Barking up the right tree: an approach to search over molecule synthesis DAGs

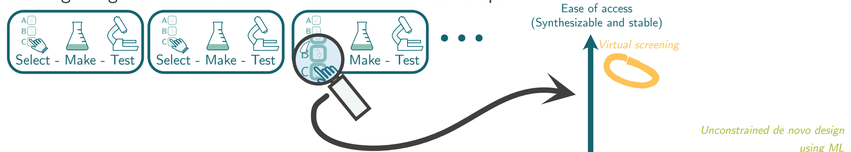
John Bradshaw[1,2], Brooks Paige[3,4], Matt J. Kusner[3,4], Marwin H.S. Segler[5,6], José Miguel Hernández-Lobato[1,4,6]

1 University of Cambridge, 2 MPI for Intelligent Systems, 3 University College London, 4 The Alan Turing Institute, 5 WWU Münster, 6 Microsoft Research Cambridge UK

Aim: design an approach for generating and searching over stable and (multi-step) synthesizable molecules (e.g. for drugs).

1. We want to find the best synthesizable and stable molecule

1. Drug design consists of a series of **select-make-test** steps

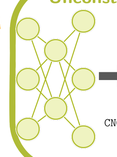


2. Recent ML approaches have traded off **ease of access** for **chemical space coverage**

Virtual screening

Virtual screening methods [1] pick from a dataset of pre-specified molecules.

Unconstrained de novo design using ML

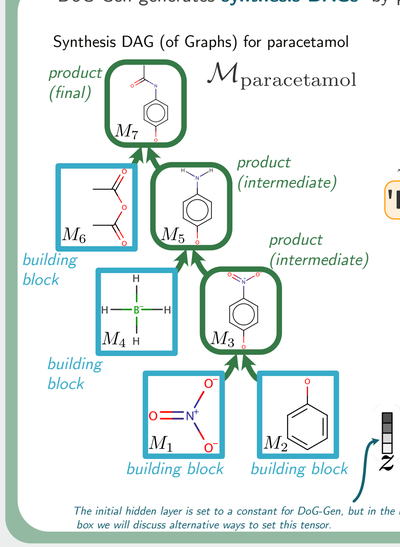


ML methods can generate new molecules by decoding to text (e.g. [2,3,4]) or graph representations (e.g. [5,6,7])

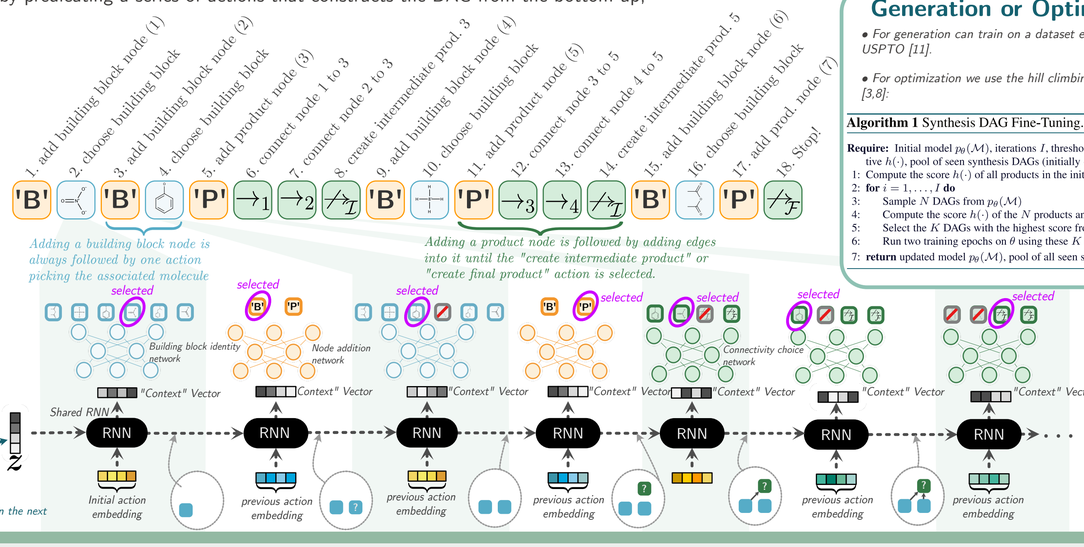
3. Can we choose suitable inductive biases to have **both,** i.e. **generate and search over** a wide range of molecules whilst ensuring that the molecules we generate are **synthesizable**.

2. We therefore propose a model, DoG-Gen, a generative model of synthesis DAGs

DoG-Gen generates **synthesis DAGs** by predicating a series of actions that constructs the DAG from the bottom up,



The initial hidden layer is set to a constant for DoG-Gen, but in the next box we will discuss alternative ways to set this tensor.



Generation or Optimization?

- *For generation can train on a dataset extracted from USPTO [11].*
- *For optimization we use the hill climbing technique from [3,8]:*

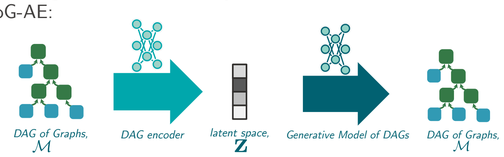
Algorithm 1 Synthesis DAG Fine-Tuning.

