Atom-Level Optical Chemical Structure Recognition with Limited Supervision

Martijn Oldenhof[1] Edward De Brouwer[1,2] Adam Arany[1] Yves Moreau[1]
[1]ESAT - STADIUS, KU Leuven, Belgium
[2]Yale University, USA
{martijn.oldenhof, edward.debrouwer, adam.arany, yves.moreau}@esat.kuleuven.be

Abstract

Identifying the chemical structure from a graphical representation, or image, of a molecule is a challenging pattern recognition task that would greatly benefit drug development. Yet, existing methods for chemical structure recognition do not typically generalize well, and show diminished effectiveness when confronted with domains where data is sparse, or costly to generate, such as hand-drawn molecule images. To address this limitation, we propose a new chemical structure recognition tool that delivers state-of-the-art performance and can adapt to new domains with a limited number of data samples and supervision. Unlike previous approaches, our method provides atom-level localization, and can therefore segment the image into the different atoms and bonds. Our model is the first model to perform OCSR with atom-level entity detection with only SMILES supervision. Through rigorous and extensive benchmarking, we demonstrate the preeminence of our chemical structure recognition approach in terms of data efficiency, accuracy, and atom-level entity prediction.

1. Introduction

Molecules and chemical reactions represent the tokens of the language of chemistry, which underlies applications such as drug or new materials discovery. Molecules can be represented by a molecular formula (*e.g.*, $C_8H_{10}N_4O_2$), or preferably by a more detailed structural formula—a graphical representation showcasing the spatial arrangement of atoms in the molecule. Isomers, molecules sharing the same molecular formulas but differing in spatial atom arrangement, typically exhibit distinct chemical and physical properties (as illustrated in Supplementary Material (SM) Section 8). Structural molecular formulas are thus ubiquitous in chemistry publications, lab notes, patents, or text books. This prevalence motivates the development of automatic pipelines to perform chemical structure recognition, parsing structural formulas from images. Such ability promises more efficient scientific literature browsing, automatic lab notes transcription, or chemical data mining, among others.

Recent advances in computer vision have allowed the development of several chemical structure recognition tools [5, 19, 24]. These tools can be classified into molecular graph predictions methods and atom-level entity prediction methods. Molecular graph prediction methods only use limited image annotation, such as SMILES, a serial notation of a molecule [5, 24, 33], and only predict the molecular graph. In contrast, atom-level entity prediction methods leverage richer image annotations for training the model, such as atom-level entity localization (*i.e.*, individual atoms and bonds are annotated in the original image) [19, 23]. These methods predict the molecular graph as well as the localization of the different components of the molecule in the original image. Figure 1 illustrates the different types of predictions for these two categories of models.

Previous research has shown that atom-level entity prediction methods typically enjoy better training sample efficiency, requiring less images for achieving the same level of performance [14]. This class of methods is also more interpretable. Atom-level entity annotation can indeed help identify the atoms that will be part of new chemical bond in a reaction, and can also facilitate human evaluation and correction when necessary, opening the way for synergistic human-in-the-loop training strategies [23]. Nevertheless, these advantages are compensated by the necessity to provide rich image annotation in the training data. Unfortunately, such supervision is often unavailable in many data domains, such as hand-drawn images. Yet, hand-drawn images represent a prevalent format in chemical notations and sketches. The strict dependency of existing atom-level entity prediction methods on rich image annotation thus prevents their deployment to crucial data domains.

Our research addresses these limitations by introducing a state-of-the-art chemical structure recognition tool, which (1) predicts a molecular graph from images, (2) provides atom-level localization in the original image, and (3) adapts to new data domain with a limited number of data samples and supervision. Our architecture relies on a object detection backbone coupled with a graph construction strategy

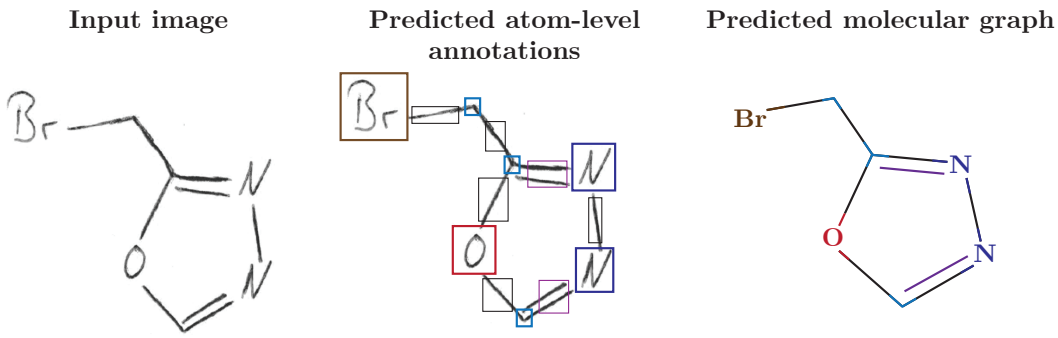


Figure 1. **Problem setup.** Our chemical structure recognition method takes an input image and predict atom-level entities predictions (atoms, bonds, charges, and stereocenters). This rich annotation can then be used to construct the molecular graph. Atom-level entity prediction models, like ours, predict both the atom-level annotations and molecular graph from the image. In contrast, molecular graph predictions models only predict the molecular graph and do not provide any localization in the original image.

that is pretrained on synthetic data, where the localization of each atom-level entity is known. We then leverage the atom-level entity localization, coupled with a efficient self-relabeling strategy, to aptly transfer to new domains where no localization is available (typically only image-SMILES pairs are available). This results in a state-of-the-art and highly data efficient architecture, as demonstrated by our rigorous benchmarking.

