Title
68 million natural product-like compound database generated via molecular language processing

Authors
Dillon W. P. Tay¹, Naythan Z. X. Yeo¹, Krishnan Adaikkappan¹, Yee Hwee Lim¹, Shi Jun Ang²

Affiliations
1. Institute of Sustainability for Chemicals, Energy and Environment, A*STAR, Singapore.

corresponding author(s): Dillon W. P. Tay (dillon_tay@isce2.a-star.edu.sg) and Shi Jun Ang (ang_shi_jun@ihpc.a-star.edu.sg)

Abstract
Natural products are a rich resource of bioactive compounds for valuable applications across multiple fields such as food, agriculture, medicine. For natural product discovery, high throughput in silico screening offers a cost-effective alternative to traditional resource-heavy assay-guided exploration of structurally novel chemical space. In this data descriptor, we report a characterized database of 68,113,839 natural product-like molecules generated using a recurrent neural network trained on known natural products, demonstrating a significant 167-fold expansion in library size over the currently estimated 406,919 natural products known. This study highlights the potential of using deep generative models to uncover novel natural product chemical space for high throughput in silico screening toward natural product discovery.

Background & Summary
Nature possesses an immense diversity of chemical compounds collectively known as natural products.¹² Natural products reflect a vast reservoir of molecular scaffolds naturally selected by organisms to interact with their environments and to engage in chemical warfare with each other. This natural diversity has in turn been leveraged for wide-ranging applications to benefit human society. Some examples include agricultural pesticides to increase food production,³ food preservatives to facilitate distribution and storage,⁴,⁵ and most prominently therapeutic agents to treat diseases.⁶⁻⁸ Indeed, it has been estimated that approximately 80% of all clinically used antibiotics can have its origins traced to a natural product.⁶

Despite nature’s potential for providing valuable molecules however, assay-guided discovery has been a low yielding investment since the golden age of natural product discovery in the early 1960s.⁹ After the initial wave of uncovering structurally unique and accessible natural product chemical space, subsequent efforts to venture into less accessible chemical space or to “rediscover” known natural product classes for novel applications has been met with limited success.¹⁰ Tremendous effort must be invested for the biosynthesis, curation and characterization of natural product libraries, resulting in only approximately 400,000 fully characterized natural products known to-date.¹¹ The significant financial and resource investment requirements of assay-guided investigations have also resulted in a broad dampening of commercial interest around natural product discovery.¹² With the advent of deep generative modelling¹³ and high throughput in silico screening¹⁴ however, there is an opportunity to circumvent traditional time-consuming, costly, and experimentally-driven natural product discovery to reformulate it as a computational inverse design problem.¹⁵ The potential of such an approach would scale with the increasing size of available natural product libraries.¹⁶ In this data descriptor, we report an expansive, curated database of >68 million natural product-like molecules generated via an in silico pipeline (Figure 1), representing a 167-fold expansion from the roughly 400K known natural products.¹¹
Figure 1 – Workflow to generate and characterize a natural product-like compound database using a recurrent neural network trained on known natural products.

Compared to manually curated natural product libraries, deep generative models can transcend the boundaries of human imagination-dependent molecular design to significantly expand chemical search space by orders of magnitude while concurrently reducing financial and resource requirements. Some examples of deep generative architectures that have been employed for molecular design include variational autoencoders (VAE), recurrent neural networks (RNN), and generative adversarial networks (GAN), each adopting a different strategy with their own strengths and weaknesses. In this work, we have trained a RNN-based long short-term memory (LSTM) network architecture on tokenized SMILES (Simplified Molecular Input Line Entry System) from 325,535 (80%) out of the 406,919 known natural products in COCONUT, the collection of open natural products (https://coconut.naturalproducts.net/, accessed 1 Aug 2022). The RNN model was able to establish a molecular language vocabulary from known natural product examples, and subsequently used it to recombine and assemble novel molecules following the same “grammar” as known natural products. In this way, 100 million natural product-like molecules were generated. Thereafter, steps were taken to analyse, validate, and characterize the generated library to ensure its quality and usability.

