Digitization of Chemical Reactions Schemes

Mark Martori Lopez
IBM Research Europe - Zurich
Rueschlikon, Switzerland
mark.martori.lopez@ibm.com

Daniel Probst*
IBM Research Europe - Zurich
Rueschlikon, Switzerland
daniel.probst@epfl.ch

Amol Thakkar†
IBM Research Europe - Zurich
Rueschlikon, Switzerland
tha@zurich.ibm.com

Teodoro Laino†
IBM Research Europe - Zurich
Rueschlikon, Switzerland
teo@zurich.ibm.com

Abstract

The success of machine learning models trained for reaction prediction and retrosynthesis is heavily influenced by the quality of the chemical data sets used in their training. While many datasets have been compiled with the help of text mining algorithms, only a few have been curated by humans. Because of the various ways in which chemical reactions can be depicted, the extraction of chemical reaction records through processing of chemical reaction schemes remains one of the most challenging tasks and is largely untapped. However, in recent years, machine-learning methods have demonstrated remarkable performance in converting molecular images to textual representations of single molecule structures, such as SMILES. In this work, we present a twofold approach. First, vision transformers (DETR) recognize key graphical elements in the images contained within chemical reaction schemes such as molecules, text, arrows, and symbols. Then, a variety of digitization techniques convert the recognized elements into machine-processable formats. Furthermore, to compensate for the scarcity of annotated chemical reaction schemes, we developed an artificial training data set with balanced intra-variability to mimic the distribution of real-world illustrations. This research opens new avenues in digital chemistry for facilitating the creation of on-demand data sets to enrich chemical spaces with insufficient or limited available data.

1 Introduction

Over the past few decades, a wealth of information on new chemical reaction schemes has been attained. However, only a fraction of these most recent scientific findings are included in commercial or publicly available data sets[1]. Even though it would be ideal to compile every single chemical reaction scheme from published article's data into a searchable database, doing so manually is tedious and prone to mistakes. An automatic knowledge extraction system would rapidly and automatically incorporate cutting-edge discoveries into chemistry. This tool must return chemical reaction schemes in the literature in a computer-friendly format to create new public comprehensive databases of chemical reactions.

ChemDataExtracter[2] and ChemSchematicResolver[3] are two open-source programs that have contributed expanding the field of chemical data mining[4]. Few technologies[5, 6, 7, 8, 9] have

*current address: EPFL, Lausanne, Switzerland
†National Center for Competence in Research-Catalysis (NCCR-Catalysis), Switzerland

Preprint. Under review.

been developed to automatically recognize chemical structure illustrations and convert them to machine-readable formats like SMILES[10] or InChI[11] in an effort to improve the knowledge extraction process. If the conversion of images of chemical structures into a format that can be read by computers is a relatively new field of study[1], the conversion of chemical reaction scheme images is a completely unexplored territory.

Chemical reaction schemes illustrate chemical processes, characterised by arrows displayed in between the molecules. More specific details are given in a textual representation, such as the conditions used to conduct the reaction, and these are frequently placed above the illustrated arrows. However, the complexity of the notation of chemical reaction schemes can rapidly increase, resulting in a wide range of variability with no standardised representation rules. This lack of format definition prevents easy machine readability and conversion[12]. In this paper, we investigate a few leading architectures providing insightful qualitative observations and quantitative scoring metrics to detect objects of high interest in chemical reaction schemes. Vision transformers scored better accuracy than convolutional neural networks (CNNs) 1. The objects detected from the most accurate models are then converted to a machine-readable reaction schema. Our technology represents the first attempt to eliminate time-consuming, labor-intensive manual annotations of chemical imagery. This approach will hasten the incorporation of new data into machine learning models for chemical reactivity.

2 Visual Recognition Systems

The directions and the textual information added to chemical reaction schemes provide key information for the scientific community to understand the applicable uses that molecules can have. In order to locate high-interest elements in schemes, object detectors aim at labeling them with rectangular bounding boxes to show their confidence of existence while extracting meaning out of the images[13]. Overall, two-stage detectors such as Faster-RCNN have high localization and recognition accuracy but a slow inference speed[14]. Likewise, YOLO, the pioneer object detection architecture of one-stage detectors increased inference speed considerably allowing object detection in real-time scenarios [15]. More recently, *RetinaNet* obtained competitive accuracy by modifying the loss based on the large class imbalance encountered during the training of dense detectors[16]. However, all one-stage and two-stage approaches such as RetinaNet, and FasterRCNN present post-rectification steps, like anchors, which slow down the inference time and trigger errors when dealing with duplicated bounding box predictions. Instead, in 2020, the DETR architecture[17] with the self-attention mechanism A.1 became the first direct set predictor via a bipartite matching technique based on the Hungarian Loss[18], eliminating the need for time-consuming post-rectification steps for duplicated predictions and thereby speeding up inference.

