Supporting Information for

Computer Vision for High-Throughput Materials Synthesis: A Tutorial for Experimentalists

Madeleine A. Gaidimas,[1,†] Abhijoy Mandal,[2,†] Pan Chen,[2] Shi Xuan Leong,[3,4] Gyu-Hee Kim,[1] Akshay Talekar,[5] Kent O. Kirlikovali,[1] Kourosh Darvish,[2,6] Omar K. Farha,[*,1,7,] Varinia Bernales,[*,3,5,6,] and Alán Aspuru-Guzik[*,2,3,6,8,9,10,11,12]

[1] Department of Chemistry and International Institute for Nanotechnology, Northwestern University, Evanston, IL 60208, United States

[2] Department of Computer Science, University of Toronto, Toronto, ON M5S 2E4, Canada

[3] Department of Chemistry, University of Toronto, Toronto, ON M5S 2E4, Canada

[4] School of Chemistry, Chemical Engineering and Biotechnology, Nanyang Technological University, Singapore 637371

[5] Materials Discovery Research Institute, UL Research Institutes, Skokie, IL 60077, United States

[6] Acceleration Consortium, Toronto, ON M5S 3H6, Canada

[7] Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208, United States

[8] Vector Institute for Artificial Intelligence, Toronto, ON M5G 1M1, Canada

[9] Department of Chemical Engineering and Applied Chemistry, University of Toronto, Toronto, ON M5S 3E5, Canada

[10] Department of Materials Science and Engineering, University of Toronto, Toronto, ON M5S 3E4, Canada

[11] Senior Fellow, Canadian Institute for Advanced Research (CIFAR), Toronto, ON M5G 1M1, Canada

[12] NVIDIA, Toronto, ON M5V 1K4, Canada

[†] Equal contribution

[*]Correspondence: o-farha@northwestern.edu, varinia@bernales.org, and alan@aspuru.com

Table of Contents

S1. Hardware Information

S1.1 Automated Synthesis Platform

Automated syntheses were performed in a Chemspeed Flex platform equipped with liquid handling, solid handling, vial capping, shaking, and heating capabilities. MOF samples were prepared in 2-dram screw-top glass vials housed in an 84-vial heater and shaker plate. After reagents were transferred to each vial, the vials were capped with the automated screw capping tool and then heated for 18 hours.

S1.2 Image Capture

A USB webcam (Logitech MX Brio) was integrated into our robotic synthesis platform by mounting it within the Chemspeed enclosure and connecting it to the control computer. Using Chemspeed's Autosuite software, we triggered image capture with an executable file that saves the sample vial number, date, and time of image capture within the image filename. As the sample vials are housed in a heating block, we use a gripper tool to sequentially pick up one vial at a time from the block and briefly hold it in place while the camera is triggered before returning the vial to continue heating.

S2. Model Training

S2.1 Detailed Phase Definitions and Examples

We divide our material class labels into 3 layers: headspace, liquid, and solid.

Headspace (two labeling options): "Empty" refers to clear glass above the liquid level, while "residue" refers to material adhered to the vial walls above the liquid level (**Figure S1**). The headspace labels are mutually exclusive; each vial image is labeled as either "empty" or "residue," but not both.

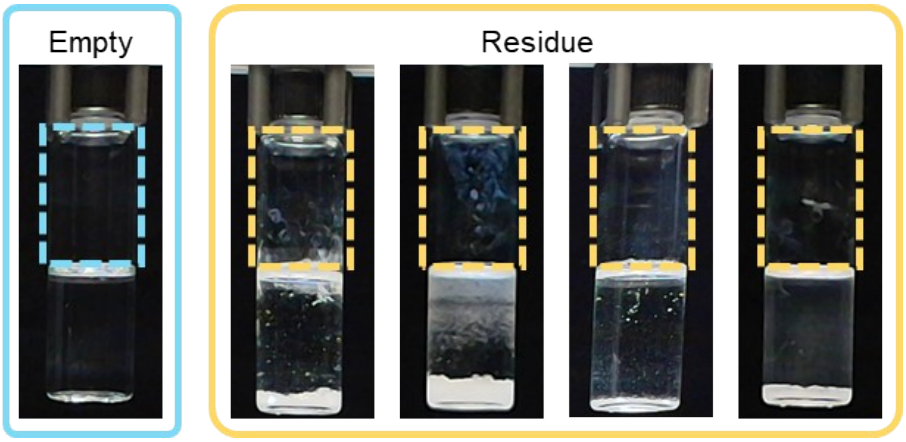

Figure S1. Examples of sample vials containing "empty" and "residue" phases.

Liquid (two labeling options): "Homogeneous liquid" refers to a clear, uniform liquid with no cloudiness or separation (**Figure S2**). "Heterogeneous liquid" refers to cloudy liquids, separated liquids, or liquids containing floating material or suspensions of particles (**Figure S3**). For the liquid layer, the two options are not exclusive; a single sample is allowed to contain both a homogeneous liquid phase and a heterogeneous liquid phase. Multiple homogeneous liquid phases are possible (for instance, layered immiscible clear liquids); however, we do not observe these cases as all solvents used in this study are miscible with each other.

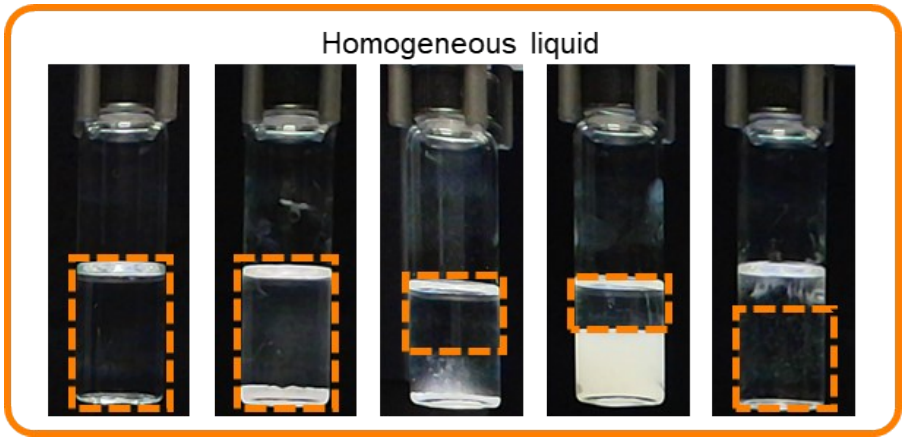

Figure S2. Examples of sample vials containing "homogeneous liquid" phase.

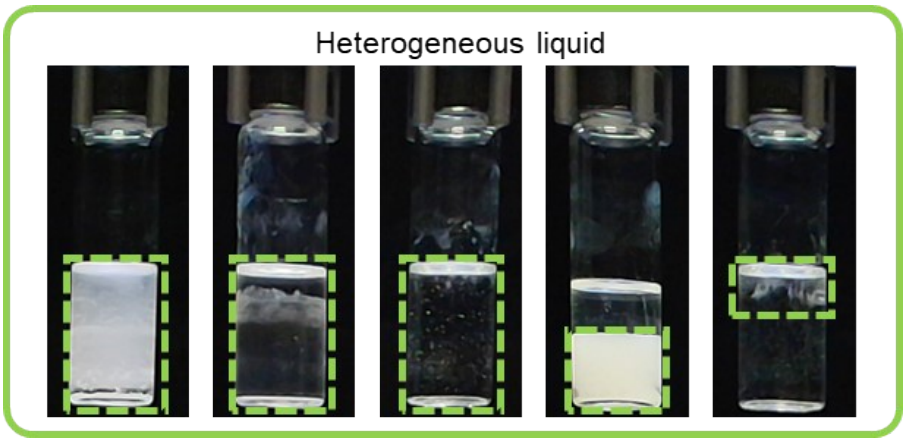

Figure S3. Examples of sample vials containing “heterogeneous liquid” phase.

Solid (one labeling option): The “solid” phase consists of aggregated solid powder, crystallites, or single crystals located at the bottom of the sample vial (**Figure S4**). While it is physically possible to form solid material floating on a liquid phase, here we restrict our definition to material located at the bottom of the sample vial, so that it may be collected for use in a downstream chemistry task.

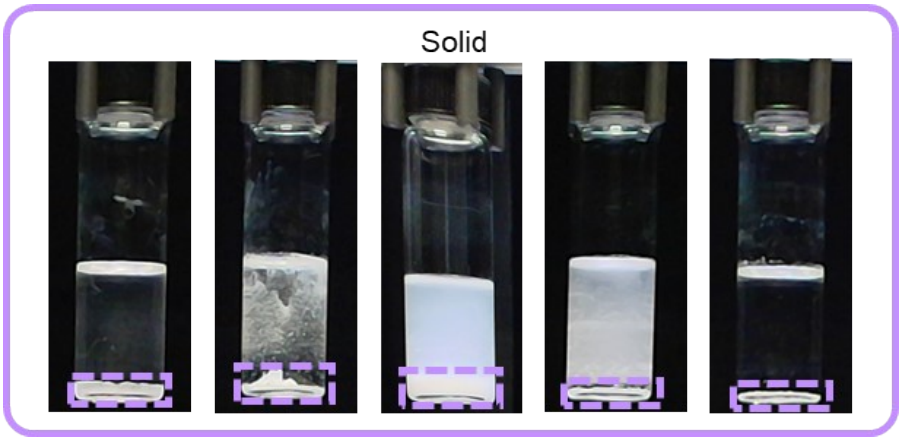

Figure S4. Examples of sample vials containing “solid” phase.

The liquid and solid layers may overlap: for instance, a layer of solid powder can be contained within a liquid phase (as illustrated in the second image of **Figure S2**, or the first image in **Figure S4**). The headspace layer is distinct and does not overlap with either of these. The headspace refers specifically to the region above the liquid level in the vial. Therefore, any material observed in this region, such as adhered powder, is labeled as ‘residue’ rather than ‘solid’ contained within an “empty” phase, even if it resembles solid-phase material.

