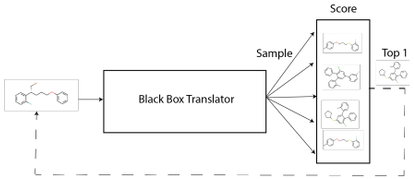
Black Box Recursive Translations for Molecular Optimization

Farhan Damani[1,2], Vishnu Sresht[2], Stephen Ra[2]

[1]Princeton University, [2]Pfizer Simulation & Modeling Sciences

Summary

- We propose a new inference method for molecular optimization called **Black Box Recursive Translation (BBRT)**.
- When generated molecules are iteratively fed back into the translator, properties improve with each step -- this finding is invariant to the choice of model.
- BBRT demonstrates results competitive with state-of-the-art for property optimization tasks using simple drop-in replacements with well-established models.
- We show BBRT generates samples with better properties relative to its non-recursive peers across different decoding strategies and is highly interpretable.



Motivation

Molecular optimization as a translation problem. As in [1], we learn a mapping between sequence pairs $(x, y) \in (X, Y)$ with high chemical similarity, and where y scores higher on a prespecified property compared to x.

- We utilize Seq2Seq with an encoder-decoder framework [2] , which learns parameters θ that estimate a conditional probability model $\mathrm{P}(y|x,\theta)$. θ is estimated by maximizing the log likelihood:

$$L(\theta) = \sum_{(x,y)\in(X,Y)} \log P(y|x,\theta)$$

- We explore both deterministic (beam search) and stochastic decoding strategies that sample from the model at generation time, $y_t \sim q(y_t|y_{<t}, x, p_\theta)$.
- We consider a top-k sampler [3] which restricts sampling to the k-most probable tokens at time-step t: a subset of vocabulary $U \subset V$ where U maximizes $\sum_{y\in U} p_\theta(y_t|y_{<t}, x)$:

$$q(y_t|y_{<t}, x, p_\theta) = \begin{cases} \frac{p_\theta(y_t|y_{<t},x)}{Z} & y_t \in U \\ 0 & otherwise \end{cases}$$

Recursive Inference

- We recursively infer new sequences, where $\{y_i^k\}_{k=1}^K$ is a set of K outputs generated from $p(y_i|x)$ at $i = 0$.
- We introduce functions S (right) to score K output sequences to determine subsequent iteration's source sequence $\hat{y}_i$. We infer K outputs from $p(y_i|\widehat{y_{i-1}})$ for n iterations.
- After n iterations, we ensemble the outputs $\{y_0, y_1, \ldots, y_n\}_{k=1}^K$ and score the sequences on a desired objective. For property optimization, we return the argmax.

Penalized logP: *Choose compound with max. logP value*

QED. *Choose compound with max. QED value.*

Max Δ Sim. *Choose compound with highest chemical similarity to previous iteration's source compound.*

Max Init Sim. *Choose compound with highest similarity to initial seed compound.*

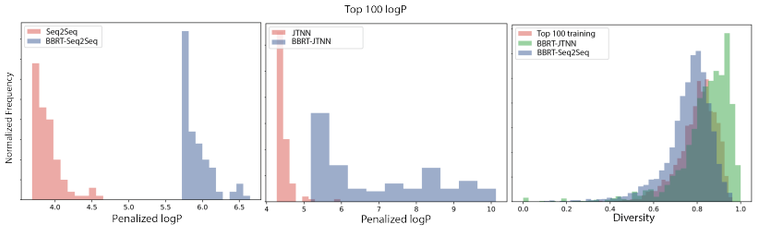
Min Mol Wt. *Choose compound with min. molecular weight.*

Experiments

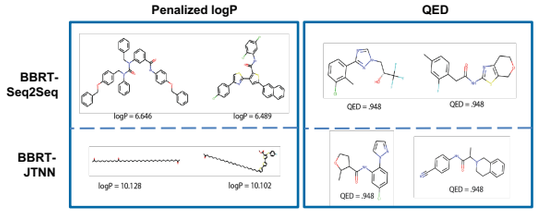
- We apply recursive translation to both sequence- and graph-based translation models and report top 3 property scores on penalized logP and QED optimization tasks against baseline models.

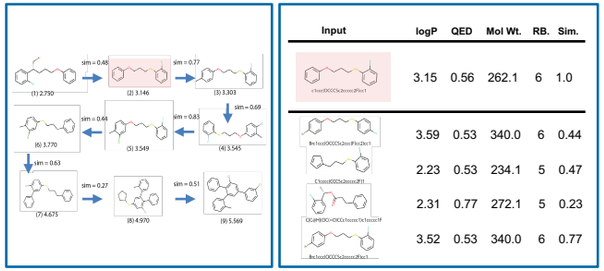
	Penalized logP			QED		
Method	1st	2nd	3rd	1st	2nd	3rd
ZINC-250K	4.52	4.30	4.23	0.948	0.948	0.948
ORGAN	3.63	3.49	3.44	0.896	0.824	0.820
JT-VAE	5.30	4.93	4.49	0.925	0.911	0.910
GCPN	7.98	7.85	7.80	0.948	0.947	0.946
JTNN	5.97	4.96	4.71	0.948	0.948	0.948
Seq2Seq	4.65	4.53	4.49	0.948	0.948	0.948
BBRT-JTNN	10.13	10.10	9.91	0.948	0.948	0.948
BBRT-Seq2Seq	6.74	6.47	6.42	0.948	0.948	0.948

- We find stochastic decoding methods outperformed deterministic methods on average logP scores and average pairwise diversity for generated compounds as a function of recursive iteration.
- We report top 100 logP generated compounds under both BBRT models, and non-recursive counterparts **(left and center)** as well as diversity of top 100 generated compounds and training set **(right)**.



- Top scoring compounds for logP and QED under both BBRT models below; for logP, BBRT-JTNN produces compounds with higher property values and BBRT-Seq2Seq generates compounds with a richer molecular vocabulary.
- BBRT generates interpretable paths of optimization **(left)** that facilitate understanding of design trade-offs We highlight alternative translations for compound (2) below **(right)**:





References

[1] Wengong Jin, Kevin Yang, Regina Barzilay, and Tommi Jaakola. Learning multimodal graph-to-graph translation for molecule optimization. In *International Conference on Learning Representations*, 2019.

[2] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems – Volume 2.*, pp. 3104-3112, 2014.

[3] Angela Fan, Mike Lewis, and Yann N. Dauphin. Hierarchical neural story generation. *ACL*, pp. 889-898, 2018.

@FarhanD7 @stephenrra

Read our pre-print on arXiv: