JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

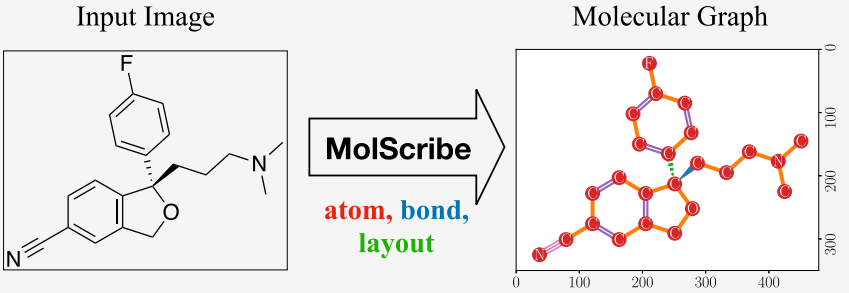
pubs.acs.org/jcim Article

MolScribe: Robust Molecular Structure Recognition with Image-to-Graph Generation

Yujie Qian,* Jiang Guo, Zhengkai Tu, Zhening Li, Connor W. Coley, and Regina Barzilay*

Cite This: J. Chem. Inf. Model. 2023, 63, 1925−1934 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

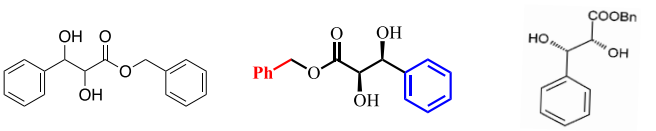


ABSTRACT: Molecular structure recognition is the task of translating a molecular image into its graph structure. Significant variation in drawing styles and conventions exhibited in chemical literature poses a significant challenge for automating this task. In this paper, we propose MolScribe, a novel image-to-graph generation model that explicitly predicts atoms and bonds, along with their geometric layouts, to construct the molecular structure. Our model flexibly incorporates symbolic chemistry constraints to recognize chirality and expand abbreviated structures. We further develop data augmentation strategies to enhance the model robustness against domain shifts. In experiments on both synthetic and realistic molecular images, MolScribe significantly outperforms previous models, achieving 76−93% accuracy on public benchmarks. Chemists can also easily verify MolScribe's prediction, informed by its confidence estimation and atom-level alignment with the input image. MolScribe is publicly available through Python and web interfaces: https://github.com/thomas0809/MolScribe.

INTRODUCTION

Molecules are commonly drawn as 2D images in chemistry literature. Such drawings exhibit a wide variety of styles, which complicates the task of translating these images into machine-readable molecular structures. For example, Figure 1 shows three different ways to draw the same molecule. The plain image (a) can be recognized adequately by existing models, but variants such as (b) and (c) are more difficult as they involve stereochemistry, functional group abbreviations, and diverse drawing styles. While prior work has demonstrated that it is possible to train models that perform well on one style assuming labeled data is available, the diversity of possible styles is a standing challenge. We cannot assume access to training data that covers all possible styles and patterns in the chemistry literature due to the high annotation cost. Therefore, we aim to enhance the robustness of molecular structure recognition model to generalize to an arbitrary molecular image.

In this paper, we propose MolScribe, which takes as input a molecular image (e.g., a PNG or JPG file) and generates its molecular graph structure. The model relies on three complementary approaches to handle data variation robustly. First, MolScribe explicitly predicts the atoms and bonds along with their geometric layouts in the image, which together constitute a *2D molecular graph*. Second, we incorporate chemistry knowledge as symbolic constraints to the model, such that it can accurately recognize complex chemical patterns. For example, we determine the chirality of asymmetric atoms based on the predicted graph and layout and design an algorithm to parse abbreviated functional groups



(a) (b) (c)

Figure 1. Three images of the same molecule with different layouts and styles. (a) depicts the full molecular structure and does not specify chirality. (b) and (c) use different abbreviations, and (b) color-codes the phenyl groups.

Received: November 22, 2022
Published: March 27, 2023

https://doi.org/10.1021/acs.jcim.2c01480
J. Chem. Inf. Model. 2023, 63, 1925−1934

that commonly appear in molecular images. Third, we propose data augmentation strategies to synthetically generate images that cover diverse drawing styles during training. Specifically, MolScribe is implemented as an image-to-graph generation model, which abstracts the input image into hidden representations and generates the molecular graph with an autoregressive decoder. MolScribe's output can be aligned with the input image at the atom-level, allowing humans to easily interpret its predictions.

MolScribe combines the strengths of previous rule-based and neural models. Rule-based methods[1,2] can easily enforce chemistry constraints and conduct symbolic reasoning over the molecular graph, while neural models[3−5] are more robust to image styles and perturbations. MolScribe naturally extends a neural generation model to predict atoms and their coordinates as a sequence and then the bond between each pair of atoms. This design allows MolScribe to robustly recognize local atoms and bonds and flexibly integrate chemistry rules when constructing the molecular graph, such that it generalizes to different scenarios.

In our experiments, MolScribe outperforms an image-to-SMILES baseline on both in-domain and out-of-domain images and achieves strong recognition accuracy (76−93%) on five public benchmarks. We also construct a new benchmark with molecular images from journal publications, and MolScribe significantly outperforms the baseline and existing systems. Moreover, MolScribe is robust against input perturbation and low-quality images, predicts chirality more accurately, and provides confidence estimation. Finally, we conduct a human evaluation to understand how our model can help chemists to parse molecular images in a semiautomated workflow. The results show that our graph prediction significantly reduces the time needed by chemists to extract the molecular structures from images. Our model, code, and data are publicly available (https://github.com/thomas0809/MolScribe) for future research in molecular structure recognition and chemistry information extraction. We have developed a demo that allows chemists to use MolScribe from a web interface (https://huggingface.co/spaces/yujieq/MolScribe).

MOLECULAR STRUCTURE RECOGNITION

Background. Research on molecular structure recognition dates back to at least the 1990s (see a recent review by Rajan et al.[6]). The task is also known as Optical Chemical Structure Recognition (OCSR). Earlier rule-based systems[7−14] relied on traditional image processing techniques (binarization, line smoothing and thinning, vectorization) to segment the pixels into atoms and bonds and standalone optical character recognition (OCR) models to identify atom labels. Then, heuristics based on line length, width, spacing, and direction were used to determine bond types (single, double, triple, wedge, dashed) and connect all these elements into molecular graphs. These systems were usually meticulously engineered by chemists, with specific rules to handle different situations in molecular images. For example, they usually have an expert-compiled dictionary to resolve the most common functional group abbreviations. A few open-source tools have been developed,[1,2,15] and developers are continuously improving them by writing new rules to cover edge cases (e.g., bridge bonds).[1] The rule-based systems have achieved decent recognition accuracy on patent images, but there is still much room for improvement on journal article images, which are more diverse. Researchers also identified that small image perturbations can cause significant performance degradation.[4]

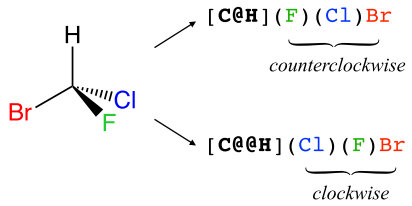


Figure 2. Chirality specification in SMILES. A chiral center may be indicated by "@" or "@@", meaning the neighbors are listed counterclockwise or clockwise, in this example when looking down the H−C bond.