Key contributions: (1) We propose a novel framework for chemical structure recognition that predicts atom-level localizations trained on a target domain with only SMILES supervision. (2) We show that our method results in state-of-the-art performance on challenging hand-drawn molecule images, with a remarkable data efficiency. (3) We release a new curated dataset containing hand-drawn molecules with atom-level annotations.

Our implementation is available on Github:
https://github.com/molden/atomlenz

2. Background

Our work builds upon the chemical structure recognition literature and takes an object detection approach for solving this task. To enable fine-tuning of the model where no atom-level annotations are present, we leverage advances in weakly supervised object detection.

2.1. Chemical structure recognition

Optimal chemical structure recognition consists in inferring the structural formulae of a chemical compound based on an image representation of it. The large majority of existing methods performing this task take the image as input, and predict the SMILES (simplified molecular-input line-entry system) representation of the molecule [33]. SMILES consists of strings of ASCII characters that are obtained by printing the chemical symbols encountered in a depth-first tree traversal of the molecular graph. This serial notation provides, at first sight, a convenient representation for training machine learning models, while encoding geometric information about the molecular graph. This justified the popularity of SMILES-based chemical structure recognition models [5, 24].

Nevertheless, SMILES do not provide a natural chemical representation and do not readily encode the geometric properties of the molecules. This hampers the trainability of the underlying machine learning model [14]. This limitation motivated the development of methods predicting the molecular graph and capable of leveraging richer image annotations, such as atom-level localizations [19, 23]. Our work belongs to this category and therefore inherits these strengths. However, we extend previous approaches by providing a mechanism to fine-tune the model to new data domains where only SMILES annotations are available.

2.2. Object detection

Our architecture draws heavily on the literature on objection detection in images [9, 10, 26, 27, 37], which underlies a wide array of high-level machine learning applications [3, 11, 12, 31]. We refer to [38] for a recent review of the field. Training object detection models typically requires comprehensive image annotations, such as the precise coordinates of bounding boxes and the associated labels for every object contained within each image. However, and crucially for our application, these annotations are not consistently accessible within certain domains of interest. This scarcity of detailed annotations has spurred the development of *weakly supervised* object detection methods.

2.3. Weakly supervised object detection

This category of methods enables the training of object detection models without the necessity for precise bounding-box annotations. As a result, these approaches can be di-

rectly applied to target domains where such annotations are unavailable. Diverse variations of Weakly Supervised Object Detection (WSOD) architectures have emerged, relying on a range of implementations, including Multiple Instance Learning (MIL) [16, 29] and Class Activation Map (CAM) [2, 36] approaches. Other advanced WSOD techniques incorporate knowledge transfer from a source domain [6, 15, 21, 32, 35]. Among these approaches, ProbKT [21] distinguishes itself as a versatile method, relying on probabilistic reasoning, which offers the capacity to train atom-level localization models using chemical background information obtained from SMILES and logical reasoning. Our architecture leverages this approach.

3. Method

Our architecture is composed of four high-level modules: (1) an object detection backbone, which is trained on richly annotated images with atom-level entities, (2) a molecular graph constructor that assembles a molecular graph from the set of atom-level predictions, (3) a weakly supervised training scheme that enables fine-tuning the model on new domains without rich annotations. Additionally, we design a chemically informed combination of experts, ChemExpert, that can further boost the prediction performance. The weakly supervised training scheme of the object detection backbone is visualized in Figure 2.

3.1. Object detection backbone

At the core of our architecture lies an object detection model that is responsible for detecting and labeling atom-level entities in the image. The objects in the image are therefore the atom-level entities such as atoms or bonds. While many object detection methods exist and can be used interchangeably in our architecture, we used the Faster RCNN model [27] in our experiments. It is fast to train, robust, and simultaneously localizes and classifies all objects in a single step. The object detection backbone is trained by minimizing a multi-task loss $\mathcal{L}$ mixing a multi-class log loss $\mathcal{L}_{cls}$ applied on the predicted class $\hat{c}$ of the objects and a regression loss $\mathcal{L}_{reg}$ applied to the predicted bounding box coordinates $\hat{b} = (\hat{b}_x, \hat{b}_y, \hat{b}_w, \hat{b}_h)$ of the objects.

$$\mathcal{L} = \mathcal{L}_{cls}(c, \hat{c}) + \mathcal{L}_{reg}(b, \hat{b}) \tag{1}$$

Bonds, atoms, or charges are graphically very different. To account for this heterogeneity, we train four distinct object detection models, each tailored to a specific atom-level entity: atoms ($\mathbf{O}^a$), bonds ($\mathbf{O}^b$), charge objects ($\mathbf{O}^c$), and stereocenters ($\mathbf{O}^s$). A stereocenter, also known as a stereogenic center, refers to an atom within a molecule that carries groups in a way that exchanging any two of these groups results in a stereoisomer [17]. More details about stereochemistry can be found in SM Section 8.