S2.2 Vessel Detection Model

The vessel detection model was trained on the LabPics dataset[1] (7900 images) supplemented with 168 images of vials captured within our automated synthesis platform. The images from the LabPics dataset comprise various types of chemistry equipment, including flasks, vials, test tubes, and Erlenmeyer flasks, among others. These vessels also contain a variety of materials, including solids, liquids, foam, and bubbles, in various laboratory settings. While this dataset is diverse, it does not contain images in a setting similar to our hardware setup. This causes the model to not perform as well on images from our setup. We address this issue by incorporating images from our setup, which include bounding boxes around vials, to enhance model performance. This supplement contains images of vials in various positions within the setup, each containing varying amounts of liquids and solids. We split the Lab-Pics dataset and the supplementary datasets, 80:20 (train:test), and combined them to get the augmented dataset.

Table S1. Performance metrics for the vessel detection model. We assess the mean average precision (mAP), which measures the overlap between predicted bounding boxes and ground-truth objects. For mAP50, a detection is considered correct if the intersection over union (IoU) is at least 0.5 (50%). A stricter measure, mAP50-95, averages across multiple IoU thresholds from 0.50 to 0.95. Precision refers to the proportion of predicted detections that are correct, and recall refers to the proportion of actual objects that are successfully detected.

Number of images	mAP50	mAP50-95	Precision	Recall
168	0.995	0.826	1.000	1.000

S2.3 Phase Detection Model

The phase detection model was trained on 168 images taken of 56 unique MOF sample vials within our automated synthesis enclosure. While this represents 3 replicate images of each sample vial, the images were taken at different timepoints throughout the synthesis reaction, so the contents of the vial are not necessarily the same: for example, a vial with solid product forming may have additional solid material formed over time. This dataset was randomly split into 80% training and 20% validation. Each image was manually annotated by a domain expert (MOF chemist). To improve dataset diversity, we enabled YOLOv8's built-in data augmentation by setting `augment = True` in the training function. This activates a range of standard augmentation techniques, including horizontal flipping, median blurring, cropping, and adjustments to contrast, hue, saturation, and brightness.[2] Importantly, YOLOv8 applies these augmentation techniques internally during training, allowing users to rely on its default mechanism without the need for separate preprocessing.

Table S2. Distribution of phases present among 168 total images used in the training dataset for the phase detection model.

Phase	Instances
Empty	103
Residue	77
Homogeneous liquid	91
Heterogeneous liquid	77
Solid	85

S2.4 Dataset Annotation

The in-house data used for model training was manually annotated using the online CV platform Roboflow.[3] This tool allows users to upload images, draw bounding boxes around objects of interest, and classify them according to custom labeling schemes. Once annotation is complete, the dataset is exported as individual `.txt` files, one per image, for subsequent YOLOv8 model training.

S3. Model Performance

S3.1 Equations

Equation S1. Precision, defined as the proportion of true positives among all predicted positives.

$$precision = \frac{TP}{TP + FP}$$

Equation S2. Recall, defined as the proportion of true positives among all actual positives.

$$recall = \frac{TP}{TP + FN}$$

Equation S3. F1 score, the harmonic mean of precision and recall.

$$F1 = \frac{(2 \cdot precision \cdot recall)}{precision + recall}$$

S4. User Study

S4.1 Participant Statistics

Table S3 provides a summary of the number of participant responses received across different sections of the user study. Here, it is important to clarify that participants were allowed to skip questions and respond only to those with which they felt comfortable. As a result, a total of 111 participants attempted either the labelling accuracy task (**Section S4.3**) or the speed task (**Section S4.4**). Among these, 90 participants voluntarily provided information about their scientific training background, and 86 responded to at least one question reflecting on the model (**Section S4.5**, **Figure S16**). The post-task survey, which addressed participants' experiences with conducting multiple experiments in parallel (**Figure S17**), was presented exclusively to those who reported performing five or more experiments simultaneously. A total of 33 participants responded to at least one question in this section.

Table S3. Summary of responses received across different survey sections.

Number of Participants who attempted at least one main task	**Scientific background**	**Accuracy task**	**Speed task**	**Post-task survey** Multiple experiments in parallel	**Post-task survey** Reflection on the model
111	90	106	90	33	86

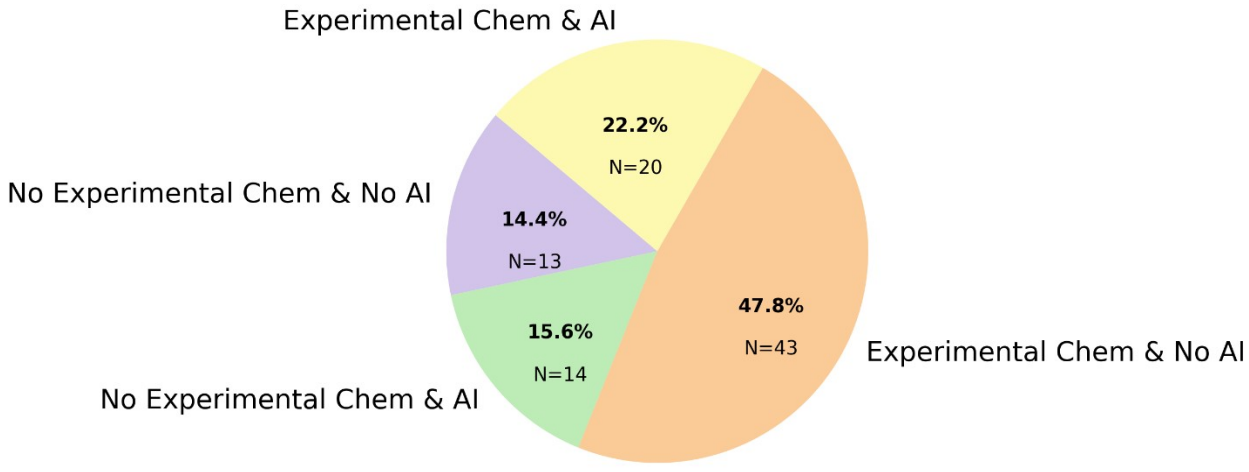

Figure S5. Self-reported backgrounds of 90 user study participants in experimental chemistry and artificial intelligence. N = number of participants.

S4.2 Dataset Description and Annotation

The final annotated dataset, with all disagreements resolved, is summarized in **Table S4**. Two experimental MOF chemists independently annotated the user study dataset to minimize potential labeling biases. A third subject matter expert then reviewed the images with conflicting annotations to make a final decision. These disagreements were categorized into four image type groups, as shown in **Table S5** and **Figure S6**.

The first category of disagreement (**Figure S6a**) involved labeling the headspace layer, where one annotator labeled it as “empty” and the other as “residue.” This occurred in 34 images (9.0%). The second category (**Figure S6b**) concerned the labeling of the primary liquid layer, with one annotator labeling it as “homogeneous” and the other as “heterogeneous.” These disagreements were found in 24 images (6.3%). The third category (**Figure S6c**) involved disagreement over the presence of the “solid” phase, where one annotator identified solid material and the other did not. This occurred in 20 images (5.3%). Finally, eight images (2.1%) had discrepancies regarding the number of liquid phases annotated (**Figure S6d**).

Table S4. Distribution of phases present among 378 total images in the user study dataset. A total of 376 images were used at least once across the accuracy and speed activities.

Phase	Instances
Empty	52
Residue	326
Homogeneous liquid	255
Heterogeneous liquid	141
Solid	279

Table S5. Disagreement in annotation between two expert annotators.

Disagreement category	Images
(a) Whether a phase is empty vs. residue	34
(b) Whether a liquid is homogeneous or heterogenous	24
(c) Whether solid is present	20
(d) Number of liquid phases present	8

