JOURNAL OF
EMERGING INVESTIGATORS

Article

Quantitative definition of chemical synthetic pathway complexity of organic compounds

Tanish Baranwal[1,8], Howard Huang*[2,8], Udbhav Avadhani*[3,9], Anya Goyal[4,8], Akhil Samavedam**[5,8], Timothy Hu**[6,8], Mihir Kale[7,8], Tvisha Nepani[7,9], Vishak Srikanth[6,8], Robert Downing[8], Edward Njoo[9]**

[1]Dublin High School, Dublin, CA
[2]Saratoga High School, Saratoga, CA
[3]Leigh High School, San Jose, CA
[4]Monta Vista High School, Cupertino, CA
[5]Westlake High School, Austin, TX
[6]BASIS Independent Silicon Valley, San Jose, CA
[7]Milpitas High School, Milpitas, CA
[8]Department of Computer Science & Engineering, Aspiring Scholars Directed Research Program, Fremont, CA
[9]Department of Chemistry, Biochemistry, & Physics, Aspiring Scholars Directed Research Program, Fremont, CA
* Asterisk denotes equal contribution
** Double Asterisk denotes equal contribution

SUMMARY

Irrespective of the final application of a molecule, synthetic accessibility is the rate-determining step in discovering and developing novel entities. However, synthetic complexity is challenging to quantify as a single metric, since it is a composite of several measurable metrics, some of which include cost, safety, and availability. Moreover, defining a single synthetic accessibility metric for both natural products and non-natural products poses yet another challenge given the structural distinctions between these two classes of compounds. Here, we propose a model for synthetic accessibility of all chemical compounds, inspired by the Central Limit Theorem, and devise a novel synthetic accessibility metric assessing the overall feasibility of making chemical compounds that has been fitted to a Gaussian distribution. Our approach utilizes a Gaussian mixture model (GMM) and Autoencoder, which rank synthetic complexity for natural products. This model can inform total synthesis of natural products, process chemistry in pharmaceutical contexts, materials science, and chemical engineering. Based on our findings, we conclude that the Autoencoder model is better suited to model the true probability distribution of synthetic complexity for natural products.

INTRODUCTION

While many motivations exist for the usage of machine learning in the process of molecular synthesis, the most prominent reason is often to conserve and optimize available resources. The development of small molecule drugs from discovery, preclinical research, clinical trials, and finally FDA approval can take up to 15 years and between $600 million to $1.4 billion (1). However, more recent approaches using machine learning have accelerated the potential for drug discovery. These approaches use virtual screening and parameterization, provide robust methods of augmentation and computational simulation for preclinical research, and improve the quality of patient selection and clinical trial optimization (2, 3). In our study, we aimed to develop a novel metric for assessing synthetic complexity of organic medicines, including small molecules and natural products, through a different machine learning methodology. Several research groups have previously defined metrics for synthetic accessibility using a variety of data about a molecule, such as a graph of the molecule or starting material complexity (4,5,6,7,8). However, the application of machine learning to this domain remains relatively nascent with cheminformatic algorithms dominating previous approaches.

Studies of molecular complexity for synthetic analysis and theories on convergence involving multicomponent reactions date as far back as 1982 (4). Previous computational scores used to define difficulty of synthesis have utilized algorithms to analyze aspects of graph theory, a molecular connectivity index, molecular graph symmetry, or frequency of structural features as factors of synthetic accessibility (5, 6). However, the constantly evolving field of chemical synthesis outpaces the rigid metrics previously established and thus motivates development of a more flexible metric.

To compensate for metric rigidity, algorithms have been tuned for optimization, although differences in experiences between researchers can result in differences between algorithms (7). Due to the constructive nature of molecules, a more complex molecule can be split up into fragments when calculating synthetic accessibility as well. This allows for the detection of significant moieties either by a) using complexity scores for each fragment that we combined to yield a comprehensive metric or b) analysis of the frequencies of the fragments in databases (8,9). A popular metric, SAScore by Ertl and Schuffenhauer, utilizes a fragment-based approach that combines both methodologies (8). The claim is that fragments that are more frequently present in databases are more easily accessible, and therefore rarer fragments do not appear as much due to difficulty in synthesis. A shortcoming of this method is that the fragments analyzed are typically those from small molecules, thereby rendering it difficult to apply

the same complexity score or fragment search algorithm for larger natural products that may not have the same available fragments. As such, our work seeks to address this issue by developing a metric capable of accounting for variations in the total number of atoms – ranging from small molecules to larger molecules with multiple conformers – therefore yielding results that are better optimized and targeted towards both synthetic and natural product molecules.

In this work, we propose a novel synthetic accessibility metric that encompasses reported total cumulative reaction time, number of steps, extremities of temperature, and the yield of each step for a synthetic pathway. After extracting this data, we then used these parameters as the input features for machine learning algorithms that were trained to deliver a synthetic complexity score for a molecular compound. We trained two separate models, a GMM and an Autoencoder, and evaluated the performance of both models. Based on our findings, we conclude that the Autoencoder model is better suited to model the true probability distribution of synthetic complexity for natural products.