Each type of atom-level entity comprises multiple classes c that the object detection backbone aims at labeling (via $\mathcal{L}_{cls}$). For instance, the bond object encompasses categories such as single bond, double bond, triple bond, aromatic bond, dashed bond, and wedged bond. Illustrations of different atom-level entity types and classes can be found in SM Section 10.

3.2. Molecular graph constructor

The output of the object detection backbone is a list of detected atom-level entities in the image, along with their predicted label and position. The objective of the molecular graph constructor is to assemble a chemically sound molecular graph from this list of predictions. This graph can then be easily converted to a SMILES format.

The output graph $G(V, E)$ is composed of a set of vertices V, corresponding to atoms, and edges E corresponding to bonds. Each vertex and edge has a label (*e.g.*, a node can be a carbon atom with a positive charge, an edge can be a double bond). Algorithm 1 outlines the series of steps involved in constructing the molecular graph from atom-level entity predictions $\mathbf{O}^a$ (atoms), $\mathbf{O}^b$ (bonds), $\mathbf{O}^c$ (charges), and $\mathbf{O}^s$ (stereocenters). It proceeds in four steps: (1) a filtering step, (2) a node creation step, (3) an edge creation step, and (4) a validation step.

The **filtering step** filters atoms from $\mathbf{O}^a$ that are severely overlapping on the image. When multiple atom objects show an Intersection over Union (IoU) score exceeding a specified threshold, only the object with the highest score is retained.

In the **node creation step**, we first attach charges to atom objects. Overlapping atom and charge objects exceeding a specific IoU threshold are then merged. The function checkCharges is responsible for determining which atom objects should carry a charge. A similar procedure is subsequently applied to identify atoms functioning as stereocenters, utilizing checkStereoChem for this purpose. The list of all atom objects, along with their potentially assigned charges or stereocenters are then added to the list of graph vertices.

In the **edge creation step**, we iterate over all bond objects and evaluate which vertices (atoms) overlap with these bonds, with the function checkEdge. If only two candidate atoms are identified, the algorithm proceeds to add the edge to the graph. However, when more than two overlapping candidates emerge, the algorithm selects the two most probable ones, factoring in the orientation of the edge and the atoms involved.

Lastly, the **validation step** identifies potential chemistry-related issues through ChemistryProblems and endeavors to resolve them via SolveChemistryProblems to ensure the prediction of a chemically valid molecular graph. For instance, each chemical element is assigned a valence

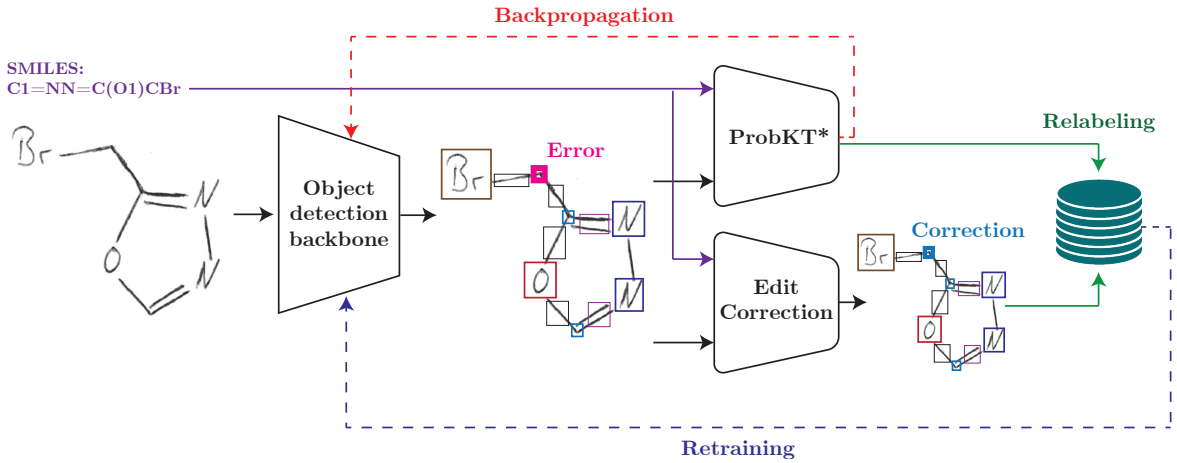


Figure 2. **Weakly supervised training**. The weakly supervised training data set consists of images of molecular depictions paired with SMILES. The input image is used by the object detection backbone to predict the atom-level entities while the SMILES is used by ProbKT* and the edit-correction scheme. In the first phase, ProbKT* will perform backpropagation to update the object detection backbone using probabilistic reasoning. In the second phase, both ProbKT* and the edit-correction mechanism will generate pseudo-labels for the atom-level entities, which are used to retrain the object detection backbone.

number, indicating the atom's capability to establish bonds with other atoms. If our algorithm detects within the output graph containing atoms with more bonds than their valence numbers permit, the SolveChemistryProblems function would attempt to remove bonds iteratively until a graph is formed without valence errors.

