SMARE — Structure Matching and Recognition Engine for Hand-Drawn Chemical Formulas

Mitra Purandare[1], Tobias Rothlin[1], Frieder Loch[1], Johannes Huwer[2,3], and Lars-Jochen Thoms[2,3]

[1] Department of Computer Science, Eastern Switzerland University of Applied Sciences, Rapperswil, Switzerland
{mitra.purandare,tobias.rothlin,frieder.loch}@ost.ch
[2] Chair of Science Education, Thurgau University of Teacher Education, Kreuzlingen, Switzerland
{johannes.huwer,lars.thoms}@phtg.ch
[3] Chair of Science Education, University of Konstanz, Konstanz, Germany

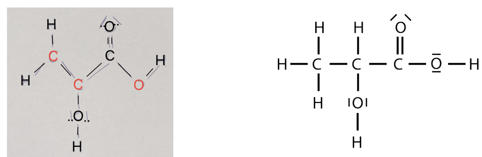
Abstract. Expressing chemical compounds in various representations is challenging. This is especially true for novices, since the task demands extensive domain-specific knowledge and spatial visualization skills. To address this challenge, we propose SMARE, our Structure Matching and Recognition Engine for chemical formulas. It interprets hand-drawn molecular structures and identifies and highlights errors and thereby is a fundamental component of educational applications. SMARE leverages a YOLO (You Only Look Once) model to recognize fundamental entities in chemical structures such as atoms or bonds. The dataset for training, validating, and testing the model consists of 1,844 hand-drawn chemical molecular images collected from students. SMARE processes the identified entities to construct an abstract molecular graph. The engine compares the identified molecular graph against a database of known molecules and detects errors such as incorrect bonding, and valency violations. Our fine-tuned YOLO model achieves an accuracy of 93.8 % in recognizing chemical entities in hand-drawn molecules. SMARE was tested on 7,909 hand-drawn chemical structures from 519 school students under real-world conditions, successfully identifying numerous errors in their chemical drawings. This demonstrates effectiveness of SMARE as a powerful and practical tool for chemistry education.

Keywords: Chemical education · Image recognition · YOLO · Error analysis · Spatial visualization.

1 Introduction

Chemical compounds can be represented in various ways, such as molecular formulas, skeletal formulas, or perspective drawings, each emphasizing different aspects of chemical bonding. Understanding the representations, choosing the appropriate one for a given purpose, and writing chemical formulas require specific skills. While experts can use and mentally translate between these formats, novices must first learn the underlying rules.

Translating between representations can be practiced with a dedicated application [19]. Such an application needs an algorithm to identify hand-drawn chemical formulas and detect errors (see Fig. 1). The task is challenging due to the diversity of student illustrations and handwriting styles. This paper presents a system that identifies and analyzes hand-drawn molecules to detect and highlight errors when drawing Lewis structures.



(a) Hand-drawn (b) Correct

Fig. 1: (a) Hand-drawn Lewis structure of R-lactic acid with overlay of recognized entities and errors (red, left to right): Missing bonds and (H) atoms at the first and second carbon atoms, two missing lone electron pairs at the oxygen atom. (b) Correct drawing of the R-lactic acid on the right.

Fig. 2 depicts our solution. The contributions are threefold: (1) a dataset of 1,844 hand-drawn chemical molecular images collected from students in Switzerland and Germany [20], (2) YOLOv8 model pipeline to detect molecular entities, which runs locally on consumer devices without cloud processing, and (3) an error detection algorithm to identify errors in the hand-drawn chemical formulas.

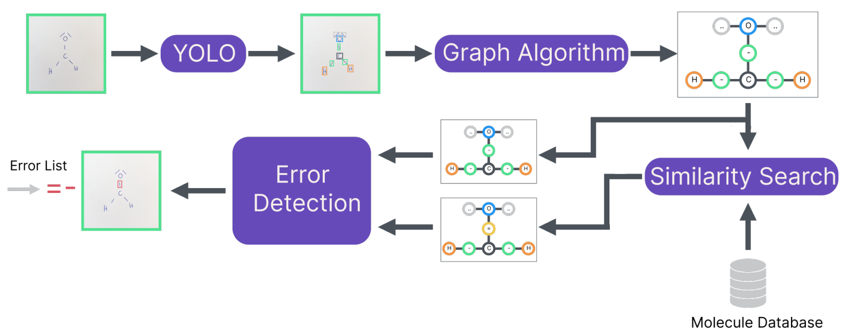


Fig. 2: SMARE for hand-written molecular recognition and error analysis.