Neural models have been proposed to improve the robustness against image variations. Staker et al. presented an image-to-SMILES generation model:[3] a convolutional neural network encodes the input molecular image into a hidden representation, and a recurrent neural network decodes the SMILES string of the molecule. A few variants have been proposed since, exploring multiple model architectures, such as Inception network,[16] Transformer,[17,18] Swin Transformer,[19] pretrained decoder,[4] graph decoder,[5] and also the application to hand-drawn molecules.[20,21] Neural models enjoy the simplicity of end-to-end training and the strength of handling different image styles, but operating on the SMILE strings instead of explicitly recognizing atoms and bonds makes it difficult to incorporate chemistry constraints. For example, as we show in the experiments, predicting stereochemistry presents a challenge for traditional neural networks due to the need for geometric reasoning over the molecular graph. Figure 2 shows two equivalent SMILES strings for a chiral molecule. Whether the chiral center should be written as "[C@H]" or "[C@@H]" depends on the relative order of its connecting bonds, which cannot be easily determined from the local pattern. Due to the lack of explicit notion of atoms and bonds, we cannot easily inject chemistry rules into these models.

Besides the image-to-SMILES formulation that directly generates the SMILES string, ChemGrapher[22] and MolMiner[23] train separate modules to detect atoms, bonds, and texts, based on image segmentation or object detection; and then construct the graph with heuristics. Such systems can also incorporate chemistry constraints during their graph construction process. Compared to these models, we simplify the method with a single end-to-end model to generate the molecular graph, so that our model does not rely on heuristics to connect the local predictions.

Task Formulation. Molecular structure recognition is the task of translating single-molecule images into corresponding molecular structures. In this paper, we formulate it as *image-to-graph generation*. Given an image I of molecule M, we translate it into a 2D molecular graph $G = (A, B)$, where $A = \{a_1, a_2, ..., a_n\}$ is the set of atoms, $B \subset A \times A \times T$ is the set of bonds, and T is the set of bond types (e.g., single, double, triple, solid wedge, dashed wedge). We define each atom as $a_i = (l_i, x_i, y_i)$, where l_i is the atom's corresponding SMILES (sub)string (e.g., "C", "*N*", "[OH−]", and "[235U]"), and x_i and y_i are the 2D coordinates of the atom in the image. Finally, the molecular graph can be converted and stored in standard data formats, such as a SMILES string[24] or a MOLfile.[25]

https://doi.org/10.1021/acs.jcim.2c01480
J. Chem. Inf. Model. 2023, 63, 1925−1934

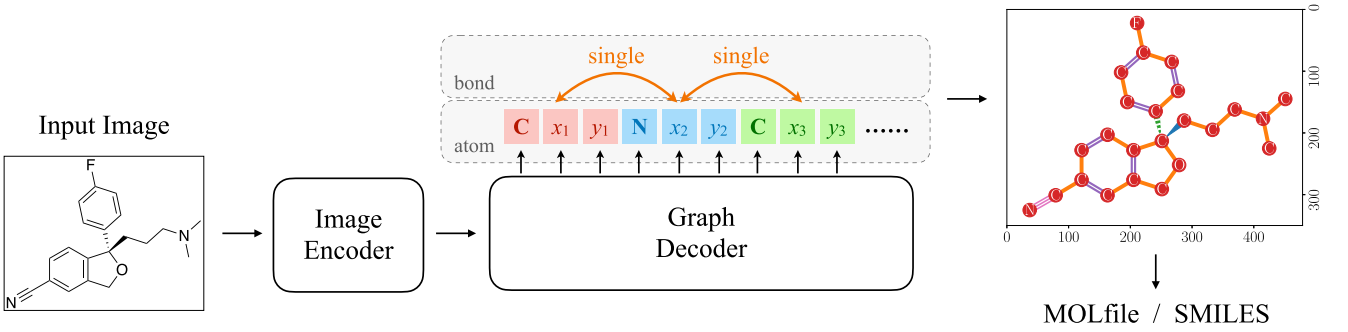


Figure 3. MolScribe model architecture. The input image is encoded with an image encoder, and the graph decoder predicts the atoms and bonds. A molecular graph is constructed from the predictions and converted to a MOLfile or a SMILES string.

Model. We propose MolScribe, an image-to-graph model that translates an input image into a 2D molecular graph. The model follows an encoder-decoder architecture, as shown in Figure 3. We first use an image encoder to encode the input image into a hidden representation and then use a graph decoder to generate the molecular structure. Unlike traditional image-to-SMILES models, which output a sequence of tokens, our decoder predicts atoms, bonds, and their geometric layouts jointly, such that the molecular structure can be reconstructed.

MolScribe formulates image-to-graph translation as a conditional generation process

$$P(G \mid I) = P(A \mid I)P(B \mid A,I) \tag{1}$$

where $P(A|I)$ and $P(B|A,I)$ are parametrized as an atom predictor and a bond predictor, respectively.

As shown in the middle of Figure 3, the atom predictor is an autoregressive decoder that generates the atoms in a sequence, i.e.,

$$P(A \mid I) = \prod_{i=1}^{n} P(a_i \mid A_{<i},I) \tag{2}$$

where $A_{<i}$ stands for the atoms before a_i. Inspired by the object detection model Pix2Seq,[26] our atom predictor simultaneously predicts atom labels and coordinates. Specifically, we construct a sequence of discrete tokens as the output format of the atom predictor

$$S^A = [l_1, \hat{x}_1, \hat{y}_1, l_2, \hat{x}_2, \hat{y}_2, ..., l_n, \hat{x}_n, \hat{y}_n] \tag{3}$$

where each atom a_i corresponds to three tokens: l_i, $\hat{x}_i$, and $\hat{y}_i$. l_i is the atom's SMILES specification, including the element identity, isotope, formal charge, and implicit hydrogen count. $\hat{x}_i$ and $\hat{y}_i$ are discrete values representing the atom's coordinates defined by binning, i.e.,

$$\hat{x}_i = \left\lfloor \frac{x_i}{W} \times n_{\text{bins}} \right\rfloor, \quad \hat{y}_i = \left\lfloor \frac{y_i}{H} \times n_{\text{bins}} \right\rfloor \tag{4}$$

where H and W are the height and width of the input image, respectively, and n_{bins} is a hyperparameter for the number of bins. This discretization is designed to simplify the model architecture. As the sequence S^A contains all the atom labels and coordinates, we implement the atom predictor to generate it autoregressively

$$P(A \mid I) = P(S^A \mid I) = \prod_{i=1}^{|S^A|} P(S_i^A \mid S_{<i}^A, I) \tag{5}$$

In practice, we split the molecule's SMILES string with an atom-wise tokenizer[27] into atom labels and append the coordinate tokens to them. The other tokens in the SMILES string, such as parentheses, bond symbols and digits, indicate connections among the atoms and are helpful to the model, so we keep them in the output sequence S^A. These tokens are not associated with coordinates.