3 Methods

Data Generation Given the scarce availability of publicly available depictions of annotated chemical reactions in databases, a fine-tuning approach has been followed by using pre-trained weights on the COCO data set from[17] and retraining the classification head on a custom data set. Thus, the invention of an artificial chemical data set was needed to avoid manually annotating thousands of images.

Artificial training data set This data set is made up of 50,000 randomly generated samples. All of the molecules (around 500k) added to the images of the data set were randomly selected and downloaded from the 109 million compounds stored in PubChem[19] to simulate the distribution of real-world chemical images as represented in the literature. In addition, arbitrary text and arrows, similar to those used in articles, have been inserted to represent the direction and additional information of each stage of the reaction.

Real-world test data set This dataset, with $approx$ 2000 samples, is used to test the accuracy of the models. Due to a lack of annotated reactions depictions in databases, the evaluation data set has a smaller number of samples than is typical for deep learning architectures[13]. The data was collected from Thieme[20] and annotated together with the help of colleagues from IBM Research in a small-scale annotation effort using the open-source MakeSense tool [21].

4 Results and Discussion

4.1 Fine-tuning details

All experiments were conducted with the pre-trained weights taken from the COCO dataset, and the models provided by detectron2 [22], ADAM[23] for FRCNN and Retinanet, and ADAMW[24] for DETR were used. The images were pre-processed by an absolute random crop, random horizontal and vertical flip, and a resize-shortest-edge technique. A more detailed summary of the fine-tuning for each architecture is referred to in the Appendix.3

4.2 Evaluation Metrics - mAP with generalized intersection over union.

The standard precision and recall metrics can be used to evaluate detection and classification tasks. The bounding box, predicted by the model to locate each object, is compared to the actual coordinates of each element (ground truth). The generalized over union (GIoU) evaluation metric measures the intersection between two rectangles. A prediction is considered *true positive* if it intersects the coordinates of the object more accurately than a previously set threshold. If the GIoU threshold is set to 0.95, predictions overlapping less than 95% of the ground truth areas are set as *false positives*. The area under the precision-recall curve is averaged into the mean average precision metric.

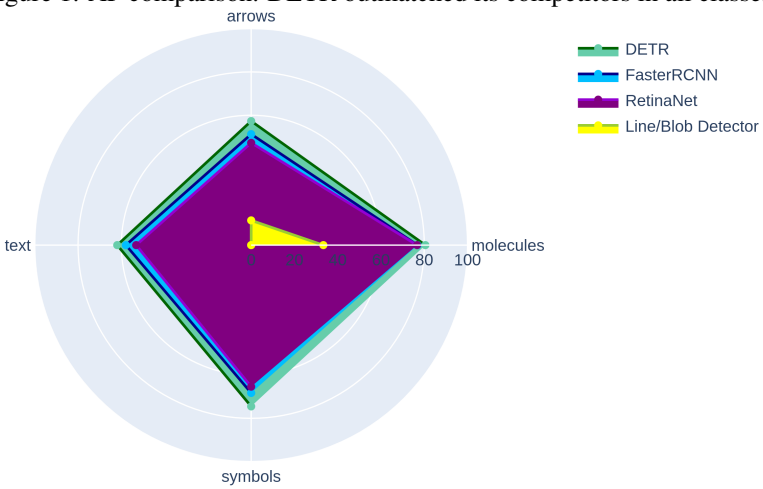
Table 1: Average Precision metrics comparison between architectures.

Model	$AP_{IoU=0.95}$	$AP_{IoU=0.75}$	$AP_{IoU=0.50}$	AP_{small}	AP_{medium}	AP_{large}
DETR	**0.678**	**0.730**	0.792	**0.507**	**0.739**	**0.559**
Faster-RCNN	0.636	0.708	0.827	0.408	0.716	0.548
RetinaNet	0.606	0.667	0.813	0.333	0.699	0.542

4.3 Quantitative Results

To evaluate the performance of the attention mechanism, neural networks kernels, and feature detectors, we computed the average precision (AP) per class (text, arrows, molecules and symbols), focusing on how well the models detect each category individually. In the appendix, we present the mean average precision (mAP) at each epoch and compare the architectures 8 as well as qualitative outputs with a confidence level of 95 % to analyze the efficacy of the aforementioned methods.

