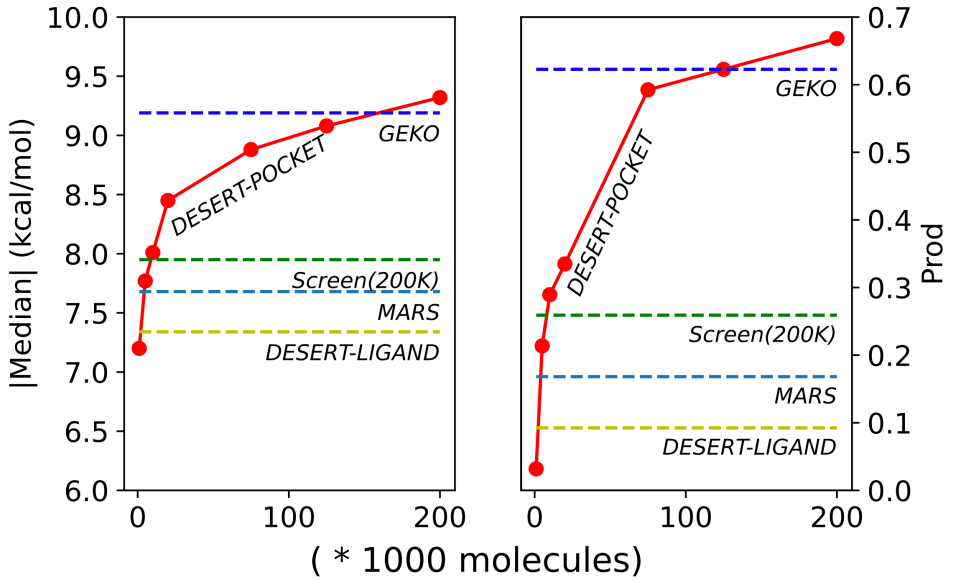
Sampling Method	Div	Prod	Median
Beam Search (Beam=10)	0.70	0.00	-5.87
Greedy Decoding	0.65	0.00	-5.83
Top K (K=10)	0.91	0.04	-5.99
Top P (P=0.95)	0.93	0.06	-6.01
+ post-processing	0.92	0.17	-7.34

I. Beam Search and Greedy Decoding are poor at providing diverse molecules

II. Top K and Top P can provide relatively diverse results

Ablation Study (Sketching)

Sampling Space Size



I. Increasing sampling space size leads to better performance

II. The shape can effectively prune the sampling space for screening

Seed Shape Type

Seed Shape	Succ	Prod	Median
None	82.0	0.74	-8.85
Sphere	75.5	0.68	-9.13
Molecules	61.0	0.55	-9.32

I. Directly using the pocket as the shape may contribute to a suboptimal result

Chemical Information

Method	Prod	Median
DESERT-POCKET	0.55	-9.32
+ chemical(Weak)	0.53	-9.19
+ chemical(Strong)	0.52	-9.03

I. *DESERT does not achieve better performance when we increase the effect of chemical information*

Conclusion

- We propose a zero-shot drug design method DESERT
- DESERT utilizes a large-scale molecular database to reduce the dependence on experimental data and docking simulation
- Experiments show that DESERT achieves a new state-of-the-art at a fast speed