Deep Generative Model for Pharmaceutical Process Design

*D. Alvarado**, B. Johnston, C. Brown

Future Continuous Manufacturing and Advanced Crystallisation Research Hub (CMAC), University of Strathclyde

diego.alvarado-maldonado@strath.ac.uk

Background

Deep generative models (DGMs) are neural networks capable of generating realistic samples and learning hidden information. Most popular developments in this area include GPT, Dall-E and midjourney applied to generate text and images.

DGMs have been employed in fields such as drug discovery to generate new drug candidates with desirable biological and chemical properties.

Applications in pharmaceutical manufacturing have not been fully explored.

Potential Benefits of DGMs

- Assist process design by producing plausible manufacturing routes for APIs/dosage forms and targeting the design space.
- Improve process understanding through the utilisation of latent variables that may be correlated to process parameters.

Challenges for their development

- Thousands of data are required to develop a model.
- No database with information on primary and secondary processing.

Aim

Develop DGMs for the design of pharmaceutical manufacturing processes taking molecular descriptors as inputs and outputting a plausible chain of unit operations.

① Collect data for model training → ② Dataset analysis and curation → ③ Design and development of DGM

Data ①

- Patents from 1971 to 2021.
- 385K procedures for synthesis and purifcation of more than 359K molecules.
- Database contains information on materials, conditions, and operations obtained using natual language processing (NLP).
- Training set split (80:10:10).

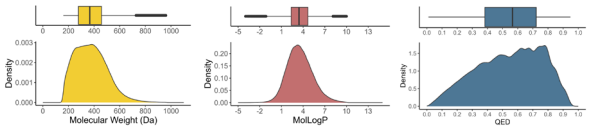


Fig 1. From left to right, distributions of molecular weight, ClogP, and drug-likeness measured as QED for the synthesis products obtained.

Processing ②

- More than 20K terms identified to refer to operations.
- Multiple ways to refer to the same materials.
- Materials were standardised using InChI, while terminology for operations was unified into 60 terms using as a reference the work conducted by [1] and [2].

Patent No. US8859535B2

Intermediate 178.2 (4-Bromo-phenyl)-(2-oxa-bicyclo[2.2.2]oct-4-yl)methanone

To a solution of (4-Bromo-phenyl)-(2-oxa-bicyclo[2.2.2]oct-4-yl)methanol (224 mg, 0.821 mmol) in DCM (13 ml) was *added* a total of manganese dioxide (1.414 g, 16.264 mmol) at RT and the reaction mixture was *stirred* for 20 hrs. The suspension was *filtered* through a pad of hyflo and the residue was *washed* with DCM, *dried in vacuo* to yield the title intermediate (220 mg, 0.745 mmol, 91%) as a colourless solid.

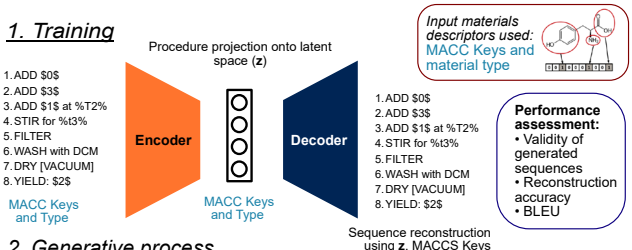
Input *Output*

Index	Type	InChIKey
0	REACTANT	MYEXTBQDVDPSHS-UHFFFAOYSA-N
1	REACTANT	NUJOXMJBOLGQSY-UHFFFAOYSA-N
2	TARGET	RZFLEKKWQKWKHJ-UHFFFAOYSA-N
3	SOLVENT	YMWUJEATGCHHMB-UHFFFAOYSA-N

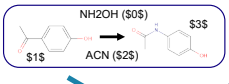
1. ADD 0
2. ADD 3
3. ADD 1 at %T2%
4. STIR for %t3%
5. FILTER
6. WASH with DCM
7. DRY [VACUUM]
8. YIELD: 2

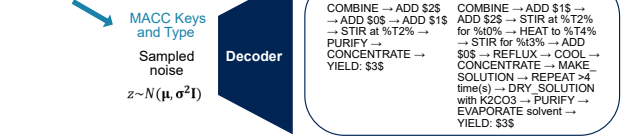
Model: Variational Autoencoder (VAE) ③

1. Training



2. Generative process





Results and Discussion

As a reference, a model developed by [1] to predict action sequences was employed. This model was retrained with our dataset.

Manual revision of sequences indicates the VAE tends to generate consistent sequences, although errors can still be observed which explains a lower validity compared to reference model.

Future work will explore:

- Changes in architecture to improve performance.
- Manual assessment by experts to check plausibility.
- Adaptation of these models for secondary processing.

Table 1. Performance metrics. * refers to prediction accuracy. Model [1] validity criteria includes additional rules

Model	BLEU	Reconst. Accuracy	Valid samples
Transformer [1]*	55%	3.6%	99%
Retrained*	48%	2.4%	95%
CVAE	99%	94%	65%
Invariant-CVAE	97%	86%	77%

Conclusion

A DGM for the generation of likely manufacturing sequences within the primary domain has been developed, showing their potential in manufacturing design. Further research is ongoing to enhance performance and applicability in other domains.

Acknowledgments

EPSRC Centre for Innovative Manufacturing in Continuous Manufacturing and Crystallisation

UKRI Engineering and Physical Sciences Research Council

University of Strathclyde Glasgow

References

1. Vaucher, A. C., Schwaller, P., Geluykens, J., Nair, V. H., Iuliano, A., & Laino, T. (2021). Inferring experimental procedures from text-based representations of chemical reactions. Nature Communications, 12(1), 2573.
2. Wang, Z., Cruse, K., Fei, Y., Chia, A., Zeng, Y., Huo, H., He, T., Deng, B., Kononova, O., & Ceder, G. (2022). ULSA: unified language of synthesis actions for the representation of inorganic synthesis protocols. Digital Discovery, 1(3), 313–324.