To maintain conciseness in our experiments and results, we refer to the combination of the object detection backbone and the molecular graph constructor as AtomLenz (ATOM-Level ENtity localiZer). Further information regarding the subroutines used in the molecular graph constructor is available in SM Section 9.

3.3. Weakly supervised training

Our architecture uses an object detection backbone to predict atom-level entities, which requires rich image annotations, such bounding boxes for every object type, including atoms, bonds, charges, and stereocenters within the images. While such a level of supervision can be obtained synthetically with tools like RDKit [1], it is usually not available in real-world target domains, such as hand-drawn images. In such domains, only SMILES are typically available. To enable the fine-tuning of the object detection backbone with only SMILES information, we use a weakly supervised training mechanism that combines (1) a probabilistic logical reasoning module that allows to differentiate through the object detection backbone with only weak supervision, and (2) a graph edit-correction mechanism that allows fine-tuning on less frequent atoms and bonds. A graphical outline of the weakly supervised training procedure is given in Figure 2.

Algorithm 1: Molecular graph constructor

Input: Atom-level predictions $\mathbf{O}^a$,$\mathbf{O}^b$,$\mathbf{O}^c$,$\mathbf{O}^s$
Result: Graph $G(V, E)$, vertices V and edges E

$\tilde{\mathbf{O}}^a$ = filterAtoms($\mathbf{O}^a$) ; // filtering step
$V = []$; // vertices of graph
/* node creation step */
for $\mathbf{o}^a$ ***in*** $\tilde{\mathbf{O}}^a$ **do**
  $\mathbf{o}^a_c$ = checkCharges($\mathbf{O}^c, \mathbf{o}^a$)
  $\mathbf{o}^a_{c,s}$ = checkStereoChem($\mathbf{O}^s, \mathbf{o}^a_c$)
  V.appendAtom($\mathbf{o}^a_{c,s}$)
end
$E = []$; // edges of graph
/* edge creation step */
for $\mathbf{o}^b$ ***in*** $\mathbf{O}^b$ **do**
  candAtoms = checkEdge($V, \mathbf{o}^b$)
  if len(candAtoms) == 2 **then**
    E.appendBond($\mathbf{o}^b$, candAtoms)
  end
  if len(candAtoms) > 2 **then**
    filteredAtoms = filterCands(candAtoms)
    E.appendBond($\mathbf{o}^b$, filteredAtoms)
  end
end
/* validation step */
if ChemistryProblems($G(V, E)$) == True **then**
  $G(V, E)$ = SolveChemistryProblems($G(V, E)$)
end

Backpropagation with weak supervision

To update the weights of the object detection backbone with only SMILES supervision, we use the ProbKT [21] framework. This weakly supervised domain adaption technique uses probabilistic programming for fine-tuning object detection models with a wide range of supervision signals, and is thus particularly suited for our application. In our experiments, we used ProbKT*, a computationally efficient variant of ProbKT that relies on Hungarian matching.

ProbKT* allows differentiating through the object detection backbone with only SMILES supervision. For better performance, it also includes a relabeling mechanism, where confident predictions are used as new atom-level annotation of the target domain images. This strategy effectively creates a richly annotated dataset that can be used to fine-tune the object detection backbone directly.

Edit-correction mechanism

While ProbKT* is generally effective at performing weakly supervised domain adaptation, it fails when dealing with rare atoms or bonds types. We therefore combine ProbKT* with a new edit-correction mechanism [20] designed to detect and rectify minor errors in model predictions. Based on the SMILES, one can generate a reference true graph, although not aligned on the original image. The edit-correction mechanism solves an optimization problem that aims at finding the smallest edit on the predicted graph such that the true and corrected graphs are isomorphic. While this optimization would be intractable in general, focusing on small edits makes it computationally feasible. If such a correction is found, it is used to annotate the image which can then be used to fine-tune the object detection backbone.

Combined weakly supervised training

When fine-tuning on a new target domain, we proceed by iteratively applying ProbKT* and the edit-correction scheme. In practice, we start with a few iterations of ProbKT*. We then use multiple iterations of the edit-correction scheme until the validation performance stops improving. For sake of conciseness in our experiments results, we abbreviate the combination of both approaches as EditKT*.

3.4. ChemExpert: Combination of experts

For the final prediction of our architecture, we propose to use a combination of experts, which is constrained by chemical soundness of the model predictions. We call this module ChemExpert. It relies on a list of chemical structure recognition tools, ordered by the user's preference in terms of the predictions. The first tool serves as the most trusted model. At inference time, ChemExpert iteratively checks the validity of the prediction of each model in the list. If a chemical issue is identified, the agent evaluates the next model in the list. The module returns the prediction of the first model with no chemical issues detected. This strategy enables us to incorporate predictions from additional tools alongside those generated by our core model, thereby improving predictive performance. In practice, we use a combination of DECIMER [25] and our approach.