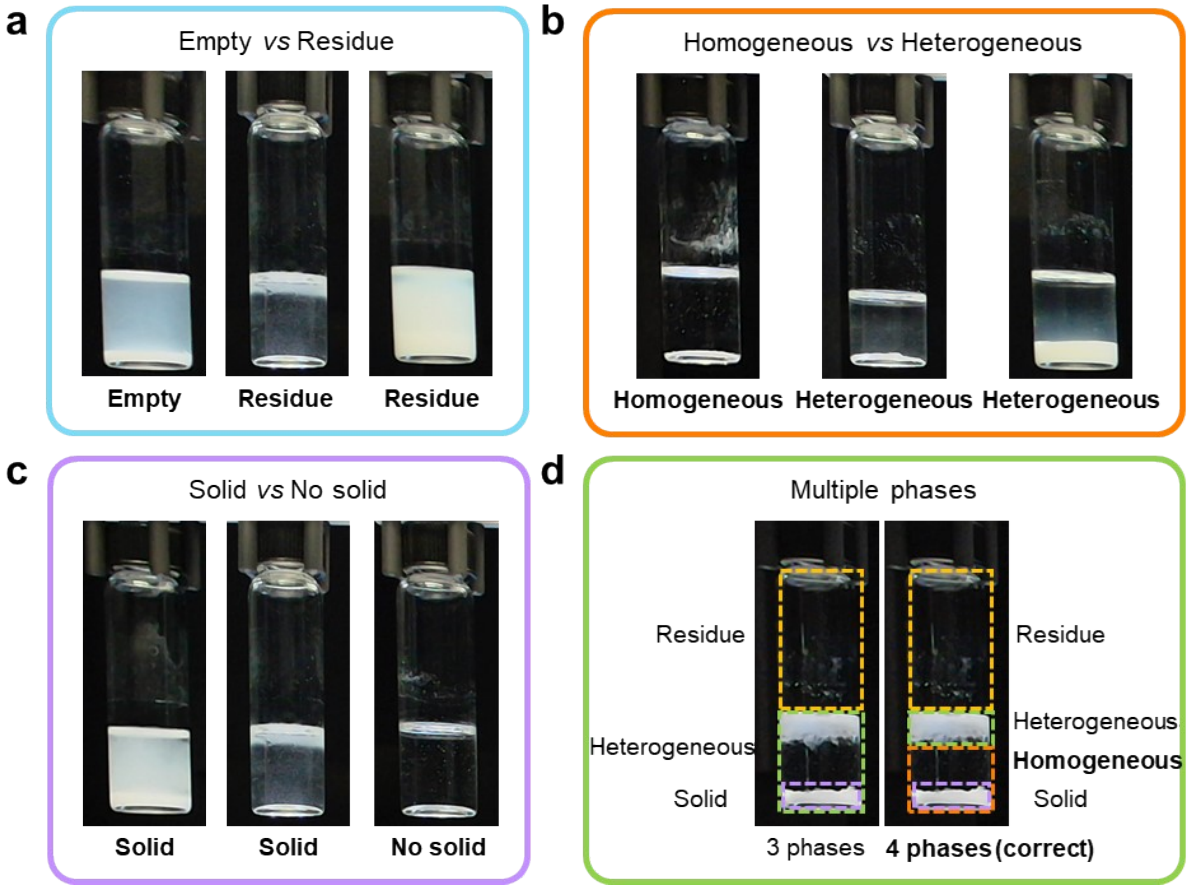

Figure S6. Examples of annotation disagreements between two expert annotators. For the first three categories, the phase shown beneath each image represents the final annotation determined by a third expert. For the multiple phases category, the image on the right, which shows four phases, reflects the final annotation made by the third expert.

S4.3 Accuracy Task

S4.3.1 Task Overview

In this activity, participants were first shown examples and definitions of five phase categories: empty, residue, homogeneous liquid, heterogeneous liquid, and solid, as shown in **Figure S7**. The participants then received detailed instructions on how to annotate images by labeling the phases present in each vial, as shown in **Figure S8**. The goal of the instructions is to teach participants to use the tool; we did not have a dedicated stage of practice and feedback.

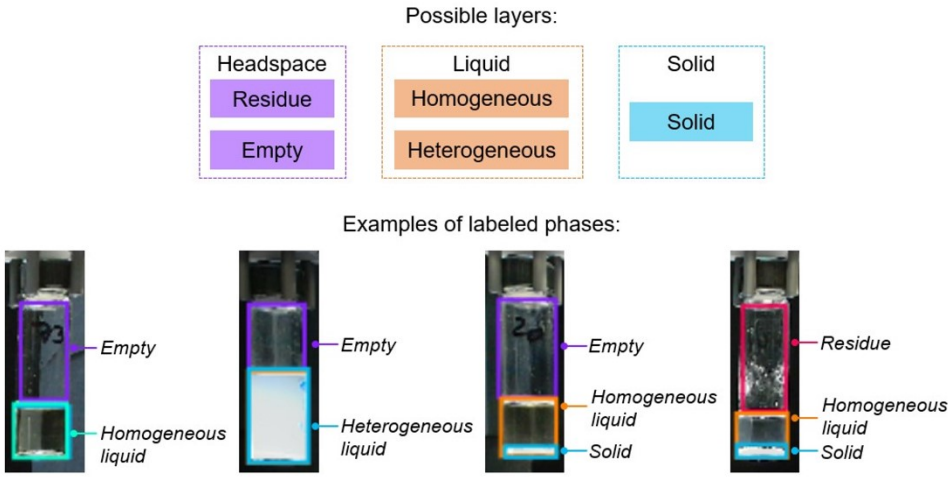

Figure S7. Examples of five phase categories that were shown to the participants: empty, residue, homogeneous liquid, heterogeneous liquid, and solid.

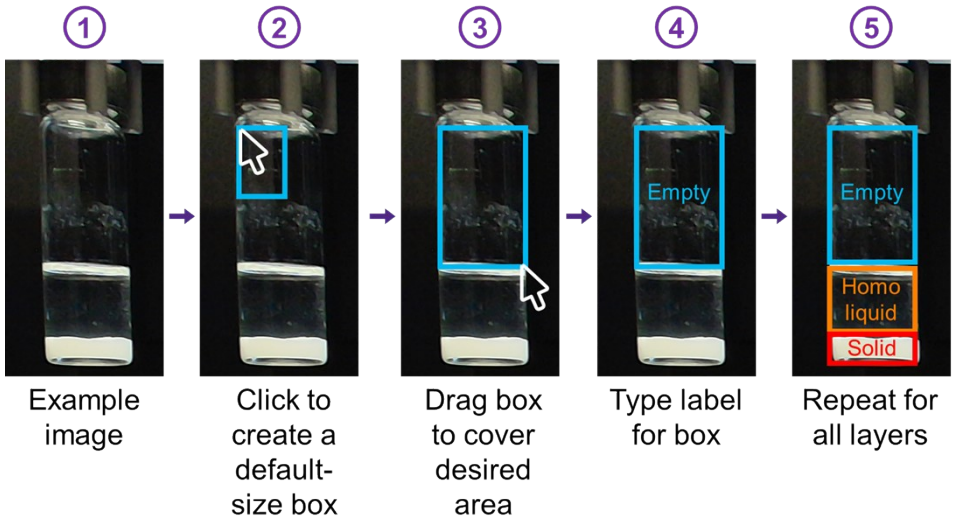

Figure S8. Step-by-step instructions for the labeling accuracy task provided to participants.

S4.3.2 Response Statistics

For the accuracy task, each participant was expected to annotate up to five images—a total of 106 unique participants completed at least one annotation. In total, they labeled 453 images, corresponding to 294 unique images, as shown in **Table S6**.

Table S6. Participant and image assignment distribution for the accuracy task.

Number of participants who attempted the task	Total number of expected annotated images (5 per participant)	Total number of unique images assigned	Actual number of images annotated by participants
106	530	294	453

Among the 294 unique images, 41 images contained an *empty* category, 253 *residue*, 194 *homogeneous liquids*, 112 *heterogeneous liquids*, and 218 *solid* phases, as shown in **Table S7**.

Table S7. Distribution of phases from unique images assigned to the accuracy task.

Empty	Residue	Homogeneous liquid	Heterogeneous liquid	Solid
41	253	194	112	218

A True Positive (TP) occurs when a phase is present and correctly observed, as shown in **Table S8**. A False Negative (FN) occurs when a phase is present but missed. Conversely, a True Negative (TN) refers to correctly identifying the absence of a phase, whereas a False Positive (FP) occurs when a phase is absent but is mistakenly detected. Representative examples of labeled images considered as FN and FP are shown in **Figure S9**.

Table S8. Definitions of TP, FN, TN, FP in the context of the accuracy task.

	Whether the phase exists	
	Ground truth	**Participant response**
True Positive (TP)	✓	✓
False Negative (FN)	✓	X
True Negative (TN)	X	X
False Positive (FP)	X	✓

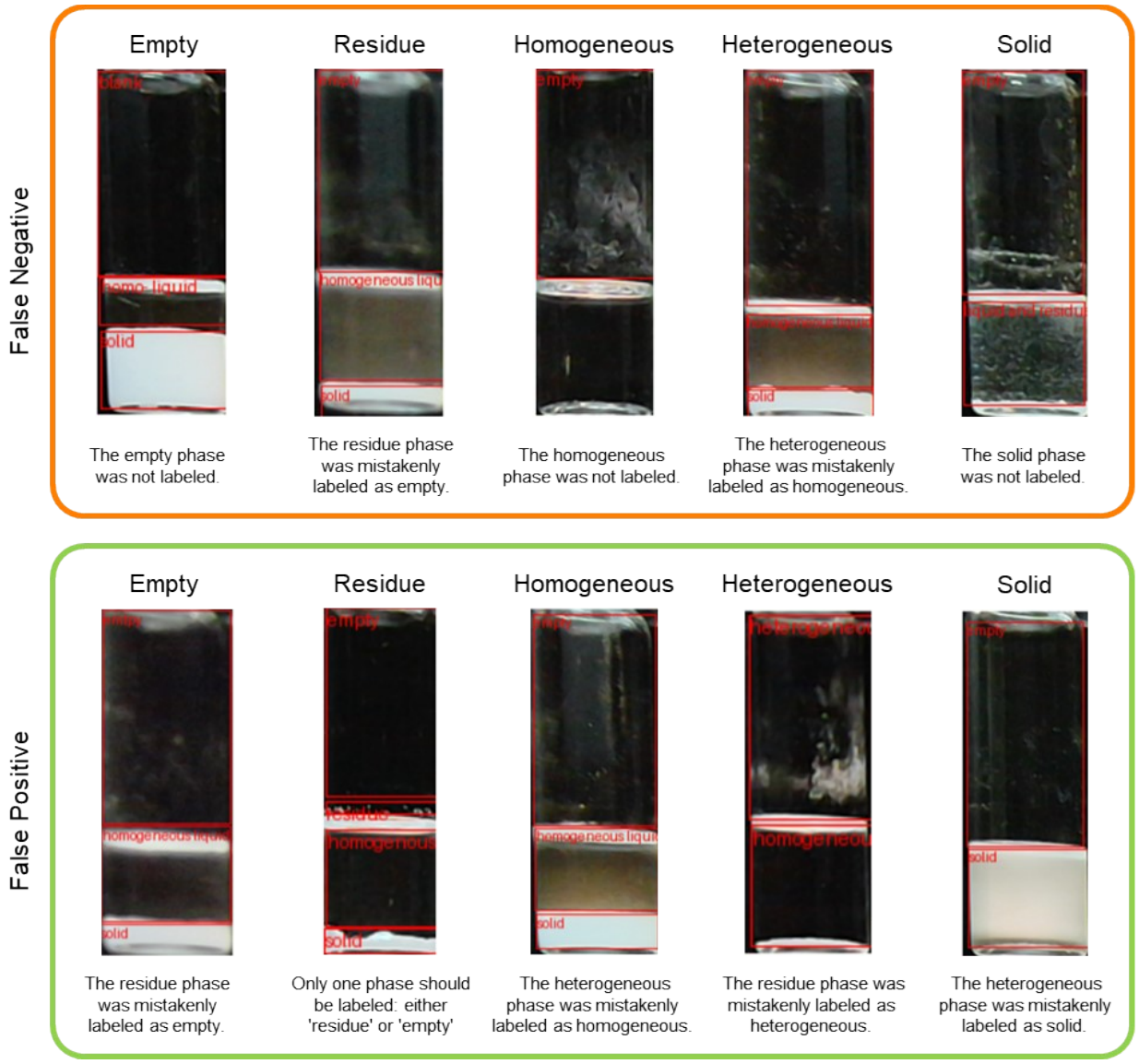

Figure S9. Representative examples of false negatives and false positives obtained from mistakes made by human participants when labeling the phases in the accuracy task.

