Terray Therapeutics • 2026
Source: https://www.terraytx.ai/news-and-research/enhanced-parsing-of-chemical-reaction-diagrams-with-Nvidia-Nemotron

Enhanced Parsing of Chemical Reaction Diagrams

Fine-tuning NVIDIA Nemotron 3.5 Super VL and NVIDIA Nemotron 3 Nano Omni to achieve state-of-the-art performance for reaction-diagram parsing.

Oct 15, 2026
Authors
Ben Kaufman
Owen Howell
Adrian Layer
Narbe Mardirossian, Ph.D.

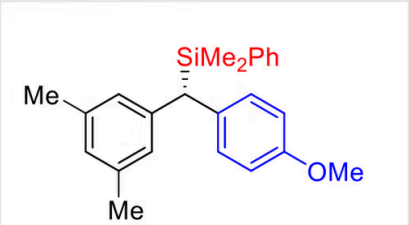
Extracting synthetic routes directly from chemical literature is a critical unlock for modern drug discovery workflows. This capability not only empowers LLMs to dynamically parse and reason over chemical documents on-the-fly, but also generates the vital ground-truth data required to continuously improve the route quality of synthesis planners like TerraSynth via reinforcement learning and other post-training approaches. At Terray Therapeutics, we've spent years building the EMMI (Experimentation Meets Machine Intelligence) platform -- including purpose-built AI models like COATI, TerraSynth, and TerraBind -- with the ultimate goal of optimizing every decision in the DMTA cycle to rapidly advance molecules toward the clinic. To orchestrate this ecosystem, we recently built a dedicated drug discovery agent that utilizes these proprietary tools daily to help medicinal chemists accelerate our pipeline programs. Unlocking the full potential of agents and models like TerraSynth ultimately relies on access to high-quality chemical data, which is why we partnered with NVIDIA to fine-tune Nemotron 3.5 Super VL and Nemotron 3 Nano Omni to achieve best-in-class parsing of reaction diagrams.

Building off the modeling approach established by RxnCaption, we show in this collaboration state-of-the-art performance on the Reaction Diagram Parsing task by fine-tuning NVIDIA Nemotron Omni, improving the previous state of the art HybridMatch F1 from 59.8 to 63.4 on RxnCaption's own benchmark test set, and clearing every previously published number across sets it was benchmarked against. We also report an early result from NVIDIA Nemotron 3.5 Super-VL, where a single run beats the matched Omni arm across the benchmarks. Additionally, we discuss data augmentations for improving performance in practical settings. Finally, we highlight how these models fit into end-to-end processing of chemical data.

Intro: The Challenge of Chemical Representations

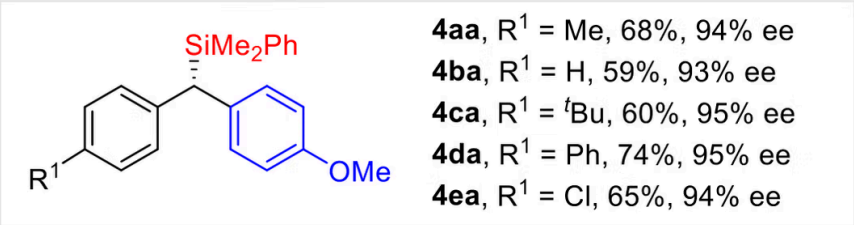
Parsing and reasoning in chemistry pose a variety of challenges due to the multitude of unique text and visual representations used to depict molecules, as shown in Figure 1.

VISUAL



Fully specified

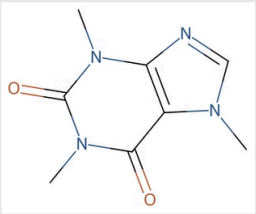
every atom and substituent is drawn



Markush scaffold

R¹ is a placeholder; the drawing stands for all five molecules

TEXTUAL



caffeine ($C_8H_{10}N_4O_2$)

SMILES `Cn1c(=O)c2c(ncn2C)n(C)c1=O`

IUPAC `1,3,7-trimethylpurine-2,6-dione`

SMARTS `[#7]1~[#6](=[#8])~[#7]~[#6]2~[#7]~[#6]~[#7]~[#6]~2~[#6]~1=[#8]`
a query: matches the whole xanthine family

InChI `InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3`

Figure 1. Both visual drawings come from a single diagram in the RxnCaption-15k corpus. A Markush scaffold is not an underspecified molecule but a deliberate statement about a family, so reading the picture alone is not enough. The bottom portion of the figure shows different textual representations of or related to caffeine.

On top of this difficulty, there is the additional complexity of representing synthetic routes, which are usually depicted as visual graphs detailing a series of reactions, shown in Figure 2.