RESULTS

In order to generate our SAScores, we first identified specific factors regarding both the required environmental conditions of the total synthesis and qualities of the synthesis itself. We inputted this information into both the Autoencoder and GMM models. We then analyzed the outputs for predictive accuracy.

Factors

The time and number of steps in the synthetic route of a complex molecule differ greatly from a relatively simple molecule. In most instances, a reaction with a greater number of steps tends to take more time than one with fewer and can indicate a more complex molecule; for example, synthesizing a more complex molecule can involve adding more functional groups and moieties to obtain the product, and thus more steps. These two factors, while significant, cannot encompass the entire intrinsic complexity of the compounds. Therefore, we examined two additional attributes of synthesis: temperature, and total product yield. The model did not penalize temperatures in the range of 0-100°C, as they are well within the range of water baths with conventional technology. We used a penalty that increases as a function of the absolute distance from the 0-100° C range, as temperature ranges in the extremes beyond 100°C and below 0°C are difficult to achieve with conventional technology available in the lab. Thus, we imposed a penalty that varies proportionally to the difference beyond each of the upper and lower bounds. We also extracted the yield from the synthetic routes of the entire library of molecules. As low reaction yields imply difficulty in scale-ups and loss of product in purification, necessitating more product to be synthesized in each prior step to continue the synthesis, low reaction yields also lead to longer reaction time and a cumulative number of steps that are not necessarily reported on the synthesis itself. Hence, high yields have lower penalty while low yields point towards more complex compounds.

Justification of Gaussian Distribution Rationale

Given that many situations in the real world can be modeled by a normal distribution, it becomes important for our data and therefore synthetic complexity to mimic the characteristics of naturally occurring distributions. This idea is an extension of the Central Limit Theorem: given that nature is mostly normally distributed, a large enough dataset should also exhibit a normally distributed synthetic complexity score (10). In this case, by using t-SNE for the synthetic complexity factors, we visualized the resulting nonlinear manifold in a single curve—analogous to the majority of the variation lying in one dimension—when utilizing an Autoencoder. Additionally, the homoscedasticity of the data revealed that the t-SNE curve was an appropriate fit, with data exhibiting very low variance from the prediction model curve. As such, we can be confident that the one-dimensional latent embedding of the data fully represents all the information contained in the data, and that it is normally distributed (**Figure 1**).

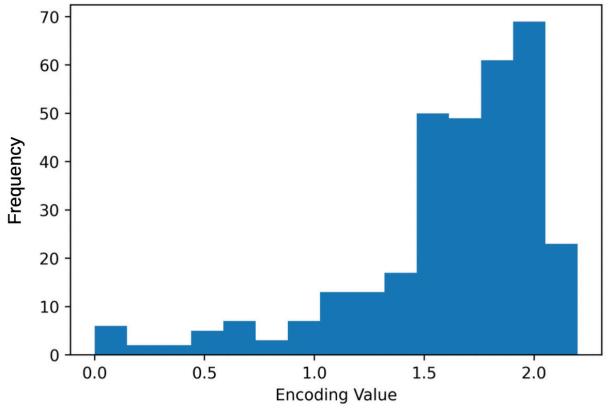

Figure 1: Diagram of Generated Encodings. Histogram of generated encodings, displaying an approximately Gaussian distribution.

Comparison of GMMs and Autoencoders

We used the t-SNE plot to determine whether an Autoencoder would be a suitable model for the multidimensional data. When plotted, the t-SNE visualization revealed a smooth and clean curve, confirming that an Autoencoder could be trained to accurately represent the synthesis of each molecule in our dataset (**Figure 2**).

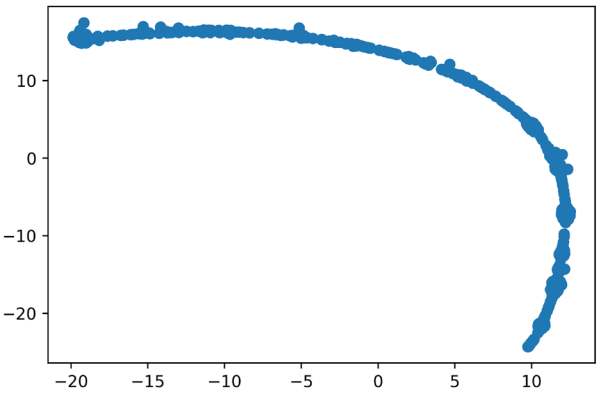

Figure 2: Graph of t-SNE Dimensionality Reduction. When plotted, the t-SNE dimensionality reduction of the five synthetic complexity factors discussed above produces a clean and nonlinear curve. Since the curve is smooth, we were able to use an Autoencoder to reduce the dimensionality of the data.

