Application Note

# SYNTHIA™ Retrosynthesis Software Application Pogramming Interface

Assess synthetic accessibility scores (SAS) of thousands of virtual molecules in minutes


Pioneer Your Pathway
**SynthiaOnline.com**



MK_AN14021EN Ver. 1.0
11/2024

The Life Science business of Merck operates as MilliporeSigma in the U.S. and Canada.

Sigma-Aldrich®
Lab & Production Materials

Sigma-Aldrich®
Lab & Production Materials



# Application Programming Interface

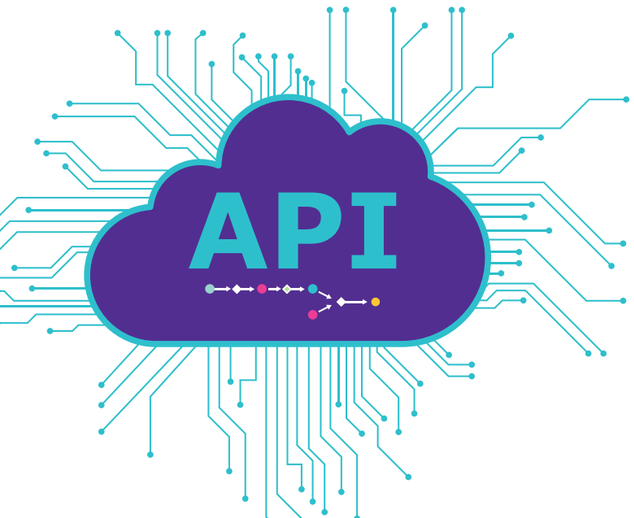
## Interconnect Your Cheminformatics Tools with SYNTHIA™ Retrosynthesis Software

### API Access

API Access (Application Programming Interface) is available for organizations that would like to interconnect other Cheminformatics tools with SYNTHIA™ for a customized experience.

### Benefits Include:

- Access full retrosynthesis or Synthetic Accessibility Score (SAS) API
- View data side-by-side to improve insights into molecule selection
- Create robust visualizations using multiple data sources
- Inform molecule selection upstream from the synthesis step
- Analyze thousands of pathways in minutes with the SAS API




### Harness the power of the Synthetic Accessibility Score (SAS)

The ability to differentiate between 'easy-to-make' and 'difficult-to-make' molecules is a hard, but widely useful task, e.g., for prioritizing compounds in virtual screening pipelines. By combining the modern deep-learning model, and data collected with our renowned retrosynthetic planning software, we deliver SYNTHIA™ Synthetic Accessibility Score (SAS) service, a tool applicable to high-throughput in-silico compounds processing.

At present, combinatorial chemistry and generative modelling are used for constructing gigantic compounds datasets [1]. However, the actual synthesis of many molecules obtained with such methods may be challenging. To address this problem, synthetic accessibility measures are used to determine molecule feasibility as early as possible in drug discovery pipeline.

SYNTHIA™ SAS API service provides the predictions for such 'molecular complexity' in terms of number of synthetic steps from small, commercially available building blocks. The machine learning model underpinning SAS has been pre-trained on synthetic scenarios obtained with algorithms from SYNTHIA™ Retrosynthetic Planning Tool [2], [3], [4]. Finally, our cloud hosted and ISO-27001 certified product offers the ability to easily process millions of molecules daily and up to a thousand molecules in a single query, enabling SYNTHIA™ SAS service prediction to be more commonly used in drug design process.

### Input/output for SAS model

Input molecules need to be provided in the widely used SMILES text format [5] and the API endpoint supports batch requests. The input SMILES consist of single fragment molecule.

The returned measure, here defined as Synthetic Accessibility Score (SAS), is a single float number from range 0-10, assigned for each corresponding input molecule. Returned score approximates how many steps it takes to synthesize the molecule using commercially available building blocks. The lowest numbers (values close to 0) are returned to chemicals that are predicted to be easy to make (or even can be commercially available). The higher numbers are returned when the model forecasts more synthetic steps to obtain the requested compound. For scores close to maximal value (10), synthesis is predicted to be either extremely complex (many reaction steps) or even unfeasible, e.g., due to exotic structural motifs in the molecule. In general, the lower the score the easier it should be to synthesize the molecule.