The bond predictor is a feedforward network that predicts the bond between each pair of atoms (see Figure 3). Each atom a_i is represented as a vector $\boldsymbol{h}_{a_i}$, the hidden state of its last token in the decoder output. For each pair a_i and a_j, we concatenate their representations to form the input to the bond predictor, which classifies the bond type. Formally,

$$P(B \mid A,I) = \prod_{i=1}^{n} \prod_{j=1}^{n} P(b_{i,j} \mid A,I) \tag{6}$$

where $b_{i,j}$ indicates the bond between a_i and a_j. We use "None" to indicate no bond exists between the atom pair, and other possible bond types include single, double, triple, aromatic, solid wedge, and dashed wedge. For single, double, triple, and aromatic bonds, it is expected that $b_{i,j} = b_{j,i}$, but wedge bonds are not symmetric. To account for this, we predict both directions $b_{i,j}$ and $b_{j,i}$ independently and merge their predicted probabilities at inference time. For symmetric bonds, $b_{i,j}$ and $b_{j,i}$ are expected to be the same, so the probabilities are averaged, i.e.

$$\hat{P}(b_{i,j} = t) = \frac{1}{2}(P(b_{i,j} = t) + P(b_{j,i} = t)),$$
$$t \in \{\text{"single"}, \text{"double"}, \text{"triple"}, \text{"aromatic"}\} \tag{7}$$

For asymmetric bonds (wedges), as a solid wedge ("s.w.") is equivalent to a dashed wedge ("d.w.") in the opposite direction, we set $b_{i,j}$ = "s.w." and $b_{j,i}$ = "d.w." if there is a solid wedge from a_i to a_j and vice versa. At inference time, we merge the probabilities by

$$\hat{P}(b_{i,j} = \text{"s.w."}) = \frac{1}{2}(P(b_{i,j} = \text{"s.w."}) + \mathrm{P}(\mathrm{b_{j,i}} = \text{"d.w."})),$$
$$\hat{P}(b_{i,j} = \text{"d.w."}) = \frac{1}{2}(P(b_{i,j} = \text{"d.w."}) + \mathrm{P}(\mathrm{b_{j,i}} = \text{"s.w."})) \tag{8}$$

$\hat{P}(b_{i,j})$ is the bond predictor's final prediction during inference. (The conditional parts are omitted.)

Finally, a molecular graph is constructed from the predicted atoms and bonds. We use the RDKit[28] toolkit to save the

https://doi.org/10.1021/acs.jcim.2c01480
J. Chem. Inf. Model. 2023, 63, 1925−1934

molecular structure as a MOLfile (with 2D coordinate information) or a SMILES string (without 2D coordinate information).

Stereochemistry. As explained in Figure 2, it can be challenging for a neural model to recognize stereochemistry. In SMILES, chirality is indicated as an atomic property, i.e., @ or @@ after an atom label specifies the relative spatial orientation of the bonds connected to this atom in the order they are listed. There is no direct correspondence between the local image pattern and whether the atom symbol is @ or @@. Understanding stereochemistry requires geometric reasoning, which is not a strength of traditional neural networks.[29] In our model, we apply chemistry rules to explicitly determine the stereochemistry, including both chirality for atoms and cis/trans isomerism for double bonds, based on the predicted molecular graph and coordinates. For example, to predict chirality, we first find the bonds connected to each chiral center, infer their relative order by the predicted atom coordinates, and determine the chiral type explicitly (with RDKit's implementation). Since the atoms and bonds are easier to predict from the image, this rule-based approach recognizes stereochemistry more accurately than a vanilla neural model.

Abbreviated Structures. Another challenge in molecular structure recognition is the parsing of abbreviated structures. Chemists often simplify their drawings of molecules by using abbreviations or condensed formulas, such as "Me" for methyl, "Et" for ethyl, and "CHO" for formyl. The model cannot understand their underlying structures without external knowledge. The abbreviated structures can be treated as "superatoms" in the molecular graph, but additional processing is required if we want to derive the complete molecular structure. Both rule-based and machine learning models in previous works typically compile a list of common abbreviations and their functional group structures and substitute the superatoms at inference time. Given the combinatorial nature of abbreviations, this solution may not be sufficient. There is a large amount of combinations of element symbols and functional group abbreviations, e.g., "OMe", "CO_2Et", and "$P(O)(OEt)_2$". While they can all be considered as superatoms, it is infeasible to enumerate all possible combinations in a list.

Algorithm 1 Expand Abbreviated Structures

Input: G — molecular graph;
Input: f — abbreviated structure formula
Input: a_s — the superatom.
$L \leftarrow$ `expand_subscripts`(f) ▷ e.g. $CO_2CH_3 \rightarrow$ COOCHHH
$a_{\text{cur}} \leftarrow L$`[0]` ▷ current atom
G_s.`add_atom`(a_{cur})
for a **in** L`[1:]` **do**
  if a_{cur}.implicit_valence $= 0$ **then**
    return FAILURE
  end if
  G.`add_atom`(a)
  if a.valence $\leq a_{\text{cur}}$.implicit_valence **then**
    G_s.`add_bond`$(a_{\text{cur}}, a,$ order $= a$.valence$)$
  else
    G_s.`add_bond`$(a_{\text{cur}}, a,$ order $= a_{\text{cur}}$.implicit_valence$)$
    $a_{\text{cur}} \leftarrow a$
  end if
end for
G.`replace`(a_s, G_s)

We propose a more flexible method to parse abbreviated structures. We first split the superatom symbols into characters, such that our model is not limited to the prediction space of patterns in the training data and can generalize to unseen patterns. During inference, we design a greedy algorithm to expand the abbreviated structures. This task is nontrivial because the abbreviations do not follow a rigorous grammar (for example, a carboxyl group can be written as either "COOH" or "CO_2H"), and there is no deterministic way to parse them. In Algorithm 1, we derive the list of atoms and greedily connect them until their valences are full, based on assumptions of how these abbreviations are typically written to reflect the connections between atoms. We adopt this simple solution instead of using another learned model, as our algorithm has already addressed most abbreviations we observe and is extensible. More details are available in the Supporting Information.

Data. We train MolScribe on publicly available data, following previous works.[3,5] Our training data comes from two sources:

- **Synthetic data.** We collect 1 million molecules from the PubChem database[30] and automatically render their images using the Indigo toolkit.[31] Atom labels and bond types can be easily obtained from the toolkit, and we modify the source code to access atom coordinates. Previous works have combined other toolkits to generate a diverse set of training data,[4,32] but many of them do not provide atom coordinates. In this work, the molecules are randomly sampled from PubChem without special constraints. More advanced sampling strategies[33] could be used to ensure a diverse coverage of the chemical space. We encourage future work to analyze this problem.
- **Patent data.** We collect 680 K examples from patent grants released by the United States Patent and Trademark Office (USPTO),[34] which contain molecular images and structure labels. As some of our benchmarks were collected from the same source, we do not include the images from the same patents as those in the benchmarks. This dataset is noisy, and exact atom coordinates in the image are not available. We use the relative coordinates that are available in the MOLfiles and normalize them according to the image size. Details about data processing and examples of the noise in this data can be found in the Supporting Information.

Data Augmentation. We design data augmentation strategies for both molecules and images, so that the training data can cover diverse chemical patterns and drawing styles.