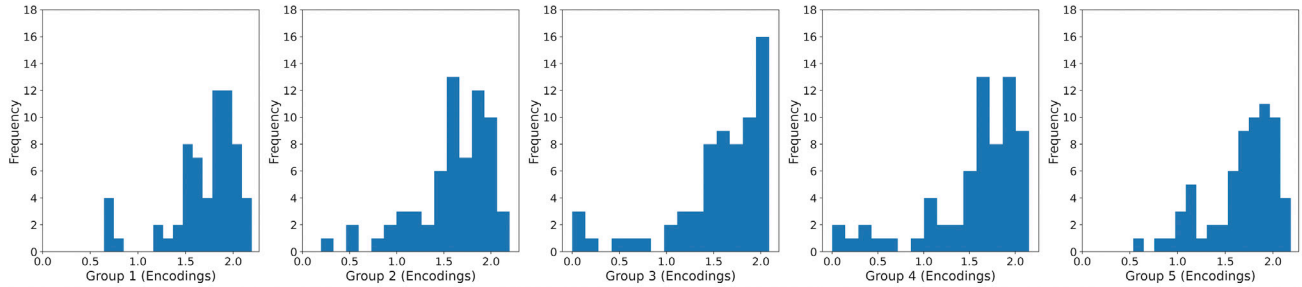

Figure 3: Five Groups of Encodings. The distributions for each of the five subsets of encodings generally follow a Gaussian curve.

The smooth curve observed in the t-SNE shows that the data is highly correlated across some nonlinear manifolds, motivating the use of the Autoencoder. We used an Autoencoder to compress the synthesis factors into a single dimension and verified that this latent representation of the factors was normally distributed.

Our GMM model further built on the Autoencoder algorithm by incorporating reconstruction loss and the encodings into a neural network based GMM.

The GMM showed a skewness to the right whereas the Autoencoder model had a general symmetric normal distribution (**Figures 3-4**). To quantitatively compare the normality of both datasets, we used the Kolmogorov-Smirnov test for normality and analyzed the Fisher Kurtosis of the outputs of the Gaussian mixture model and the Autoencoder model (11). We also calculated the Jensen Shannon (JS) distance between the output of each model and a Gaussian distribution with the same mean and standard deviation (12). Because the JS distance for probability distributions is analogous to the Euclidean distance for points in space, analyzing the JS distance allows us to calculate how similar the outputs of both models are to a normal distribution. A lower JS Distance corresponds to a model closer to the normal distribution, and similarly, a lower Fisher Kurtosis corresponds to a model closer to the normal distribution. These two metrics are calculated based on the probability value generated by the GMM and the encodings from the encoder portion of the Autoencoder model.

The Kolmogorov-Smirnov test for normality indicated that the GMM is non-normal but failed to reject the null hypothesis for the Autoencoder model with a significance level of 0.001. This was empirically verified as the GMM was visibly asymmetrical and thus non-normal whereas the Autoencoder appears symmetrical.

To further analyze the distributions, we analyzed the JS Distance and Kurtosis to generate numerical comparisons between the two approaches. We found that the GMM has a higher kurtosis (33.2293 for the GMM and 2.5491 for the Autoencoder) and is farther from the normal distribution than the Autoencoder, with a JS Distance of 0.7663 versus 0.4866 respectively. This numerically verifies the fact that the Autoencoder outputted SA scores that were more normally distributed than the GMM.

Regression Model

To create a model that could predict synthetic complexity in the absence of previous data about a compound's synthetic pathway, we fitted the generated encodings from our Autoencoder, along with the chemical descriptors and SMILES of each molecule, to a regression model. This enabled us to create a model which can predict synthetic complexity given only basic chemical information about a compound. For this, we attempted two approaches.

Based on the SMILES encodings for each chemical compound in SynArchive, we used the molecular descriptor library Mordred to generate molecular descriptors for each molecule. We subsequently trained the regression model to fit the molecular descriptors from Mordred to the synthetic accessibility we computed from the Autoencoder. This ensured that the regression model can predict synthetic accessibility scores for future synthetic and natural product structures.

Alternatively, we used a one-hot encoding algorithm to generate SMILES embeddings for each molecule. This approach greatly reduced the validation loss and produced a wider range of synthetic accessibility score outputs. We employed various dropout layers to prevent overfitting. However, we noticed persistent problems with still relatively high validation losses and difficulty of the model generalizing to the SMILES embeddings.

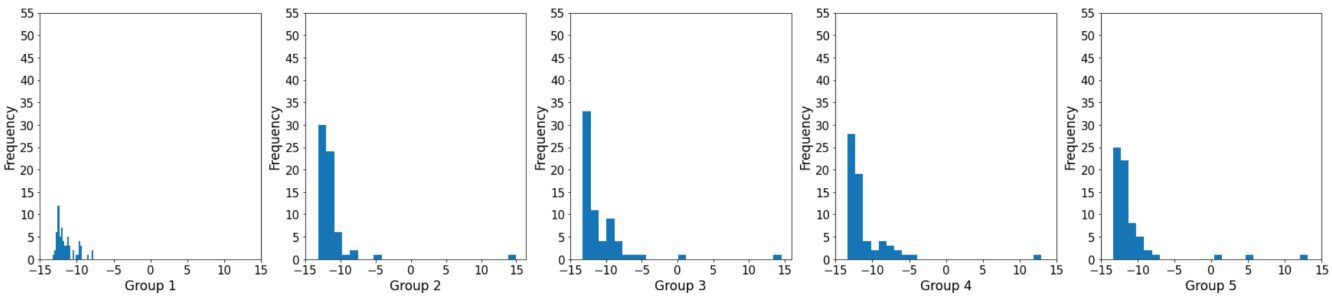

Figure 4: Evaluation of the GMM. The skewness of the GMM data distribution is reflected in the distributions of five randomly selected subsets of the same generated values.

