A Model to Search for Synthesizable Molecules

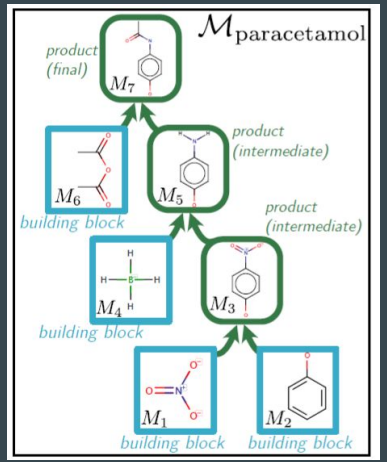
Authors: John Bradshaw, Brooks Paige, Matt J. Kusner, Marwin H. S. Segler, José Miguel Hernández-Lobato

Presenter: Jun Chen

Recap

- This paper : (2019.)

- Barking up the right tree: an approach to search over molecule synthesis DAGs (2020.)



Introduction:

- Deep generative models allows one to generate molecules with desirable properties, but they give no guarantees that the molecules can be synthesized in practice.

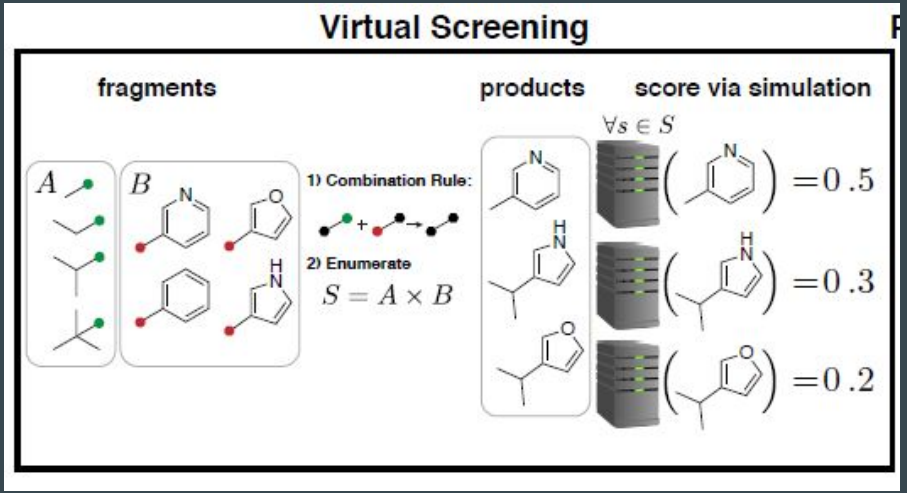
Background: Finding molecules with desirable properties

Approach 1 : Virtual Screening

- Find novel molecules by enumeration over all possible combinations of fragment.

Problems:

- Computational expensive
- Does not search for desirable properties.



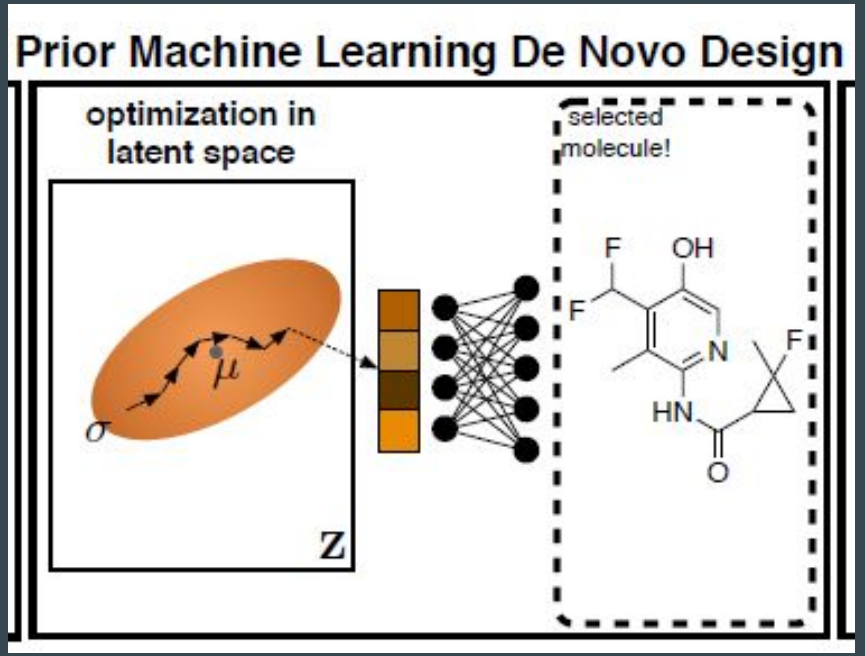
Finding molecules with desirable properties

Approach 2 : ML approaches

- Find novel molecules by optimizing in a continuous latent space.