Functional Groups and R-Groups. Synthetic images generated by Indigo never contain functional group abbreviations or R-groups. We dynamically augment the molecules to cover such patterns. Figure 4 illustrates the augmentation process. For functional groups, we construct a list of common

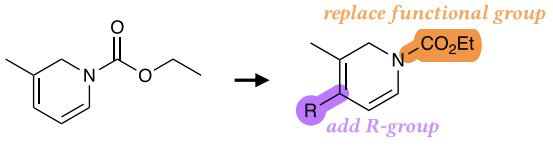


Figure 4. Examples of replacing a functional group with its abbreviation and adding an R-group during synthetic data generation.

https://doi.org/10.1021/acs.jcim.2c01480
J. Chem. Inf. Model. 2023, 63, 1925−1934

abbreviations, where each item consists of a SMARTS pattern for the functional group, an abbreviated label, and a substitution probability. During training, if a functional group exists in a molecule, we randomly replace it with the abbreviation according to the substitution probability. Specifically, the functional group branch is removed from the molecular graph, and a *superatom* with the abbreviation label is added. For R-groups, we also have a list of common R-group labels (R, R_1, R_2, R′, etc.) and randomly add R-groups as superatoms to the molecule. Furthermore, in order to generalize to abbreviations that are not covered in these lists, we add superatoms consisting of random characters. Although they are not necessarily chemically meaningful, this augmentation helps the model acquire the capability of optical character recognition (OCR), so that the model can recognize unseen labels. It does not hurt the model, as the molecular images should be valid at inference time. The functional group and R-group superatoms are associated with labels and coordinates, so our model predicts them in the same way as the other atoms.

The data augmentation allows our model to handle basic Markush structures,[35] e.g., an R-group attached to a particular atom or within a ring. However, positional variation (a substituent that could be attached to any atom within a ring) and frequency variation (brackets and variables indicating repetition of a substructure) are not covered. Such structures can not be represented in SMILES strings or MOLfiles; thus, we leave them as future work.

Drawing Styles. We use image augmentation to improve the robustness of our model with respect to drawing styles. Specifically, we vary the rendering options from Indigo when generating the synthetic data, such as font, bond width, bond length, etc. Furthermore, we apply random perturbations to the molecular images, such as rotation, padding, cropping, and Gaussian noise. Details about the augmentation strategy are available in the Supporting Information. The image augmentation strategy guarantees our model is trained with diverse image styles and quality, such that it can generalize better in the real world.

EXPERIMENTS

Experimental Setup. *Model Implementation.* Our image encoder is a Swin Transformer,[36] a state-of-the-art model in many computer vision tasks. We use the Swin-B model,[37] which has 88 M parameters in total and pretrained on ImageNet-22K. A recent work, SwinOCSR,[19] used the same image encoder as MolScribe. We resize the input image to 384 × 384 resolution for both training and inference. Our decoder is a 6-layer Transformer[38] with 8 attention heads, a hidden dimension of 256, and sinusoidal positional encoding. We apply dropout with probability 0.1. The bond predictor is a 2-layer feedforward network with ReLU activation on top of the decoder and has the same hidden dimension.

The model is trained by teacher forcing, i.e., the decoder is fed with the ground truth token at each step, and predicts the next step conditioned on the previous tokens. The bond predictor takes the hidden states of the decoder as input and predicts the bond between each pair of atoms. All modules of the model are fully differentiable and trained jointly. (Note that during training, the input to the bond predictor does not depend on the atom predictions.) We use a maximum learning rate of 4e-4 with a linear warmup for 5% steps and a cosine function decay. We use a batch size of 128 and train the model for 30 epochs. We use label smoothing with $\epsilon = 0.1$. The atom coordinates are converted to discrete tokens with $n_{bin} = 64$ in both the x and y directions. During inference, we use greedy decoding to generate the atoms and then predict the bonds. This procedure is a little different from that during training time, as we have to first autoregressively decode the atoms and then run the bond predictor.

Benchmarks. Table 1 lists the benchmarks in our evaluation. First, we construct two synthetic test sets to evaluate MolScribe's performance on in-domain and out-of-domain images. The in-domain dataset is generated by Indigo (same as training data), and the out-of-domain dataset is generated by ChemDraw. The two datasets share the same set of 5719 molecules. We also evaluate our model on five public benchmarks of realistic images: CLEF, JPO, UOB, USPTO,[6] and Staker.[3]

Table 1. Summary of the Benchmarks for Evaluation

Dataset	Source	No. of Images	% of chiral
Indigo	synthetic	5,719	20.2%
ChemDraw	synthetic	5,719	20.2%
CLEF	patent (US)	992	32.7%
JPO	patent (Japanese)	450	0%
UOB	catalog	5,740	0%
USPTO	patent (US)	5,719	20.2%
Staker	patent (US)	50,000	17.3%
ACS	journal article	331	19.3%

We create a new dataset with 331 molecular images collected from American Chemistry Society (ACS) Publications and manually annotate their SMILES strings. This dataset is a complement to existing benchmarks as they do not contain molecular images from journal articles, which are more diverse in terms of drawing style and use of abbreviations.

Evaluation Metric. We evaluate the model's recognition performance with exact matching accuracy as the evaluation metric, i.e., the prediction is considered correct only if the entire molecular graph structure matches the ground truth. This metric has been used by previous works[3−5] but with small differences. Specifically, we use RDKit to convert both prediction and ground truth into canonical SMILES, a unique molecular representation, and then compute the string exact match. Regarding stereochemistry, we require the prediction to match the tetrahedral chirality of the ground truth but ignore other forms of stereoisomerism (such as cis−trans) because such information is often not available in the ground truth. Regarding R-groups, as the their symbols are not allowed in SMILES, we replace them with wildcards (*) during evaluation. For numbered R-group "R_d" where d is a integer, we replace them with "[d*]". We make our evaluation script publicly available and encourage future work to follow the same metric for fair comparison. In the Supporting Information, we present other evaluation metrics, such as accuracy without considering chirality and Tanimoto similarity.

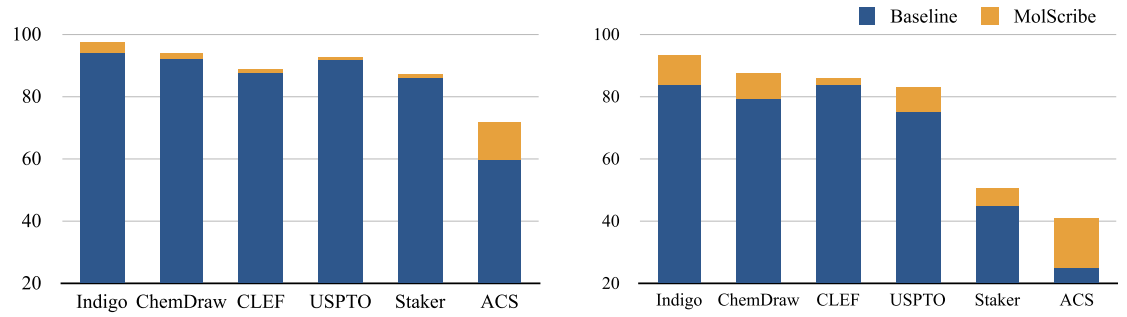
Compared Methods. We compare MolScribe with state-of-the-art molecular structure recognition methods. For direct comparison, we train an image-to-SMILES baseline model with the same data and the same encoder-decoder architecture. We also run existing systems for comparison, including rule-based MolVec[2] and OSRA,[1] and machine learning-based models Img2Mol,[4] DECIMER (version 2.1.0),[39] and SwinOCSR.[19] For systems that are not publicly available, such as MSE-

https://doi.org/10.1021/acs.jcim.2c01480
J. Chem. Inf. Model. 2023, 63, 1925−1934