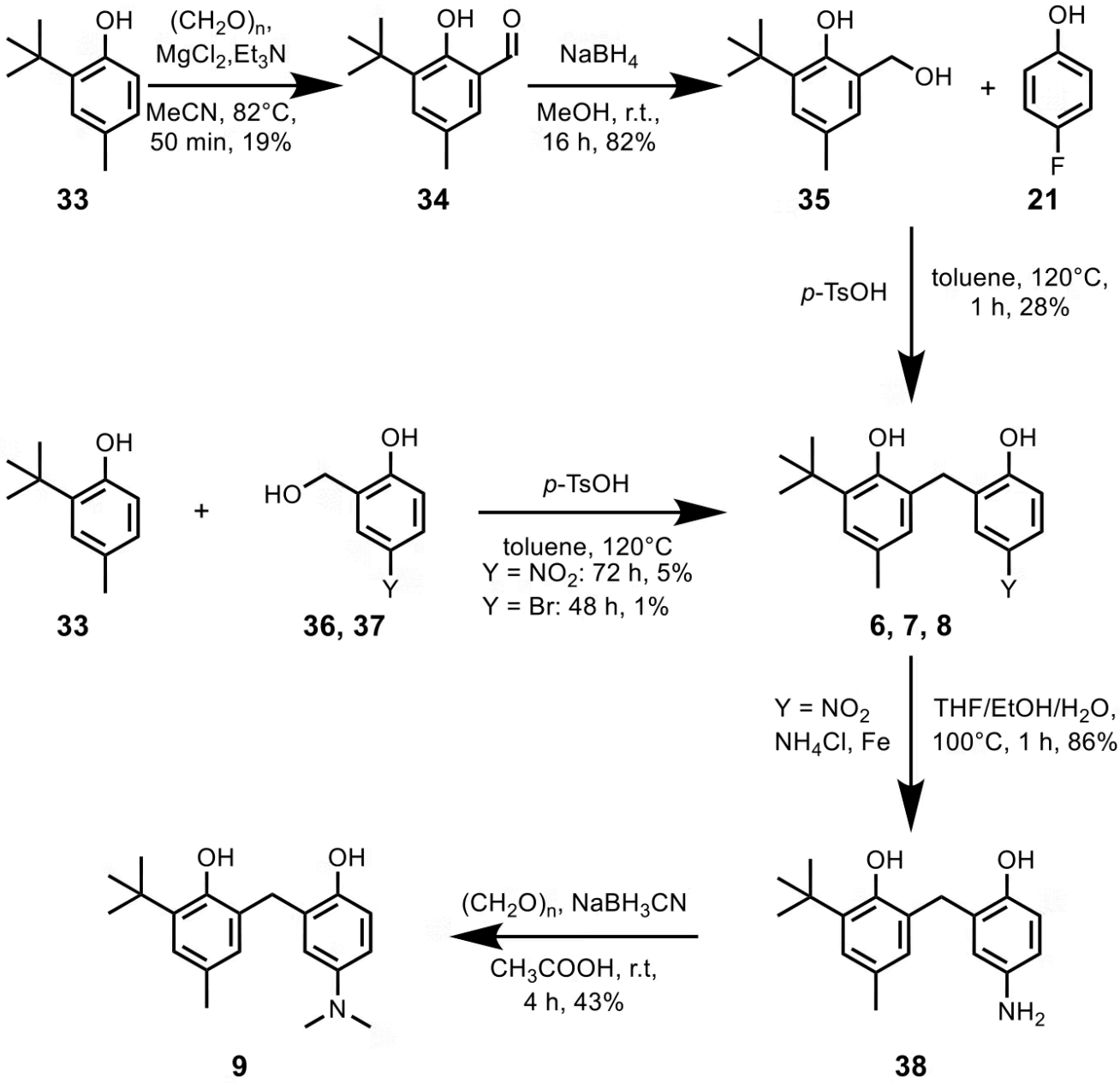


Figure 2. A reaction diagram from the RxnCaption-15k corpus. Six reactions, nine structures, two routes converging on one intermediate, and every condition written as free text beside an arrow.

While headway has been made in synthesis-aware models (read about Terray's state-of-the-art synthesis model, TerraSynth), holistic coverage will only be obtained if these synthesis diagrams can be parsed effectively.

Setup: Fine-tuning Nemotron

Earlier reaction-diagram parsers (RxnScribe, RxnIM) make the model emit coordinates. The output sequence contains a bounding box for every molecule it refers to, so localization and reaction-graph construction are learned by the same head. VLMs are weak at regressing coordinates, and an error there corrupts everything downstream.

The previous state of the art came from RxnCaption, which argued that this pairing does not play to a VLM's strengths, and supported it by showing zero-shot gains for frontier models on diagram parsing once an upstream detector supplied coordinates. The RxnCaption setup externalizes detection by having a dedicated molecular detector run first, drawing its boxes onto the diagram numbered in reading order, and training the model on the reaction parsing task given that input. The VLM never predicts a coordinate. It captions the diagram, referring to molecules by the numbers it can see, and spatial parsing becomes a language task. We adopt this approach and use the same MolYOLO detector released with RxnCaption throughout.