Figure 1: AP comparison. DETR outmatched its competitors in all classes.



5 Digitization

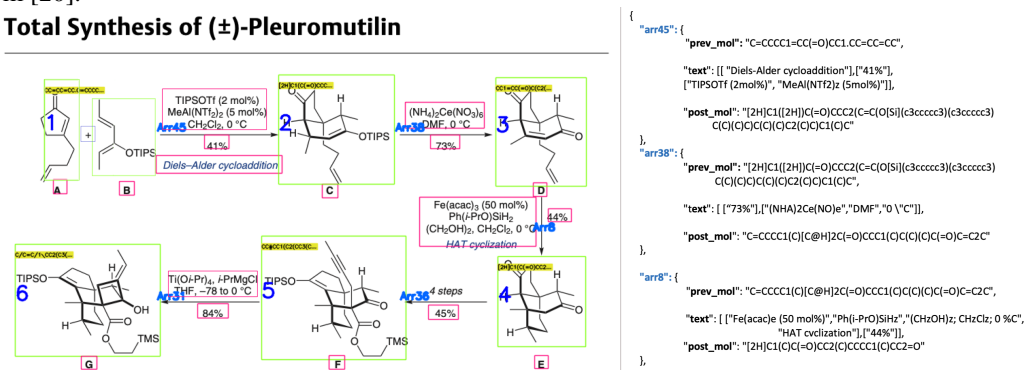
Molecule Class We used the open-access tool DECIMER[5] to convert objects identified as belonging to the molecule class into the SMILES format.

Text Class Since the segmented text is related to chemistry, we used the Text Recognition Data Generator[25] to train the document text recognition powered by Tensorflow 2 and PyTorch (DocTr)[26] and on five iterations of sample data drawn from US patents[27], relevant to the task of identifying text containing chemical information. With the newly learned weights, we were able to recognize more accurately the data belonging to the *text* label objects.

Arrow Class The reconstruction of the reaction graph was accomplished by ordering the detections according to their relative positions and distances between segmentations, which, for the majority of benchmark samples, results in the correct reaction direction.

Symbols Class In order to aggregate molecules that participate as reactants in a sub-reaction, the symbol "+" has been located and translated to a SMILES format ".", concatenating the SMILES structures of the corresponding molecules.

Figure 2: Digitization of a shortened chemical reaction with the presented baseline. Image extracted from [20].



6 Conclusions

While many high-quality experiments have been published on object detection, most of them dealing with the recognition of commonly-seen objects, very few have dealt with the recognition of chemical reaction schema. We presented a machine learning approach that provides high performance in automatically recognizing the most common elements used in the depiction of chemical reaction schema. We compared well-established two-stage (Faster Region-Based CNN) and one-stage (RetinaNet) object detectors against the DEtection TRansformer, which outperformed its competitors, scoring a total mAP of 67.8% with a 95% IoU threshold. In addition, DETR's inference time was half that of CNNs due to the elimination of post-rectification steps for duplicated predictions using the bipartite matching technique. Since all models were fine-tuned with an artificial data set but evaluated with public images, DETR also generalized the learning more accurately. Finally, we developed a completed baseline that converts imagery of chemical reaction scheme from literature in a machine-readable representation. The architecture takes advantage of an OCR model fine-tuned with chemical reaction-related textual information to deal with text objects and the DECIMER open-source tool to translate molecule objects into SMILES. Our method enables digital tools for on-demand data set creation, thereby supporting chemistry-related machine learning tasks to use reaction schema not yet available in commercial/public data sets. Given recent advances in object detectors training speed, detection accuracy, and segmentation, the extraction of data from graphical chemical reaction data has plenty of room to grow.

Acknowledgments and Disclosure of Funding

This publication was created as part of NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation.

References

[1] Rajan Kohulan, Otto Brinkhaus Henning, Zielesny Achim, and Steinbeck Christoph. A review of optical chemical structure recognition tools. 2020.

[2] Matthew C. Swain and Jacqueline M. Cole. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *Journal of Chemical Information and Modeling*, 56(10):1894–1904, 2016.