Table 2. Molecule Structure Recognition Accuracy on Synthetic, Realistic, and Perturbed Benchmarks[a]

		Synthetic		Realistic						Perturbed			
Models		Indigo	ChemDraw	CLEF	JPO	UOB	USPTO	Staker	ACS	CLEF$_p$	UOB$_p$	USPTO$_p$	Staker$_p$
Rule-based	MolVec	95.4	87.9	82.8	67.8	80.6	88.4	0.8	47.4	43.7	74.5	29.7	5.0
	OSRA	95.0	87.3	84.6	55.3	78.5	87.4	0.0	55.3	11.5	68.3	4.0	4.6
Machine learning-based	Img2Mol[c]	58.9	46.4	18.3	16.4	68.7	26.3	17.0	23.0	21.1	74.9	29.7	51.7
	DECIMER	69.6	86.1	62.7	55.2	**88.2**	41.1	40.8	46.5	70.6	**87.3**	46.4	47.9
	SwinOCSR[d]	74.0	79.6	30.0	13.8	44.9	27.9	–	27.5	32.2	–	–	–
	MSE-DUDL[b]	–	–	–	–	–	–	77.0	–	–	–	–	–
	ChemGrapher[b]	–	–	–	–	70.6	–	–	–	–	–	–	–
	Image2Graph[b]	–	–	51.7	50.3	82.9	55.1	–	–	–	–	–	–
Ours	Baseline	94.1	92.2	87.4	74.8	**88.2**	91.5	86.1	59.8	88.0	87.1	91.4	**65.9**
	MolScribe	**97.5**	**93.8**	**88.9**	**76.2**	87.9	**92.6**	**86.9**	**71.9**	**90.4**	86.7	**92.5**	65.0

[a]Scores are exact matching accuracy in %. [b]Results from the original papers; – means not available. [c]Img2Mol does not predict chirality. Additional evaluation results with chirality ignored can be found in the Supporting Information. [d]We omit the results of SwimOCSR on the large datasets that take it more than 1 day to complete.



(a) Overall accuracy (b) Accuracy on chiral molecules

Figure 5. MolScribe explicitly determines chirality from the predicted graph and coordinates, thus improving recognition accuracy on chiral molecules.

DUDL,[3] ChemGrapher,[22] and Image2Graph,[5] we use the reported results in their papers.

RESULTS

Recognition Accuarcy. Table 2 shows the recognition accuracy of different models. Our model MolScribe achieves consistently higher scores than existing systems on most benchmarks, demonstrating its robust performance in molecular structure recognition. MolScribe outperforms the baseline image-to-SMILES model, validating the benefits of our graph generation formulation and the integration of symbolic chemistry knowledge. The JPO dataset involves Japanese characters, which are sometimes recognized by MolScribe as extra fragments. We apply a postprocessing step to keep the largest molecule if there are multiple fragments in the prediction. The rule-based systems, MolVec and OSRA, achieve decent performance on CLEF, JPO, UOB, and USPTO but degrade severely on Staker, in which the images are blurry and low-resolution. Compared with previous neural models such as MSE-DUDL,[3] ChemGrapher,[22] Img2Mol,[4] DECIMER,[39] and Image2Graph,[5] MolScribe achieves stronger performance despite being trained on less data. (We use 1.68 M examples in total, while MSE-DUDL uses 68M, ChemGrapher uses 1.5M, DECIMER uses 400 million, Image2Graph uses 7.1M, and Img2Mol uses 11M.) DECIMER performs slightly better than MolScribe on UOB. The images in this dataset are close to DECIMER's training distribution and relatively simple (no abbreviations, R-groups, or chirality), and thus a neural model trained with more data works well.

Following the setup of Clevert et al., we further evaluate the model on perturbed datasets with slight image rotation and shearing. MolScribe continues to exhibit robust performance on these datasets, outperforming other methods by large margins. MolVec and OSRA's performance drops significantly as compared to the performance on original images, showing that rule-based systems are sensitive to input perturbations. We note that our baseline model performs slightly better than MolScribe on UOB$_p$ and Staker$_p$. It is because these two datasets do not contain chiral molecules and the strength of MolScribe is concealed. The evaluation in Table 2 requires the chirality in the predicted molecules to be correct. It can be unfair for Img2Mol, which does not predict chirality. In the Supporting Information, we present additional evaluation results when chirality is ignored.

Chirality. Figure 5 compares the performance of our baseline image-to-SMILES model and MolScribe on chiral molecules. MolScribe predicts chirality more accurately than the baseline. As discussed in the Stereochemistry section, end-to-end neural models make mistakes more frequently on chiral molecules, as it requires geometric reasoning over the graph structure to determine the chirality. MolScribe, on the other hand, predicts all the atom coordinates and bond types and explicitly determines the chiral types. This design leads to significant improvement on chiral molecules.

Ablation Study. Table 3 shows the ablation study of our model design. In this experiment, we use a subset of our synthetic training data (200 K images) and train three model variants to validate the effects of several components of

https://doi.org/10.1021/acs.jcim.2c01480
J. Chem. Inf. Model. 2023, 63, 1925−1934

Table 3. Ablation Study of MolScribe[a]

	Indigo (in-domain)	ChemDraw (out-of-domain)
Baseline	88.7	82.5
- without data augmentation	86.0	57.2
MolScribe	95.9	89.7
- without data augmentation	91.4	72.1
- continuous coordinates	89.3	79.5
- remove nonatom tokens	95.3	89.4

[a]We train model variants with 200 K synthetic images generated by Indigo and evaluate their in-domain and out-of-domain performance.

MolScribe. First, the model trained without data augmentation achieves much worse performance on the out-of-domain ChemDraw dataset, which consists of images generated by a different software than the one used to generate training data (Indigo). This demonstrates that our data augmentation strategy is the key for the model to generalize to different styles. Second, we train a model that predicts continuous atom coordinates with an additional layer on top of the decoder as a regression task.[5] MolScribe performs better by converting the coordinates to discrete tokens and generating the atoms and coordinates in a sequence. Third, we remove the nonatom tokens (parentheses, digits, equal signs, etc.) from the decoder output, and the performance drops slightly. The nonatom tokens in SMILES indicate branching and connections, thus helping the decoder to better model the graph structure.

Figure 6 shows the contributions of the two datasets used to train MolScribe. The model trained with patent data alone performs well on CLEF, USPTO, and Staker, which are also collected from patent images, but performs unsatisfactorily in other domains. Our synthetic data and data augmentation strategies contribute to MolScribe's robust performance across different domains.

Model Confidence. MolScribe can provide the confidence of its atom and bond predictions. Figure 7 shows a qualitative analysis of the model-assigned probabilities of atoms and bonds and their correctness. Usually, lower probabilities are assigned to incorrect atoms/bonds, indicating the model is uncertain about certain parts of the image.