S4.3.3 Human vs. Model Performance

As mentioned in the main text, the model outperformed humans based on any of the following criteria:

- **Analyzing whether humans detected a phase correctly:** As shown in **Figure S10**, humans exhibited low precision in detecting the 'empty' phase, indicating a high False-Positive rate. According to the confusion matrix in **Figure S11**, among all 453 annotations by humans, 294 annotations included the 'empty' phase when the ground truth did not. Human participants also had low recall on the 'solid', 'heterogeneous liquid', and 'Residue' phases, indicating their relatively lower performance in detecting these phases when present. These common mistakes by humans contributed to the overall low performance reported in **Figure S12**.
- **Detecting all phases in an image correctly**: As shown in **Figure S12**, humans were more likely than the model to make at least one mistake in their annotations, regardless of the number of phases presented in the image, except in the case where there were four phases present in the image. We acknowledge that in a real experimental setting, where the human researcher would have a better contextual background knowledge of the images (e.g., the reaction involved or the expected phases), accuracy might be higher than what we observed in this study.
- **IoU metrics were omitted for this analysis and comparisons**: We omitted the IoU metrics for the human participants to acknowledge the difficulty in annotating with perfect positions especially when phases overlap. To make a fair comparison, when comparing a model's performance, we also omitted the IoU for the model.

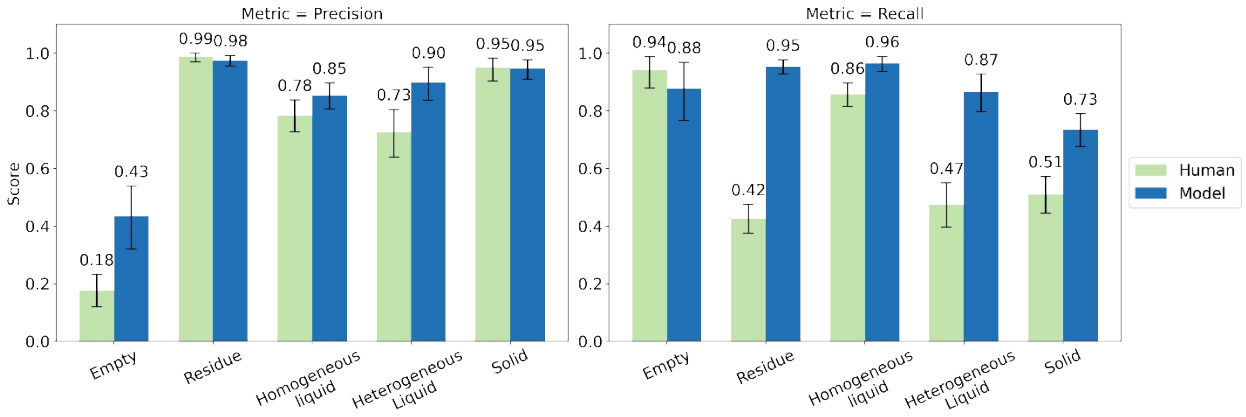

Figure S10. Comparison of human participants' and model precision (left) and recall (right) metrics for each phase in the accuracy task. The bar chart illustrates the percentage of correctly labeled images for each number of phases present, for both human participants (green) and the model (blue). The error bars represent the 95% bootstrap confidence intervals of the precision and recall estimates, computed independently for human participants and the model by resampling images with replacement (N = 1000).

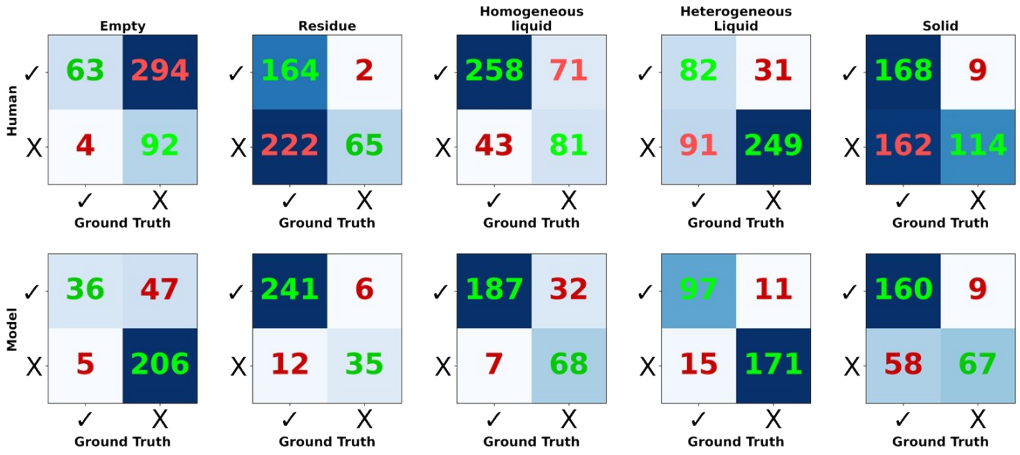

Figure S11. Confusion matrices by phase for responses from human participants (top row) and model predictions (bottom row). Note that red values indicate the number of mistakes, while green values represent the number of correct responses. Human participants annotated a total of 453 images for the top five confusion matrices, whereas the model annotated 294 unique images from this set for the bottom five confusion matrices. See **Table S6** for details.

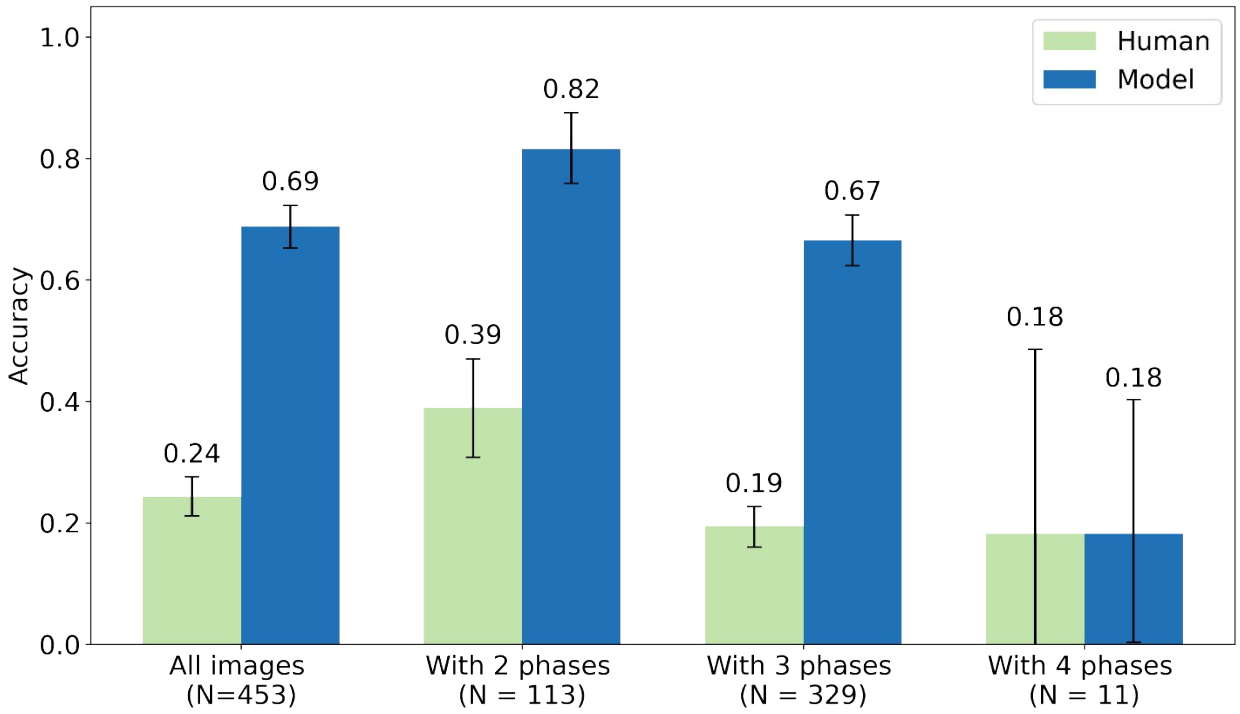

Figure S12. Comparison of human (green) and model (blue) accuracy as a function of the number of phases present in the ground truth annotations. A response was considered correct only when all ground truth phases in an image were accurately identified. N is the actual number of images annotated by participants per category. Note that an image could be annotated by more than one participant (**Table S6**). The model annotated the full set of images (**Table S6**). The error bars represent the 95% bootstrap confidence intervals of the accuracy estimates, computed independently for human participants and the model by resampling images with replacement (N = 1000).