Require: Initial model $p_\theta(\mathcal{M})$, iterations I, threshold K, sample size N, objective $h(\cdot)$, pool of seen synthesis DAGs (initially empty), P.

Compute the score $h(\cdot)$ of all products in the initial training set and add to P
for $i = 1, \ldots, I$ **do**
  Sample N DAGs from $p_\theta(\mathcal{M})$
  Compute the score $h(\cdot)$ of the N products and add to pool, P
  Select the K DAGs with the highest score from P
  Run two training epochs on θ using these K DAGs as training data
return updated model $p_\theta(\mathcal{M})$, pool of all seen synthesis DAGs (ranked) P

3. We also propose an autoencoder variant, DoG-AE

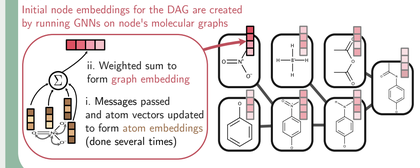
- We can use our generative model of synthesis DAGs as the decoder in an autoencoder structure DoG-AE:



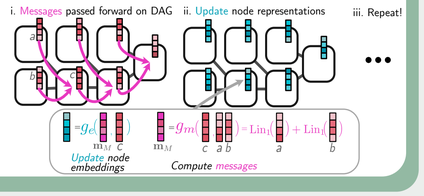
- As the encoder we propose a hierarchical message passing procedure:

(1) Molecular graph message passing

Initial node embeddings for the DAG are created by running GNNs on node's molecular graphs



(2) Synthesis graph message passing.



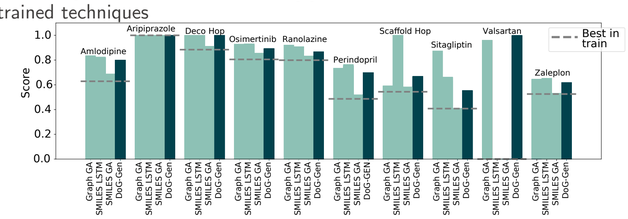
4. We can generate new molecules (& DAGs)

We generate **20k samples** from our model and look at validity (whether they can be parsed by RDKit). Conditioned on validity we look at uniqueness, novelty, quality (train-dataset-normalized proportion of molecules that pass the quality filters proposed in [8]), and finally Fréchet ChemNet Distance (FCD) [9].

Model Name	Validity (↑)	Uniqueness (↑)	Novelty (↑)	Quality (↑)	FCD (↓)
DoG-AE	100.0	98.3	92.9	95.5	0.83
DoG-Gen	100.0	97.7	88.4	101.6	0.45
Training Data	100.0	100.0	0.0	100.0	0.21
SMILES LSTM [3]	94.8	95.5	74.9	101.93	0.46
CVAE [2]	96.2	97.6	76.9	103.82	0.43
GVAE [4]	74.4	97.8	82.7	98.98	0.89
GraphVAE [5]	42.2	57.7	96.1	94.64	13.92
JT-VAE [7]	100.0	99.2	94.9	102.34	0.93
CGVAE [6]	100.0	97.8	97.9	45.64	14.26
Molecule Chef [10]	98.9	96.7	90.0	99.0	0.79

5. We can optimize for GuacaMol score, whilst generating synthesizable molecules

When optimizing over 10 GuacaMol tasks [8] we find we can obtain **comparable** scores to unconstrained techniques



However, we also generate more **stable** and **synthesizable** molecules:

References

[1] Brian K Shoichet. *Virtual screening of chemical libraries.* Nature, 2004;

[2] Gómez-Bombarelli et al. *Automatic chemical design using a Data-Driven continuous representation of molecules.* ACS Cent Sci, 2018;

[3] Segler et al. *Generating focused molecule libraries for drug discovery with recurrent neural networks.* ACS Cent. Sci., 2018;

[4] Kusner et al. *Grammar variational autoencoder.* ICML, 2017;

[5] Simonovsky and Komodakis. *GraphVAE: Towards generation of small graphs using variational autoencoders.* ICANN, 2018;

[6] Liu et al. *Constrained graph variational autoencoders for molecule design.* NeurIPS, 2018;

[7] Jin et al. *Junction tree variational autoencoder for molecular graph generation.* ICML, 2018;

[8] Brown et al. *GuacaMol: benchmarking models for de novo molecular design.* Journal of Chemical Information and Modeling, 2019;

[9] Preuer et al. *Fréchet ChemNet Distance: A metric for generative models for molecules in drug discovery.* Journal of Chemical Information and Modeling, 2018 ;

[10] Bradshaw et al. *A model to search for synthesizable molecules.* NeurIPS, 2019 ;

[11] Lowe. *Extraction of chemical structures and reactions from the literature.* PhD Thesis, University of Cambridge, 2012.

github.com/john-bradshaw/synthesis-dags

jab255@cam.ac.uk, b.paige@ucl.ac.uk, m.kusner@ucl.ac.uk, marwin.segler@wwu.de, jmh233@cam.ac.uk