Figure 2 – Comparison overview of generated and COCONUT natural product databases. (A) RDKit sorting of 100 million natural product-like Simplified Molecular Input Line Entry System

(A) Generated SMILES

22% Duplicated
10% Invalid
68% Valid & Unique

(B) Density

(C) Density

18% Shikimates and Phenylpropanoids
15% Fatty acids
5% Polyketides
4% Alkaloids
6% Terpenoids
8% Amino acids and Peptides
1% Carbohydrates
(SMILES) strings generated with recurrent neural network (RNN). (B) Natural product-likeness score (NP Score)\textsuperscript{31} distributions and (C) NPClassifier\textsuperscript{32} pathway classifications of valid unique generated natural product-like molecules versus known natural product molecules from COCONUT (collection of open natural products) database.\textsuperscript{11}

Python-based cheminformatics packages such as RDKit\textsuperscript{30}, NP Score,\textsuperscript{31} and NPClassifier\textsuperscript{32} were employed to analyze and characterize the generated 100 million natural product-like compounds (Figure 2).

Firstly, open-source cheminformatics software RDKit\textsuperscript{30} was used to canonicalize and filter out 9,595,376 invalid and 22,287,375 duplicated SMILES from the 100 million generated natural product-like SMILES to leave 68,117,249 unique and valid canonical SMILES (Figure 2A). 3,410 of these were found to be ionized molecules with no specified counterion and hence were also filtered out to allow uniform comparison between the remaining 68,113,839 neutral generated SMILES and 406,919 neutral known natural product SMILES from the COCONUT database.

Secondly, natural product-likeness scores (NP Score)\textsuperscript{31} of the known natural product SMILES and generated SMILES were calculated (Figure 2B). NP Score employs atom-centred fragments (HOSE codes)\textsuperscript{33} and bonding information to characterize structural features and calculate a Bayesian measure of how similar a molecule is to known natural product structural space.\textsuperscript{31} The NP Score distribution of the generated natural product-like SMILES bears good resemblance to that of known natural products in the COCONUT database (Figure 2B), suggesting the generation of natural product-like structures.

Thirdly, NPClassifier\textsuperscript{32} was used to classify both generated natural product-like SMILES as well as known natural products from the COCONUT database. NPClassifier\textsuperscript{32} is a deep learning tool that considers structural features (counted Morgan fingerprints),\textsuperscript{34,35} taxonomy of the producing organism, nature of the biosynthetic pathway, and biological activity to characterize molecules in a holistic natural product classification framework. However, 8,006,367 (11.8%) of the generated and 35,708 (8.8%) of the known SMILES received no pathway classification via NPClassifier. It has been reported\textsuperscript{32} that deficiencies in NPClassifier can be traced back to limitations in its training data as the model relies on existing knowledge of natural products to classify molecules based on structural similarities. The comparatively higher percentage of generated SMILES with no NPClassifier pathway outputs suggests the presence of either synthetic structural features unrelated to, or novel natural product class(es) distinct from those in the NPClassifier training dataset. However, similarities between the natural product-likeness score distributions of the generated and known datasets suggests promising potential toward the latter. The remaining 60,110,882 generated SMILES were annotated with a comparable distribution of biosynthetic pathways as known natural products from the COCONUT database (Figure 2C).

Finally, to describe the physiochemical space covered by the known natural products and the generated database, the following 10 physiochemical RDkit molecular descriptors were calculated for each molecule:

1. Number of aromatic rings
2. Number of aliphatic rings
3. Wildman-Crippen LogP (partition coefficient)\textsuperscript{36}
4. Molecular weight
5. Number of hydrogen bond acceptors
6. Number of hydrogen bond donors
7. Number of heteroatoms
8. Topological polar surface area (TPSA)
9. Number of rotatable bonds
10. Number of valence electrons

T-distributed stochastic neighbour embedding (t-SNE) dimensionality reduction of the 10 calculated descriptors into two dimensional space was performed to visualize physiochemical space coverage by the generated natural product-like SMILES and known natural product SMILES from the COCONUT database (Figure 3A).