Problems:

- No guarantee the molecules can be synthesized



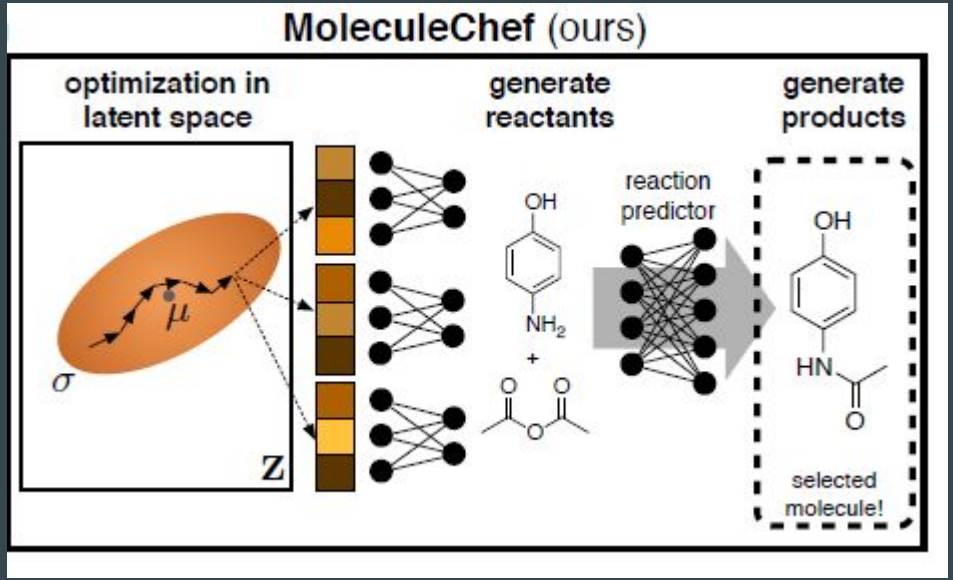
Reaction predictors

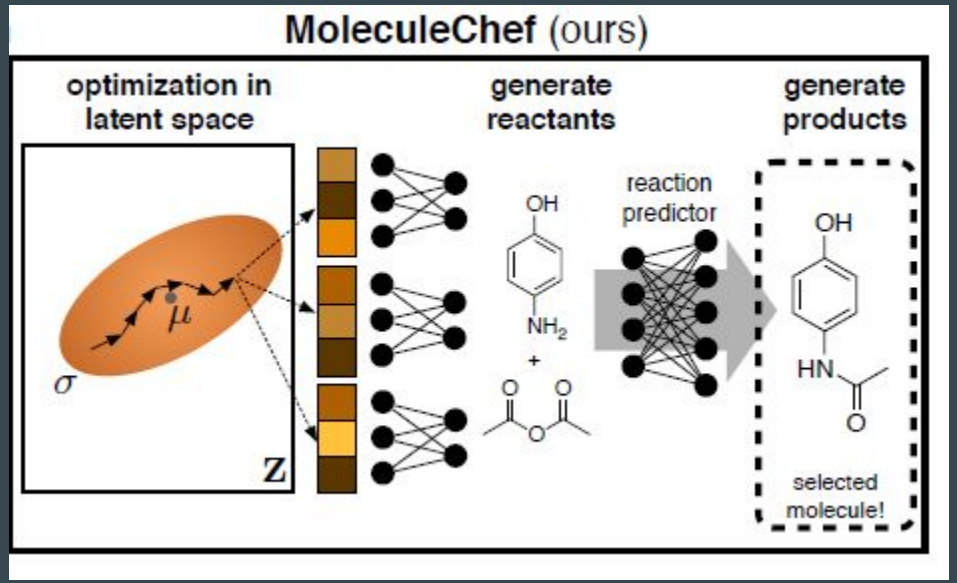
- works by recursively deconstructing a molecule. This deconstruction is done via (reversed) reaction predictors: models that predict how reactant molecules produce a product molecule.