4. Datasets

4.1. Synthetically generated dataset

For the pretraining of the object detection models, we generate images synthetically using RdKit [1] and Indigo [22] paired with bounding boxes delineating all objects within, including atoms, bonds, charges, and stereocenters, similarly to what is used in other chemical structure recognition tools [19, 23]. Specifically, we collect approximately 214,000 chemical compounds in SMILES format from the ChEMBL [8] database. To enhance the method's resilience to stylistic variations, we introduce variability in elements such as fonts, font sizes, line widths, and the spacing between multiple bonds during image generation. More details on this dataset can be found in SM Section 7.

4.2. Hand-drawn images datasets

To facilitate the training, fine-tuning, and testing of our models on hand-drawn images, we meticulously curate multiple datasets. We begin with the dataset introduced by Brinkhaus et al. [4], which consists of hand-drawn chemical depictions matched with their corresponding SMILES representations. This dataset is partitioned into 4,070 samples for training and validation purposes, along with an additional 1,018 samples for testing. These sets are referred to as the *hand-drawn training set* and the *hand-drawn test set*.

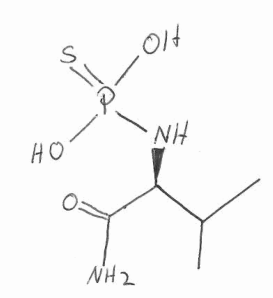
In addition to this primary test set, we incorporate an extra test dataset of 614 hand-drawn chemical depictions sourced from Weir et al. [34], which we call the *ChemPix test set*, to further evaluate the performance of the models on hand-drawn images.

4.3. Atom localization dataset

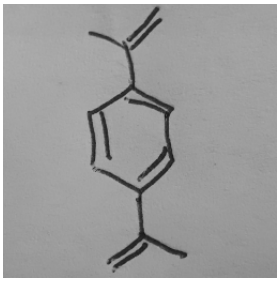
To assess our models' capability for object localization, we employ a synthetically generated dataset using RdKit [1] provided by Oldenhof et al. [21]. This dataset encompasses 1,000 images depicting chemical structures, each meticulously annotated with bounding boxes outlining the positions and corresponding classes of all the atoms within the molecules. Example images for each test dataset are shown in Figure 3 and SM Section 7.

5. Experiments and Results

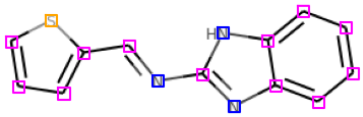
Our experiments investigate the performance of our approach on hand-drawn images of chemical structures, as



(a) Hand-drawn test set



(b) ChemPix test set



(c) Atom Localization test set, labeled with bounding boxes, here illustrated with different color for every atom type.

Figure 3. **Image samples from the dataset.** Different example samples for the different datasets used in experiments. The hand-drawn and ChemPix datasets are used to assess the domain adaptation and out-of-domain performance. The atom localization dataset is used for testing object localization.

this domain suffers from limited data availability and has been shown to be a weak point of existing tools. We evaluate our architecture and state-of-the-art baselines on four distinct fronts: (1) molecular recognition on the new target domain (*i.e.,* only predicting the SMILES), (2) atom-level entity localization, (3) training efficiency (when retrained from scratch), and (4) model evaluation per atom and bond type.

Baselines We compare our architecture with the following baselines. **DECIMER** [24] is an image-transformer approach trained on more than 400 million synthetically generated data samples. The authors of DECIMER have also introduced a version specifically tailored for hand-drawn images (DECIMER fine-tuned [25]). Although it is trained on synthetically generated images, the training dataset of this version mimics the style of hand-drawn images more closely. **Img2Mol** [5] integrates a deep convolutional neural network trained on molecule depictions (11 million synthetically generated images) with a pretrained decoder. **MolScribe** [23] and **ChemGrapher** [19] employ atom-level entity localization annotations in their training process on synthetically generated images. These are the only baselines that can predict atom-level annotations, alongside the SMILES predictions. ChemGrapher is trained on 114,000 generated images and MolScribe on 1 million generated images. Lastly, **OSRA** [7] is a non-trainable, rule-based approach.

5.1. Performance of molecular recognition on hand-drawn images

We compare the performance of our approach with the baselines on the hand-drawn and ChemPix dataset. Results are given in Table 1. To assess the impact of our fine-tuning strategy, we evaluate three versions of our architecture. The first is a version trained on the synthetic dataset but not fine-tuned to the new hand-drawn dataset (AtomLenz). The second is fine-tuned to the hand-drawn dataset using EditKT*. The third is ChemExpert, combining DECIMER fine-tuned and AtomLenz. Performance of other combinations in ChemExpert are reported in SM Section 11.

We assess the molecular structure prediction performance using accuracy and Tanimoto similarity. Tanimoto similarity (T) [30], a widely used metric for quantifying molecular similarity, to assess the resemblance between the model's predictions and the actual molecular graphs. Tanimoto similarity values range from 0 to 1, with higher values indicating greater similarity. A Tanimoto similarity of 1 indicates that the structural descriptors are identical or that they are matching 'on-bits' in a binary fingerprint. The binary fingerprint employed to measure the Tanimoto similarity is the Extended-connectivity fingerprint [28] with radius 3 (ECFP6) and fingerprint length of 2048. More details on the calculation of the ECFP6 fingerprint and other fingerprints can be found in SM Section 11.