[3] Edward J. Beard and Jacqueline M. Cole. ChemSchematicResolver: A Toolkit to Decode 2D Chemical Diagrams with Labels and R-Groups into Annotated Chemical Named Entities. *Journal of Chemical Information and Modeling*, 60(4):2059–2072, 2020.

[4] Martin Krallinger, Obdulia Rabal, Anália Lourenço, Julen Oyarzabal, and Alfonso Valencia. Information Retrieval and Text Mining Technologies for Chemistry. *Chemical Reviews*, 117(12):7673–7761, 2017.

[5] Rajan Kohulan, Zielensny Achim, and Steinbeck Christoph. Decimer: towards deep learning for chemical image recognition. *Journal of Cheminformatics and 12 and 65*, 2020.

[6] Djork-Arné Clevert, Tuan Le, Robin Winter, and Floriane Montanari. Img2Mol – accurate SMILES recognition from molecular graphical depictions. *Chemical Science*, 12(42):14174–14181, 2021.

[7] Peryea T, Katzel D, Zhao T, Southall N, and Nguyen D-T. Molvec: Open source library for chemical structure recognition. 2019.

[8] Martijn Oldenhof, Adam Arany, Yves Moreau, and Jaak Simm. ChemGrapher: Optical graph recognition of chemical compounds by deep learning. *Journal of Chemical Information and Modeling*, 60(10):4506–4517, sep 2020.

[9] Staker J, Marshall K, Abel R, and McQuaw CM. Molecular structure extraction from documents using deep learning. Feb 2019.

[10] David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28(1):31–36, 1988.

[11] Stephen R Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics*, 7(1):23, 2015.

[12] Volker D. Hähnke, Sunghwan Kim, and Evan E. Bolton. Pubchem chemical structure standardization. 2018.

[13] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review, 2018.

[14] Ross Girshick. Fast-rcnn. 2015.

[15] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified and real-time object detection. 2015.

[16] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2017.

[17] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. 2020.

[18] H. W. Kuhn. The hungarian method for the assignment problem. 1955.

[19] Cheng T Kim S, Chen J. Pubchem in 2021: new data content and improved web interfaces. 2021.

[20] Thieme. Thieme-chemistry. https://www.thieme.de/en/thieme-chemistry/home-51399.htm. Accessed: 2022-02-10.

[21] Piotr Skalski. Make Sense. https://github.com/SkalskiP/make-sense/, 2019.

[22] Wu Yuxin, Kirillov Alexander, Massa Francisco, Lo Wan-Yen, and Girshick Ross. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017.

[25] Belval Edouard. Text recognition data generator. https://github.com/Belval/TextRecognitionDataGenerator, 2020.

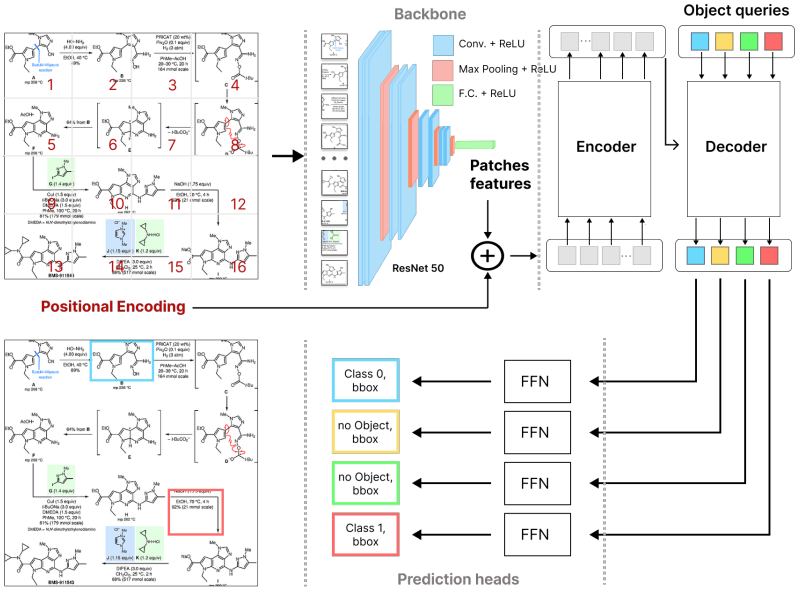
[26] Mindee. doctr: Document text recognition. https://github.com/mindee/doctr, 2021.

[27] Daniel Lowe. Chemical reactions from us patents (1976-sep2016), 2017.