MoleculeChef

- First, a mapping from continuous space to a set of known, reliable, easy-to-obtain reactant molecules.

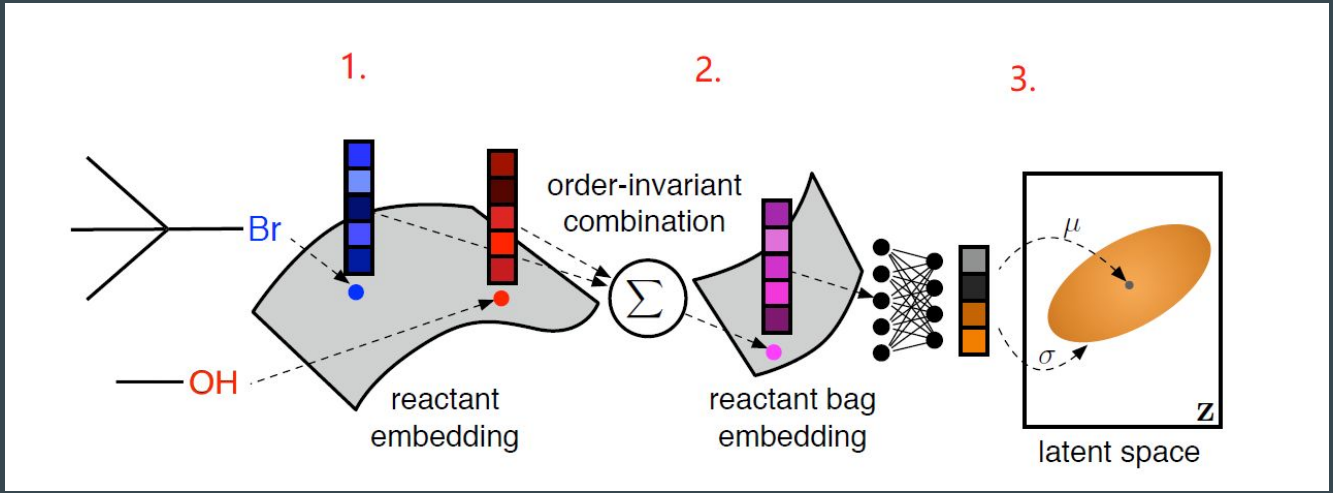
- Second a mapping from this set of reactant molecules to a final product molecule, based on a reaction prediction model.