2 Related Work

Many students struggle to reason about abstract representations of scientific concepts that require a high level of spatial visualization [14]. While experts can mentally translate between multiple representations [6,8,15], novices must first develop the necessary mental models [1,15,16], often supported by interactive applications [7,21].

Deep neural networks can recognize printed or handwritten chemical structures [2,9,11,12,23,24,4,25], and even take the context into account [10]. Large databases with (partly already decoded) images of chemical formulas are available (e.g., CASIA-CSDB [3]). Most of the above methods convert a molecular image to SMILES [22] or InChI [5] format which assume correct molecular structures. They lack the ability to encode errors like missing atoms, incorrect bonds, or lone pair errors explicitly. Since SMARE focuses on identifying and correcting structural errors, a molecular graph representation is necessary.

Molscribe[11] and Molminer[23] generate molecular graphs from images but assume error-free input. Molscribe employs a complex encoder-decoder architecture trained on correct drawings. Molminer[23] employs YOLOv5 for distance-based graph construction without error detection.

3 SMARE: Methods and Approach

SMARE must recognize hand-drawn chemical formulas despite variations in handwriting and overlaps between atoms and bonds. Errors occur when the detected structure deviates from the expected molecule. We apply YOLO (You Only Look Once) for real-time recognition. Our approach has three steps: Object Detection, Graph Construction, and Graph Matching with Error Detection.

SMARE is designed for handheld devices. Local processing enhances privacy by not having to transmit images for analysis and ensuring independence from internet connection. This limits AI model and error detection choices, requiring efficiency. To maintain a smooth user experience, SMARE imposes an upper limit of 50 ms for end-to-end processing, from image capture to error display.

Object Detection We train a YOLOv8 model to detect individual chemical components in an image, such as atoms, bonds, or free electrons. Our full training dataset comprising 1,844 hand-drawn chemical molecules is openly available [20].

These curated examples reflect hand-writing variations and practical learning scenarios. Additionally, 32 examples were incorporated into the training data to address images devoid of molecular structures such as scratched-out drawings. Data augmentation such as rotation, scaling, noise addition, and contrast adjustments are performed in Roboflow. The model is fine-tuned using the usual 80-20 training-validation split. The natural imbalance of elements cannot be mitigated as it reflects the inherent bias present in real molecules.

Graph Construction The output of the YOLO Model comprises all relevant bounding boxes and their class designations. The graph construction method selects a random bond as the initial point of reference and establishes an edge between its two nearest atoms. This process is applied iteratively to all the bonds to construct the core structure of the graph. The free electrons and charges are then associated with their proximal atoms.

Graph Matching and Error Detection If the reference molecule is known, no matching is required but the reference still needs to be aligned with the hand-drawn structure for error detection. Aligning the two graphs even if they differ in orientation and positioning is crucial. Fig. 3 shows two molecules to be compared. The hand-written molecules in our dataset are small, averaging only 14 nodes with a maximum of 39 nodes, justifying a brute-force approach shown in Fig. 4. A random starting point is selected from the reference graph, and every atom in the hand-drawn graph is tested as a potential starting point. The alignment with the fewest errors is chosen. Due to molecular symmetry, multiple optimal alignments may exist. The alignment with the minimal number of errors is selected as a chosen alignment for displaying errors (see Fig. 5).

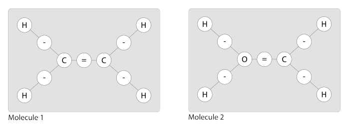
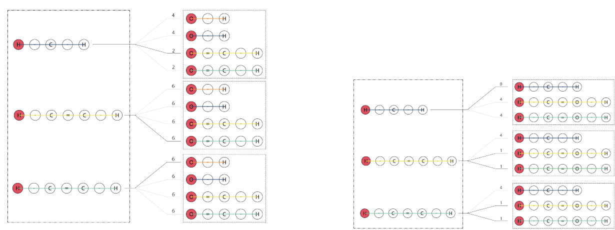


Fig. 3: Compare molecular graphs: hand-drawn (left) vs reference (right).