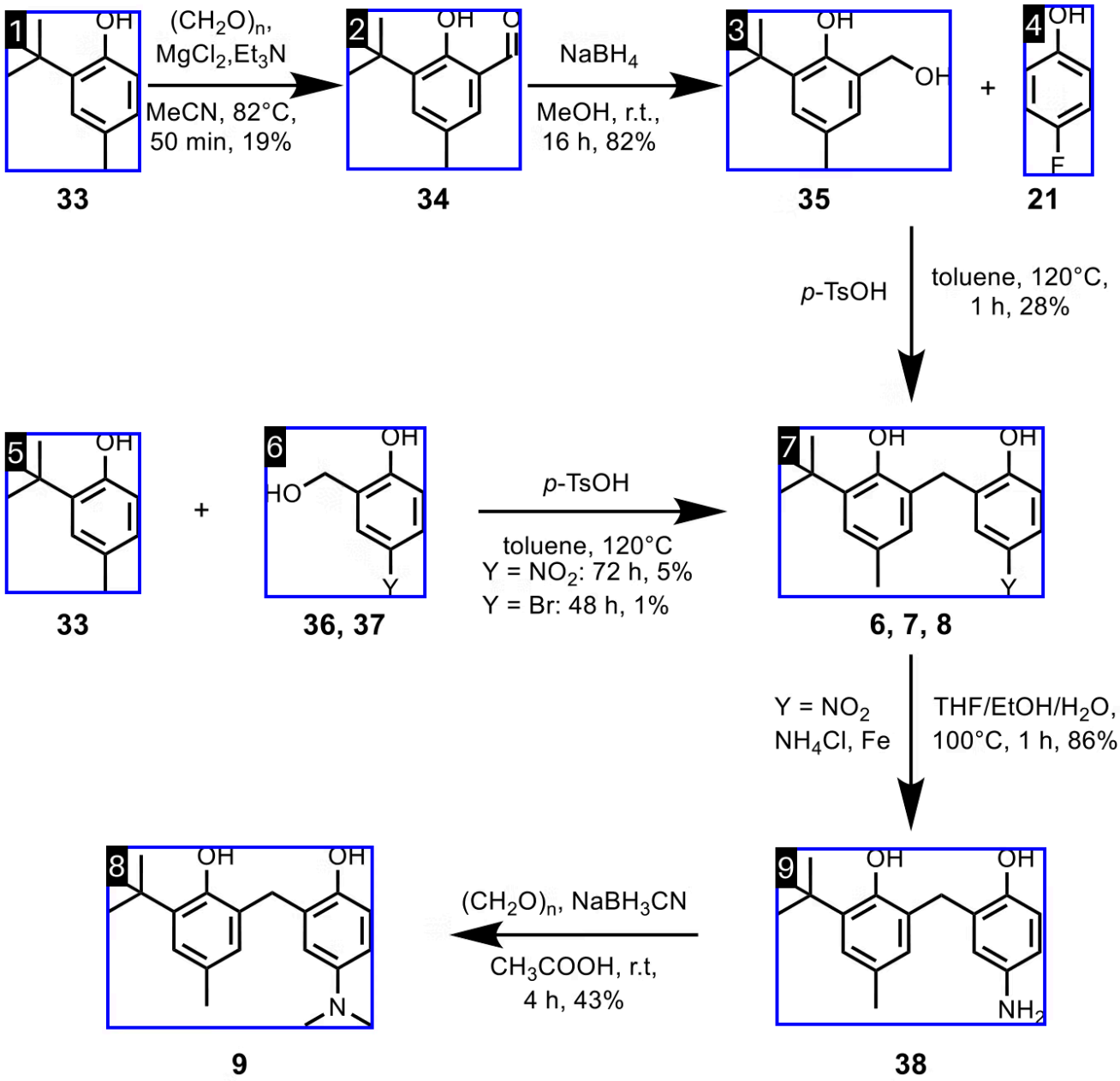


Figure 3. The diagram from Figure 2, after the detector. Every structure is boxed and numbered in reading order.

The target is a JSON list of reactions. Each reaction contains reactants, conditions, and products, whose elements may be either text transcribed directly from the diagram (e.g. “MeOH”) or numeric indices referring to labeled molecular structures in the diagram.

```
[{"reactants":  [{"structure": 1}],
  "conditions": [{"text": "(CH₂O)n, MgCl₂, Et₃N"},
                 {"text": "MeCN, 82 °C, 50 min, 19%"}],
  "products":   [{"structure": 2}]}, …]
```

Many reaction diagrams have molecules that are not part of the reaction, which the upstream detector should also capture. We follow RxnCaption's setup, where the training data consist of ground-truth geometry for the molecules in a reaction plus the MolYOLO detections for the remainder of the diagram. For evaluation, we focus mainly on the practical setup, where the eval set is labeled only with MolYOLO detections, but we also present some numbers with oracle boxes for the ground truth. This separates practical performance today from what it would be if every reaction component were identified by the detector.

We train on their dataset, RxnCaption-15k, containing 15,000 real reaction diagrams, and apply the same upweightings the original paper uses for rare layouts (reactions that run right-to-left, reactions that run bottom-to-top, and reversible pairs), duplicating a figure once for each awkward property it carries.

The paper trains on a second corpus as well: the 1,240 figures of the RxnScribe training split, an older set drawn from a different slice of the literature. We triple it as the paper does, to 3,720 figures, and apply the same rare-layout upweighting used on the RxnCaption-15k set. We ablate this corpus in and out rather than fixing it. The training arms without it never see a RxnScribe diagram, so their scores on the RxnScribe test split are a zero-shot read.

Everything to this point follows the RxnCaption paper, on Nemotron models. Our own addition is a synthetic data generation process for reaction diagrams.

Real, labeled figures are scarce, and their layouts repeat. A synthetic data generator is appealing because it can scale and diversify the data at a multiple of whatever labeling goes into building the generator. That should make pre- or post-training data easier to scale, and performance with it. We label the components of 413 figures from the RxnCaption-15k training split, keeping each one's *geometry* as a template (arrow endpoints and waypoints, text zones with roles, Markush regions, dividers), then extract a connected reaction subgraph, compose two or three subgraphs from *different* source figures onto one canvas, renumber every box, and render. Molecule slots are filled from real multi-step Pistachio routes, so each region is chemically coherent, and rendering goes through ChemSema, an open-source engine that follows ChemDraw's depiction conventions rather than RDKit's. There is additional noise that allows molecules, bounding boxes and other figure components to be dropped out. This, along with bond-length and text-size jitter, results in additional layouts and topologies to expand the dataset. In all we generated 5,000 synthetic diagrams for training.