A Appendix

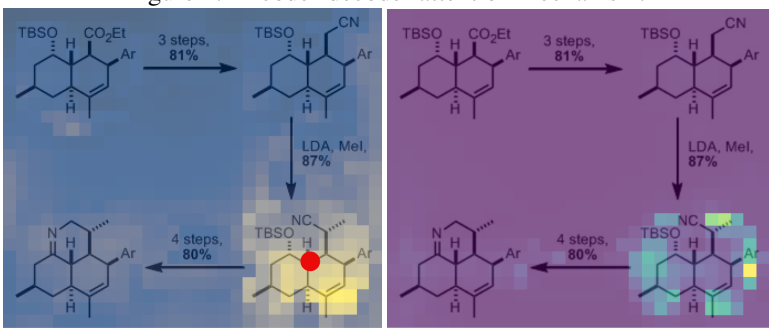
A.1 Approach - DETR

Figure 3: A sketch of the DETR architecture.



Encoder-decoder transformer To facilitate the understanding of the attention mechanism that backs up the encoder-decoder transformer, we provide an attention plot of a test sample.

Figure 4: Encoder-decoder attention mechanism.



To condition the object queries, the multi-head attention structure in the encoder provides, for each value of the input vector (represented as the red dot in the image), a measured correlation plotted as yellow, to inform to which other values does the red point attend. The intensity of the yellow indicates the strength of the correlation. With the aside information from the encoder and the learned and conditioned object queries, the decoder shares with the prediction heads enough information to understand which pixels belong to the object that is being evaluated. A *molecule* object class in this case.

Prediction heads This block predicts the normalized center point of the anticipated bounding box by using a 3-layer perceptron with ReLU as the activation function. The computed output w.r.t. the input image not only contains the central coordinates of the predicted box, but also the height, width, and a class label. Given that N , denoted as the fixed number of maximum possible true positive objects in the image, is usually higher than the actual amount of real true positives in the image, the class label accepts also $\emptyset$ as background object.

A.2 Extended results

Figure 5: Total Synthesis of (+)-Ribostamycin. Image extracted from Thieme [20]

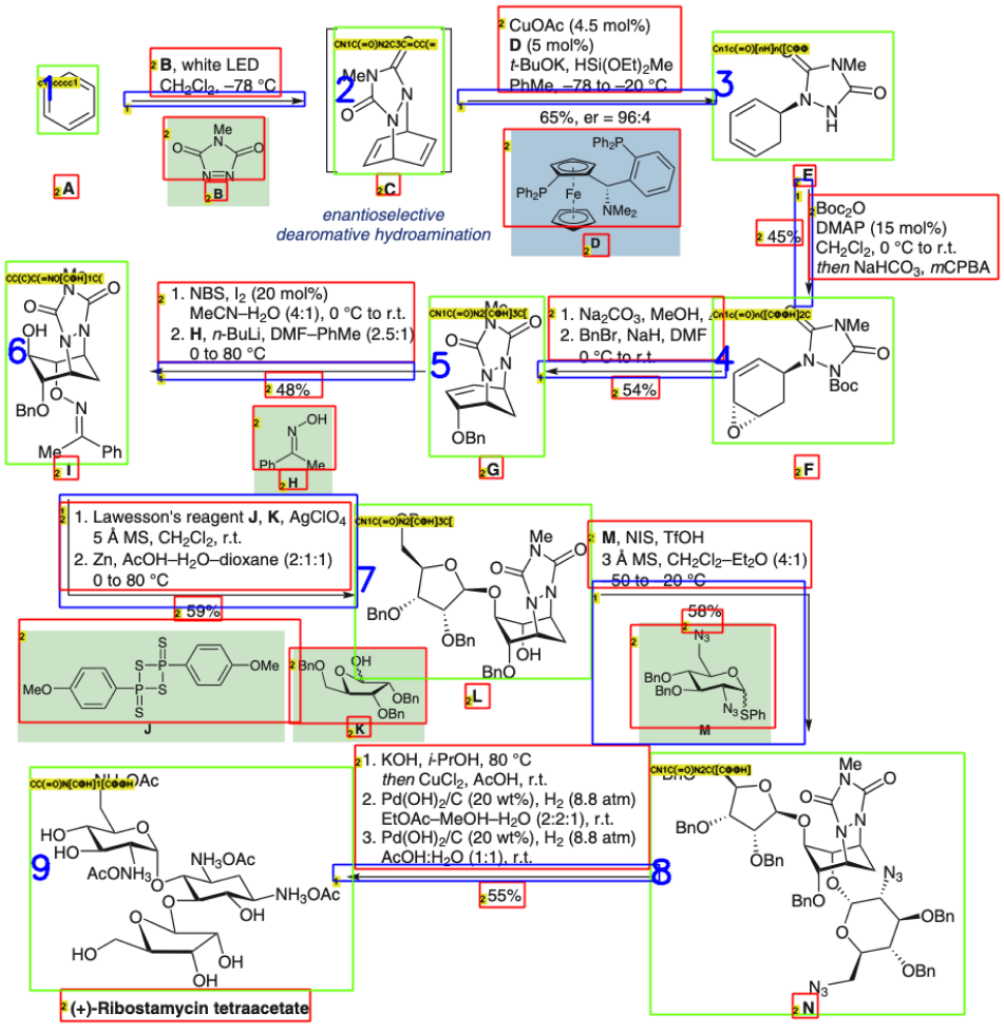
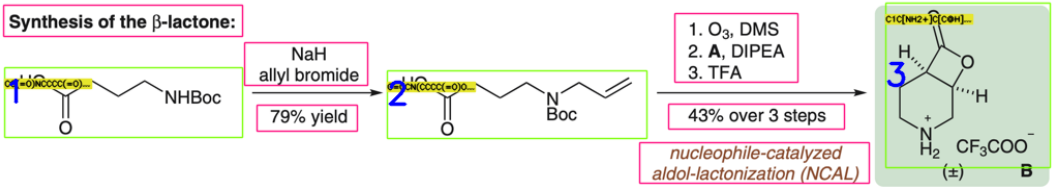


