A Model to Search for Synthesizable Molecules

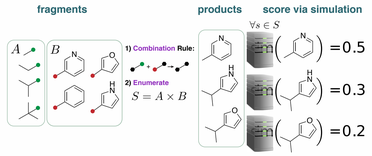
John Bradshaw[1,2], Brooks Paige[1,4], Matt J. Kusner[3,4], Marwin H. S. Segler[5,6], José Miguel Hernández-Lobato[1,4,7]

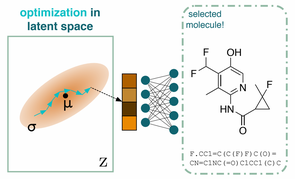
[1]University of Cambridge, [2]MPI for Intelligent Systems, [3]University College London, [4]The Alan Turing Institute, [5]BenevolentAI, [6]Westfälische Wilhelms-Universität Münster, [7]Microsoft Research Cambridge.

Aim: to design a generative model that enables the searching for useful molecules (eg for drugs) over its continuous latent space, whilst having a decoder which produces both stable chemical products and their synthetic routes.

1. We don't just want to know what molecule to make...

Virtual screening can be used to narrow down the number of candidates when searching for desirable drug molecules. However, this often consists of expensive **combination and enumeration** steps.

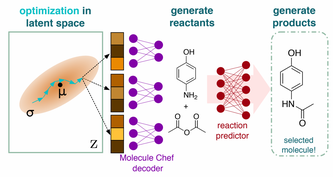




Recent machine learning approaches [1–3] have proposed learning an autoencoder to encode and decode molecules to a continuous latent space. The costly enumeration above can instead be replaced by **search** (eg **local optimization**).

But outstanding questions for these ML methods remain: (1) (how) are the proposed molecules **synthesizable**, and (2) how can we better imbue our models with inductive biases that result in **semantically valid (chemically stable) molecules**?

2. ...we also want to know how to make it! Therefore, our model, MOLECULE CHEF, generates reactant bags!



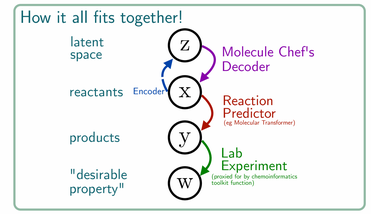
Our model, Molecule Chef, decodes using a **two step process**.

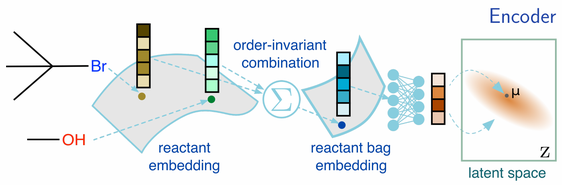
1. The decoder first maps from the latent space to a *reactant bag*.
2. This is then fed through a **reaction predictor** model (we use the Molecular Transformer [4]) to predict a final product.

By using stable reactant building blocks, we hope that our model proposes more semantically valid molecules, ie molecules that are non toxic or not about to break down.

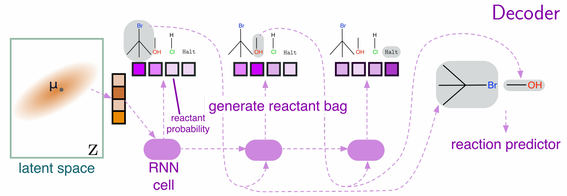
We train the model (on a dataset derived from USPTO [8]) using the WAE objective [7], which involves minimizing:

$$L = \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\left[c(\mathbf{x}, p(\mathbf{x}|\mathbf{z}))\right] + \lambda D\left(\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[q(\mathbf{z}|\mathbf{x})\right], p(\mathbf{z})\right)$$





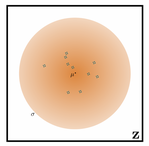
The **encoder** embeds each reactant using a graph neural network. These embeddings are summed to produce an order-invariant embedding.



The **decoder** uses a RNN to sequentially output reactants. Reactants are selected from a fixed set of 3180 easily obtainable and common molecules.

3. MOLECULE CHEF generates a wide range of stable molecules

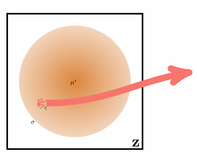
Following previous work we **sample 20000 molecules from the prior**. We assess these molecules for validity (whether they can be parsed by chemoinformatics software), and conditioned on that: uniqueness, novelty wrt training set, and quality (normalized proportion of molecules that pass the quality filters proposed in [13]).