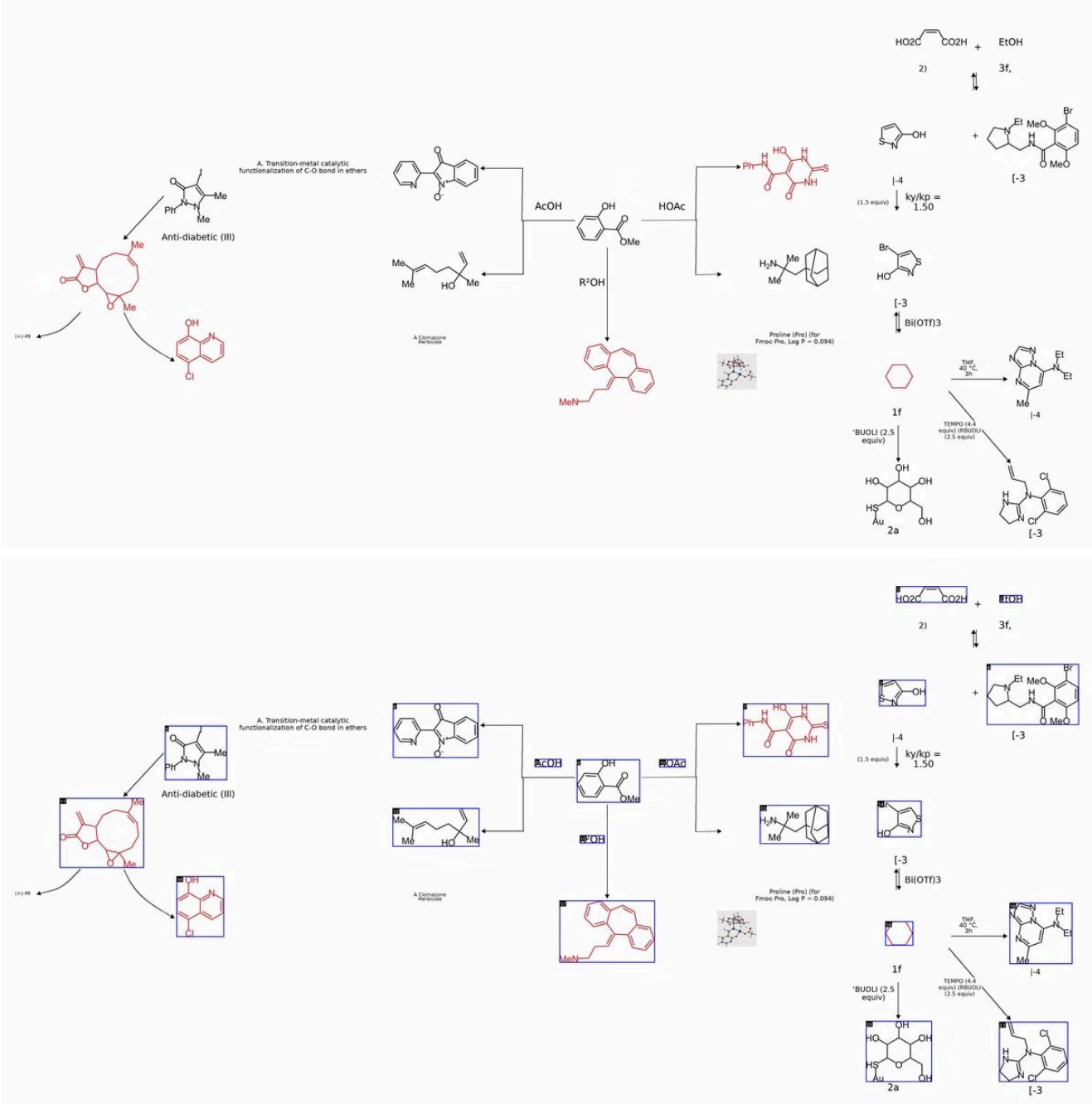


Figure 4. One spliced synthetic diagram. Three subgraphs from three different source figures share this canvas; the gray raster patch near the center is a filler crop of real pixels cut from a source figure.

Nemotron Omni is a sparse mixture of experts, 30B parameters in total with about 3B active per token, and we train it on a single 8 × NVIDIA H100 node. All runs are a full fine-tune, trained with NVIDIA NeMo AutoModel using FSDP2 and bf16 at sequence length 8192, for one epoch, and we take the final checkpoint. Hyperparameters are held constant with cosine decay to zero after 5% linear warmup, AdamW (β 0.9/0.95, weight decay 0.01) and global batch 8. We ablate freezing the RADIO vision tower, the vision transformer that turns diagram pixels into tokens the language model can read, to get a sense of how well Nemotron Omni's base visual features transfer to a new task without tuning. Every dataset ablation is reported with the tower frozen and unfrozen. The trainable vision tower runs use a lower learning rate, 5e-5, as 1e-4 was too hot for the pretrained visual features; the frozen runs keep 1e-4. Every Omni configuration was trained four times with different seeds, and every number we report is the mean across them.

After making substantial progress with Nemotron 3 Nano Omni, we received early access to Nemotron 3.5 Super-VL and extended our approach, building on NVIDIA's post-training recipe, obtaining promising initial results. The configuration carries over almost unchanged: the w/ RxnScribe mix for one epoch, NeMo AutoModel with FSDP2, bf16 at sequence length 8192, global batch 8, and AdamW with the same betas, weight decay and warmup schedule. The learning rate drops to 2.5e-5, we run it with the RADIO vision tower frozen only, and at four times the parameter count it has to be configured to train across four 8 × H100 nodes.

Results: Achieving State-of-the-Art Extraction

We evaluate on the 400 held-out diagrams and roughly 2,800 annotated reactions of the RxnCaption-15k test set, using the authors' own scoring code and the two metrics they report. **SoftMatch** counts a predicted reaction as correct when its molecules match the ones annotated, ignoring the text on the diagram entirely. **HybridMatch** keeps that requirement and adds the text: the reagents, solvents, temperatures and yields beside the arrows, plus any reactant or product written as a name rather than drawn, compared with the tolerance their scorer allows. Both are reported on two box sets: **detector boxes**, MolYOLO's real output and the deployment setting, and **oracle boxes**, which measure the parser when no ground truth boxes are missing. HybridMatch is the more reflective of practical use, since the conditions and reagents are crucial to actually running the chemistry, so it is what the figures below report; SoftMatch is left to the appendix table.