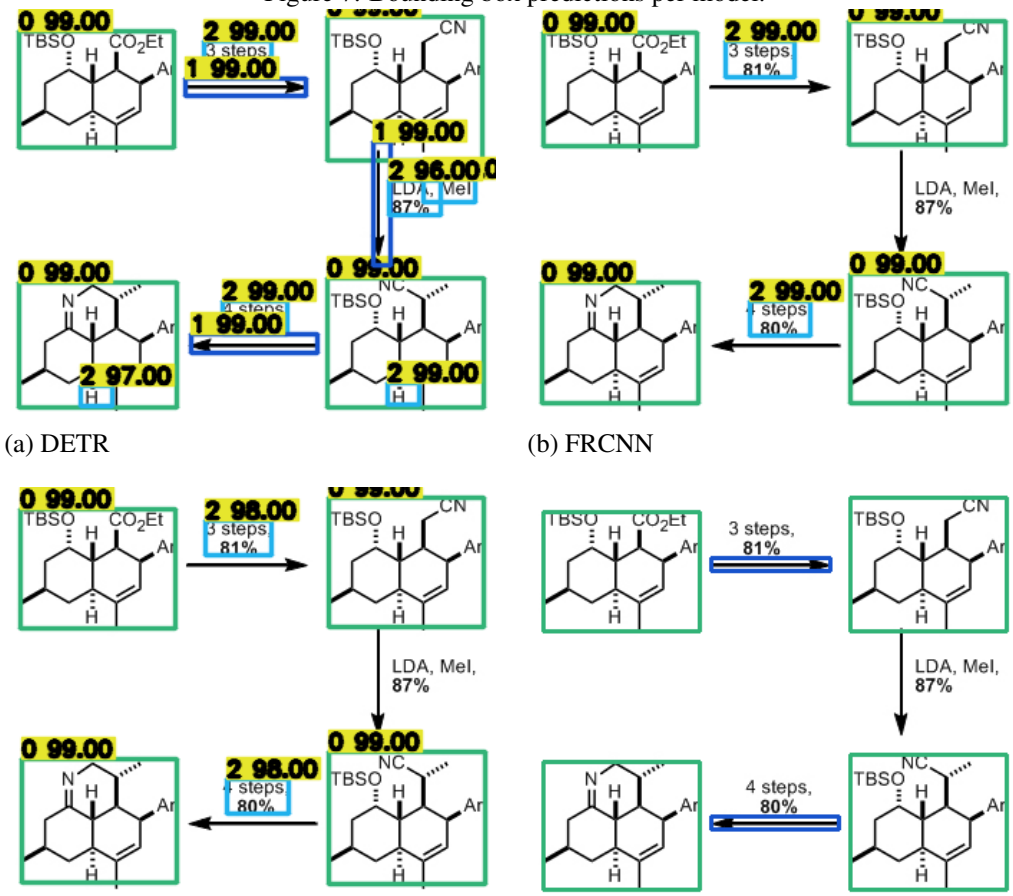
Figure 6: Synthesis of the β-lactone. Image extracted from Thieme[20]