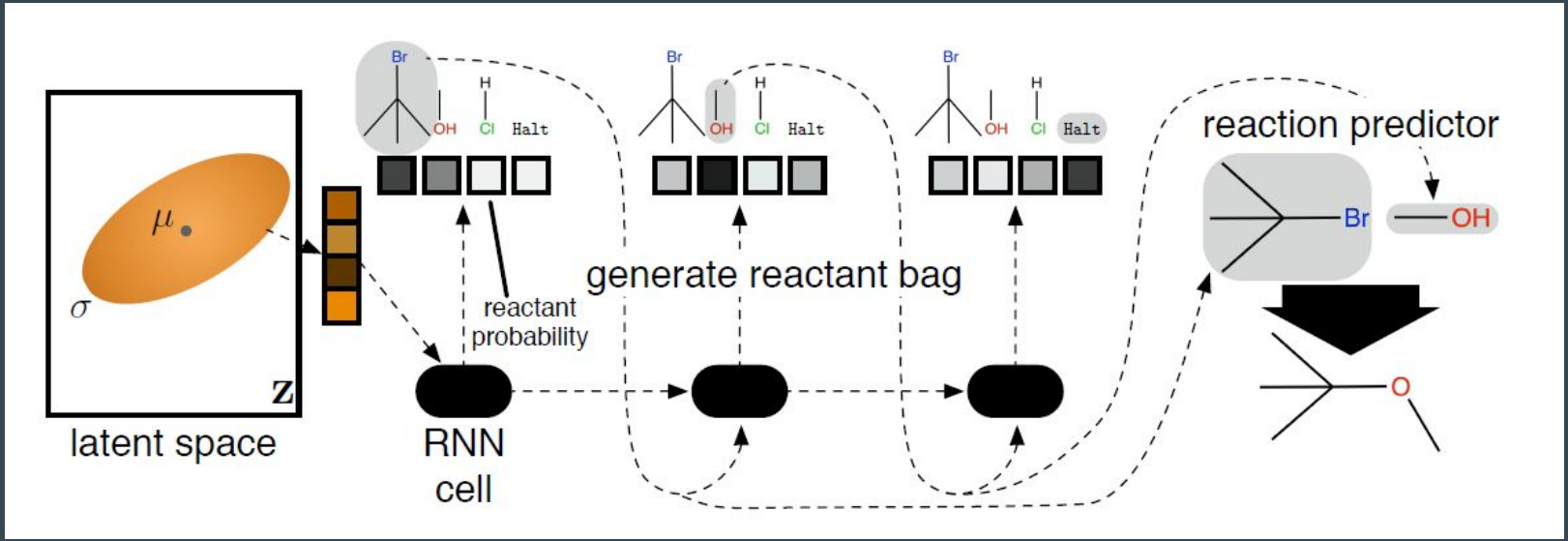
1. Decoder: a decoder from a continuous latent space, to a multiset of easily obtainable reactants.
2. A reaction predictor: to transforms the multiset of easily obtainable reactants into a multiset of product molecules. (Molecular Transformer)

Encoder



1. The reactants molecules are embedded into a continuous space by using GGNNs to form molecule embeddings.
2. The molecule embeddings in the multiset are summed to form one order-invariant embedding for the whole multiset
3. Used it as input to neural network which parameterized a Gaussian distribution over z.

Decoder



1. Generates the multiset of reactants in sequence through calls to a RNN.
2. Latent vector z is used to parameterize the initial hidden layer of the RNN.
3. Selected reactants are fed back into RNN next step.
4. The reactant multiset is later fed through a reaction predictor to form a final product.

Results

Table 1: Table showing the validity, uniqueness, novelty and normalized quality (all as %, higher better) of the products/or molecules generated from decoding from 20k random samples from the prior $p(\mathbf{z})$. Quality is the proportion of valid molecules that pass the quality filters proposed in Brown et al. [7, §3.3], normalized such that the score on the training set is 100. FCD is the Fréchet ChemNet Distance [36], capturing a notion of distance between the generated valid molecules and the training dataset (lower better). The uniqueness and novelty figures are also conditioned on validity. MT stands for the Molecular Transformer [44].

Model Name	Validity	Uniqueness	Novelty	Quality	FCD
MOLECULE CHEF + MT	99.05	95.95	89.11	95.30	0.73
AAE [23, 35]	85.86	98.54	93.37	94.89	1.12
CGVAE [30]	100.00	93.51	95.88	44.45	11.73
CVAE [14]	12.02	56.28	85.65	52.86	37.65
GVAE [27]	12.91	70.06	87.88	46.87	29.32
LSTM [46]	91.18	93.42	74.03	100.12	0.43

Optimization

- Property predictor network is train when training the Moleculechef simultaneously, mapping from latent space of Moleculechef to QED score of the final product.
- 250 bags of reactants encode into the latent space of Moleculechef, and repeatedly moving in the latent space using the gradient direction of the property predictor
- Decoded 10 different reaction bags compare 10 from random walk.

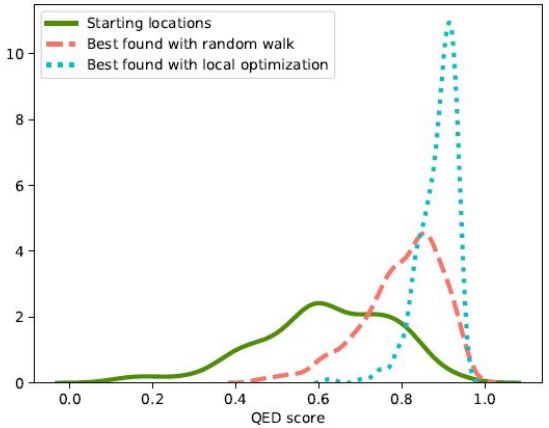


Figure 4: KDE plot showing that the distribution of the best QEDs found through local optimization, using our trained property predictor for QEDs, has higher mass over higher QED scores compared to the best found from a random walk. The starting locations' distribution (sampled from the training data) is shown in green. The final products, given a reactant bag are predicted using the MT [44].