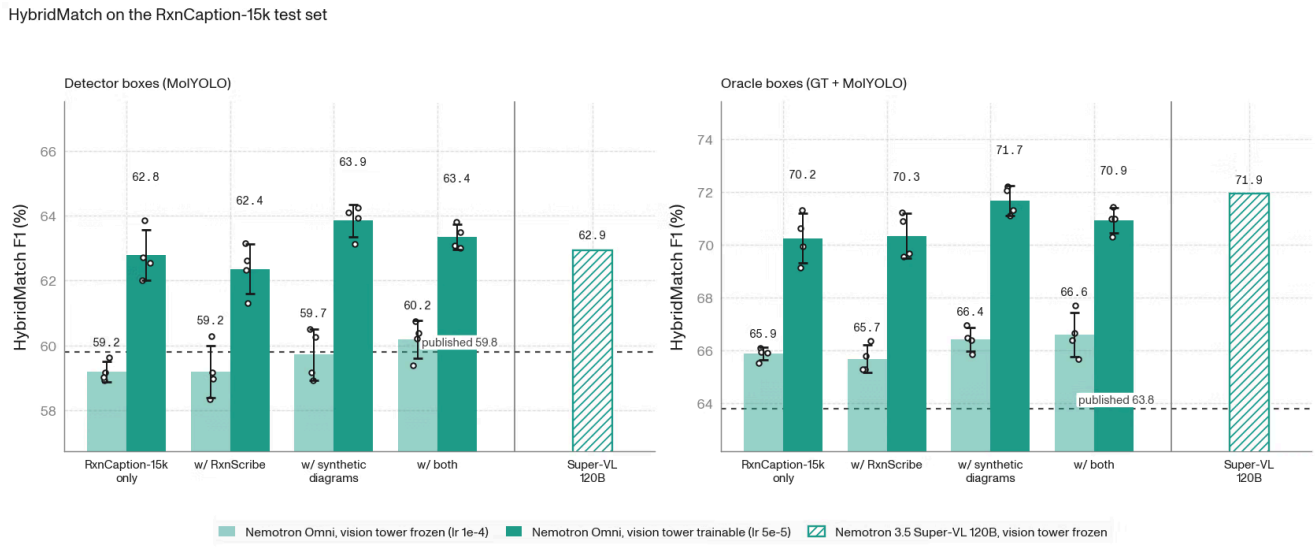


Figure 5. Nemotron Omni training ablations and resulting HybridMatch scores. Unfreezing the vision tower has the largest positive impact across Nemotron Omni ablations, and all ablations with that setting surpass RxnCaption-VL's previous state of the art on this test set. Adding synthetic diagrams also boosts model performance. The hatched bar past the divider is the readout for Nemotron 3.5 Super-VL 120B, which was trained on the w/ RxnScribe setup and outperforms the Nemotron Omni equivalent despite having a frozen vision tower. For the Nemotron Omni ablations bars are means over four seeds, error bars one standard deviation, open circles individual runs.

We achieve state-of-the-art HybridMatch performance with several of the dataset setups. The w/ RxnScribe arms replicate the dataset used to train RxnCaption-VL, and on detector boxes they clear the published numbers by +2.6 on the RxnCaption-15k test set and +1.8 on RxnScribe's when the vision tower is trainable, with wider margins on oracle boxes. Frozen vision tower setups perform

worse across the board than their matched trainable ones, yet the frozen w/ RxnScribe arm still lands roughly level with RxnCaption-VL, which suggests Nemotron Omni's pretrained representations are already reasonably suited to the task. Additionally, the arms that didn't train on RxnScribe data show strong zero-shot performance and come within 1.3 of RxnCaption-VL's HybridMatch F1 score (Figure 6).

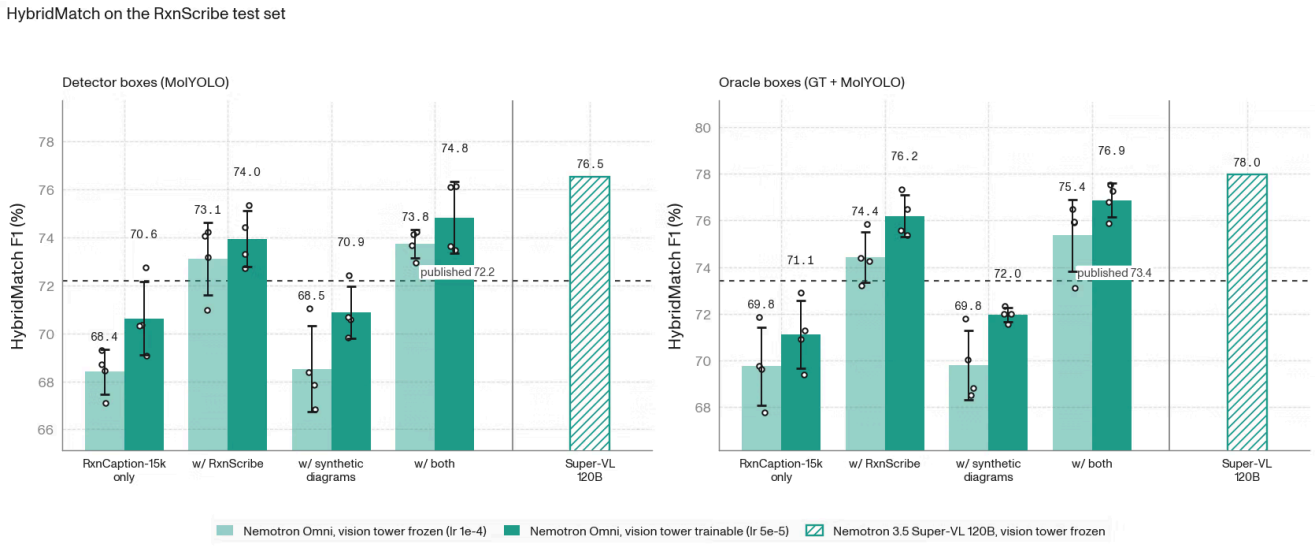


Figure 6. The same ablations scored on the RxnScribe test set, 138 diagrams drawn from a different slice of the literature. Every arm trained with RxnScribe data clears its published state of the art (dashed line). The arms that never saw a RxnScribe diagram land just below it, within 1.3 points at best, as a zero-shot read on this set. The hatched bar past the divider is the readout for Nemotron 3.5 Super-VL 120B, which was trained on the w/ RxnScribe setup and outperforms the Nemotron Omni equivalent here too, despite having a frozen vision tower. For the Nemotron Omni ablations bars are means over four seeds, error bars one standard deviation, open circles individual runs.

Moving from detector boxes to oracle boxes gains Nemotron Omni 6-8 HybridMatch F1 points on the RxnCaption-15k test set and 0.5-2.2 for RxnScribe. For both sets this gain is close to the maximum gain if the detector didn’t miss any ground truth boxes.

Synthetic diagrams derived from the RxnCaption training data give a small +1.1 over the model trained without them, and a further +1.0 when added on top of RxnScribe, each positive in all four seed pairs. That second gain takes the arm trained on both corpora to 63.4 HybridMatch, +3.6 over the published 59.8. The diagrams also trend positive against the equivalent setups without them on the RxnScribe test set, though less consistently. Increasing the number of synthetic diagrams beyond 5000 did not scale performance in any cell we measured. Increasing the number of base templates, to broaden the diversity of what the generator can produce, is another axis worth exploring, and we leave that for future work.

On SoftMatch (appendix table) every ablation outperforms RxnCaption-VL on the RxnCaption-15k test set, with or without a trainable vision tower. On the RxnScribe test set only the trainable arms with RxnScribe data do better on detector boxes.