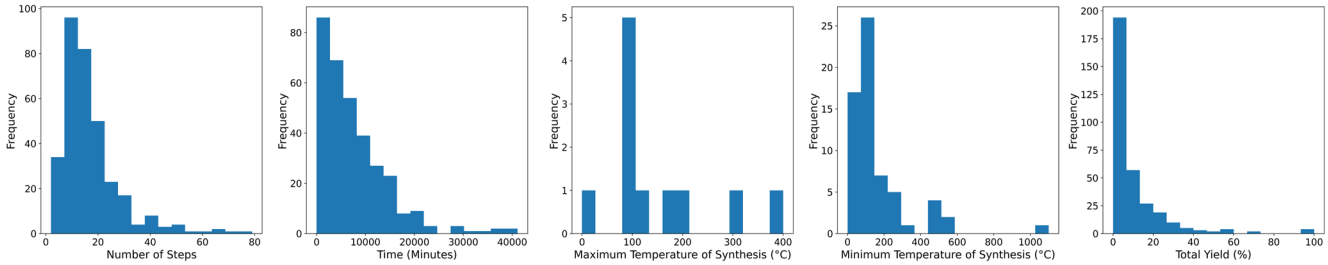

Figure 5: Distributions of Synthetic Complexity Factors. The distributions for each synthetic complexity factor are generally unimodal and skewed to the right. This means that a trivial linear combination of the data will not yield a gaussian distribution, and the task requires a more complex model. Since a measurable metric should be normally distributed, the GMM is less suitable for our metric.

Assessing Predictive Accuracy

The distribution of data for each factor individually was generally similar; most of the distributions were generally unimodal and skewed right, excluding maximum temperature, which contained some outliers (**Figure 5**). After putting this data through an Autoencoder, we observed that the distribution of our generated encodings was roughly symmetric and formed a Gaussian curve (**Figure 1**). This distribution had a mean of approximately 2.059 and a standard deviation of 0.311.

To further prove that the synthetic pathway data was indeed Gaussian, we divided the dataset from SynArchive into five randomly generated groups. We put each group into the Autoencoder and then generated the histograms for each dataset shown below. The histograms for each group conformed to the definition of the Normal Distribution with any combination of datasets (**Figure 3**).

For the GMM, we found that the distribution generated is slightly skewed to the right (**Figure 5**). We ran our evaluation strategy on this model as well, and the skewness of the GMM's outputs were apparent when split into random subgroups (**Figure 4**). This qualitatively supports the numerical analysis of the two models.

DISCUSSION

The wide variations in the amount of structural information that can be obtained from various parameters in cheminformatics make the development of a synthetic complexity score highly controversial. For example, certain molecular structures can be more complex based on different parameters such as cyclic structures or heteroatoms, or even fragmentation of molecular structure and need to evaluate combinatorial groups of other descriptors. With scale-up for mass production remaining a primary concern for a pharmaceutical industry that handles a wide variety of molecules, our synthetic complexity score should be precise and widely applicable given that organic compounds in any given pharmaceutical pipeline can range from small molecule synthetic drugs to natural products and close derivatives. Therefore, we put forth a novel unified synthetic accessibility score that considers both structural- and reaction-related factors to handle the versatility demanded by the pharmaceutical industry.

Because we do not have any labels for this task, we look to a heuristic interpretation of the scores outputted by both models. We tasked synthetic chemists to interpret the outputted scores and determine what a bigger or smaller score means in context of the situation for both the Autoencoder and the GMM. The SAScores from our Autoencoder, when scaled from 0-10 (1 being "easy" to make and 10 a "high difficulty synthesis"), are roughly accurate in predicting a molecule's synthetic accessibility score. While there is some variability at the lower extremes, a higher score from the Autoencoder corresponds to an easier molecule to synthesize, whereas a lower value represents harder molecules. For example, Penitrem D was given a scaled score of 0, and has over 56 steps, a total estimated time of about 587 hours, and with a yield of only 0.17%, which was categorized by a synthetic chemist within a category of upper-level difficulty (**Table 1**). Molecules that were scored high, and are therefore simpler, were accurate. For example, triquinacene has a low number of steps and requires relatively simple reagents in our third collected synthetic route. Given that this molecule had a scaled score of 9.9, the predicted score matches its actual synthetic accessibility. On the other hand, the Autoencoder SAScores are not entirely accurate for certain molecules that have scores less than five. For example, Adamantane, which has less than 10 steps, total time of less than 24 hours in the simplest synthesis, and two reported syntheses yielding upwards of 10% according to data collected from Synarchive, has a scaled SAScore of 0.016 (**Table 1**) but is a molecule that would likely be rated as relatively simple and straightforward to make.