MolScribe also estimates how likely the entire molecular graph is correct by aggregating atom and bond confidence. We compute the molecule confidence with the average log probabilities of its atoms and bonds. Figure 8 shows the molecule-level confidence-accuracy and precision-recall curves. The precision-recall curve slopes downward, and model accuracy increases with the confidence. By setting thresholds on confidence scores, we can obtain more confident and accurate predictions.

Human Evaluation. MolScribe also has the advantage of being more interpretable due to the atom-level alignment between the predicted graph and the molecular image. Given the molecular graph predicted by our model, chemists can easily determine whether it is the same molecule as depicted in the image and make corrections if there are mistakes. To validate this function, we asked three students with a chemistry background to conduct an experiment of converting molecular images to SMILES strings. We compare three setups:

(1) Image-only: the students are given only the image and reconstruct the molecule;
(2) Predicted SMILES: the students are given the image and the predicted SMILES;
(3) Predicted graph: the students are given the image and the predicted molecular graph.

As it is difficult to manually write the SMILES, the students use ChemDraw to edit the molecule's structure. In (2) and (3), the students import the prediction into ChemDraw, judge whether the prediction is correct, and edit it if there are any mistakes. Setup (2) (predicted SMILES) does not preserve the molecular layout, making it harder to compare. Each setup contains 20 images of similarly sized molecules (34−36 atoms on average). We record the time taken by the students to make the judgment and edit.

Figure 9 shows the experimental results. In the image-only setup, it took these students 137 s on average to draw the depicted molecule in ChemDraw. (Note that these students are not power users of ChemDraw who can take advantage of the large number of keyboard shortcuts.) When the predicted SMILES is provided, the average time reduces to 39 s. When the predicted graph is provided, it further reduces to 20 s. The results clearly demonstrate the benefits of our graph generation method in visually inspecting the correctness of the model-generated structure.

CONCLUSION

In this paper, we propose MolScribe, an image-to-graph generation model for molecular structure recognition. MolScribe is built on an encoder-decoder architecture and jointly predicts atoms and bonds, along with their geometric layout. We design data augmentation strategies such that the model is robust to the domain shift and diverse patterns in real-world molecular images. Our model's flexibility allows us to incorporate symbolic chemistry constraints such as chirality verification and abbreviated structure expansion. Evaluations on both synthetic and realistic benchmarks show strong

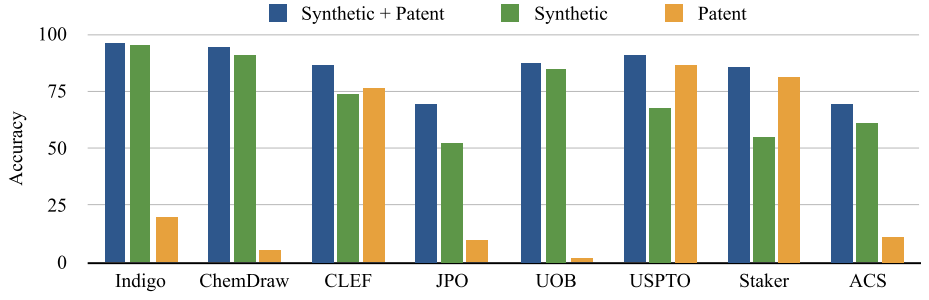


Figure 6. MolScribe is trained with both synthetic (PubChem) and patent (USPTO) data and achieves better performance than training on each of them alone.

https://doi.org/10.1021/acs.jcim.2c01480
J. Chem. Inf. Model. 2023, 63, 1925−1934

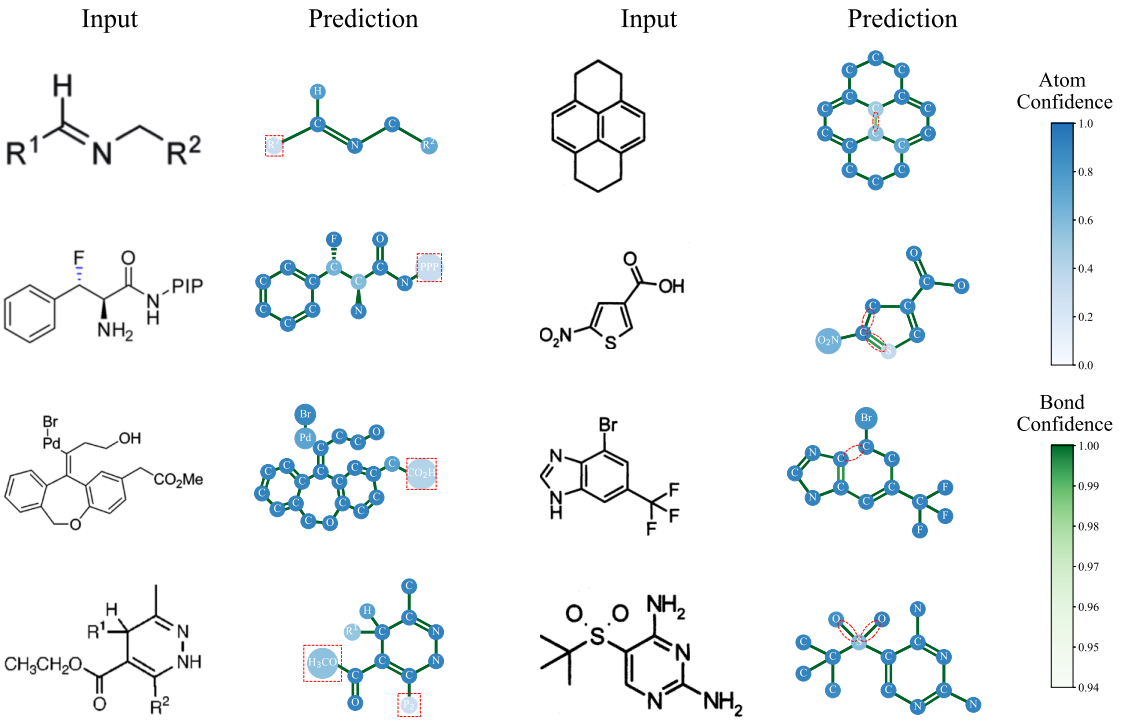
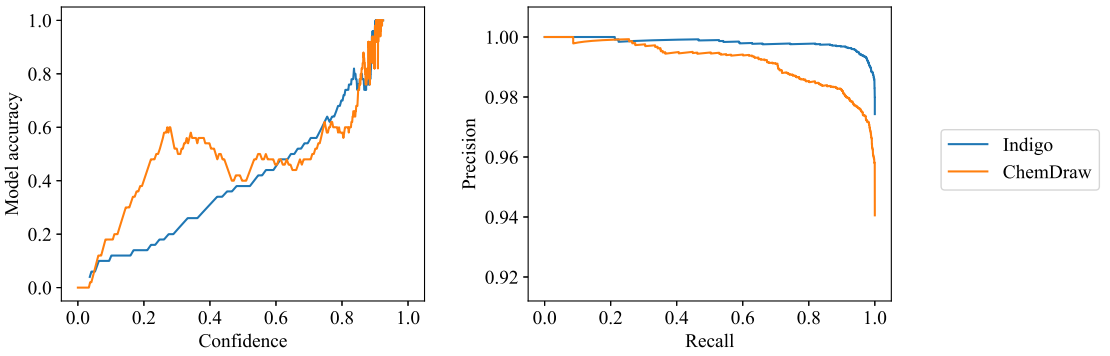
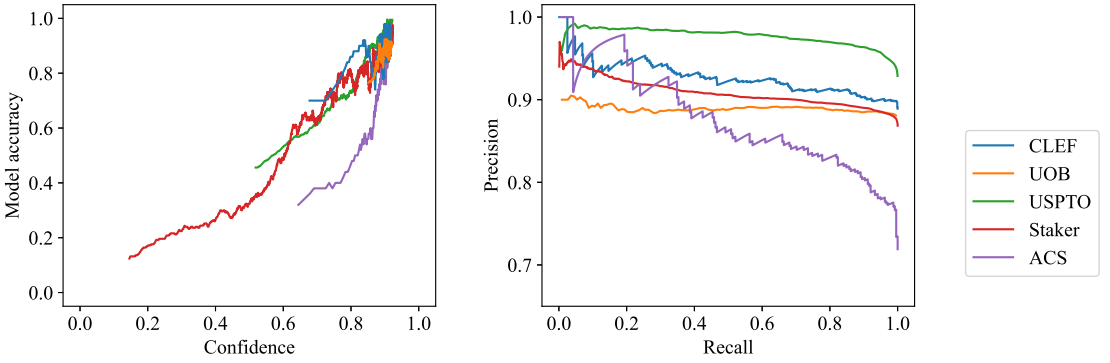