To see where an off-the-shelf model lands on the same task, we ran four frontier models zero-shot with the paper's own prompt. Those runs are on detector boxes only, since that is the setting a practical pipeline would use. We also scored the pre-fine-tune Nemotron models as a baseline for the

impact of the fine-tune. The figure below sets all of them against our two trained models and the published specialist systems.

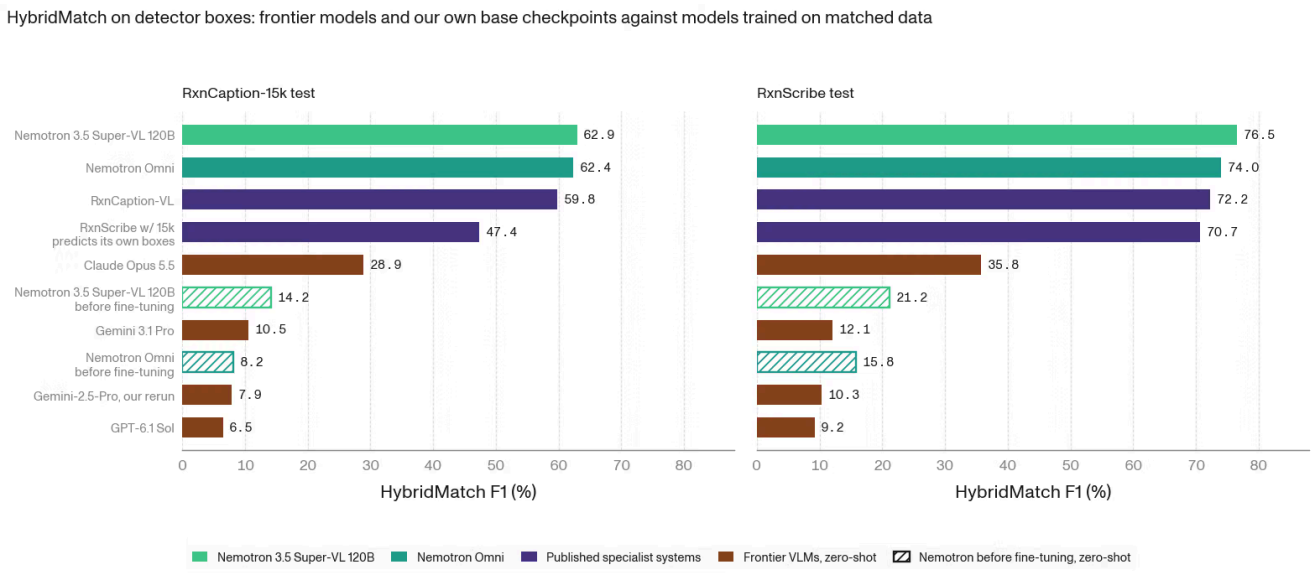


Figure 7. Both of our models trained on the same dataset as the published specialist systems. Nemotron 3.5 Super-VL leads on both test sets with its vision tower frozen, and the matched Omni arm with a trainable tower is next; both clear RxnCaption-VL's previous state of the art. The four frontier models were run zero-shot and fall well short on this measure. The two hatched bars are the two Nemotron models before any fine-tuning, a baseline showing that the pickup on these tasks comes from the fine-tune.

Gemini 3.1 Pro and Claude Opus 5.5 are level at the top on SoftMatch over the RxnCaption-15k detector boxes, and all the latest frontier models have strong zero-shot SoftMatch (appendix table). On HybridMatch they struggle. One of those four is the paper's own Gemini-2.5-Pro baseline, rerun here with its prompt and scorer. That rerun is within a few points of the published SoftMatch on both test sets, but not of the published HybridMatch. We have not established why, and the appendix table carries both rows.

Looking at the frontier-model outputs, they struggle to attach precise text to the right reaction and tend to add extra items, which hurts the scores. It is possible the published zero-shot rows were scored with a more permissive text match that does not penalize extra text, but neither the paper nor the released code describes one. The substantial performance boost of the fine-tuned Nemotron models over their base versions, +54.1 HybridMatch F1 for Nemotron Omni and +48.8 for Nemotron 3.5 Super-VL on the RxnCaption-15k test set, highlights the benefit of trainable open-weight models.

The initial results with Nemotron 3.5 Super-VL are strong. On the same w/ RxnScribe mix it outperforms the published RxnCaption-VL numbers on both test sets and both box conditions: +3.1 and +4.3 HybridMatch F1 on detector boxes, and +8.1 and +4.6 on oracle boxes (Figure 5 and Figure 6). It clears the equivalent frozen Omni setup by +3.4 to +6.2, and matches or exceeds the unfrozen one. It is the top bar on both panels of Figure 7. We are running more experiments to see whether an unfrozen setup gives a further pickup, as it did with Omni.

Altogether, both Nemotron models achieve state-of-the-art results on this task, on both test sets and both box conditions. Three of the Omni arms clear every HybridMatch number we compare against, and we will release the strongest of them, synthetic diagrams and RxnScribe with a trainable vision tower, on Hugging Face in the near future.

Looking Forward: Building an End-to-End Pipeline

Improved performance on reaction parsing is a key component that must be integrated with other tools for practical use. To pull all the information out of figures in the literature, the component molecules of a reaction need to be processed into a standardized format, and additional reasoning is needed to map information about the reaction that lives elsewhere in the paper.

Take this example, a single scheme from a J. Med. Chem. paper on irreversible inhibitors of the SARS-CoV-2 papain-like protease.

