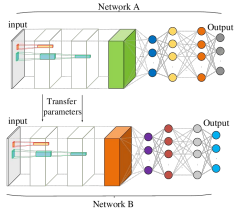
Yasonik, J. (2020). Multiobjective de novo drug design with recurrent neural networks and nondominated sorting. *Journal of Cheminformatics*, *12*(1). https://doi.org/10.1186/s13321-020-00419-6

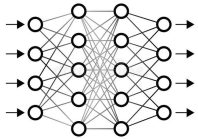
Lane, H., Hapke, H., & Howard, C. (2019). *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python* (1st ed.). Manning Publications.

Multiobjective de Novo Drug Design with Recurrent Neural Networks and Nondominated Sorting

Introduction



6 Proposes new approach for optimizing for chemical properties collectively using **Transfer Learning**



4 Proposes new approach for generating **new chemical compounds** using **Recurrent Neural Networks**

5 Optimizing for **specific chemical properties** is **hard**

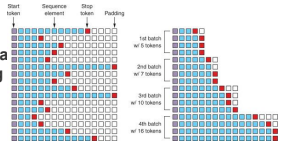
2 **Higher costs**, **longer timelines**, and **lower success rates**

1 Research productivity in pharmaceutical industry **decreased** in recent decades

3 Rise of **de novo drug design**

19 Further exploration of **ML in drug discovery** offers enormous **potential** to reduce the **cost and time** needed for the development of drugs.

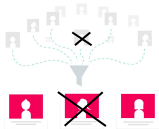
18 **Additional data preprocessing techniques** (encodings, paddings) to increase efficiency and accuracy



17 **Further Improvements**

Large quantities of data are **not required** to train the model

Unrealistic or inferior molecules with one good property can **go through the selection**



16 **Limitations**

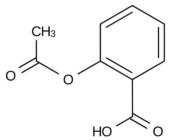
15 **Advantages**

This cycle of de novo drug design provides **scalable generation** of molecules with **multiobjective optimization**.

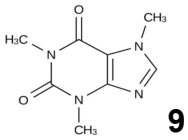
Conclusion

Methods

7 Molecules are represented using the **SMILES string notation**

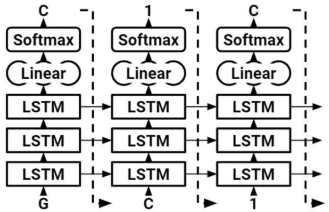


Acetylsalicylic Acid (Aspirin)
CC(=O)Oc1ccccc1C(=O)O



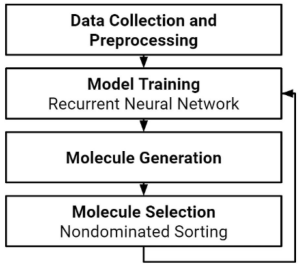
Epinephrine
CNC[C@H](O)c1ccc(O)c(O)c1

8 The network used composed of **three stacked LSTM layers**, each of size 1024, regularized with a 0.2 dropout ratio.
21 million trainable parameters.
Sequence length = 75 time steps



9 Generates molecules and then selects the **best ones based on nondominated sorting** then **train with the newly generated**

10 This process **simulates the traditional design–synthesis–test** cycle far more rapidly.

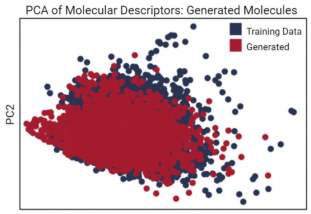


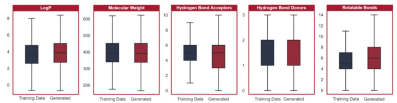
Results

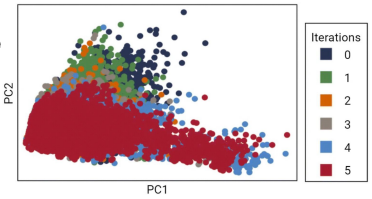
11 One million characters were sampled from the LSTM model, yielding **19,722 molecules, none of which were in the original training data.**

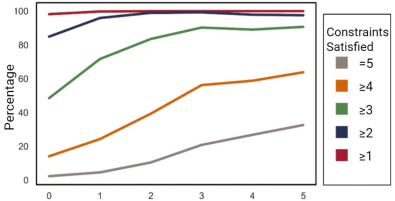
12 **77% were valid** and 6,295 were duplicates. Filtering left **9,415 unique, novel, and valid molecules.**

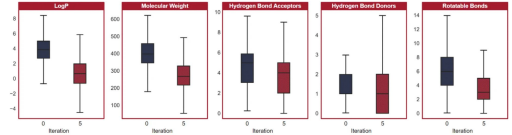
13 Molecules were evaluated based on five chemical properties then their **properties were visualized using PCA**











14 After **retraining the model with selected molecules**, the molecules properties were visualized again.