Figure 7. Examples of incorrectly predicted atoms/bonds and their confidence. Atom/bond confidence is represented by shade (darker means more confident), and mistakes are indicated with dashed boxes.



(a) Synthetic Benchmarks



(b) Realistic Benchmarks

Figure 8. MolScribe's confidence estimation on synthetic and realistic benchmarks. We show the correlation between model accuracy and confidence and the precision-recall curve by setting thresholds on confidence.

https://doi.org/10.1021/acs.jcim.2c01480
J. Chem. Inf. Model. 2023, 63, 1925−1934

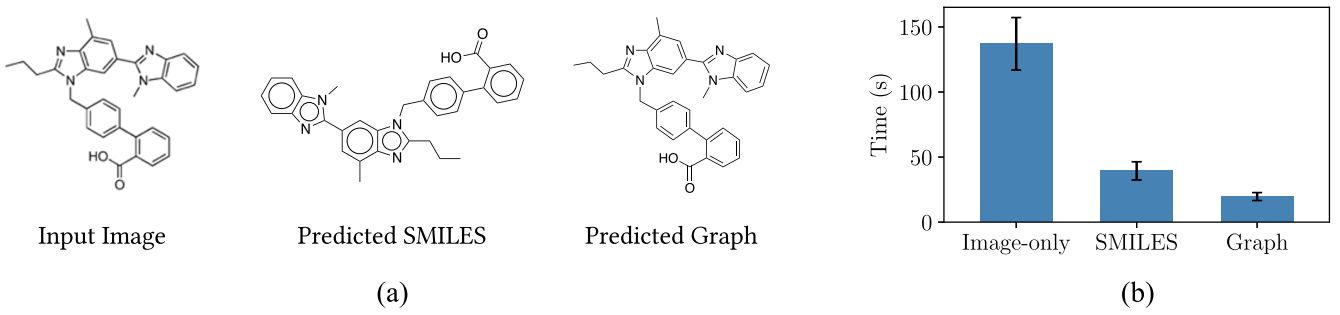


Figure 9. Human evaluation. (a) shows an example image with its predicted SMILES and molecular graph (imported into ChemDraw). (b) shows the average time taken by a student to convert a molecular image into SMILES, with or without the model prediction.

performance, surpassing existing rule-based and learned systems. Finally, our model produces more interpretable output, allowing chemists to easily examine the prediction and correct the mistakes. We release our code, data, and model for reproducibility and provide interfaces so that chemists can easily use MolScribe in their research.

The scope of this work is limited to the recognition of single-molecule images. MolScribe achieves strong and robust performance on this task, which is an essential component in building a chemistry information extraction system.[40−42] An important future direction is to extend MolScribe to handle hand-drawn molecules. MolScribe can recognize basic Markush structures with R-groups, but more complicated cases, such as the R-group that could be attached to any atom of a ring or a variable number of repetitions of a substructure, are not covered. Such cases cannot be represented as SMILES strings or MOLfiles, and future work may need to design a specialized format to handle Markush structures.

DATA AND SOFTWARE AVAILABILITY

Our code, data, and model checkpoints are publicly available at https://github.com/thomas0809/MolScribe. We have also developed a web interface for MolScribe: https://huggingface.co/spaces/yujieq/MolScribe. More details about the data sources can be found in the Supporting Information.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c01480.

Data processing details, example molecular images in the benchmarks, implementation details of data augmentation, the algorithm for expanding abbreviations, additional evaluation results, and example predictions on hand-drawn molecules (PDF)

AUTHOR INFORMATION

Corresponding Authors

Yujie Qian − *Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0003-3747-4552; Email: yujieq@csail.mit.edu

Regina Barzilay − *Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* Email: regina@csail.mit.edu

Authors

Jiang Guo − *Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-9816-805X

Zhengkai Tu − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*

Zhening Li − *Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States*

Connor W. Coley − *Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States;* orcid.org/0000-0002-8271-8723

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c01480

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The author thanks the members of Regina Barzilay's Group at MIT CSAIL for helpful discussions and feedback. This work was supported by the DARPA Accelerated Molecular Discovery (AMD) program under contract HR00111920025 and the Machine Learning for Pharmaceutical Discovery and Synthesis Consortium (MLPDS).

REFERENCES

(1) Filippov, I. V.; Nicklaus, M. C. Optical structure recognition software to recover chemical information: OSRA, an open source solution. *J. Chem. Inf. Model.* **2009**, *49*, 740.

(2) Peryea, T.; Katzel, D.; Zhao, T.; Southall, N.; Nguyen, D.-T. MOLVEC: Open source library for chemical structure recognition. *Abstracts of papers of the American Chemical Society*; 2019.

(3) Staker, J.; Marshall, K.; Abel, R.; McQuaw, C. M. Molecular Structure Extraction from Documents Using Deep Learning. *J. Chem. Inf. Model.* **2019**, *59*, 1017−1029.

(4) Clevert, D.-A.; Le, T.; Winter, R.; Montanari, F. Img2Mol − accurate SMILES recognition from molecular graphical depictions. *Chem. Sci.* **2021**, *12*, 14174−14181.

(5) Yoo, S.; Kwon, O.; Lee, H. Image-to-Graph Transformers for Chemical Structure Recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23−27 May 2022*; 2022; pp 3393−3397, DOI: 10.1109/ICASSP43922.2022.9746088.

(6) Rajan, K.; Brinkhaus, H. O.; Zielesny, A.; Steinbeck, C. A review of optical chemical structure recognition tools. *J. Cheminf.* **2020**, *12*, 60.