Retrosynthesis

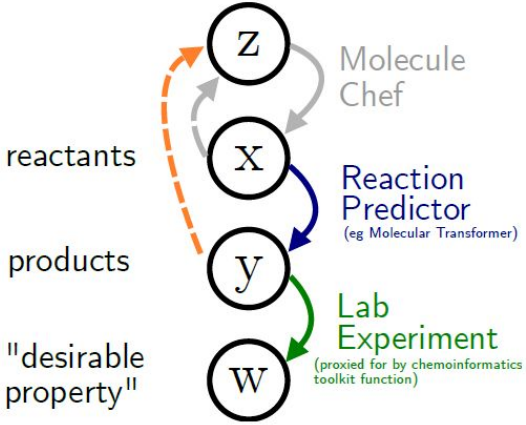


Figure 5: Having learnt a latent space which can map to products through reactants, we can learn a regressor back from the suggested products to the latent space (orange dashed –– arrow shown) and couple this with MOLECULE CHEF's decoder to see if we can do retrosynthesis – the act of computing the reactants that create a particular product.

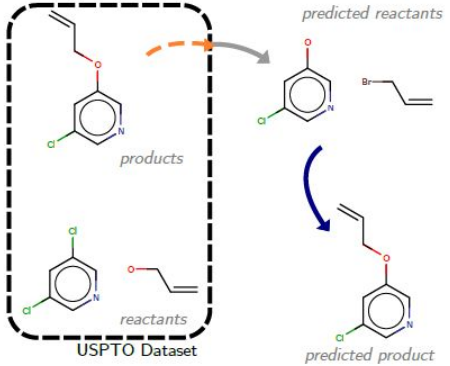
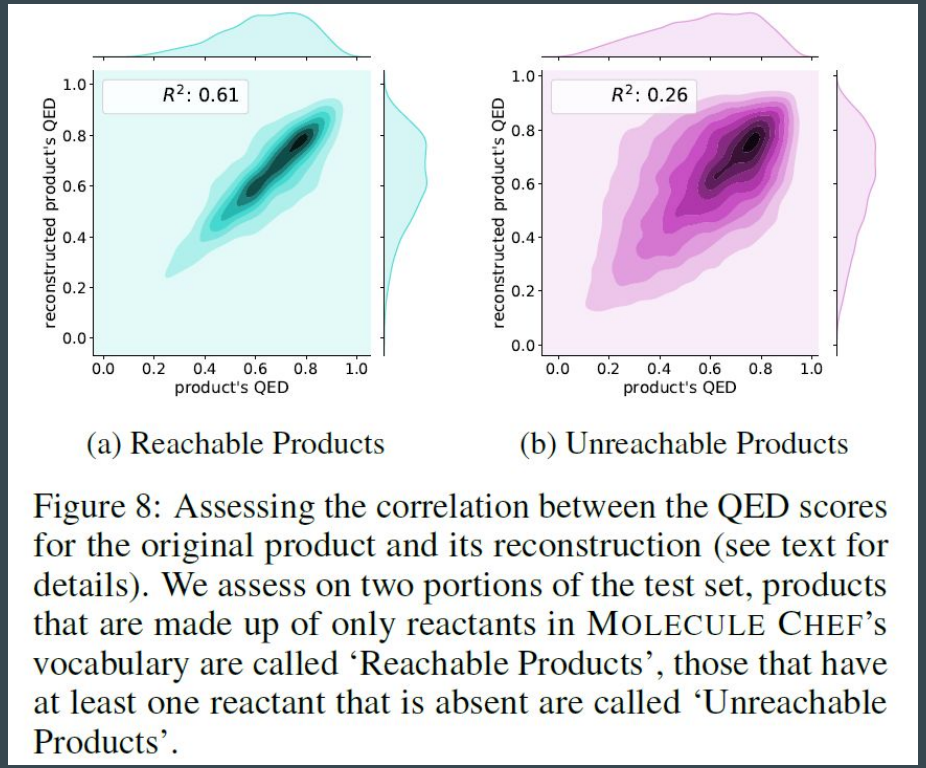


Figure 6: An example of performing retrosynthesis prediction using a trained regressor from products to latent space. This reactant-product pair has not been seen in the training set of MOLECULE CHEF. Further examples are shown in the appendix.