Our tables report both the accuracy, computed by counting the instances where the predicted structures have identical structural ECFP6 descriptors (denoted by a Tanimoto similarity of 1) and the average Tanimoto similarity. Additional measured metrics can be found in SM Section 11.

For both datasets *hand-drawn test set* and *ChemPix test set*, our ChemExpert performs best. This demonstrates that the combination of our approach with other baselines results in state-of-the-art performance. We further appreciate a significant increase in performance from EditKT*, compared to the non-fine-tuned version, highlighting the effectiveness of our fine-tuning approach.

5.2. Performance of atom-level localization

In Table 2, we assess the performance of the different methods in terms of their atom-level localization abilities. We employ a test set from Oldenhof et al. [21], which comprises images of chemical representations along with the corresponding atom objects. We use two evaluation met-

	hand-drawn test set		ChemPix test set	
Method	Acc.($T=1$)	$\overline{T}$	Acc.($T=1$)	$\overline{T}$
DECIMER (v2.2.0)[24]	0.295	0.451	0.05	0.1
DECIMER fine-tuned(v2.2.0)[25]	0.622	0.727	0.508	0.643
Img2Mol[5]	0.084	0.275	0.015	0.084
MolScribe[23]	0.102	0.288	0.269	0.417
ChemGrapher[19]	0.002	0.065	0.187	0.286
OSRA[7]	0.006	0.065	0.047	0.071
AtomLenz	0.009	0.087	0.054	0.064
AtomLenz+EditKT*	0.338	0.484	0.484	0.605
ChemExpert([25],AtomLenz+EditKT*)	**0.635**	**0.749**	**0.518**	**0.655**

Table 1. Benchmark results on target domain (hand-drawn images test set) and out of domain ChemPix test set. Both the accuracy, computed by counting the instances where the predicted structures have identical structural ECFP6 descriptors (denoted by a Tanimoto (T) similarity of 1) and the average Tanimoto similarity ($\overline{T}$) are reported.

rics: the count accuracy (which can be evaluated without bounding box predictions), and the mean average precision (mAP) localization of the bounding boxes. The atoms count accuracy measures the ability to predict the correct number of atom types in each image. The average precision is computed as the weighted mean of precisions at various Intersect over Union (IoU) thresholds, with the weight reflecting the increase in recall from the previous threshold. Mean Average Precision represents the average of AP values across each class. We use the rather low IoU thresholds of $[0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35]$ in our experiments, considering the relatively small size of the bounding boxes of interest (see Figure 5c), where significant overlap with the true bounding boxes is not anticipated. Methods that do not provide any form of localization are marked with n/a in the table.

In Table 2, we observe the commendable localization performance of our pretrained backbone model (AtomLenz) and also note the high counting accuracy of DECIMER. We posit that both outcomes may be attributed to the nature of the test dataset, aligning with the characteristics of the images used for training both DECIMER and our core pretrained model (AtomLenz). Additionally, we note that DECIMER was trained on 2000 times more images than our architecture, and does not provide any localization.

5.3. Training efficiency

Baseline architectures were trained on significantly larger number of images than our model. Molscribe uses 4 times more samples, while DECIMER uses a staggering 2000 times more. In Table 3, we evaluate the sample complexity of the different methods by retraining them from scratch on the same small training dataset, which mimics limited data availability scenarios. We use the *hand-drawn training set* (4,070 data samples), enriched with atom-level entity localization annotations generated using EditKT* as the training dataset. We observe that the methods that leverage these atom-level entity annotations tend to fare better (ChemGrapher [19] and MolScribe [23]) than the ones using SMILES as only supervision signal (DECIMER [24] and Img2Mol [5]). Our approach significantly outperforms all baselines at this task, highlighting the remarkable data efficiency of our architecture. These findings align with those in the work of Hormazabal et al. [13], where the author concluded that the use of atom-level entity annotations can enhance data efficiency during training. The hand-drawn images utilized in this experiment, along with the corresponding bounding box labels for 1417 images, we release as a novel annotated dataset. More info in the SM Section 7.

Method	Count Acc.	mAP
DECIMER (v2.2.0)[24]	**0.973**	n/a
DECIMER fine-tuned (v2.2.0)[24]	0.97	n/a
Img2Mol[5]	0.929	n/a
MolScribe[23]	0.829	0.008
ChemGrapher[19]	0.248	0.002
OSRA[7]	0.255	n/a
AtomLenz	0.602	**0.801**

Table 2. Benchmark results on object (atom) detection test set to compare localization performance.