Figure 3 – Visualization of expanded chemical search space from generated database. (A) T-distributed stochastic neighbour embedding (t-SNE) 2D projection of 10 RDkit physiochemical descriptors for 68,113,839 natural product-like structures generated from our trained model and 406,919 known natural product structures from COCONUT, the collection of open natural products. (B) Density plot of known natural product structures in t-SNE 2D projected space. (C) Density plot of generated natural product-like structures in t-SNE 2D projected space.

The t-SNE 2D comparison shows a significant increase in physiochemical structural space covered by the generated SMILES (Figure 3A), suggesting the presence of structurally novel natural product-like molecules in the generated database. Density plots (Figures 3B, 3C) showing the concentration of structures across the t-SNE 2D projected space also highlight the significantly expanded structural space offered by the generated database even in overlapping physiochemical space (Figure 3C). Overall, this workflow has enabled us to generate a significantly expanded database of 68,113,839 characterized natural product-like molecules, greatly increasing accessible chemical space by >167-fold over the currently estimated 400,000 natural products known. As an indication of its cost efficiency, the total computation time for training and sampling was less than 24 hours on a NVIDIA V100 GPU card with 192 GB of memory. In comparison, a commercially available 2,576 natural product library can be purchased for USD$33,513 (https://www.selleckchem.com/screening/natural-product-library.html, accessed 11 Jan 2023). This broad chemical search space may serve as starting points for high throughput in silico discovery of functional natural products.

Aside from discovery of bioactive natural products as therapeutics, natural product alternatives to synthetic ingredients have also been gaining attention amidst increasing consumer demand for their perceived health and wellness benefits. Natural products can also be produced sustainably via biosynthetic and synthetic biology approaches, adding to their attractiveness as environmental regulation steadily expands in response to address global issues of climate change, pollution, and resource depletion. Computationally generated structural databases of natural product-like molecules like the one reported here
are well positioned to push the boundaries of known natural product chemical structures, provide expanded search spaces, and act as key enabling resources to progress next generation in silico high throughput screening methods for novel functional natural product discovery.

**Methods**

**Molecule generation with recurrent neural network.** The generative model was trained with a recurrent neural network (RNN) architecture using long-short-term-memory (LSTM) units (https://github.com/skinnider/low-data-generative-models). To assemble the training and held out datasets, the COCONUT collection of open natural products (https://coconut.naturalproducts.net/)\(^{11}\) was filtered to remove invalid SMILES and take away stereochemistry to yield 325,535 unique canonical SMILES for training and 81,384 unique canonical SMILES as a held-out dataset for validation. The training dataset was augmented by 10 times with their respective non-canonical SMILES prior to the RNN training using SmilesEnumerator (http://github.com/EBjerrum/SMILES-enumeration). This has been shown to improve the validity of the sampled SMILES after the model is trained.\(^{21}\) Determination of the vocabulary of the RNN was carried out by deconstructing SMILES strings into elemental tokens. The network consists of 3 layers of RNN with hidden layer dimension of 512, no dropout was used. Training of the network was done with a batch size of 128, learning rate of 0.001 and the Adam optimizer. Early stopping patience of 10,000 minibatches was employed using 10% of the training set held-out for validation. A total of 100,000,000 SMILES strings were sampled from the trained model after completion of model training.

**Data processing and characterization with cheminformatics tools.** Generated SMILES strings were converted to its canonical form using RDKit\(^{30}\) (Chem.MolFromSmiles then Chem.MolToSmiles) to yield 68,117,249 valid, unique, canonical SMILES. SMILES strings were considered invalid if “none” was returned from this process. 3,410 Further SMILES describing ionized molecules with no specified counterion were then filtered out to leave 68,113,839 valid, unique, canonical, neutral SMILES. 10 RDkit molecular descriptors (Number of aromatic rings, number of aliphatic rings, Wildman-Crippen LogP,\(^{36}\) molecular weight, number of hydrogen bond acceptors, number of hydrogen bond donors, number of heteroatoms, topological polar surface area, number of rotatable bonds, and number of valence electrons) were calculated for each valid, unique canonical SMILES string. Natural product likeness scores (NP Score)\(^{31}\) were calculated using npscorer (https://github.com/rdkit/rdkit/tree/master/Contrib/NP_Score) and natural product pathway, superclass, and class classifications were assigned using NPClassifier API (https://npclassifier.ucsd.edu/).\(^{32}\) Queries without outputs from NPClassifier were assigned the value “none” in the Pandas dataframe.