Another important finding from our generated SAScores is that the GMM scores have a skewed distribution, as all but five of the molecules had a score under 5, with most of the scores concentrated between 0 and 1. As a result, it is not possible to compare the GMM scores directly with the Autoencoder scores. However, it is possible to compare the GMM scores of the molecules to each other in a trimmed

Chemical	# of Steps	Max Temp (°C)	Min Temp (°C)	Time (hr.)	Yield	SAScore from Autoencoder	SAScore from GMM	Synthetic Chemist Ranking
Penitrem D	56	120	-78	586.55	0.0017	0	-6.98	8
Triquinacene	4	25	25	5	0.019	2.33	-12.54	1
Adamantane	6	400	105	23.5	0.19	0.0037	-8.89	3
Taxol[6]	61	240	25	400	0.040	0.32	-7.43	9
[4,5]Coronane	5	3180	20	80	11.00	1.79	-12.31	2

Table 1: Molecules within the dataset that the Autoencoder and GMM assigned synthetic accessibility scores for. The table contains information for five molecules and each of their synthetic route's number of steps, maximum and minimum temperatures, total time, and yield, as well as raw unscaled rankings of the machine learning models for synthetic accessibility and a chemist's approximate ranking.

dataset that excludes seven outlier molecules ranked above 4, as their scores were disproportionately higher above the other molecules' max score of 3.1. In addition, a noticeable finding is that the Autoencoder and GMM had similar rankings for the molecules in the dataset, i.e., molecules that one model ranked as higher difficulty also tended to be higher ranked by the other model. For example, Penitrem D is given a scaled score of 2.25, which is on the upper score end of the trimmed datasets, and thus the Autoencoder and GMM both agreed with each other in terms of synthetic accessibility scoring. Another molecule of high synthetic difficulty is Taxol, with 61 total steps, 5% yield, and upwards of 20,000 minutes of synthetic time (**Table 1**). This molecule had an Autoencoder score of 0 as well, and a GMM score of 2.24, showing that both models demonstrated consistency when ranking high difficulty molecules.

The GMM also appeared to be accurate for molecules of easier synthetic accessibility, such as for (4,5) Coronane. The synthesis of this molecule has only five steps, relatively mild temperature extremes, and an overall yield of 11% (Table 1). Thus, chemists would likely rank it at around 3, which corresponds with the scaled GMM score of 0.36. The Autoencoder score for this molecule is 7.6, and so also matches the synthetic chemist and GMM rankings.

Ultimately, in the field of molecular generation, complexity is vital in selecting synthetically viable compounds. No matter how novel and promising the compound is, synthetic chemists must always consider the synthetic accessibility of the molecule. Therefore, a quantitative metric for synthetic accessibility must be incorporated as part of any machine learning for drug discovery efforts involving Generative Adversarial Network (GAN), Variational Autoencoder (VAE), Reinforcement Learning (RL), or any other methods.

By comparing two separate approaches, the Autoencoder and the GMM, to obtain a synthetic complexity score, we showed that an Autoencoder is more effective. We expect our algorithm to drastically reduce the computation time and processing power needed to calculate the complexity of a drug molecule without relying on any arbitrary assignments or subjective weighting assumptions. Our model is objective and generates weights for a molecule's complexity solely based on intrinsic factors. This can be generalized to the synthesis of many compounds by enabling rapid high-throughput virtual screening of drug molecules that can greatly shorten the whiteboard to clinical timeline.

The use of the Autoencoder framework in both models in compressing the data is possible because of the Autoencoder 's ability to compress information for a nonlinear dimensionality reduction. If the activation function used within the Autoencoder is linear within each layer, the latent variables present at the bottleneck (the smallest layer in the network) directly correspond to the principal components from Principal Component Analysis (PCA). When nonlinear activation functions are used, such as in this case the Rectified Linear Unit (ReLU), Autoencoders work as a nonlinear dimensionality reduction algorithm, thus being able to capture the smooth curve that is displayed in the t-SNE, like how a linear Autoencoder would match the PCA. The main benefit added with the Autoencoder is that while the t-SNE must be recalculated every time there is a new datapoint added to the dataset, the Autoencoder generalizes well and can be applied to new datapoints without the need for retraining.

A central limitation to our current model is the lack of sufficient data to train the GMM. We believe that this limitation can be addressed in future work by compiling a larger dataset of small molecule compounds. A more diverse dataset of synthetic molecules, natural products, and natural product-inspired structures can further improve the classification of synthetic accessibility among a larger manifold of molecules. Our synthetic accessibility score can also be combined and stacked via ensemble machine learning with other metrics for synthetic accessibility to provide a more comprehensive and holistic assessment of the synthetic complexity of chemical entities. Nevertheless, the synthetic accessibility score we put forth in this paper is a novel metric that can help determine the viability of not just synthetic and natural product drugs, but also the feasibility of synthesizing any chemical substance, including antibiotics, chemical cells, and chemical electronics.