5.4. Fine-grained model evaluation

Additionally, we conduct a detailed performance analysis of the most effective models from Table 1, presented in Figure 4. This figure showcases the count accuracies per atom or bond type. For each specific type, we identify im-

Method	Acc.($T = 1$)	$\overline{T}$
DECIMER (v2.2.0)[24]	0.001	0.039
Img2Mol[5]	0.0	0.0867
MolScribe[23]	0.013	0.0865
ChemGrapher[19]	0.004	0.067
AtomLenz	**0.338**	**0.484**

Table 3. All methods are retrained from scratch on same training dataset (4070 samples of hand-drawn images) to asses data efficiency. Benchmark results on hand-drawn images test set. Both the accuracy, computed by counting the instances where the predicted structures have identical structural ECFP6 descriptors (denoted by a Tanimoto (T) similarity of 1) and the average Tanimoto similarity ($\overline{T}$) are reported.

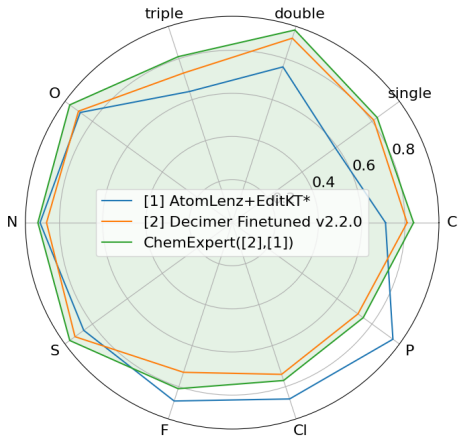


Figure 4. Count accuracies per type over images if type is present in image for hand-drawn test set. We observe errors of 'AtomLenz+EditKT*' and 'DECIMER fine-tuned' tend to occur on different samples. Combining both approaches in ChemExpert improves performance.

ages featuring that particular atom or bond type, then examine the predictions made by the methods on these images. Subsequently, we calculate the count accuracies for the predicted objects of the specific type within these images. For instance, when analyzing the 'triple bond' type, we select test images where at least one triple bond is depicted in the molecule and evaluate whether the method accurately predicts the correct number of triple bonds in the resulting molecular graph.

The plot in Figure 4 exhibits distinct patterns between 'AtomLenz+EditKT* ' and Decimer fine-tuned [25]. For example 'AtomLenz+EditKT* ' performs better on images with Chlorine (Cl), Fluorine (F), and Phosphorus(P) compared to Decimer fine-tuned but worse on bonds. This variability may clarify why combining both predictions into ChemExpert leads to improved performance, as errors tend to occur on different samples, and the two approaches complement each other. The same analysis is performed for the ChemPix dataset in the SM section 11 in Figure 11.

6. Conclusion

This study has undertaken a comprehensive evaluation of various methods for chemical structure recognition, with a primary focus on the challenging domain of hand-drawn images. Our findings reveal insights into the strengths and limitations of existing tools and provide a compelling case for the efficacy of our approach. We showed that our method fares competitively despite a lower number of training samples, and resulted in state-of-the-art performance when combined with previous approaches. Our experiments highlighted our method's proficiency in precisely localizing atom-level entities, a feature notably lacking in many existing tools. Importantly, we showed that our architecture is remarkably more data-efficient than previous models

Despite these improvements in chemical structure recognition, reliably predicting the molecular structure from hand-drawn remains a challenge, and higher prediction performance would be required for a wide adoption of these tools. We hope that the release of our curated hand-drawn molecules images dataset, with detailed atom-level annotations, to the community will contribute to the development of more efficient and reliable tools.

Acknowledgments

AA, MO and YM are funded by (1) Research Council KU Leuven: Symbiosis 4 (C14/22/125), Symbiosis3 (C14/18/092); (2) Federated cloud-based Artificial Intelligence-driven platform for liquid biopsy analyses (C3/20/100); (3) CELSA - Active Learning (CELSA/21/019); (4) European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 956832; (5) Flemish Government (FWO: SBO (S003422N), Elixir Belgium (I002819N), SB and Postdoctoral grants: S003422N, 1SB2721N, 1S98819N, 12Y5623N) and (6) VLAIO PM: Augmenting Therapeutic Effectiveness through Novel Analytics (HBC.2019.2528); (7) YM, AA, and MO are affiliated to Leuven.AI and received funding from the Flemish Government (AI Research Program). Computational resources and services used in this work were partly provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government – department EWI.

References

[1] Rdkit: Open-source cheminformatics. accessed on 01.02.2022. 4, 5, 1

[2] Wonho Bae, Junhyug Noh, and Gunhee Kim. Rethinking class activation mapping for weakly supervised object localization. In *European Conference on Computer Vision*, pages 618–634. Springer, 2020. 3

[3] Aseem Behl, Omid Hosseini Jafari, Siva Karthik Mustikovela, Hassan Abu Alhaija, Carsten Rother, and Andreas Geiger. Bounding boxes, segmentations and object coordinates: How important is recognition for 3d scene flow estimation in autonomous driving scenarios? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2574–2583, 2017. 2

[4] Henning Otto Brinkhaus, Achim Zielesny, Christoph Steinbeck, and Kohulan Rajan. Decimer - hand-drawn molecule images dataset, 2022. 5, 1

[5] 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. 1, 2, 6, 7, 8

[6] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *International journal of computer vision*, 100 (3):275–293, 2012. 3