S4.4 Speed Task

S4.4.1 Task Overview

In the speed task, we evaluated phase identification speed by comparing the response times of human participants with those of the model. The goal was to assess how quickly each could determine the presence of a specific phase in an image without compromising accuracy.

Participants of the user study received detailed instructions on how to address the speed task, as shown in **Figure S13**.

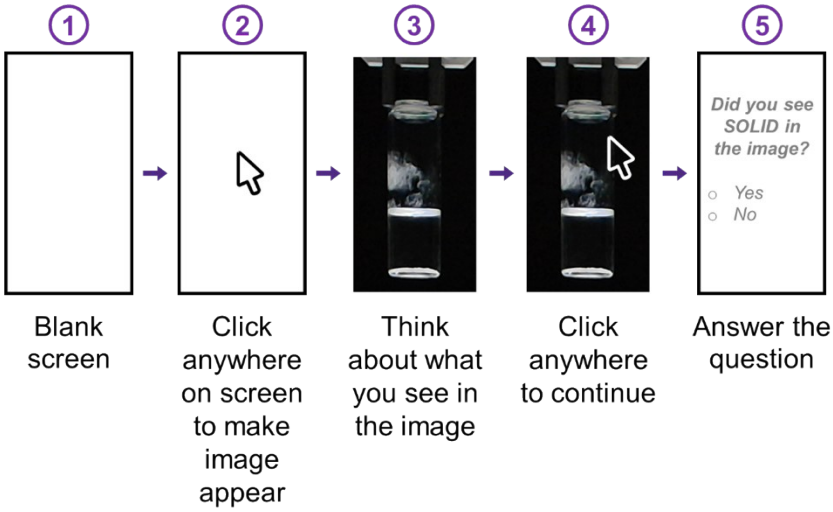

Figure S13. Step-by-step instructions provided to user study participants for the speed task.

S4.4.2 Response Statistics

For each question in the speed activity, participants were randomly assigned an image along with a yes-or-no question, such as "*Is* `<phase class>` *present?*", where `<phase class>` refers to one of the following phases: residue, homogeneous liquid, heterogeneous liquid, and solid (excluding empty).

We noted that the total number of responses (449) exceeded the number of unique image-question pairs (405), since multiple participants were asked the same image-question pair.

Table S9. Participant and image assignment distribution for the speed task.

Participants who attempted the task	Expected responses considering 5 questions per participant	Responses collected	Unique images assigned	Unique image/question pairs from responses collected
90	450	449	289	405

Among the 289 unique images, there were 42 empty, 247 residues, 195 homogeneous liquids, 107 heterogeneous liquids, and 214 solid phases, as shown in **Table S10**.

Table S10. Distribution of phases from unique images assigned to the speed task.

Empty	Residue	Homogeneous liquid	Heterogeneous liquid	Solid
42	247	195	107	214

As mentioned above, for each image, participants were asked whether it contained one of the four phases: residue, homogeneous liquid, heterogeneous liquid, or solid. This assignment was independent of the actual distribution of these four phases in the images.

Table S11. Distribution of queried phase classes across the images selected for the speed task.

Residue in <Image>?	Homogeneous liquid in <Image>?	Heterogeneous liquid in <Image>?	Solid in <Image>?	Unique image/question pairs
104	103	96	102	405

S4.4.3 Human vs. Model Performance

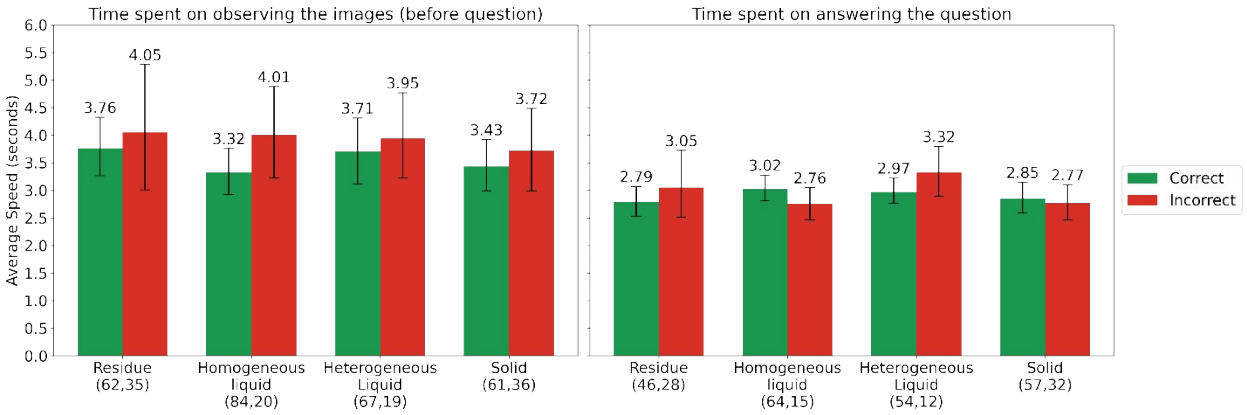

Figure S14. On the left: Average time participants spent viewing each image before receiving a question (left). Due to technical limitations, viewing times were unavailable for 36 out of 449 responses. Outliers (29 out of 413) were excluded using the interquartile range (IQR) method. On the right: average time participants spent answering yes-or-no questions. Response times were missing for 110 out of 449 cases, and 31 outliers (out of 339) were removed using the IQR method. The number of responses per category is shown on the *x*-axis in the following format: (*correct* responses, *incorrect* responses). On average, regardless of correctness, participants spent 3.65 seconds on each image. The error bars indicate the 95% bootstrap confidence intervals of the mean speed, obtained by resampling participant responses with replacement (N = 1000).

Incorrect responses were associated with longer image observation times, even though participants were unaware of the questions they would be asked (left panel of **Figure S14**). In contrast, the actual response times recorded after the image-viewing phase (i.e., the exact time spent responding to the question) showed minimal differences between correct and incorrect answers (right panel of **Figure S14**). This may suggest that certain images were inherently more difficult to interpret, requiring greater visual effort regardless of the participant's eventual accuracy. However, we acknowledge that factors such as mouse movement speed and image loading time may have influenced the recorded response times. On average, regardless of correctness, participants spent 3.54 seconds on each image. In terms of accuracy, the model outperforms humans in all phase categories. Overall, it achieves 17% higher accuracy, highlighting its strong potential for deployment in phase detection tasks.

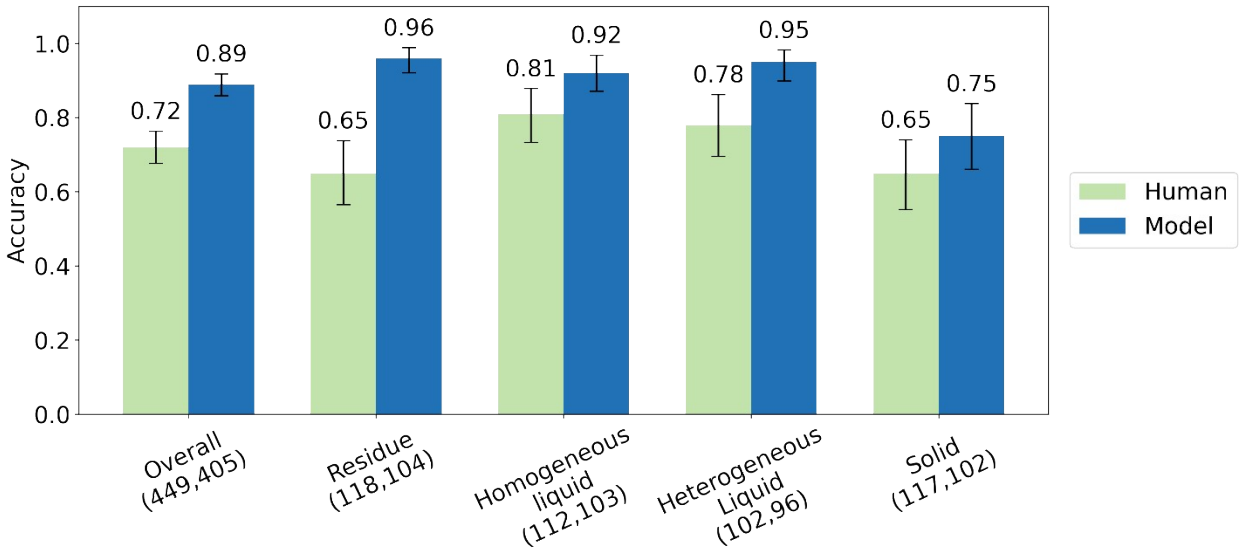

Figure S15. Accuracy in detecting the presence of each phase during the speed task, across different phase types. The absolute number of image-question pairs is shown on the *x*-axis in the following format: (human participant responses, model responses). Note that some image-question pairs were duplicated, as multiple participants may have been shown the same pair. Please refer to **Table S11** for additional details. The error bars represent the 95% bootstrap confidence intervals of the accuracy estimates, computed independently for human participants and the model by resampling pairs of image and phase with replacement (N = 1000).

S4.5 Post-Task Survey

In this part of the user study, participants were asked to evaluate general statements related to conducting parallel experiments with a three-point agreement scale (**Figure S16**) and using computer vision analysis for phase labeling with a five-point agreement scale (**Figure S17)**.

Note that this question was only visible to participants who mentioned that they often conducted multiple (5 or more) reactions or experiments in parallel.

Question 1:

"Imagine that you are manually labeling materials from many simultaneous experiments. Consider whether the presence of many experiments would have a negative or positive effect on the following perspectives."