In an event that some of the molecules in request are invalid (e.g., hypervalent, incomplete rings, improper protonation of aromatic atoms, multi-fragment) the request will still be processed. Scores for such entries will be null and appropriate comments will be returned alongside in the response structure.

### Predictive model characteristics

SYNTHIA™ SAS v1.0 is based on a regressor that includes graph convolutional neural network (GCNN). Such architecture allows for learning an internal representation of each molecule by operating on its graph structure rather than pre-computed molecular descriptors [6]. In particular, the model consists of bond-level directed message passing neural network (D-MPNN) followed by feedforward neural network (FNN) The implementation was adapted from Chemprop open-source project [7].

Machine learning model was trained using SYNTHIA™ automatic retrosynthesis module results as a target value. Specialized and normalized SYNTHIA™ score was used to reflect the number of steps, e.g., not penalizing non-selective reactions, implicit protections strategy, minimal price contribution to the score, and only small building blocks were used as SYNTHIA™ search settings. Additionally, a smoothing function was applied to better build gradient for high scores, aimed for better resolution of hard to synthesize molecules (see also Fig. 1).

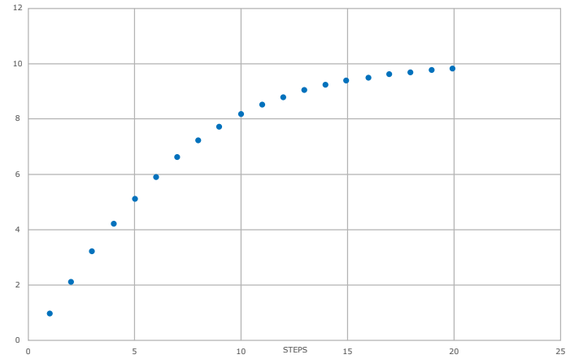



**Figure 1.** Depiction of smoothing function applied to scores. Note, that over small and moderate values (x-axis), the synthetic accessibility score (y-axis) behaves close to linear. In other words, the returned score corresponds to the number of synthetic steps predicted by the model. For a higher number of predicted synthesis steps (around 10 or above), the related score is smoothened such that the returned value is still close to (and not greater than) 10. This allows to re-scaling all considered cases to [0, 10] interval.

The data used for training of machine learning models has 33306 molecules in total. It is composed of known molecules (ChEMBL database) [8] and combinatorically generated small molecules (GDB) [9]. The composition of data before train/test split:

- GDB subset: 16081, including:
  - compounds with 1-7 heavy atoms (C, N, O, Cl, S): 7198
  - compounds with 8-9 heavy atoms (C, N, O): 8883
- ChEMBL subset: 17225, including:
  - randomly selected synthetic small compounds: 15449
  - randomly selected natural product-derived compounds: 1776

Training and evaluation of machine learning model required splitting the data into training and testing sets (common 80/20 train/test split was used). Further, the internal validation set was extracted using 9:1 ratio from the training set and was used for network parameters optimization.

Predicted score (SYNTHIA™ SAS model) correlates with target value based on SYNTHIA™ scores with R2 = 0.726 and MAE = 1.1497. Scatter plot with fitted line and box plot showing density/distribution of data points, are presented on Fig. 2.

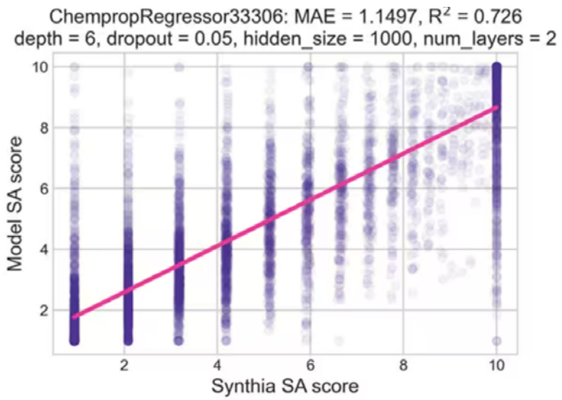



**Figure 2.** Scatter and box plot showing the correlation between SA scores calculated with SYNTHIA™ vs. scores learned by the model.

