RNN-driven Approaches to Self-healing Compound Synthesis

Amitesh Badkul [1] Ashif Iquebal [2]

[1]*Department of Chemistry, BITS-Pilani, Hyderabad, India* [2]School of Computing and Augmented Intelligence, Arizona State University, Tempe, USA

Motivation

3D printing technology has revolutionized manufacturing processes, offering enhanced precision and versatility in product design. However, the materials commonly used in this domain often exhibit brittleness, leading to concerns about their durability. The frequent and irreversible damage to these materials necessitates a solution to enhance their longevity and reduce maintenance.

Self-healing materials, characterized by their ability to recover from damage autonomously, present a promising avenue to address this challenge. Hydrogen bonding, a fundamental atomic interaction, plays a pivotal role in facilitating the self-healing properties of materials. Yet, systematically exploring the chemical space to identify compounds with optimal hydrogen bonding for self-healing remains a complex task.

This research aims to employ ***Recurrent Neural Networks***-based (RNNs) algorithms to navigate this vast chemical space, striving to design compounds that harness the potential of hydrogen bonding for enhanced self-healing properties.

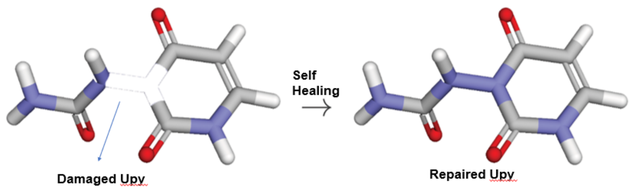

Figure 1. Self-healing

Background

The convergence of computational design and molecular science, empowered by the advancements in machine learning, has primarily focused on the realm of drug discovery. Architectures such as Recurrent Neural Networks (RNNs) and autoencoders have demonstrated notable precision in generating drug molecules. However, their utilization has been predominantly confined to this specific domain.

The table presented herein offers a comprehensive overview of the prevalent methods within this field, emphasizing their architectural selections and primary applications. Notably, the predominant emphasis has been on the generation of small drug molecules, harnessing the sequential nature of SMILES representations and the exploration of latent molecular features.

Method	Architecture	Dataset	Primary Application
REINVENT	RNN, LSTM GRU	GDB-13 ChEMBL	Drug molecule generation
CharRNN	LSTM	ChEMBL	Drug molecule generation
LatentGAN	AutoEncoder & GAN	ChEMBL	Drug molecule generation
ORGAN	GAN and RL	GDB17 (Subset)	Drug molecule generation
GGM	GNN	IBS	Organic compound generation

Table 1. Summary of Methods in Drug Molecule Generation

Nevertheless, the potential of these generative models extends beyond drug synthesis. This work explores the relatively uncharted territory of designing compounds tailored for 3D printing applications, with a particular focus on those possessing self-healing properties.

Exploratory Data Analysis

Dataset Selection and Sampling: The GDB-13 dataset, an extensive repository comprising 970 million small organic molecules, was selected for the work. To conduct a comprehensive exploration, we performed random sampling, extracting 10,000 molecules on ten separate occasions, creating a representative subset for in-depth analysis.

Refinement for Self-Healing Compounds: Given the research's primary focus on self-healing compounds tailored for 3D printing applications, a meticulous refinement process was imperative. Compounds were primarily chosen based on their hydrogen bonding capabilities, a fundamental aspect of self-healing mechanisms.

Analysis and Findings:

1. Hydrogen Bonding Metrics: On average, the sampled compounds displayed approximately 3 Hydrogen Bonding (HB) Acceptors and 1 HB Donor. This distribution suggests a substantial potential for self-healing within the compounds.
2. Aliphatic Carbocycles & Heterocycles: The dataset exhibited a higher prevalence of aliphatic heterocycles compared to aliphatic carbocycles. This distinction implies an increased likelihood of hydrogen bonding sites in the compounds.
3. Heteroatoms: The dataset compounds had an average count of approximately 3 heteroatoms, further emphasizing their versatility and applicability in the domain of self-healing materials.

The EDA has revealed that while the GDB-13 dataset is extensive, a refined subset of compounds with strong hydrogen bonding attributes is pivotal for achieving the research objectives. This subset, characterized by its inherent potential for hydrogen bonding and subsequent self-healing compound generation.

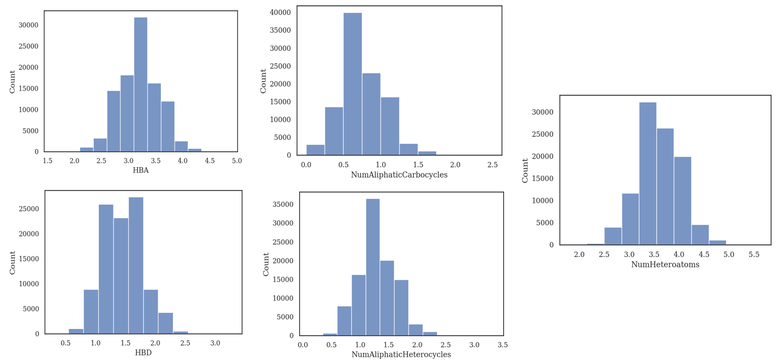

Figure 2. Exploration of GDB-13

Methods

Choice of Model - RNN-based REINVENT: In this study, we employed the RNN-based REINVENT model due to its proficiency in handling sequential data. The representation of chemical compounds using the SMILES notation inherently involves sequential information.