**T-distributed stochastic neighbor embedding (t-SNE) 2D projection of chemical space.** T-distributed stochastic neighbor embedding (t-SNE) dimensionality reduction was performed on the 10 RDkit descriptors (Number of aromatic rings, number of aliphatic rings, Wildman-Crippen LogP, molecular weight, number of hydrogen bond acceptors, number of hydrogen bond donors, number of heteroatoms, topological polar surface area, number of rotatable bonds, and number of valence electrons) using scikit-learn\(^{43}\) (sklearn.manifold.TSNE) with the following parameters: n_components=2, init=“pca”.

**Data Records**

All relevant datasets and experimental parameters have been uploaded and are available on figshare.

<table>
<thead>
<tr>
<th>Filename</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
coconut_smiles_nostereo_training80.txt | Training dataset of 325,535 unique canonical SMILES without stereochemistry from COCONUT database, January 2022 version (Accessed on 1 August 2022)
---|---
coconut_smiles_nostereo_heldout20.txt | Held-out validation dataset of 81,384 unique canonical SMILES without stereochemistry from COCONUT database, January 2022 version (Accessed on 1 August 2022)
environment.yml | List of dependencies to run RNN model
rnn_settings.txt | Settings for RNN model
100million_sampled_smiles.smi | 100 million generated SMILES sampled from trained RNN model
68M_canonical_unique.txt | 68,117,249 unique, valid, canonical natural product-like SMILES sampled from RNN model
coconut_original_analysed.json | Pandas dataframe of 406,919 unique canonical SMILES without stereochemistry from COCONUT database characterized with 10 RDkit descriptors, NPScore, NPClassifier classifications (pathway, superclass, class), and t-SNE 2D descriptors
coco68M_analysed.json | Pandas dataframe of 68,113,839 unique canonical generated SMILES characterized with 10 RDkit descriptors, NPScore, NPClassifier classifications (pathway, superclass, class), and t-SNE 2D descriptors.

Table 1. List of files encompassing the datasets and parameters in this work that have been uploaded to figshare.

**Technical Validation**

Validation of generated natural product-like molecules. From the 406,919 known, valid, unique, canonical, natural product SMILES strings in the COCONUT database with stereochemistry removed, 81,384 (20%) were held-out and the remaining 325,535 (80%) were used to train the recurrent neural network to generate natural product-like molecules. Of the 81,384 known natural products that were held out from the training dataset, 30,229 (37% of held-out set) known natural products were reproduced in the 68,117,249 generated natural product-like molecule database, confirming the trained model can generate actual natural product molecules.

**Code Availability**

The RNN training was performed with [https://github.com/skinnider/NPS-generation](https://github.com/skinnider/NPS-generation). The full list of RNN settings, and dependencies used is available on figshare.

**Acknowledgements**

The authors gratefully acknowledge funding support from the Agency for Science, Technology and Research (A*STAR), Singapore (#21719). This work was supported by the A*STAR Computational Resource Centre through the use of its high performance computing facilities.

**Author contributions**

D.W.P.T. co-designed the study, performed data processing and data analysis, and wrote the manuscript with inputs from all authors. N.Z.X.Y. and K. A. performed data processing and analysis. Y.H.L. conceptualized and co-designed the study, and acquired financial support. S.J.A. conceptualized and co-designed the study, managed the overall project, acquired financial support, trained the recurrent neural network, and generated the natural product-like molecules.
Competing interests
The authors declare no competing interests.

References


Landrum, G. in *Abstracts of Papers of the American Chemical Society*. (American Chemical Society 1155 16th Street, NW, Washington, DC 20036 USA).