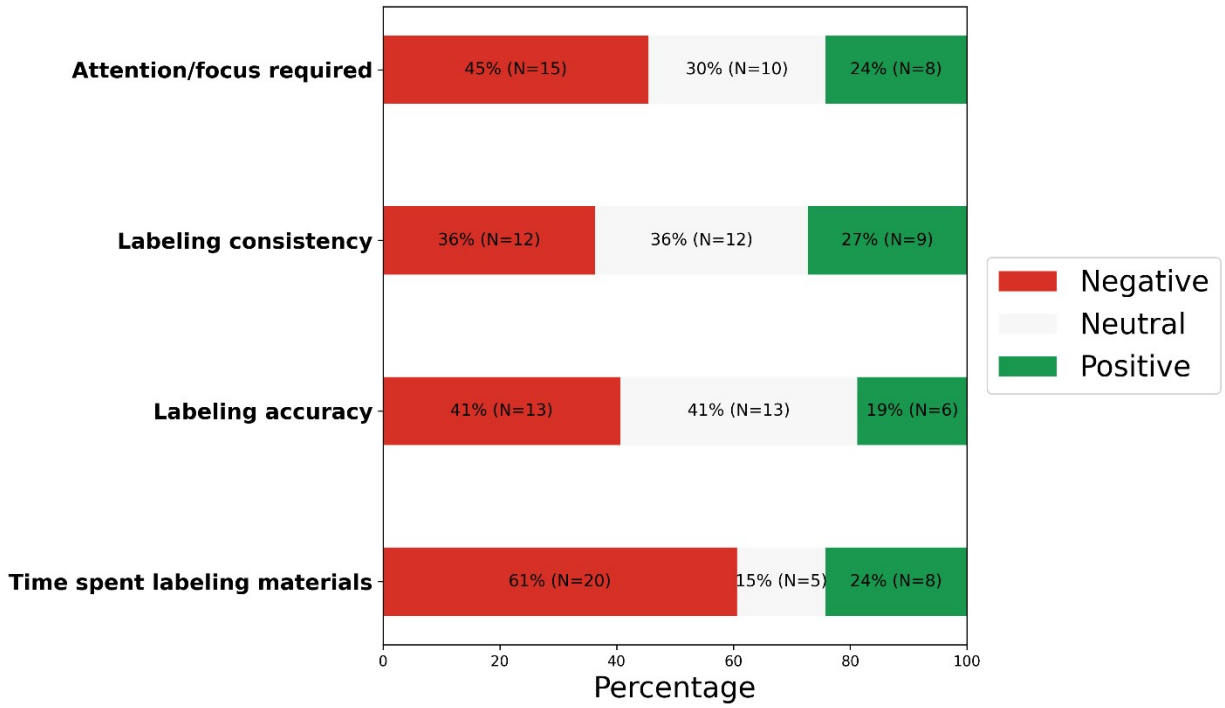

Figure S16. Participants' responses to post-survey assessment regarding their perception of conducting parallel experiments. The question also explored how increasing the number of experiments would affect attention and focus, labeling consistency, labeling accuracy, and the time spent labeling.

Question 2:

"Please answer the following general statements by indicating to what extent you agree or disagree with them."

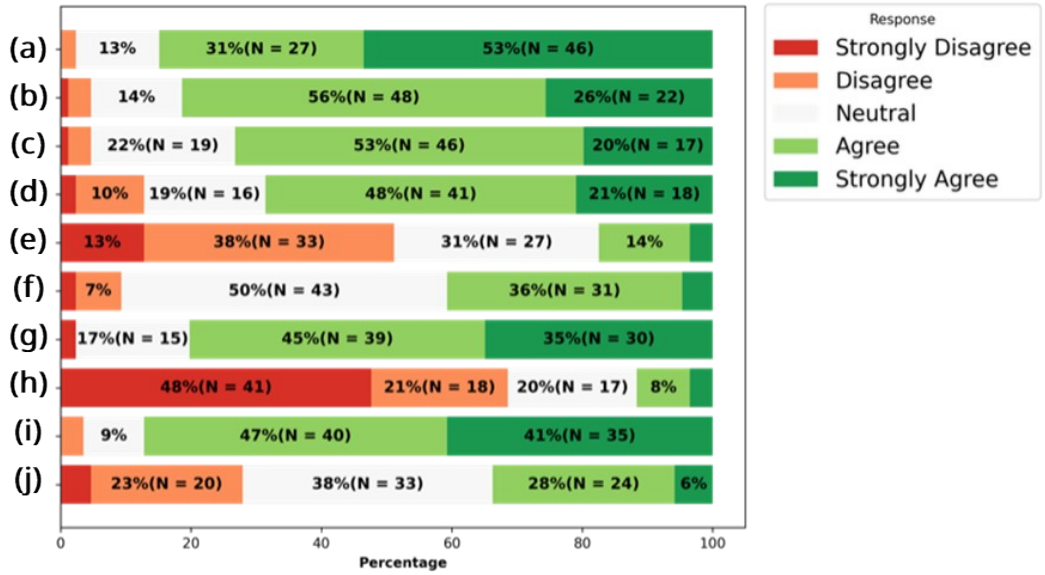

Figure S17. Participants' responses to the post-survey assessment regarding their perception of using a computer vision model for phase labeling compared to manually labeling images. Participants were asked to respond to the following statements: *a) Doing the material labelling by human would be tedious; (b) Using this model will help us select the samples that need characterization faster; (c) The model might make sporadic errors; (d) I was able to understand the logic behind the labels; (e) I would trust this model more than a human; (f) I would trust this model in my experiments; (g) Using this model can help with conducting parallel experiments; (h) I have used or known similar models; (i) I believe that this system can label phases faster than humans; (j) I believe that this system can label phases more accurately than humans.*

S4.6 Open-Ended Responses

Two open-ended questions below were included in the user study:

- **Open-ended question 1:** Please comment on how deploying this AI model can benefit your experimental workflows
- **Open-ended question 2:** Please comment on any concerns or potential issues you may have regarding the AI model we are developing

A total of 49 and 42 responses were collected for open-ended questions 1 and 2, respectively, and are reported in **Table S12.** Additionally, a word cloud was generated by extracting text from user responses to open-ended question 1 (**Figure S18**), removing stop words and punctuation, and then calculating the frequency of the remaining words. Words mentioned more frequently appear larger in the word cloud image, visually indicating the most common keywords among participants' responses. The word cloud was generated using Python and the WordCloud library, with the random state set to 100.

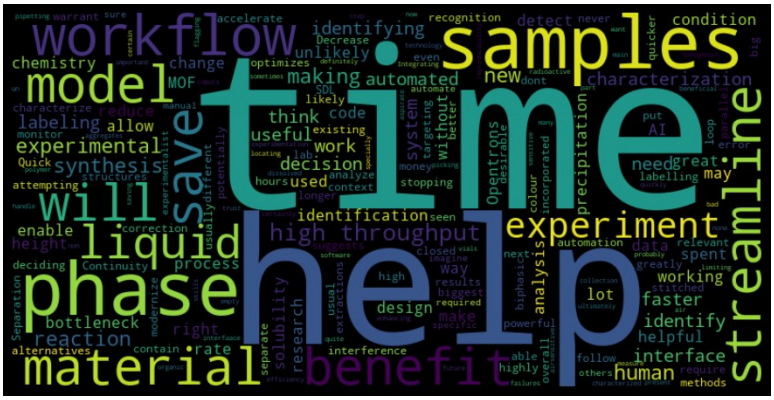

Figure S18. Word cloud created from participant feedback in the open-ended responses from the user study for the open-ended question: *Please comment on how deploying this AI model can benefit your experimental workflows*.

Table S12. Raw responses from participants to the two open-ended questions (open-ended questions 1 and 2). Please note that responses listed in the same row are not necessarily from the same individual. Both questions were optional. N indicates the number of responses collected for each question.

Open-ended question 1: Please comment on how deploying this AI model can benefit your experimental workflows (N = 49)	**Open-ended question 2:** Please comment on any concerns or potential issues you may have regarding the AI model we are developing (N = 42)
1. Making the labeling and recognition faster	1. Accuracy
2. It can accelerate the rate of labelling samples without human interference	2. The image detection may be affected by the reflection of surrounding, especially working with homogenous solution
3. Continuity without stopping the reaction that require longer time than usual working hours	3. Nature of the chemical reaction where chemical bond breaking and formation takes place by the formation of cationic, anionic and radical

Open-ended question 1: Please comment on how deploying this AI model can benefit your experimental workflows (N = 49)	**Open-ended question 2:** Please comment on any concerns or potential issues you may have regarding the AI model we are developing (N = 42)
	reactive species will be disrupted by the impurity as well as other reactions parameters.
4. it can be used for decision making in a closed-loop SDL system. Being able to detect phase change will help to monitor the overall reaction and help the system make decision on what to do next.	4. biased toward certain classes. if a human cannot distinguish from different class visually, I think it is impossible for any model to correctly identify different classes.
5. Decrease time spent on deciding which samples warrant further characterization and potentially save money and time spent on attempting to characterize samples that do not contain material.	5. I think there needs to be human oversight to catch serious mistakes.
6. I think this would be a great way to modernize lab work.	6. Are bounding boxes the best descriptor?
7. I dont have experimental workflows, I have never seen the model	7. Concerns include data accuracy, reproducibility of AI suggestions, potential bias, and over-reliance on the model, which may limit critical thinking.
8. It streamlines data analysis, optimizes designs, suggests alternatives, and saves time.	8. I'd probably still want a person to review images for crucial features, but this can be out of the loop.
9. Any automated analysis that can be stitched into the existing workflow is powerful and enable error correction.	9. False positives or negatives. Especially for clear/colourless crystals such as small needles on the bottom or side of the vial. They can be nearly invisible by eye so a camera may miss them.
10. The biggest benefit of this model will be identifying samples that may need follow-up by a human. This will greatly reduce time. If colour identification is also incorporated that would be even better.	10. would there be enough time for fine particulates to settle, so that confusion about emulsions vs. suspensions could be determined?
11. not relevant to my research	11. If the refractive index is very similar between layers it can be challenging to distinguish visually without very close inspection.
12. This would likely be useful in the context of biphasic systems (such as extractions) or in precipitations of my reactions. However, this would have to be in parallel with other automation.	12. What are the accuracy statistics? I would be concerned that it might miss solids if the amount of crystals is small
13. Useful for high through put synthesis targeting new structures or identifying different phases	13. Depending on toxicity of material, if wrong material is isolated and human comes in contact with it, potential danger
14. Quick characterization can help automated methods identify the right phase to separate. Separation is usually a big bottleneck in experimental chemistry	14. I think I still would not fully trust this if a human researcher or second look through was not involved
15. It would be a lot faster and allow me to do more at a time.	15. The quality of the image may have a large effect on the outputs of the model. An image or video with poor resolution could be improperly characterized
16. Would help streamline the process of phase identification	16. I do not have any