[7] Igor V Filippov and Marc C Nicklaus. Optical structure recognition software to recover chemical information: Osra, an open source solution, 2009. 6, 7, 8

[8] Anna Gaulton, Anne Hersey, Michał Nowotka, A Patricia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J Bellis, Elena Cibrián-Uhalte, et al. The chembl database in 2017. *Nucleic acids research*, 45(D1):D945–D954, 2017. 5, 1

[9] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 2

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2

[11] Eleonora Giunchiglia, Mihaela Cătălina Stoian, Salman Khan, Fabio Cuzzolin, and Thomas Lukasiewicz. Road-r: The autonomous driving dataset with logical requirements. *arXiv preprint arXiv:2210.01597*, 2022. 2

[12] Ibtihaal M Hameed, Sadiq H Abdulhussain, and Basheera M Mahmmod. Content-based image retrieval: A review of recent trends. *Cogent Engineering*, 8(1):1927469, 2021. 2

[13] Rodrigo Hormazabal, Changyoung Park, Soonyoung Lee, Sehui Han, Yeonsik Jo, Jaewan Lee, Ahra Jo, Seung Hwan Kim, Jaegul Choo, Moontae Lee, et al. Cede: A collection of expert-curated datasets with atom-level entity annotations for optical chemical structure recognition. 7

[14] Rodrigo Hormazabal, Changyoung Park, Soonyoung Lee, Sehui Han, Yeonsik Jo, Jaewan Lee, Ahra Jo, Seung Hwan Kim, Jaegul Choo, Moontae Lee, et al. Cede: A collection of expert-curated datasets with atom-level entity annotations for optical chemical structure recognition. *Advances in Neural Information Processing Systems*, 35:27114–27126, 2022. 1, 2

[15] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5001–5009, 2018. 3

[16] Dong Li, Jia-Bin Huang, Yali Li, Shengjin Wang, and Ming-Hsuan Yang. Weakly supervised object localization with progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3512–3520, 2016. 3

[17] Kurt Mislow and Jay Siegel. Stereoisomerism and local chirality. *Journal of the American Chemical Society*, 106(11): 3319–3328, 1984. 3

[18] H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5 (2):107–113, 1965. 6

[19] 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, 2020. 1, 2, 5, 6, 7, 8

[20] Martijn Oldenhof, Adam Arany, Yves Moreau, and Jaak Simm. Self-labeling of fully mediating representations by graph alignment. In *Benelux Conference on Artificial Intelligence*, pages 46–65. Springer, 2021. 5

[21] Martijn Oldenhof, Adam Arany, Yves Moreau, and Edward De Brouwer. Weakly supervised knowledge transfer with probabilistic logical reasoning for object detection. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 3, 5, 6

[22] Dmitry Pavlov, Mikhail Rybalkin, Boris Karulin, Mikhail Kozhevnikov, Alexey Savelyev, and A Churinov. Indigo: universal cheminformatics api. *Journal of cheminformatics*, 3(Suppl 1):P4, 2011. 5, 1

[23] Yujie Qian, Jiang Guo, Zhengkai Tu, Zhening Li, Connor W Coley, and Regina Barzilay. Molscribe: Robust molecular structure recognition with image-to-graph generation. *Journal of Chemical Information and Modeling*, 63(7):1925–1934, 2023. 1, 2, 5, 6, 7, 8

[24] Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. Decimer: towards deep learning for chemical image recognition. *Journal of Cheminformatics*, 12(1):1–9, 2020. 1, 2, 6, 7, 8

[25] Kohulan Rajan et al. Decimer.ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. *Nat Commun*, 14 (5045), 2023. 5, 6, 7, 8

[26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region

proposal networks. *Advances in neural information processing systems*, 28, 2015. 2, 3

[28] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. PMID: 20426451. 6

[29] Hyun Oh Song, Ross Girshick, Stefanie Jegelka, Julien Mairal, Zaid Harchaoui, and Trevor Darrell. On learning to localize objects with minimal supervision. In *International Conference on Machine Learning*, pages 1611–1619. PMLR, 2014. 3

[30] Taffee T Tanimoto. Elementary mathematical theory of classification and prediction. 1958. 6

[31] Matteo Tomei, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Art2real: Unfolding the reality of artworks via semantically-aware image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5849–5859, 2019. 2

[32] Jasper Uijlings, Stefan Popov, and Vittorio Ferrari. Revisiting knowledge transfer for training object class detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1101–1110, 2018. 3

[33] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988. 1, 2

[34] Hayley Weir, Keiran Thompson, Amelia Woodward, Benjamin Choi, Augustin Braun, and Todd J. Martínez. Chempix: automated recognition of hand-drawn hydrocarbon structures using deep learning. *Chem. Sci.*, 12:10622–10633, 2021. 5

[35] Yuanyi Zhong, Jianfeng Wang, Jian Peng, and Lei Zhang. Boosting weakly supervised object detection with progressive knowledge transfer. In *European conference on computer vision*, pages 615–631. Springer, 2020. 3

[36] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 3

[37] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2

[38] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 2023. 2