Results

- Product -> retorsysthiss
  = bags of reactants.
- Bags of reactants -> reaction predator(MT) = reconstructed product
- Compare the QED of original product and reconstructed product.



(a) Reachable Products (b) Unreachable Products

Figure 8: Assessing the correlation between the QED scores for the original product and its reconstruction (see text for details). We assess on two portions of the test set, products that are made up of only reactants in MOLECULE CHEF's vocabulary are called 'Reachable Products', those that have at least one reactant that is absent are called 'Unreachable Products'.

Results

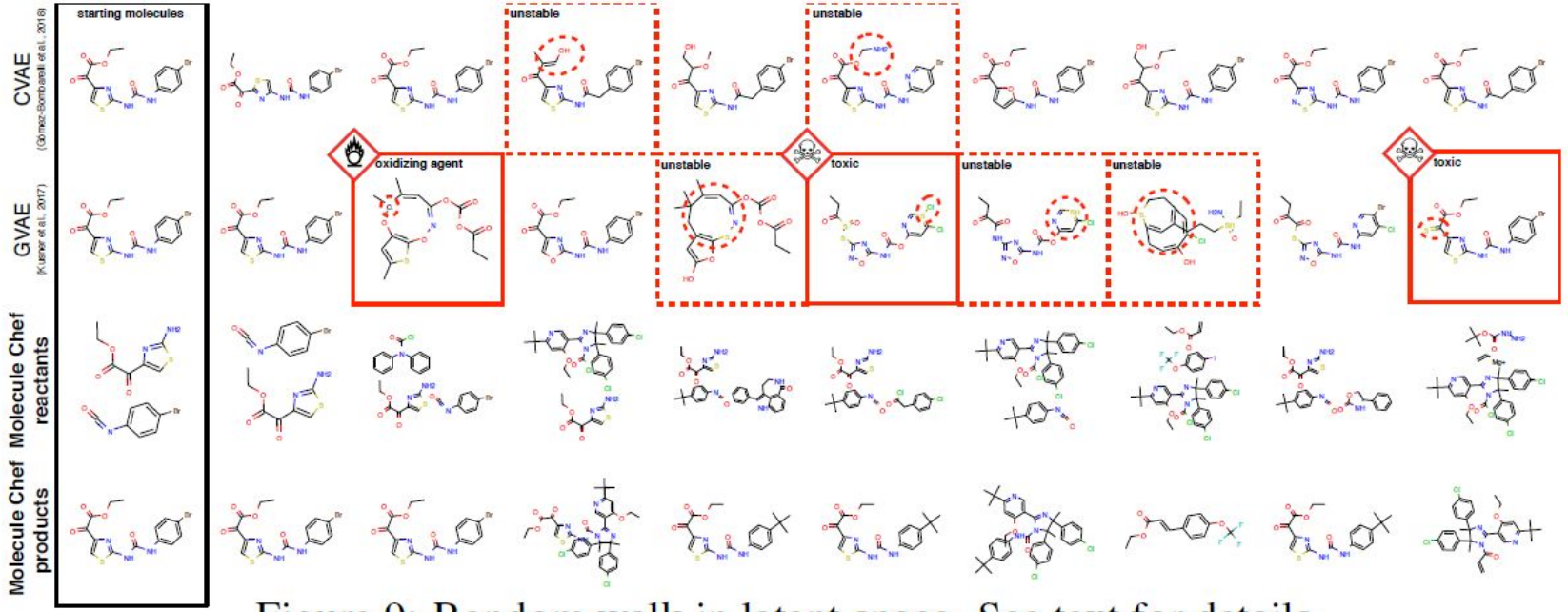


Figure 9: Random walk in latent space. See text for details.

Thank you