Open-ended question 1: Please comment on how deploying this AI model can benefit your experimental workflows (N = 49)	**Open-ended question 2:** Please comment on any concerns or potential issues you may have regarding the AI model we are developing (N = 42)
17. Having an AI model to detect phase changes or liquid-liquid interfaces would be highly desirable	17. The pictures appear to be taken from an angle that is perfectly aligned with the camera, under excellent lighting conditions, and against a black background. It would be helpful to see how the model performs in less-than-ideal lighting and with a more cluttered background, both are conditions that are more representative of a real lab setting. Understanding the model's performance in these scenarios is crucial for assessing both its reliability and safety.
18. It can't.	18. The robot may want to look at the vial from different angles. Perhaps a special source of background light can help with the characterization.
19. analyze my results quicker when I do a lot of experiments at once	19. Sometimes, identifying phases can be challenging. For example, if the vessel is dirty or reflections interfere, the model may struggle to differentiate materials. Additionally, accurately detecting the phase at the liquid-gas interface can be difficult due to subtle transitions. I am not familiar with the approach used, but perhaps there is a method to better distinguish between heterogeneous liquids and solids, such as leveraging texture or other distinguishing features.
20. Save time for high throughput experiments	20. none
21. I don't think it will benefit my specific workflows but it can for sure help others	21. Some classifications aren't clear cut so errors are definitely a concern
22. It can help automate the process	22. In some of the images, it was hard to distinguish whether there was residue present on the vial or not due to possible glare, or other slight imperfection.
23. for experimental design	23. if deployed in an experimental workflow with dangerous chemicals, and presented with an out-of-distribution sample, it could lead to unwanted outcomes
24. I'm not an experimentalist myself, but I imagine this could help streamline the manual workflow required for high-throughput experimentation.	24. How can it tell the difference between residue stuck to the walls of the vials below the liquid line in a homogeneous liquid, and a heterogeneous liquid?
25. Integrating this model into high-throughput experiments could streamline data collection and enable faster decision-making, ultimately enhancing efficiency and reproducibility in materials research.	25. N/A
26. I think the most important part of this software is locating vials which are empty/homogenous liquid and flagging them as failures quickly	26. Does color of crystal have any impact on dectection? All examples shown were white but if very clear/colorless could that be an issue?
27. Identifying phases is not the rate limiting step in organic synthesis, so i am unlikely to ever benefit from this technology.	27. consistency in the results
28. The main benefit is probably saving time	28. I believe the main concern could be the

Open-ended question 1: Please comment on how deploying this AI model can benefit your experimental workflows (N = 49)	**Open-ended question 2:** Please comment on any concerns or potential issues you may have regarding the AI model we are developing (N = 42)
	reflection on the glass. As a human, it was difficult to identify in the picture, so I'm unsure how accurately the model is detecting it
29. It would definitely save time on labeling and can be helpful in picking out which samples need to be characterized.	29. I have concerns if the model used is closed source, like OpenAI
30. i would not have to trust my bad experimental chemistry skills	30. I think this model is interesting and valuable to develop, but its utility feels limited to experiments where a successful reaction is visually distinct from unsuccessful ones. For the organic experiments I performed, I cannot tell whether it has worked based on visuals alone. So, all experiments need to be prepared for LC-MS characterization anyway, regardless of how they appear. This is especially true for high-throughput experimentation, where reaction containers are arranged in a grid, so the sideways view of the vials, which were used in this survey, is not available, unless the vials are lifted up row by row, which adds more time to the characterization process. Generally, experiments that already have efficient methods of characterization (eg LC-MS takes 2-5 minutes for one sample, and can be programmed to analyze an entire wellplate overnight) won't save much time by having failed samples pre-eliminated from the characterization queue. If anything, this AI model feels more useful for single reactions that a robot arm(s) are performing. It can check for things like where a interface is during liquid-liquid extraction, whether a filtration successfully removed all solids, whether solids have appeared in a recrystallization workflow, etc.
31. May help identify un-dissolved polymer aggregates, or other materials	31. I found the most challenging part in labeling to be distinguishing between the homogeneous and heterogeneous liquids in certain cases, especially at the top of the liquid, so I am wondering how the model deals with this issue.
32. -(*)	32. unclear differentiation between residue and solid; unclear differentiation between heterogeneous liquid and non-transparent, colored liquids.
33. It can be very helpful specially if there are many samples	33. none as long as the user can visualize the predictions and make corrections if necessary
34. At present, I handle radioactive and air-sensitive materials, so it is not beneficial for me. However, in the future, when I work with non-airsensitive materials, it will certainly be quite useful	34. It wont be 100% accurate. 95% is not good enough for an experimentalist
35. none of my workflows	35. Differentiating top of liquid from a solid might

Open-ended question 1: Please comment on how deploying this AI model can benefit your experimental workflows (N = 49)	**Open-ended question 2:** Please comment on any concerns or potential issues you may have regarding the AI model we are developing (N = 42)
	not always be clear if solid is floating on top?
36. When working up high-throughput reactions on pipetting robots such as Opentrons, sometimes I want to aspirate above a certain liquid-liquid interfaace. Right now, the only way is to measure the height of the interface from the bottom of a few wells, take the average, and hard-code the height into the Opentrons code. If this interface can be detected automatically, it would save some measuring and programming time. It would also allow high-throughput experiment workup and characterization prep to be more fully automated.	36. No concerns to-date...
37. Not applicable since I do not perform experiments.	37. Poor picture quality exaggerating AI accuracy relative to human accuracy. These pictures were a bit blurry and low-resolution, making them more difficult to identify than in real lab situations. In a real lab setting, a human would jiggle the flask and change their viewing angle to ensure they identify the vial contents correctly; this dramatically increases the accuracy of a human chemist's identifications. This survey compares human brains equipped with computer eyes and arms versus computer brains equipped with computer eyes and arms; the more critical comparison is human brains equipped with human eyes and arms versus computer brains equipped with computer eyes and arms.
38. Might help with drug solubility testing	38. Whether the picture is clear enough to recognize by AI
39. If making a new MOF using a new synthesis, could help streamline the synthetic conditions tested	39. solid/liquid might be easy to distinguish. But, e.g. how about mixture with impurities, the product color, shape and yield sometime were also considered to tell if the reaction succeed.
40. Reduce workload on precipitation. No need for human verification of precipitate.	40. We should know who take responsibility when a harmful incident caused by it.
41. Identify compounds with solubility issues before being assessed in bioactivity assays	41. If the model consider various views in various angle, the accuracy would be increased. It was hard to distinguish the labels only with one view point in some cases even by myself when doing the test in this survey.
42. Unlikely to benefit - this kind of labeling isn't a bottleneck in our work	42. What if the residue may have different color or the solvent change to dark color after reaction?
43. Yes I would like to transition to the AI model...	43. N/A
44. N/A. I don't do this in my workflows ever.	44. N/A
45. More efficient	45. N/A
46. neutral	46. N/A
47. I don't conduct experimnt	47. N/A
48. It would be great to be used for the screening	48. N/A

Open-ended question 1: Please comment on how deploying this AI model can benefit your experimental workflows (N = 49)	**Open-ended question 2:** Please comment on any concerns or potential issues you may have regarding the AI model we are developing (N = 42)
the synthesis condition of new MOFs.	
49. Can first make a raw judgement before human do	49. N/A

S4.7 Analysis of Experimental Chemist Subset

To examine the influence of a background in experimental chemistry on image classification performance, we have conducted a separate analysis on participants who were members of the Farha group at Northwestern University at the time of this user study. All the members self-reported to have a background in experimental chemistry.

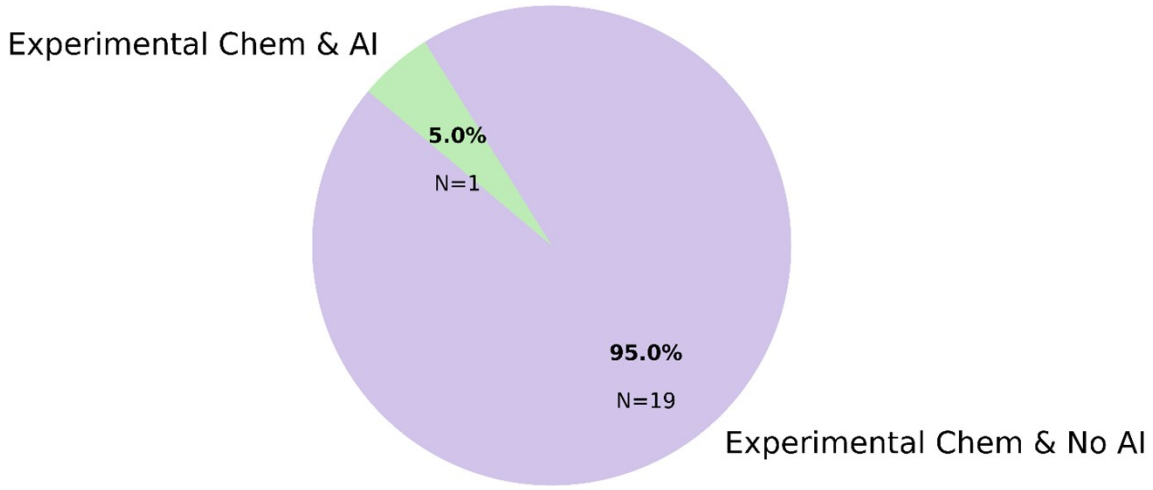

Figure S19. Self-reported backgrounds of 20 user study participants in experimental chemistry and artificial intelligence. N = number of participants who disclosed their background in the user study survey. This figure is to be compared with **Figure S5**.