(a) Alignment 1 (b) Alignment 2

Fig. 4: Different starting points in Fig. 3 result in different path alignments. Paths in the hand-drawn molecule starting with hydrogen (H) are shown on the left. Paths in the reference molecule with different starting points are shown on the right. Each path in the hand-drawn molecule is compared to the paths of the reference graphs. Error scores are mismatches on each path.

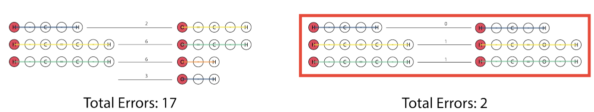


Fig. 5: From the alignments in Fig. 4, the paths with least errors are chosen and the final error count is generated. The alignment with the least count is chosen.

If the reference molecule is not known a priori, i.e., the learner scans an arbitrary hand-written molecule, a matching reference needs to be identified. The unknown reference molecule must be carefully identified resulting in a minimal number of errors to not overwhelm the learner. A simple histogram matching algorithm can be employed to reduce the number of reference candidates. Comparing the number of atoms and bond types in the molecules eliminates positional information. A molecule's histogram need not always be unique. To address this issue, error detection is applied to all the reference candidates. The reference with the least error score is chosen to display errors.

4 Evaluation

To evaluate the generalization capability of the YOLO model, a test set of 100 hand-drawn molecules created by an independent individual is used.

Laboratory testing The confusion matrix reveals that the model detected the correct atoms with an accuracy of above 90 %. There are minor instances of incorrect atom classification. The majority of the remaining 10 % errors are associated with the background class, indicating that the model detects elements that are not actually present in the image. Users rarely observe this phenomenon, because these extraneous elements get discarded during graph-construction.

Tables 1 and 2 show the confusion matrices for electrons, charges, bonds, and atoms, showing high overall accuracy. The non-zero entries in the last column in Table 2 show that background may get misidentified as an element. The *Negative* label, representing a negative charge, is the least frequent in the training data and also has the worst test performance. The training data distribution explains why underrepresented classes perform worse. Despite the issues, models exhibit high weighted average accuracy of 93.8 %. The weights reflect the prevalence of chemical entities in the training data.

Table 1: Normalized confusion matrices for electrons, charges, and bonds

Electrons ground truth column-wise and prediction row-wise			
	`FreeElectronPair`	`SingleFreeElectron`	`Background`
`FreeElectronPair`	0.99	0.09	0.67
`SingleFreeElectron`	0.00	0.82	0.33
`Background`	0.01	0.09	0.00

Charges ground truth column-wise and prediction row-wise			
	`Negative`	`Positive`	`Background`
`Negative`	0.78	0.00	0.00
`Positive`	0.11	1.00	1.00
`Background`	0.11	0.00	0.00

Bonds ground truth column-wise and prediction row-wise						
	`DashedTri.`	`DoubleB.`	`FullTri.`	`SingleB.`	`TripleB.`	`Background`
`DashedTri.`	1.00	0.00	0.00	0.00	0.00	0.05
`DoubleB.`	0.00	0.98	0.00	0.00	0.00	0.00
`FullTri.`	0.00	0.00	0.83	0.00	0.04	0.05
`SingleB.`	0.00	0.00	0.00	0.98	0.00	0.65
`TripleB.`	0.00	0.00	0.00	0.00	0.92	0.25
`Background`	0.00	0.02	0.17	0.02	0.04	0.00

Field testing To assess SMARE under real-life conditions, 519 school students tested SMARE via the iOS app OrChemSTAR [17], generating 7,909 hand-drawn scans [18]. Without prior knowledge of drawing structural formulas or organic chemistry, the students learned only through the app. Therefore, Table 3 shows a

Table 2: Normalized Confusion Matrix for Atoms (BG = Background).

	Br	C	Cl	F	H	I	N	O	S	BG
Br	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13
C	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.02	0.08	0.15
Cl	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.04
F	0.00	0.00	0.00	0.92	0.00	0.00	0.00	0.00	0.00	0.02
H	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.28
I	0.00	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.11
N	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.02
O	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.98	0.00	0.23
S	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.92	0.02
BG	0.10	0.00	0.00	0.00	0.01	0.04	0.00	0.00	0.00	0.00

high error rate. The error detection of the application helped the students refine their structural drawing skills in further exercises.