(7) McDaniel, J. R.; Balmuth, J. R. Kekule: OCR-optical chemical (structure) recognition. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 373−378.
(8) Casey, R. G.; Boyer, S.; Healey, P.; Miller, A.; Oudot, B.; Zilles, K. Optical recognition of chemical graphics. *2nd International Conference Document Analysis and Recognition, ICDAR '93, October 20−22, 1993, Tsukuba City, Japan*; 1993; pp 627−631, DOI: 10.1109/ICDAR.1993.395658.
(9) Ibison, P.; Jacquot, M.; Kam, F.; Neville, A.; Simpson, R. W.; Tonnelier, C.; Venczel, T.; Johnson, A. P. Chemical literature data extraction: the CLiDE Project. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 338−344.
(10) Valko, A. T.; Johnson, A. P. CLiDE Pro: the latest generation of CLiDE, a tool for optical chemical structure recognition. *J. Chem. Inf. Model.* **2009**, *49*, 780−787.
(11) Park, J.; Rosania, G. R.; Shedden, K. A.; Nguyen, M.; Lyu, N.; Saitou, K. Automated extraction of chemical structure information from digital raster images. *Chem. Cent. J.* **2009**, *3*, 4.
(12) Sadawi, N. M.; Sexton, A. P.; Sorge, V. Chemical structure recognition: a rule-based approach. Document Recognition and Retrieval XIX, part of the *IS&T-SPIE Electronic Imaging Symposium, Burlingame, California, USA, January 25−26, 2012, Proceedings*; 2012; p 82970E, DOI: 10.1117/12.912185.
(13) Algorri, M.-E.; Zimmermann, M.; Friedrich, C. M.; Akle, S.; Hofmann-Apitius, M. Reconstruction of Chemical Molecules from Images. *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*; 2007; pp 4609−4612, DOI: 10.1109/IEMBS.2007.4353366.
(14) Frasconi, P.; Gabbrielli, F.; Lippi, M.; Marinai, S. Markov logic networks for optical chemical structure recognition. *J. Chem. Inf. Model.* **2014**, *54*, 2380−2390.
(15) Smolov, V.; Zentsev, F.; Rybalkin, M. Imago: Open-Source Toolkit for 2D Chemical Structure Image Recognition. *TREC*; 2011.
(16) Rajan, K.; Zielesny, A.; Steinbeck, C. DECIMER: towards deep learning for chemical image recognition. *J. Cheminf.* **2020**, *12*, 65.
(17) Rajan, K.; Zielesny, A.; Steinbeck, C. DECIMER 1.0: deep learning for chemical image recognition using transformers. *J. Cheminf.* **2021**, *13*, 61.
(18) Khokhlov, I.; Krasnov, L.; Fedorov, M. V.; Sosnin, S. Image2SMILES: Transformer-Based Molecular Optical Recognition Engine. *Chem.: Methods* **2022**, *2*, No. e202100069.
(19) Xu, Z.; Li, J.; Yang, Z.; Li, S.; Li, H. SwinOCSR: end-to-end optical chemical structure recognition using a Swin Transformer. *J. Cheminf.* **2022**, *14*, 41.
(20) Weir, H.; Thompson, K.; Woodward, A.; Choi, B.; Braun, A.; Martínez, T. J. ChemPix: automated recognition of hand-drawn hydrocarbon structures using deep learning. *Chem. Sci.* **2021**, *12*, 10622−10633.
(21) Brinkhaus, H. O.; Zielesny, A.; Steinbeck, C.; Rajan, K. DECIMER—hand-drawn molecule images dataset. *J. Cheminf.* **2022**, *14*, 36.
(22) Oldenhof, M.; Arany, A.; Moreau, Y.; Simm, J. ChemGrapher: optical graph recognition of chemical compounds by deep learning. *J. Chem. Inf. Model.* **2020**, *60*, 4506−4517.
(23) Xu, Y.; et al. MolMiner: You Only Look Once for Chemical Structure Recognition. *J. Chem. Inf. Model.* **2022**, *62*, 5321.
(24) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(25) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.
(26) Chen, T.; Saxena, S.; Li, L.; Fleet, D. J.; Hinton, G. Pix2seq: A Language Modeling Framework for Object Detection. *International Conference on Learning Representations*; 2021.
(27) SmilesPE atomwise_tokenizer. https://xinhaoli74.github.io/SmilesPE/pretokenizer/#atomwise_tokenizer (accessed 2022-10-01).
(28) Landrum, G. *RDKit: Open-Source Cheminformatics Software*; 2016.
(29) Battaglia, P. W.; Hamrick, J. B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R., et al. Relational inductive biases, deep learning, and graph networks. 2018, arXiv:1806.01261. *arXiv preprint*. https://arxiv.org/abs/1806.01261 (accessed 2023-20-23).
(30) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound databases. *Nucleic Acids Res.* **2016**, *44*, D1202−D1213.
(31) Pavlov, D.; Rybalkin, M.; Karulin, B.; Kozhevnikov, M.; Savelyev, A.; Churinov, A. Indigo: universal cheminformatics API. *J. Cheminf.* **2011**, *3*, P4.
(32) Brinkhaus, H. O.; Rajan, K.; Zielesny, A.; Steinbeck, C. RanDepict: Random chemical structure depiction generator. *J. Cheminf.* **2022**, *14*, 31.
(33) Ashton, M.; Barnard, J.; Casset, F.; Charlton, M.; Downs, G.; Gorse, D.; Holliday, J.; Lahana, R.; Willett, P. Identification of Diverse Database Subsets using Property-Based and Fragment-Based Molecular Descriptions. *Quant. Struct.-Act. Relat.* **2002**, *21*, 598−604.
(34) United States Patent and Trademark Office *Grant Red Book*. https://bulkdata.uspto.gov (accessed 2021-12-31).
(35) Lynch, M. F.; Holliday, J. D. The sheffield generic structures project a retrospective review. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 930−936.
(36) Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; 2021; pp 10012−10022, DOI: 10.1109/ICCV48922.2021.00986.
(37) Swin Transformer. https://github.com/microsoft/Swin-Transformer (accessed 2022-10-01).
(38) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*; 2017; Vol. *30*.
(39) DECIMER Image Transformer V2. https://github.com/Kohulan/DECIMER-Image_Transformer (accessed 2023-1-20).
(40) Beard, E. J.; Cole, J. M. ChemSchematicResolver: A toolkit to decode 2D chemical diagrams with labels and R-groups into annotated chemical named entities. *J. Chem. Inf. Model.* **2020**, *60*, 2059−2072.
(41) Wilary, D. M.; Cole, J. M. ReactionDataExtractor: A Tool for Automated Extraction of Information from Chemical Reaction Schemes. *J. Chem. Inf. Model.* **2021**, *61*, 4962−4974.
(42) Guo, J.; Ibanez-Lopez, A. S.; Gao, H.; Quach, V.; Coley, C. W.; Jensen, K. F.; Barzilay, R. Automated Chemical Reaction Extraction from Scientific Literature. *J. Chem. Inf. Model* **2022**, *62*, 2035.

NOTE ADDED AFTER ASAP PUBLICATION

This paper was published ASAP on March 27, 2023 with errors in the footnotes of Table 2. The corrected version was reposted on March 27, 2023.

https://doi.org/10.1021/acs.jcim.2c01480
J. Chem. Inf. Model. 2023, 63, 1925−1934