In future work, we will need to address the issue of choosing the optimal molecular representation in order to capture maximal information about synthetic accessibility. A larger SMILES embedding model with denser and dropout layers is a promising solution. In future work, we want to explore how different graph-based representations of the molecular structures might enhance the SAScore. This is because certain molecular traits such as those involving a molecule's stereochemistry can only be calculated from three-dimensional spatial representations. Finally, to test the accuracy of the SAscores, synthetic chemists can physically synthesize the molecules detailed in this paper in the lab to further validate the complexity of the molecules presented.

MATERIALS AND METHODS

Dataset

We used SynArchive as the source of our chemical synthetic data (13). The database contains 194 unique natural products and their previously reported synthetic routes. Each synthetic route contains the number of total steps, the number of parts in the synthesis, and each of the steps in the synthesis, with the reactants, reagents, solvent, temperature, and the named reaction that is occurring, if any or all are applicable. We extracted information regarding the total reaction steps, cumulative reaction time, the temperature extremities, and the composite yield of the entire synthetic route for each molecule from the dataset, making use of Optical Structure Recognition (OSRA) to convert from the molecular image on the website to the molecular graph (14).

SMILES Encodings

We represented the synthetic pathway of a molecule, primarily reactants, products, and intermediary compounds, using the Simplified Molecular Input Line Entry System, or SMILES, a notation that captures a chemical's three-dimensional structure into a string of symbols that can be processed by computer software (15). While this omits some structural information about the individual molecules in the route, such as the ability to explicitly define stereochemistry for individual stereocenters, we found that incorporating the structural information is memory intensive and makes the underlying data compression problem unwieldy and the results difficult to interpret.

T-SNE Visualization

Our approach for building a synthetic complexity model

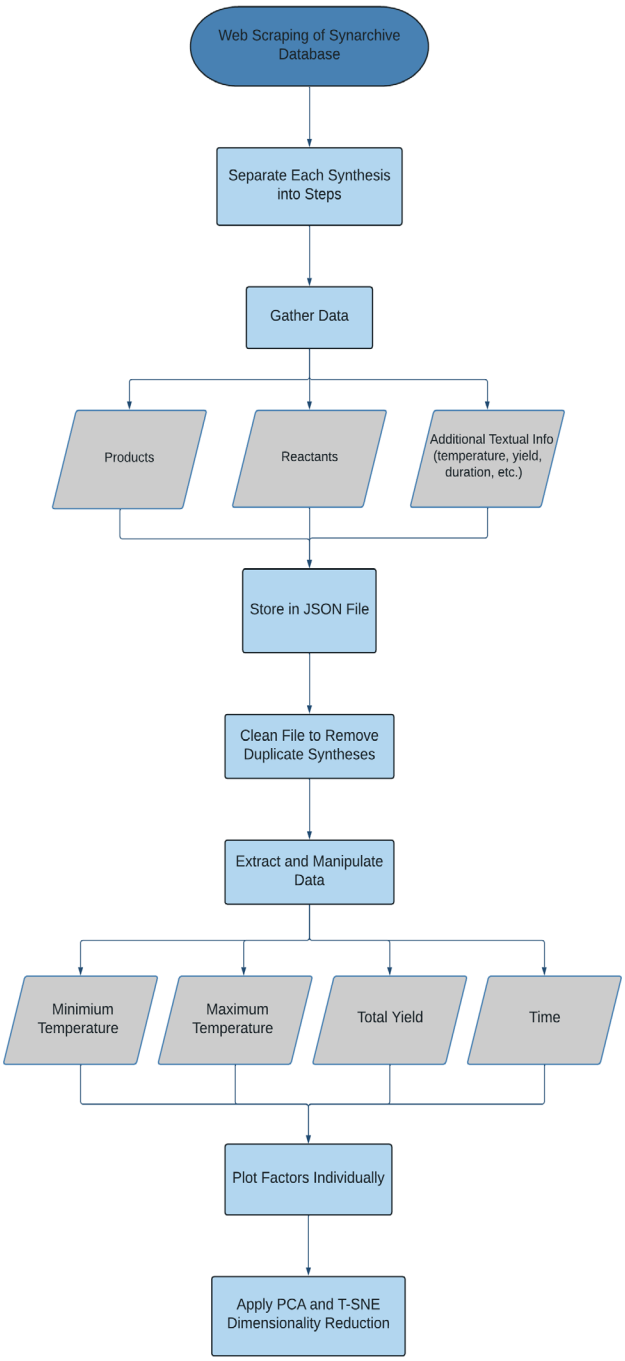

Figure 6: Schema of Synthetic Complexity Pipeline. Schema of synthetic accessibility computation with the Autoencoder. The time, steps, maximum temperature, and minimum temperature are extracted for each molecule and stored in a dictionary. The four factors are then compressed to one dimension after undergoing the t-SNE dimensionality reduction step. The 1D output becomes the label for each molecule and the labeled data is used to train the Autoencoder.

broke down chemical synthesis pathways into four influential factors. Using Python, we parsed the data into the major synthetic accessibility (SA) factors: time, minimum and maximum temperature, the number of steps, and the total yield of each synthetic pathway, and inputted them into the t-SNE algorithm (**Figure 6**). T-distributed Stochastic Neighbor Embedding, or t-SNE, is a nonlinear dimensionality reduction algorithm that has been shown to preserve both the local and global structure of higher dimensional data into two or three dimensions, allowing for visualization (16).