A.3 Qualitative comparison between models

To visually assess the performance of the aforementioned approaches, the predicted results can be shown if the bounding boxes are drawn on the test images. We provide a sample of the qualitative outputs 7 for the DETR, Faster Region-Based CNN, RetinaNet with a 95% confidence classification, and the Line-Blob detector. Bounding box predictions per model of the reaction $neosymbioimine_4$ were extracted from Thieme[20].

Figure 7: Bounding box predictions per model.



(a) DETR (b) FRCNN

(c) RetinaNet (d) Line-Blob Detector

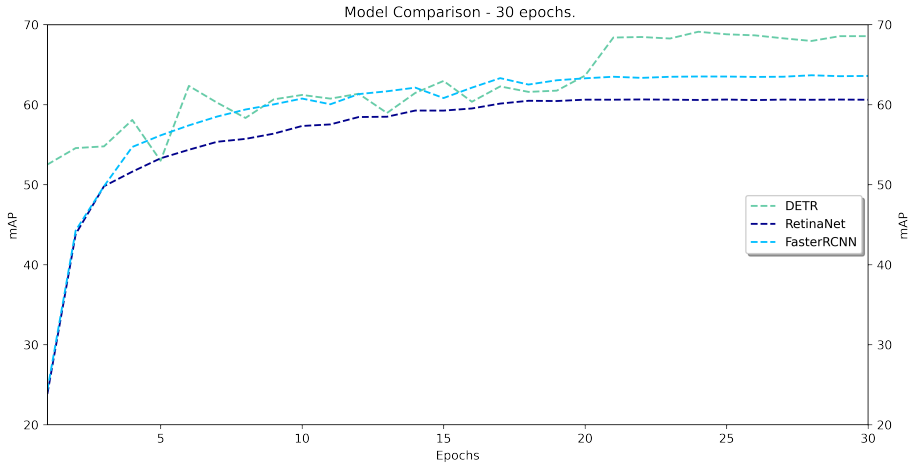
Table 2: Inference time for Figure 7

	DETR	**FRCNN**	**RetinaNet**	**Line-Blob Detector**
Inference Time (seconds)	0.103938	0.326384	0.170654	$\simeq 25$

A.4 Quantitative results - mAP per epoch between models

To further analyze the progression of the mean average precision metrics in the object detectors architectures throughout the training, we provide in Figure 8 a visualization of the mAP scores per epoch.

Figure 8: mAP comparison. The 3 compared architectures commence with an accuracy score higher than 25 since the initial weights are pre-trained with COCO. While both FRCNN and RetinaNet maintain a steep progression in the first 15 epochs, showing RetinaNet slightly fewer accuracy than FRCNN, they reach a plateau for the rest of the training schedule. Instead, DETR's accuracy strongly oscillates through the starting epochs and becomes more stabilized until epoch 22. Then, a 10 mAP score increase suddenly occurs when the learning rate is dropped from 1e-4 to 1e-5 leading to a saturated 67.8 mAP. We hypothesize that starting the experiments with a lower learning rate, (i.e. 1e-5) would avoid falling in a plausible local minimum (given the appreciated oscillations) and would accelerate the learning with better accuracy at earlier epochs.



DETR scores 67.8% mAP, after 30 epochs of training, leading its competitors performance by 4% mAP.

A.5 Traning schedules

Table 3: Detailed summary of the fine-tuning for each architecture

Architecture	Epochs	GPUs	#Params	Batch Size	Learning Rate (lr)	lr Drop
DETR	30	2	41M	8	1e-4	epoch 22
FasterRCNN	30	2	42M	8	1e-4	epoch 22
RetinaNet	30	2	36M	8	1e-4	epoch 22

A.6 Generalized intersection over union - GIoU

Metric used to threshold correct and wrong predictions against ground truth coordinates.

$$GIoU = \frac{Area\ of\ Intersection}{Area\ of\ Union} - \frac{Convex\ Hull - (Area\ of\ Union - Area\ of\ Intersection)}{Convex\ Hull}. \quad (1)$$