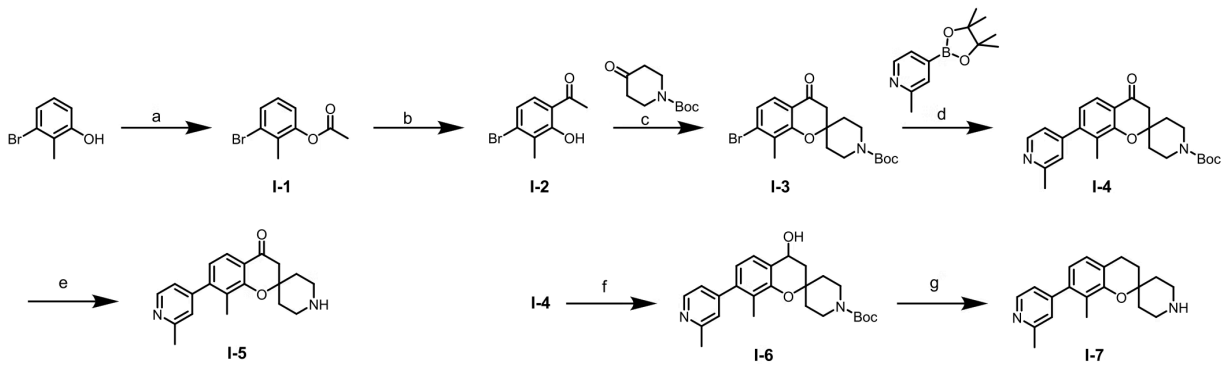


Figure 8. Scheme 1 from J. Med. Chem. 2026, 69, 3588–3608, cropped; CC BY 4.0. The arrows carry letters, not conditions: what a to g stand for is written in a footnote under the scheme, outside the page segmenter's crop. Nothing in this image says what any of these seven steps was run with, so the conditions have to be found elsewhere in the paper and matched back.

Our reaction parser model can identify the route structure, but processing the molecules and mapping labels in the diagram to other parts of the text requires additional tooling. Optical structure recognition is needed to turn the drawn molecules into SMILES, and reasoning over the rest of the paper's text is a task for an LLM. Setting this up so that it preserves precision and can reason over multi-step reactions is crucial, since even small errors can invalidate a whole reaction.

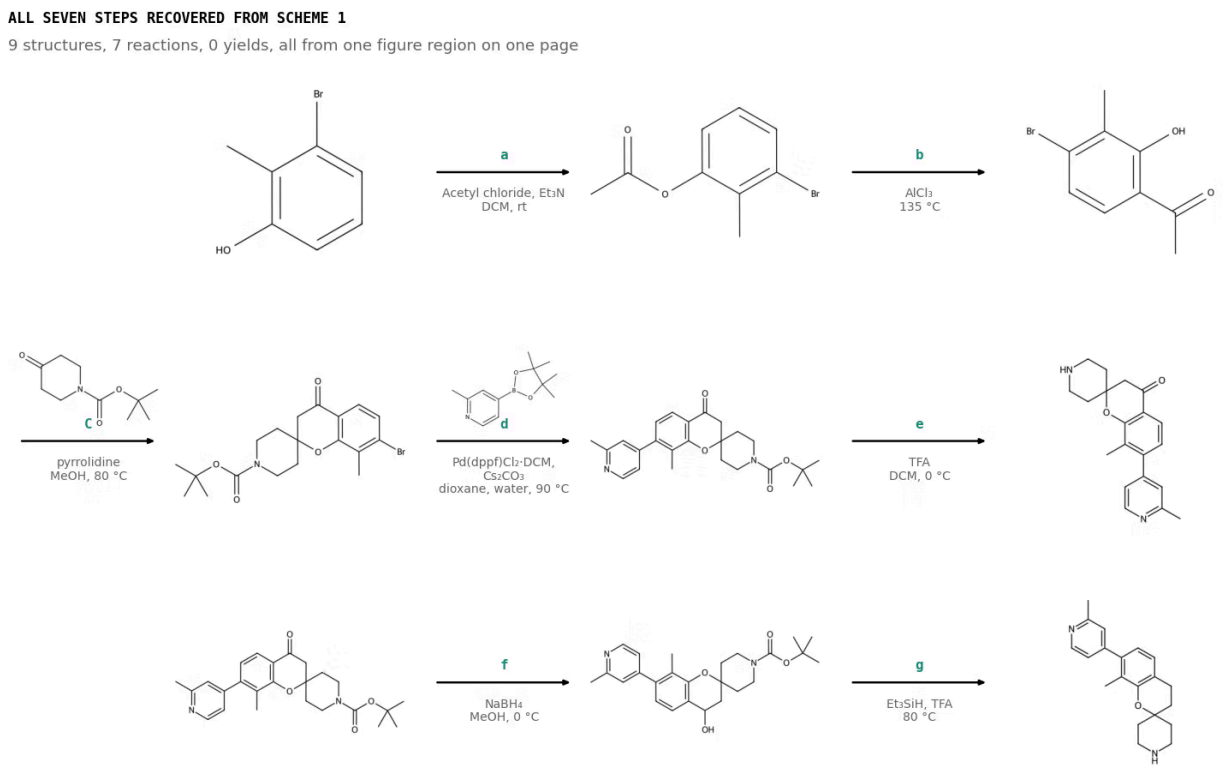


Figure 9. Results recovered from the agentic pipeline described above. Every structure here is an RDKit depiction of a SMILES string read off the page, and the arrow labels are parsed condition records; nothing in this figure is a crop of the original.

We are excited to continue building out this end-to-end pipeline—refining the parser, closing tooling gaps, and building the confidence signals that make these outputs safe to use unattended. The improvement in parsing reaction diagrams that we show here opens the door to translating published chemistry into actionable data and is a crucial step toward unlocking the full potential of synthesis planners and autonomous agents.

Appendix: The Full Grid

Every arm on both test sets, both box conditions and both metrics, with the precision and recall the figures leave out.