Addressing RNN Limitations with GRUs: While RNNs excel in processing sequential data, they encounter challenges related to short-term memory and the vanishing gradient problem during backpropagation. To overcome these limitations, we incorporated Gated Recurrent Units (GRUs).

Model Architecture: The process begins with the tokenization and one-hot encoding of SMILES. This preprocessed data undergoes a transformation in an embedding layer, converting discrete SMILES into a continuous 256-dimensional representation. The core of the model comprises three layers with 512 GRU units each. The final step involves a fully-connected linear layer, which performs a softmax operation.

Results

Training Dataset and Epochs: The model utilized a subset of 1 million compounds from the GDB-13 dataset, emphasizing those with at least 2 Hydrogen Bond acceptors and donors, resulting in a dataset of approximately 500,000 compounds. Training persisted for 50 epochs. Around 100,000 compounds were generated for evaluation, with different temperature settings to gauge the conservative nature of the model. We systematically evaluated the model on various metrics: Validity, Synthetic Accessibility Score (SA Score), Baeyer Strain, and Flexibility (Rotatable Bond Count (RBC)) to provide a comprehensive assessment of the generated compounds ensuring they align with the research's objectives of self-healing and hydrogen bonding capabilities.

1. Flexibility (RBC): Zero RB dominate among the Generated Compounds, indicating structurally rigid molecule generation. This fits into the criterion for creating durable 3D printing material.
2. Baeyer Strain: A significant peak at zero Baeyer strain for the Generated Compounds highlights the model's proficiency in generating cyclic compounds with minimal strain.
3. SA Score: The substantial clustering around an SA Score of 5 for the Generated Compounds demonstrates the model's ability to create molecules that strike a balance between novelty and practical synthetic accessibility.

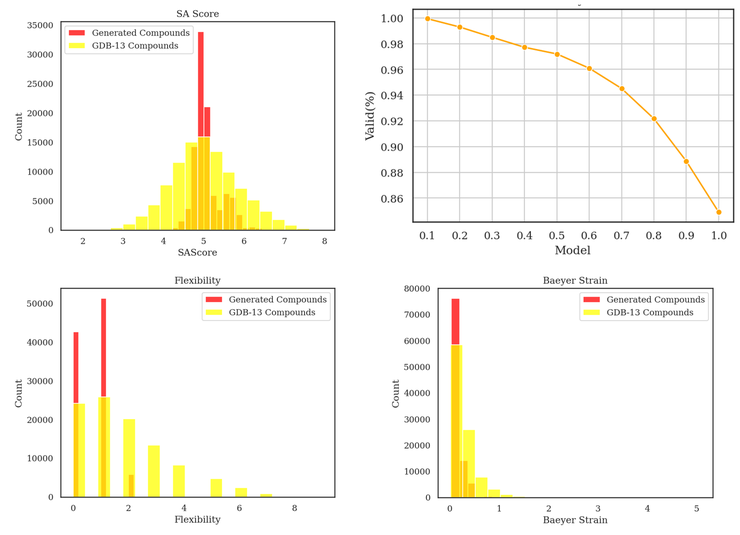

Figure 3. Comparison with the trained dataset

Acknowledgement

I would like to thank ASU SURI for giving me this wonderful opportunity and Dr. Ashif Iquebal for his constant guidance, support and the insightful conversations that shaped this project.

References

[1] Josep Arús-Pous, Thomas Blaschke, Silas Ulander, Jean-Louis Reymond, Hongming Chen, and Ola Engkvist. Exploring the gdb-13 chemical space using deep generative models. *Journal of cheminformatics*, 11(1):1–14, 2019.

[2] Lorenz C Blum and Jean-Louis Reymond. 970 million druglike small molecules for virtual screening in the chemical universe database gdb-13. *Journal of the American Chemical Society*, 131(25):8732–8733, 2009.

[3] Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1:1–11, 2009.

[4] Benjamin Sanchez-Lengeling and Alán Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018.

[5] Xiaochu Tong, Xiaohong Liu, Xiaoqin Tan, Xutong Li, Jiaxin Jiang, Zhaoping Xiong, Tingyang Xu, Hualiang Jiang, Nan Qiao, and Mingyue Zheng. Generative models for de novo drug design. *Journal of Medicinal Chemistry*, 64(19):14011–14027, 2021.