Table 3: Frequency (in %) of classes of errors in 7,909 scans by school students.

Class	*Freq.*	Class	*Freq.*
`Octet rule violated`	50.21	`Missing Bond`	49.18
`Missing Atom`	49.18	`Incorrect Atom`	28.68
`Incorrect Bond`	22.49	`Excess Bond`	21.76
`Excess Atom`	21.58	`Excess Electron`	7.75
`Missing Electron`	7.50	`Detected Natta Projection`	2.52
`Excess Charge`	1.01	`Missing Charge`	0.10
`No Errors found`	27.42		

5 Conclusion and Future Work

SMARE analyzes hand-drawn molecules accurately, achieving 93.8 % accuracy in detecting chemical entities. Its robust architecture allows integration of new molecules without retraining. Evaluations with many students confirm its effectiveness to detect errors in chemical drawings in real-world conditions.

While graph construction from YOLO output is efficient, moving beyond a greedy proximity-based algorithm could improve reliability. The error detection algorithm performs well on the molecule sizes that are typically encountered in school exercises. However, the scalability to larger molecules is limited. The role of maximum subgraph isomorphism [13] in error detection requires further exploration for error detection in larger structures.

Acknowledgements This work is part of the project "OrChemSTAR - Organic Chemistry Science Teaching and Learning with Augmented Reality", funded by the Swiss National Science Foundation (CRSK-1_221108).

Disclosure Statement The authors declare that they have no competing interests to declare that are relevant to the content of this article.

References

1. Ainsworth, S.: DeFT: A conceptual framework for considering learning with multiple representations. Learning and Instruction **16**(3), 183–198 (Jun 2006). https://doi.org/10.1016/j.learninstruc.2006.03.001
2. Bukhari, S.S., Iftikhar, Z., Dengel, A.: Chemical Structure Recognition (CSR) System: Automatic Analysis of 2D Chemical Structures in Document Images. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1262–1267. IEEE, Sydney, NSW, Australia (Sep 2019). https://doi.org/10.1109/ICDAR.2019.00-41
3. Ding, L., Zhao, M., Yin, F., Zeng, S., Liu, C.L.: A Large-Scale Database for Chemical Structure Recognition and Preliminary Evaluation. In: 2022 26th International Conference on Pattern Recognition (ICPR). pp. 1464–1470. IEEE, Montreal, QC, Canada (Aug 2022). https://doi.org/10.1109/ICPR56361.2022.9956654
4. Hagag, A., Omara, I., Alfarra, A.N.K., Mekawy, F.: Handwritten Chemical Formulas Classification Model Using Deep Transfer Convolutional Neural Networks. In: 2021 International Conference on Electronic Engineering (ICEEM). pp. 1–6. IEEE, Menouf, Egypt (Jul 2021). https://doi.org/10.1109/ICEEM52022.2021.9480627
5. Heller, S.R., McNaught, A., Pletnev, I., Stein, S., Tchekhovskoi, D.: Inchi, the iupac international chemical identifier. Journal of Cheminformatics **7**, 23 (2015). https://doi.org/10.1186/s13321-015-0068-4
6. Kohl, P.B., Finkelstein, N.D.: Patterns of multiple representation use by experts and novices during physics problem solving. Physical Review Special Topics - Physics Education Research **4**(1), 010111 (Jun 2008). https://doi.org/10.1103/PhysRevSTPER.4.010111
7. Kozma, R., Russell, J.: Students becoming chemists: Developing representationl competence. In: Gilbert, J.K. (ed.) Visualization in Science Education, pp. 121–145. Springer Netherlands, Dordrecht (2005)
8. Kozma, R.B.: Use of multiple representations by experts and novices. In: Van Meter, P., List, A., Lombardi, D., Kendeou, P. (eds.) Handbook of Learning from Multiple Representations and Perspectives. Routledge, New York, NY : Routledge, 2020., 1 edn. (Mar 2020). https://doi.org/10.4324/9780429443961
9. Li, Y., Chen, G., Li, X.: Automated Recognition of Chemical Molecule Images Based on an Improved TNT Model. Applied Sciences **12**(2), 680 (Jan 2022). https://doi.org/10.3390/app12020680
10. Musazade, F., Jamalova, N., Hasanov, J.: Review of techniques and models used in optical chemical structure recognition in images and scanned documents. Journal of Cheminformatics **14**(1), 61 (Sep 2022). https://doi.org/10.1186/s13321-022-00642-3
11. Qian, Y., Guo, J., Tu, Z., Li, Z., Coley, C.W., Barzilay, R.: Molscribe: Robust molecular structure recognition with image-to-graph generation. Journal of Chemical Information and Modeling **63**(7), 1925–1934 (2023)
12. Rajan, K., Zielesny, A., Steinbeck, C.: DECIMER: towards deep learning for chemical image recognition. Journal of Cheminformatics **12**(1), 65 (Dec 2020). https://doi.org/10.1186/s13321-020-00469-w
13. Raymond, J.W., Willett, P.: Maximum common subgraph isomorphism algorithms for the matching of chemical structures. Journal of Computer-Aided Molecular Design **16**(7), 521–533 (2002). https://doi.org/10.1023/A:1021271615909
14. Sahin, D., Yilmaz, R.M.: The effect of augmented reality technology on middle school students' achievements and attitudes towards science education. Computers