The results predicted with SYNTHIA™ SAS are based on relationships retrieved from datasets (possibly, quite complex and not straightforward to capture). This should be taken into consideration when novel molecules are queried via SYNTHIA™ SAS-API. Namely, scores for molecules that are not related to test set might fall out of so-called applicability domain, hence corresponding results might not be meaningful. This is a typical limitation for data-driven models, nevertheless it is always good to remember such a limitation to avoid misinterpretation of obtained scores.

## Case studies

### Case 1

N-acetyl derivative of sulfamethoxazole (Fig. 3, left) is a direct precursor of this drug (Fig. 3, right). Despite the more complex chemical structure, the derivative is recognized as easier to synthesize (SAS=1.038 is much smaller than SAS=4.051).

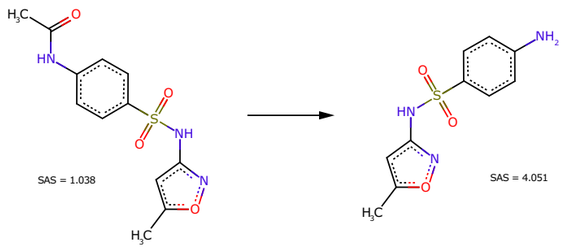



**Figure 3.** Chemical structures of molecules for sulfamethoxazole use case.

### Case 2

On the other hand N-Boc derivative of adrenaline (Fig. 4, left) is not a direct precursor of adrenaline (Fig. 4, right). In typical procedure there is no need to protect the amino group throughout the synthesis pathway. N-Boc derivative is correctly recognized as more complex in terms of synthetic accessibility (SAS=8.399 is greater than SAS = 7.631). This is in line with the fact that adrenaline is a precursor of its N-Boc derivative.

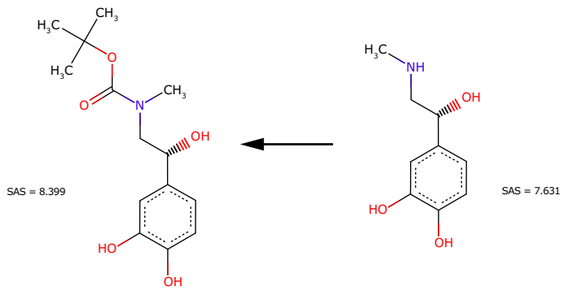



**Figure 4.** Chemical structures for adrenaline use case.

## User Dataflow

SYNTHIA™ SAS is a cloud hosted service, available for each customer via RESTful API. It is horizontally scalable and provides high throughput via a single API entry point for all customers. The end-user needs to provide a list of molecules in a SMILES format and SYNTHIA™ SAS returns a score for each of them (Fig. 5). The service is stateless and designed to scale according to the demand.

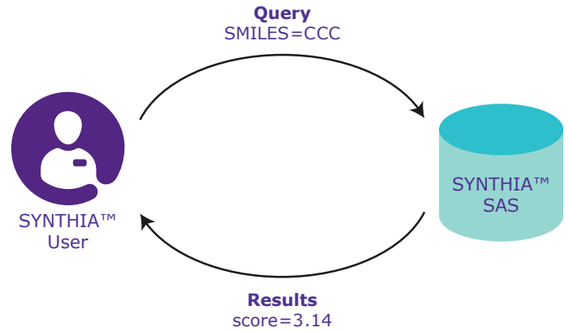



**Figure 5.** Schematic representation of SYNTHIA™ SAS service data flow.

## References


1. Joshua Meyers, Benedek Fabian, Nathan Brown, De novo molecular design and generative models, *Drug Discovery Today*, 26, **2021**, 2707-2715. DOI
2. SYNTHIA™ Retrosynthesis Software
3. Tomasz Klucznik, et al., Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory, *Chem, 4,* **2018**, 522-532. DOI
4. Mikulak-Klucznik, B., et al. Computational planning of the synthesis of complex natural products, *Nature, 588,* **2020**, 83–88. DOI
5. Daylight Chemical Information Systems, Inc.
6. Yang, K., *et al.* Analyzing Learned Molecular Representations for Property Prediction, *Journal of chemical information and modeling, 59,* **2019**, 3370-3388. DOI
7. Chemprop open-source project
8. ChEMBL database
9. GDB database