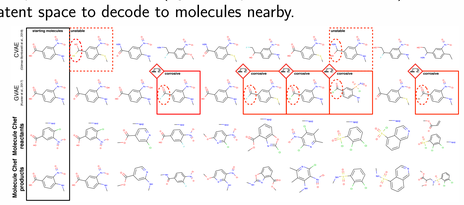


Model Name	Validity	Uniqueness	Novelty	Quality	FCD
Molecule Chef + MT	99.05	95.95	89.11	95.30	0.73
AAE [11, 12]	85.86	98.54	93.37	94.89	1.12
CGVAE [9]	100.00	93.51	95.88	44.45	11.73
CVAE [1]	12.02	56.28	85.65	52.86	37.65
GVAE [2]	12.91	70.06	87.88	46.87	29.32
LSTM [10]	91.18	93.42	74.03	100.12	0.43

Tab. 1: Table showing the validity, uniqueness, novelty, and normalized quality (all as %, with uniqueness, novelty, and quality conditioned on validity) of the products/or molecules generated from decoding from 20k random samples from the prior $p(\mathbf{z})$. MT stands for the Molecular Transformer [4]. FCD is the Fréchet ChemNet Distance between valid molecules and training set.

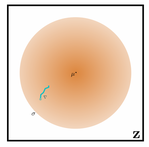
We also qualitatively evaluate the semantic validity of our molecules (eg are they stable and non-toxic). To do this we start from a training molecule and **randomly walk** in the latent space to decode to molecules nearby.

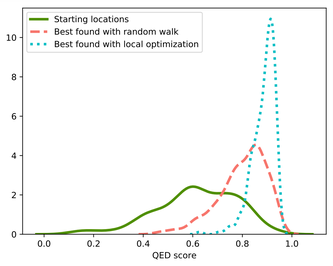




4. MOLECULE CHEF's latent space can be used for optimization

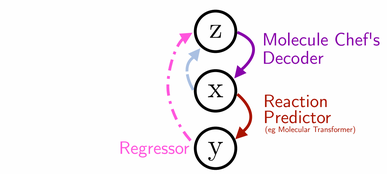
We train a regression network from the latent space to a property of interest, the QED (quantitative estimate of drug-likeness). This regression network can be used for **local optimization** and we compare the QED of the best molecules found this way to those found from a **random walk**.

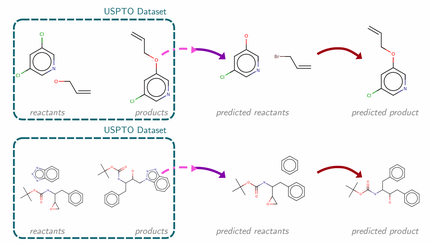




5. We can also use MOLECULE CHEF for retrosynthesis!

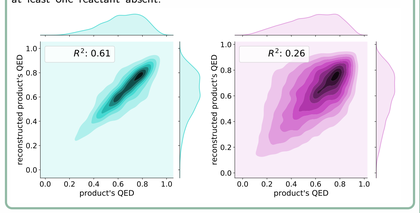
We train a **regression network from the products to the latent space**, based on graph neural networks. This allows us to do **retrosynthesis**, ie predicting what reactants created a product.





Can we find molecules that are easier to make but with similar properties?

Even if the reconstructed product is not correct we are interested in whether it has similar properties, as our model may be useful in suggesting products with similar properties but that are easier to make. We assess correlations between QEDs on two subsets of the test set below: left shows results for subset of reactions for which the reactants are all in Molecule Chef's vocabulary, right corresponds to subset of reactions with at least one reactant absent.



References

[1] Gómez-Bombarelli Rafaeli et al. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules ACS Cent. Sci, 2018.
[2] Kusner, Matt J., Brooks Paige, and José Miguel Hernández-Lobato. Grammar Variational Autoencoder. ICML, 2017.
[3] Jin, Wengong, Regina Barzilay, and Tommi Jaakkola. Junction Tree Variational Autoencoder for Molecular Graph Generation. ICML, 2018.
[4] Schwaller, Philippe, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Costas Bekas, and Alpha A. Lee. Molecular Transformer for Chemical Reaction Prediction and Uncertainty Estimation. ACS Cent. Sci., 2019.
[5] Segler, Marwin HS, Thierry Kogej, Christian Tyrchan, and Mark P. Waller. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. ACS Cen. Sci., 2017.
[6] Pyzer-Knapp, Edward O., Changwon Suh, Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, and Alán Aspuru-Guzik. What Is High-Throughput Virtual Screening? A Perspective from Organic Materials Discovery. Annual Review of Materials Research, 2015.
[7] Tolstikhin, Ilya, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein Auto-Encoders. ICLR, 2018.
[8] Lowe, Daniel M. Extraction of chemical structures and reactions from the literature PhD Thesis (University of Cambridge), 2012
[9] Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, and Alexander L Gaunt. Constrained graph variational autoencoders for molecule design. NeurIPS, 2018.
[10] Marwin HS Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Cent. Sci.*, 2017.
[11] Daniil Polykovskiyi et al. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *arXiv preprint arXiv:1811.12823*, 2018.
[12] Artur Kadurin et al. The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget*, 2017.
[13] Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. Guacamol: Benchmarking models for de novo molecular design. Journal of Chemical Information and Modeling , 2019.

arxiv.org/abs/1906.05221

jab255@cam.ac.uk, bpaige@turing.ac.uk, m.kusner@ucl.ac.uk, marwin.segler@benevolent.ai, jmh233@cam.ac.uk