& Education **144**, 103710 (2020). https://doi.org/https://doi.org/10.1016/j.compedu.2019.103710
15. Schnotz, W., Bannert, M.: Construction and interference in learning from multiple representation. Learning and Instruction **13**(2), 141–156 (Apr 2003). https://doi.org/10.1016/S0959-4752(02)00017-8, https://linkinghub.elsevier.com/retrieve/pii/S0959475202000178
16. Sunyono, S., Leny, Y., Muslimin, I.: Supporting students in learning with multiple representation to improve student mental models on atomic structure concepts. Science Education International **26**(2), 104–125 (2015), publisher: International Council of Associations for Science Education (ICASE)
17. Thoms, L.J.: OrChemSTAR [ios application]. https://apps.apple.com/de/app/orchemstar/id6636548074 (2024)
18. Thoms, L.J.: Hand-drawn chemical formulas captured during field testing of the orchemstar app with and without annotations gained using the structure matching and recognition engine smare (Apr 2025). https://doi.org/10.5281/zenodo.15191840
19. Thoms, L.J., Huwer, J.: Das projekt orchemstar – strukturformeln durch augmented reality zeichnen lernen [the orchemstar project – learning to draw structural formulas using augmented reality]. In: Huwer, J., Wilke, T., Banerji, A. (eds.) Progress in Digitalisation in Chemistry Education 2024, pp. 113–118. Waxmann (2025)
20. Thoms, L.J., Rothlin, T.: Hand-drawn chemical formulas used to train the structure matching and recognition engine smare (Jun 2024). https://doi.org/10.5281/zenodo.15191704
21. van der Meij, J., de Jong, T.: Supporting students' learning with multiple representations in a dynamic simulation-based learning environment. Learning and Instruction **16**(3), 199–212 (2006)
22. Weininger, D.: 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). https://doi.org/10.1021/ci00057a005
23. Xu, Y., Xiao, J., Chou, C.H., Zhang, J., Zhu, J., Hu, Q., Li, H., Han, N., Liu, B., Zhang, S., Han, J., Zhang, Z., Zhang, S., Zhang, W., Lai, L., Pei, J.: Molminer: You only look once for chemical structure recognition. Journal of Chemical Information and Modeling **62**(22), 5321–5328 (2022)
24. Xu, Z., Li, J., Yang, Z., Li, S., Li, H.: SwinOCSR: end-to-end optical chemical structure recognition using a Swin Transformer. Journal of Cheminformatics **14**(1), 41 (Dec 2022). https://doi.org/10.1186/s13321-022-00624-5
25. Zhang, T., Wang, Y., Jin, X., Gu, Z., Zhang, X., He, B.: An Auto-Grading Oriented Approach for Off-Line Handwritten Organic Cyclic Compound Structure Formulas Recognition. Computer Modeling in Engineering & Sciences **135**(3), 2267–2285 (2023). https://doi.org/10.32604/cmes.2023.023229