Model	Vision tower	Runs	SoftMatch			HybridMatch		
			P	R	F1	P	R	F1
RxnCaption-15k test, detector boxes (MolYOLO)								
Gemini-2.5-Pro (as published in the paper)	–	–	64.2	69.2	66.6	38.9	42.1	40.4
Gemini-2.5-Pro, our rerun (zero-shot, single run)	–	–	69.3	69.2	69.2	7.9	7.9	7.9
Gemini 3.1 Pro (zero-shot, single run)	–	–	**79.3**	75.5	**77.3**	10.8	10.2	10.5
Claude Opus 5.5 (zero-shot, single run)	–	–	*78.9*	*75.6*	*77.2*	29.6	28.2	28.9
GPT-6.1 Sol (zero-shot, single run)	–	–	60.8	**76.5**	67.7	5.8	7.3	6.5
RxnCaption-VL (published)	–	–	71.3	69.4	70.4	60.3	59.3	59.8
RxnScribe w/ 15k (published, predicts its own boxes)	–	–	72.1	44.7	55.2	61.2	38.7	47.4
Nemotron Omni, RxnCaption-15k only	frozen	4	72.4	71.3	71.9	59.1	59.3	59.2
Nemotron Omni, RxnCaption-15k only	trainable	4	75.4	74.1	74.7	62.9	62.7	62.8
w/ RxnScribe	frozen	4	72.4	71.2	71.8	59.2	59.2	59.2
w/ RxnScribe	trainable	4	75.3	73.9	74.6	62.6	62.1	62.4
w/ synthetic diagrams	frozen	4	72.5	71.1	71.8	59.9	59.5	59.7
w/ synthetic diagrams	trainable	4	76.6	74.2	75.4	**64.5**	**63.2**	**63.9**
w/ both	frozen	4	72.3	72.0	72.2	60.0	60.4	60.2
w/ both	trainable	4	76.0	74.3	75.1	*63.7*	*63.0*	*63.4*
Nemotron 3.5 Super-VL 120B	frozen	1	75.0	73.0	74.0	63.2	62.6	62.9
RxnCaption-15k test, oracle boxes (GT + MolYOLO)								
RxnCaption-VL (published)	–	–	75.9	76.7	76.3	63.8	63.9	63.8
Nemotron Omni, RxnCaption-15k only	frozen	4	79.7	79.6	79.6	66.0	65.8	65.9
Nemotron Omni, RxnCaption-15k only	trainable	4	83.2	83.3	83.2	70.4	70.2	70.2
w/ RxnScribe	frozen	4	79.4	79.8	79.6	65.6	65.8	65.7
w/ RxnScribe	trainable	4	83.0	84.0	83.5	70.0	70.6	70.3
w/ synthetic diagrams	frozen	4	79.7	80.9	80.3	66.0	66.8	66.4
w/ synthetic diagrams	trainable	4	**84.1**	**84.4**	**84.2**	**71.7**	**71.6**	**71.7**
w/ both	frozen	4	79.5	81.3	80.4	66.0	67.2	66.6
w/ both	trainable	4	*83.8*	*84.1*	*84.0*	*70.9*	*70.9*	*70.9*
Nemotron 3.5 Super-VL 120B	frozen	1	84.0	84.8	84.4	71.7	72.2	71.9
RxnScribe test, detector boxes (MolYOLO)								
Gemini-2.5-Pro (as published in the paper)	–	–	67.9	86.5	76.1	44.7	56.1	49.8
Gemini-2.5-Pro, our rerun (zero-shot, single run)	–	–	70.9	84.3	77.0	9.5	11.2	10.3
Gemini 3.1 Pro (zero-shot, single run)	–	–	86.1	84.5	85.3	12.2	12.0	12.1

Model	Vision tower	Runs	SoftMatch			HybridMatch		
			P	R	F1	P	R	F1
Claude Opus 5.5 (zero-shot, single run)	–	–	77.8	86.9	82.1	33.9	38.0	35.8
GPT-6.1 Sol (zero-shot, single run)	–	–	51.2	86.1	64.2	7.3	12.2	9.2
RxnCaption-VL (published)	–	–	85.3	**87.1**	86.2	71.6	72.7	72.2
RxnScribe w/ 15k (published, predicts its own boxes)	–	–	84.1	81.7	82.8	72.4	69.1	70.7
Nemotron Omni, RxnCaption-15k only	frozen	4	85.4	84.5	84.9	68.6	68.2	68.4
Nemotron Omni, RxnCaption-15k only	trainable	4	86.8	85.2	86.0	71.2	70.1	70.6
w/ RxnScribe	frozen	4	85.4	86.6	86.0	72.6	73.7	73.1
w/ RxnScribe	trainable	4	**87.0**	*87.0*	**87.0**	*73.9*	*74.0*	*74.0*
w/ synthetic diagrams	frozen	4	85.4	84.8	85.1	68.7	68.4	68.5
w/ synthetic diagrams	trainable	4	86.3	85.2	85.7	71.3	70.5	70.9
w/ both	frozen	4	85.5	86.0	85.8	73.5	74.0	73.8
w/ both	trainable	4	*86.8*	*87.0*	*86.9*	**74.7**	**74.9**	**74.8**
Nemotron 3.5 Super-VL 120B	frozen	1	88.6	88.4	88.5	76.5	76.5	76.5
RxnScribe test, oracle boxes (GT + MolYOLO)								
RxnCaption-VL (published)	–	–	87.2	88.1	87.6	73.3	73.5	73.4
Nemotron Omni, RxnCaption-15k only	frozen	4	86.4	86.3	86.3	69.8	69.8	69.8
Nemotron Omni, RxnCaption-15k only	trainable	4	88.1	87.6	87.9	71.3	71.0	71.1
w/ RxnScribe	frozen	4	86.8	88.7	87.7	73.7	75.2	74.4
w/ RxnScribe	trainable	4	**89.2**	*89.8*	**89.5**	*75.9*	*76.5*	*76.2*
w/ synthetic diagrams	frozen	4	87.6	87.7	87.6	69.8	69.8	69.8
w/ synthetic diagrams	trainable	4	88.5	87.4	88.0	72.5	71.4	72.0
w/ both	frozen	4	87.6	89.0	88.3	74.8	75.9	75.4
w/ both	trainable	4	*89.0*	**90.0**	*89.5*	**76.5**	**77.2**	**76.9**
Nemotron 3.5 Super-VL 120B	frozen	1	91.0	91.2	91.1	77.9	78.1	78.0

Table 1. Both test sets, both box sets, both metrics. RxnCaption-15k test is 400 diagrams and 2,829 reactions; RxnScribe test is 138 diagrams and 392 reactions. **Bold** is the best value in a column within a section, *italic* is second. The published rows are the numbers the paper reports, including its Gemini-2.5-Pro zero-shot baseline; the four rows below it are our own single zero-shot runs. RxnScribe emits its own coordinates, so the box conditions do not apply to it. Our rows are means over the run counts shown. The last row of each section is a single run of a different model, Nemotron 3.5 Super-VL 120B. It is ruled off and left out of the bold and italic ranking; where its run would have topped a column, the value is underlined instead.