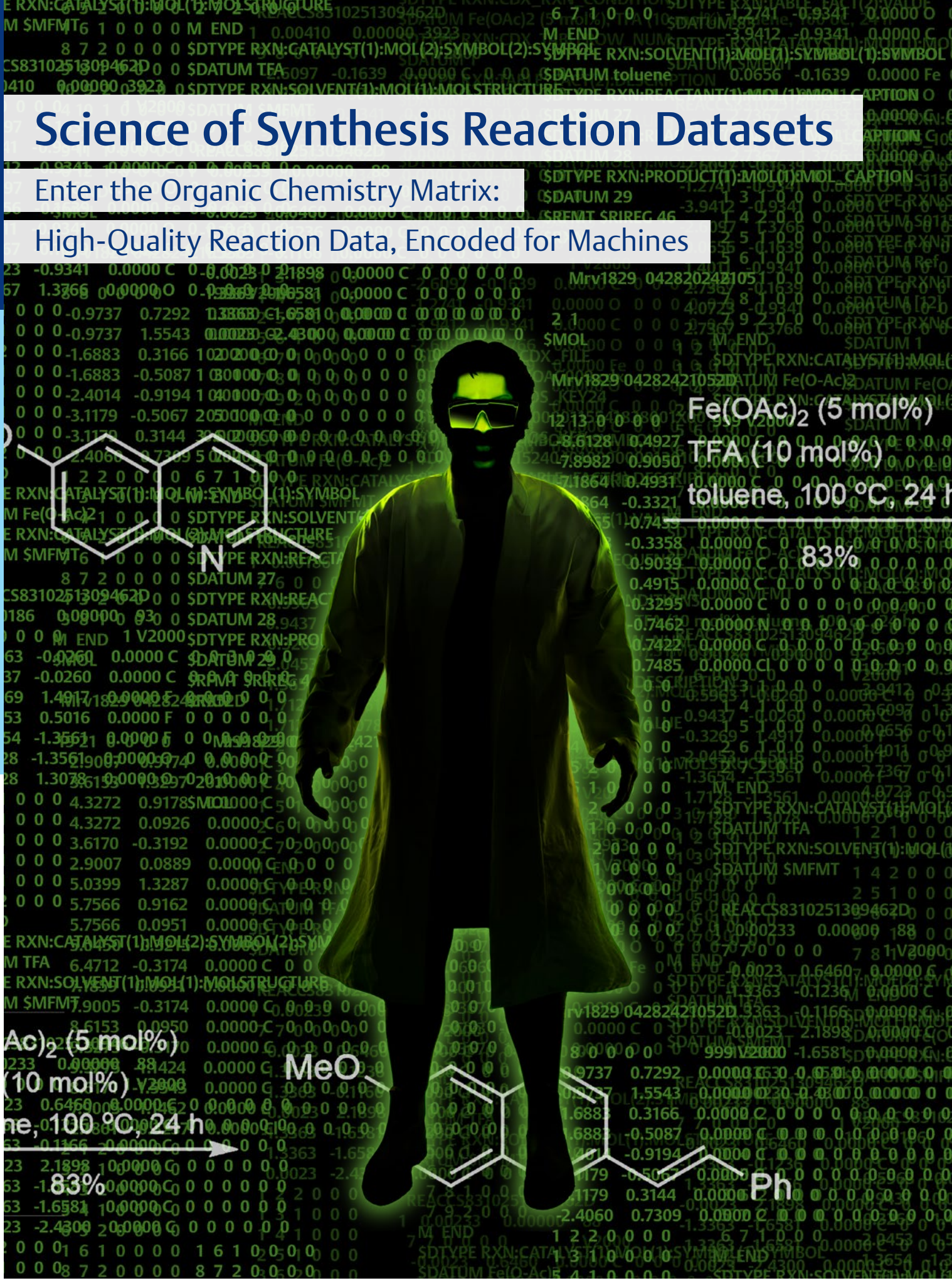
Science of Synthesis Reaction Datasets

Enter the Organic Chemistry Matrix:

High-Quality Reaction Data, Encoded for Machines



Science of Synthesis

Thieme

Science of Synthesis Reaction Datasets:

High-Quality Reaction Data in Organic Chemistry, Converted into Machine-Readable Format

Science of Synthesis (SOS), a prestigious reference work in organic chemistry, contains critical reviews of the entire field of organic and organometallic chemistry. The reactions included are those judged to be most synthetically relevant and most reliable, as selected by experts in each field. Thieme has converted this wealth of organic synthesis knowledge into machine-readable format: **Science of Synthesis Datasets!**

These highly structured datasets can provide a crucial basis for training AI-based models or rules-based algorithms for retro-synthesis and forward-reaction prediction. The consistently formatted experimental procedures potentially allow the automation of synthetic chemistry. Such use cases have application in academia as well as in industry.

Improved Results when Training Models for Reaction Prediction and Retrosynthesis

The better the quality of the data used for training your models the better the results you obtain.

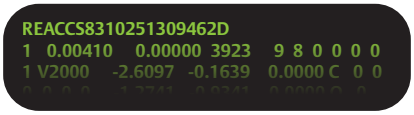
Science of Synthesis provides chemical reaction and structure data to an unprecedented level of accuracy and reliability. The abstraction of the data has been carried out with great care, predominantly manually, to ensure highest quality. Experimental procedures have been edited for clarity and checked for scientific accuracy. In addition to this, the Science of Synthesis datasets are very diverse, covering a much wider range of chemistry than that included in, for example, publicly available datasets automatically abstracted from patents.

The SOS Dataset Currently Includes

- Over 500,000 reactions (available in V2000 BIOVIA CT RD file format and SMILES format)
- Over 2.5 million molecules (available in V2000 BIOVIA CT SD file format)
- Over 2,500 full-text files in PDF format
- Over 80,000 full-text files in XML format
- Over 60,000 experimental procedures in XML format, edited for clarity and checked for scientific accuracy.

High-Quality, Highly Structured Data: Ready for Use

There is no need for time-consuming, expensive text and data mining. The SOS dataset is already machine-readable and ready for use. It can be employed alone or readily used to supplement other data, such as your own in-house data or commercially or publicly available datasets.



Advantages of the SOS Dataset

- **Consistent and highly structured:** the consistent and accurate format allows rapid integration into your system without significant cleanup needed
- **Very diverse:** it covers a very broad range of organic reactions and so allows AI models to learn from the full breadth of knowledge
- **Has a high proportion of unique reactions:** More balanced and less skewing during training
- **Hand-picked:** Extra quality – only that chemistry recommended by experts is included
- **Predominantly manually curated:** Fewer errors than machine-abstracted data
- **Regularly updated:** it is always current and up to date. New chemistry is always being added (approx. 10,000 reactions p.a.)

Choose Your Reality

Real Laboratory

Synthetic Organic Chemists, Medicinal Chemists, Process Chemists

Virtual Laboratory

Cheminformaticians, Theoretical Chemists, Data Scientists

The Solutions We Offer You

A high-performing model for retrosynthesis and forward-reaction prediction on IBM's RXN for Chemistry platform

In collaboration with IBM, we offer a pre-trained model for retrosynthesis and forward-reaction prediction on IBM's RXN for Chemistry platform.

High-quality, diverse datasets covering the broad scope of organic chemistry

We can provide you with the whole dataset, or just the part that focuses on your area of research. We offer flexible pricing models to cover academic or various commercial use cases.

Interested?

If you are interested in using the Science of Synthesis (SOS) data, please get in touch. We will be pleased to provide you with our content and further information!

Please contact: **sos-datasets@thieme.com**

24P3EP/2026

Thieme