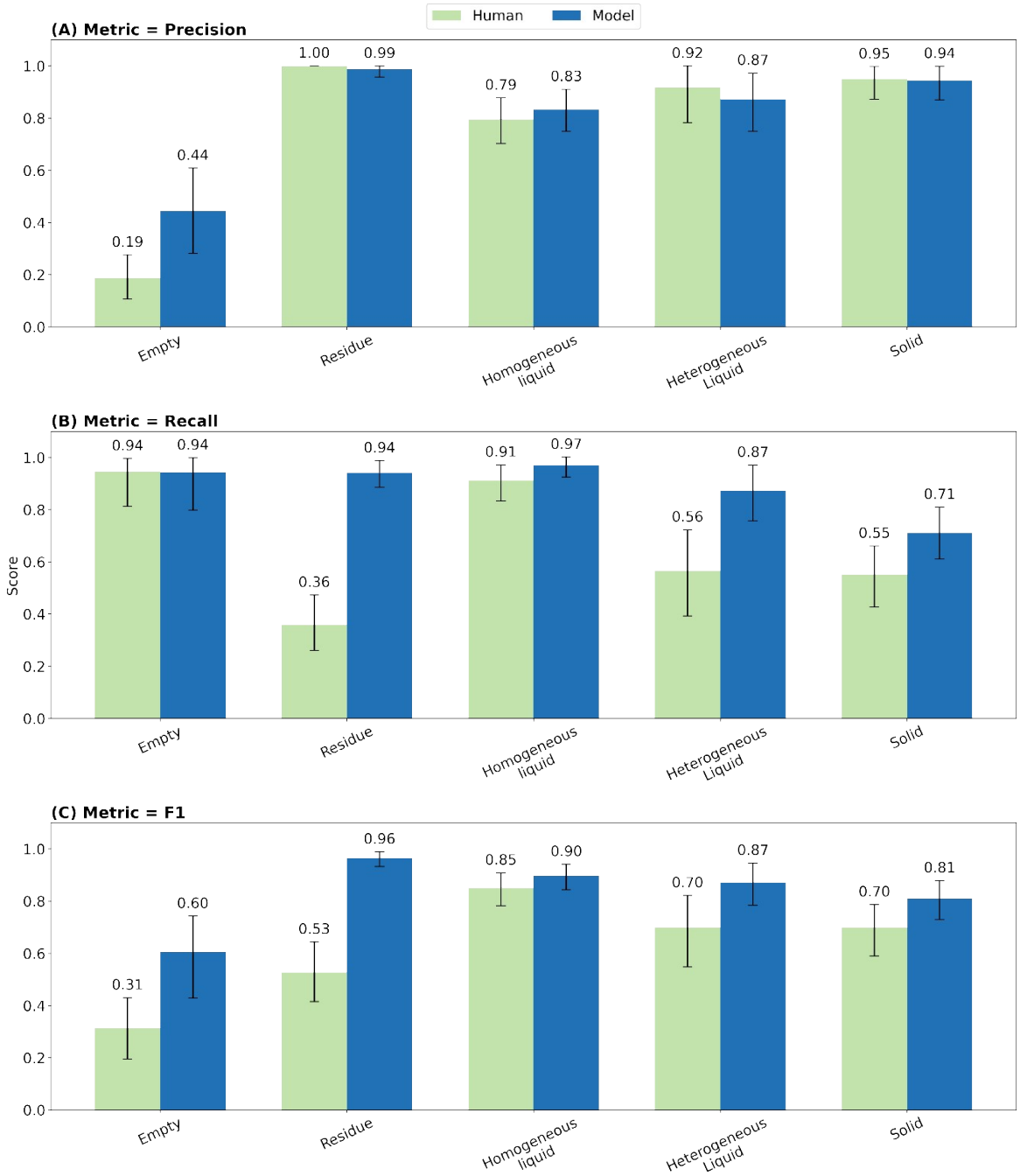

Figure S20. Precision (A), Recall (B), and F1 Score (C) for the accuracy task from Farha group members, to be compared with the results of the entire user study cohort in Figure 6 and Figure S10. Note that the model's performance was different because it was calculated only on the images that these participants saw. Here, the number of participants (from the Farha group) was 21, which is lower than the 106 participants in the whole group. Farha group members have a higher F1 score (C) for heterogeneous liquid (0.7 vs. 0.57). For the other four phases, Farha group members achieved similar F1 scores, including empty (0.31 vs. 0.30), residue (0.53 vs. 0.59), homogeneous liquid (0.85 vs. 0.82) and solid (0.70 vs. 0.66).

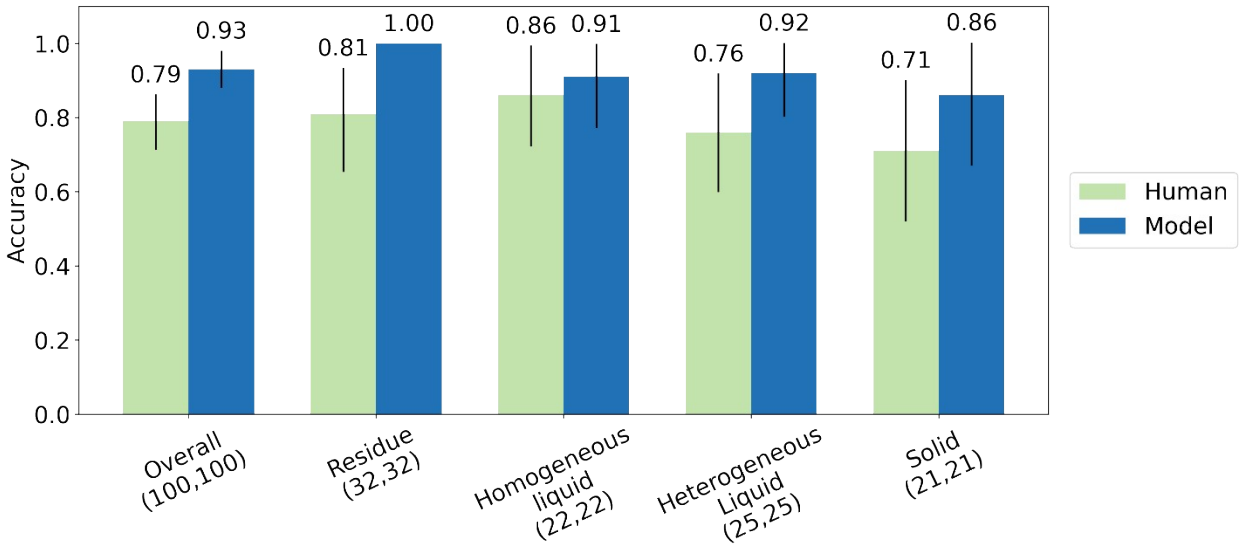

Figure S21. Accuracy of speed task results from Farha group members (20 participants). This subgroup had a higher overall accuracy across all four relevant phases (0.79 vs. 0.72). They achieved higher accuracy on residue (0.81 vs. 0.65), homogeneous liquid (0.86 vs. 0.81), and solid (0.71 vs 0.65), while performance was similar to the entire participant cohort for heterogeneous liquid (0.76 vs. 0.78). This suggests that participants with experimental chemistry experience did not necessarily perform better when directly annotating the phases, but did outperform the general participants when asked to report the existence of a specific phase, particularly for residue.

S4.8 User Study Survey Link and Consent

The user study survey demo can be accessed at: https://tiny.cc/cv-mof-userstudy-survey. The text of the consent form at the start of the survey is reproduced below:

“By participating in this study, you give us consent to:

- Collect your email address, your interactions and responses to the survey questions;
- Collect your demographic information, including but not limited to your age, your highest level of education, etc.;
- Contact you in the future for a potential follow-up interview or discussion;
- Include your interactions and responses anonymously in a research paper

Please note that:

- You may withdraw from the activity anytime and for any reason;
- You can request to have your interactions and responses removed even after you finish the survey, by contacting us at matterlabsurvey@cs.toronto.edu;
- Your email will remain confidential and not be shared with anyone. It is solely to facilitate follow-up discussions if needed;
- Your interactions and responses will be kept private and confidential;
- Your responses will not be sent to the server unless you choose to submit them at the end of the survey.

If you have any questions about this study, please contact us at matterlabsurvey@cs.toronto.edu.

If you have no other concerns, please write your email and click the "I consent" button below, and we will prepare you to start the activity.”

5. References

(1) Eppel, S.; Xu, H.; Bismuth, M.; Aspuru-Guzik, A. Computer Vision for Recognition of Materials and Vessels in Chemistry Lab Settings and the Vector-LabPics Data Set. *ACS Central Science* **2020**, *6* (10), 1743-1752.

(2) *Ultralytics YOLO (Version 8.0.0) [Software]*; https://github.com/ultralytics/ultralytics, 2023.

(3) *Roboflow (Version 1.0) [Software]*; https://roboflow.com, 2024.