To do so, the conditional probabilities were calculated for the Euclidean distances between the complexity features, which represent the similarity of datapoint x_j to datapoint x_i using the conditional probability $p_{j|i}$ that x_i would pick x_j as its neighbor. A t-distribution centered at x_i was then sampled to pick the neighbors.

$$\text{Equation 1: } p_{j|i} = \frac{\frac{exp\,(-||x_i - x_j||^2)}{2\sigma_i^2}}{\sum_{k \neq i} \frac{exp\,(-||x_i - x_k||^2}{2\sigma_i^2}}$$

A similar function $q_{j|i}$ was calculated for the low dimensional counterparts of x_i and x_j, y_i and y_j.

$$\text{Equation 2: } q_{j|i} = \frac{(\frac{exp\,(-||y_i - y_j||^2)}{2\sigma_i^2})^{-1}}{\sum_{k \neq i} (\frac{exp\,(-||y_i - y_k||^2)}{2\sigma_i^2})^{-1}}$$

For the low dimensional representations of xi and xj to reflect the structure of the high dimensional data, the conditional probabilities $p_{j|i}$ and $q_{j|i}$ must be close to equal. A measure of the difference between the q and p distributions is the Kullback-Leibler divergence, an information-based measure of disparity among probability distributions. The cost function thus consisted of the sum of the Kullback-Leibler divergences over all data points.

$$\text{Equation 3: } C = \sum_i KL(P||Q) = \sum_i \sum_i p_{j|i} log\,(\frac{p_{j|i}}{1_{j|i}})$$

This cost function was minimized using gradient descent, an iterative optimization algorithm based on first order gradients used to find the local minimum or maximum of a function.

Gaussian Mixture Model

We found that using the manually selected synthetic factors omits a significant amount of information about the synthetic pathways. Due to the curse of dimensionality (17), it was intractable to directly use all the data available on SynArchive since the dataset size was extremely limited. Therefore, we utilized a Deep Autoencoding Gaussian Mixture Model (GMM) to map the entire synthetic pathway of a molecule into a single Gaussian energy function. The GMM allowed us to train a model using our synthetic pathway data, and we used this as a second method to predict synthetic complexity to compare against our regression model. We followed the approximate model architecture developed by Zong et al., which uses latent embeddings and reconstruction error to learn a multimodal Gaussian distribution in an end-to-end manner (18). First, we trained a separate Autoencoder to generate encodings for the 41 unique characters in the model, with each character representing either an atom of the molecule or an atomic bond. This Autoencoder captures meaning about the symbol itself as it relates to the total syntheses of all the molecules. Next, we used Gated Recurrent Unit (GRU) cells to create the Autoencoder to encode the sequential data (19). This GRU-Autoencoder (GRU-AE) was then used to "score" a window of sequential data according to how close it was to normality.

We then trained the model to generate a unimodal Gaussian energy function. Finally, we plotted the energy function outputted by the model and verified that it is Gaussian given the bell curve shape of its distribution.

Autoencoder Model

After obtaining validation from the t-SNE step mentioned above that the data in fact fit a Gaussian curve, we utilized an Autoencoder to flatten the input variables into a univariate output. Autoencoders are an unsupervised deep learning model which compresses data into a latent vector in a lower-dimensional subspace by mapping the higher dimensional points onto a nonlinear manifold (20). We used the Autoencoder to compress the five factors: total time of the synthesis, number of steps, the maximum temperature in the synthesis, the minimum temperature of the synthesis, and total yield into a single one-dimensional encoding output: the synthetic accessibility score.

The Autoencoder architecture consists of encoder and decoder networks with a bottleneck layer. Layers 2, 3, and 4 take the ReLU activation function and the last layer takes a sigmoid activation function. The input layer consists of 5 nodes, one for each of the variables we considered. The Autoencoder first undergoes the encoding process and compresses the data to a 1-dimensional representation of the input. This 1D synthetic accessibility score encompasses all the information held within the chemical molecule. The model then transitions into the decoding process and tries to generate the five original variables from the input.

In all experiments, we used a batch size of 327 and trained using the Adam optimizer with a mean squared error loss function. To ensure that the Autoencoder was properly encoding the dataset, we split the data into training and test data, using an 83 to 17 percent split, respectively, and monitored the validation loss to ensure that the model was not overfitting.

We trained the Autoencoder to minimize the reconstruction error given by equation 4.

$$\text{Equation 4: } L(x, x') = \|x - x'\|^2$$

This equation represents the mean squared error between the input data and the output of the Autoencoder model, represented by x and x'. We reported the best model after 1000 epochs. The GMM and Autoencoder models were then compared to determine the best metric for synthetic accessibility.

ACKNOWLEDGEMENTS

We would like to acknowledge and thank the developers and authors of all open-source packages used in our research. The authors declare no competing conflicts of interests in the work presented. All code will be provided upon request. We gratefully acknowledge the Aspiring Scholars Directed Research Program for funding for our research.

Received: January 9, 2022
Accepted: June 14, 2022
Published: June 19, 2023

REFERENCES

1. Bellman, Richard. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1972.
2. Bertz, Steven H. "Convergence, Molecular Complexity, and Synthetic Analysis." *Journal of the American Chemical Society*, vol. 104, no. 21, 1 Oct. 1982, pp. 5801–5803., https://doi.org/10.1021/ja00385a049.
3. Cho, Kyunghyun, et al. "Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3 Sept. 2014, https://doi.org/10.3115/v1/d14-1179.
4. Coley, Connor W., et al. "SCScore: Synthetic Complexity Learned from a Reaction Corpus." *Journal of Chemical Information and Modeling*, vol. 58, no. 2, 2018, pp. 252–261., https://doi.org/10.1021/acs.jcim.7b00622.
5. Ertl, Peter, and Ansgar Schuffenhauer. "Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions." *Journal of Cheminformatics*, vol. 1, no. 1, 2009, https://doi.org/10.1186/1758-2946-1-8.
6. Filippov, Igor V., and Marc C. Nicklaus. "Optical Structure Recognition Software to Recover Chemical Information: OSRA, an Open Source Solution." *Journal of Chemical Information and Modeling*, vol. 49, no. 3, 2009, pp. 740–743., https://doi.org/10.1021/ci800067r.
7. Gupta, Anshul, et al. "Descriptive Statistics and Normality Tests for Statistical Data." *Annals of Cardiac Anaesthesia*, vol. 22, no. 1, 22 Mar. 2019, p. 67., https://doi.org/10.4103/aca.aca_157_18.
8. Johnson, Oliver. *Information Theory and the Central Limit Theorem*. Imperial College Press, 2006.
9. Laurens, Van der Maaten, and Geoffrey Hinton. "Visualizing Data Using t-SNE." *Journal of Machine Learning Research*, vol. 9, no. 11, 2008.
10. Li, Jun, and Martin D. Eastgate. "Current Complexity: A Tool for Assessing the Complexity of Organic Molecules." *Organic & Biomolecular Chemistry*, vol. 13, no. 26, 12 May 2015, pp. 7164–7176., https://doi.org/10.1039/c5ob00709g.
11. Nielsen, Frank. "On the Jensen–Shannon Symmetrization of Distances Relying on Abstract Means." *Entropy*, vol. 21, no. 5, 2019, p. 485., https://doi.org/10.3390/e21050485.
12. Office, U.S. Government Accountability. "Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development [Reissued with Revisions on Jan. 31, 2020.]." *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development [Reissued with Revisions on Jan. 31, 2020.]* | U.S. GAO, 6 Feb. 2020, https://www.gao.gov/products/gao-20-215sp.
13. Randi?, Milan, and Dejan Plav??i? "Characterization of Molecular Complexity." *International Journal of Quantum Chemistry*, vol. 91, no. 1, 2002, pp. 20–31., https://doi.org/10.1002/qua.10343.
14. Rumelhart, David E., and David E. Rumelhart. *Parallel Distributed Processing - Vol. 1: Foundations*. MIT Press, 1987.
15. "A Total Synthesis Database." *SynArchive*, http://www.synarchive.com/.
16. Voršilák, Milan, et al. "Syba: Bayesian Estimation of Synthetic Accessibility of Organic Compounds." *Journal of Cheminformatics*, vol. 12, no. 1, 2020, https://doi.

org/10.1186/s13321-020-00439-2.
17. Weininger, David. "Smiles, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules." *Journal of Chemical Information and Modeling*, vol. 28, no. 1, 1988, pp. 31–36., https://doi.org/10.1021/ci00057a005.
18. Weissler, E. Hope, et al. "The Role of Machine Learning in Clinical Research: Transforming the Future of Evidence Generation." *Trials*, vol. 22, no. 1, 2021, https://doi.org/10.1186/s13063-021-05489-x.
19. Zhu, Jingyu, et al. "Integrating Machine Learning-Based Virtual Screening with Multiple Protein Structures and Bio-Assay Evaluation for Discovery of Novel gsk3β Inhibitors." *Frontiers in Pharmacology*, vol. 11, 2020, https://doi.org/10.3389/fphar.2020.566058.
20. Zong, Bo, et al. "Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection." *International Conference on Learning Representations*, 15 Feb. 2018.