

Supplemental Information for "vIR-OLO: an open-source YOLO-based interface for functional-group detection and localization in infrared spectra"

Uriel Garcilazo-Cruz,^{a,b,†} Qichen Nie^{a,b}, Rudra Shondi^a, and Rodrigo A. Vargas-Hernández^{a,b,c,†}

The purpose of this supplemental information is to provide additional details on the results and protocols used to generate the main portion of the manuscript. Section S1 provides detailed information on the protocols used to obtain the data used in the creation of the datasets used to train the YOLO-based architectures. Section S3 provides details on the architecture and choices made during training and evaluation of the models presented in the main manuscript. Section S4 presents detailed information on the metrics obtained by every trained model. Finally, Section S5 details specific technical information on the modular design of the vIR-OLO GUI interface. Throughout this document, and consistent with the main text, the three pretrained models (vIR-OLO-10FG, vIR-OLO-12FG, and vIR-OLO-13FG) are referred to by their short names 10FG, 12FG, and 13FG, respectively.

S1 Dataset preparation

The data downloaded from the Spectral Database for Organic Compounds (SDBS) were manually annotated using an early version of the vIR-OLO software, generating a total of 2,200 image/annotation pairs. Table S1 presents the relative frequency of each functional group in the original dataset.

Table S1 Absolute and relative frequency of each functional group class in the raw dataset.

index	name	count	percentage (%)
0	aromatics	1664	29.5
1	alcohols	536	9.5
2	amines	776	13.8
3	esters	199	3.5
4	alkene	700	12.4
5	carb. acids	328	5.8
6	ketones	217	3.9
7	phenol	173	3.1
8	nitriles	171	3.0
9	amides	531	9.4
10	aldehydes	170	3.0
11	alkyne	170	3.0
	Total	5635	100.0

All models were split into three sub-datasets using an 80/10/10 ratio: test and validation both with 220 data points, whereas the training dataset included 1,760 data points, using a random seed

with value 42. The training dataset was later augmented five times for each original data point, totaling 10,560 data points.

The distribution of functional groups within the training and validation datasets for models 12FG, 10FG and 13FG is shown in Table S2, Table S3 and Table S4, respectively.

Table S2 Training, validation and test box counts for the model vIR-OLO-12FG.

index	name	train	validation	test
0	aromatics	8100	136	178
1	alcohols	2502	56	63
2	amines	3660	84	82
3	esters	984	16	19
4	alkene	3330	69	76
5	carb. acids	1542	31	40
6	ketones	1146	12	14
7	phenol	828	13	22
8	nitriles	786	22	18
9	amides	2508	56	57
10	aldehydes	828	21	11
11	alkyne	822	10	23
	Total	27036	526	603

Model 10FG: Prior to any pre-processing, each subdataset was traversed, and the functional groups ketones, aldehydes, and esters were collapsed into the label *carbonyl subgroups*. Table S3 provides the sample counts for this model.

Table S3 Training, validation and test box counts for the model vIR-OLO-10FG.

index	name	train	validation	test
0	aromatics	8100	136	178
1	alcohols	2502	56	63
2	amines	3660	84	82
3	carbonyl subgroups	2958	49	44
4	alkenes	3330	69	76
5	carb. acids	1542	31	40
6	phenols	828	13	22
7	nitriles	786	22	18
8	amides	2508	56	57
9	alkynes	822	10	23
	Total	27036	526	603

Model 13FG: The pre-processing step for this model also involved the creation of the label *carbonyl subgroups* observed in model 10FG, followed by an additional splitting of alkynes (alkynes-L, alkynes-R), alkenes (alkenes-L, alkenes-R), and amines (amines-L, amines-R). Table S4 provides the sample counts for this model.

^a Department of Chemistry and Chemical Biology, McMaster University, Hamilton, ON, Canada

^b School of Computational Science and Engineering, McMaster University, Hamilton, ON, Canada

^c Brockhouse Institute for Materials Research, McMaster University, Hamilton, ON, Canada

[†] garcilau@mcmaster.ca, vargasrh@mcmater.ca

S2 Data Driven Decisions

Although aldehydes and esters typically only show a slight overlap in their absorption frequencies (1730 cm^{-1} with additional C-H stretching bands near $2720\text{--}2820\text{ cm}^{-1}$ for aldehydes, and $1735\text{--}1750\text{ cm}^{-1}$ with coupled C=O stretching for esters), these distinctions can be obscured by variability that naturally occurs within the infrared (IR) spectrum. Ketones (1715 cm^{-1}) further complicate the discrimination task due to the continuous spectral overlap across this region. This created a notorious challenge by introducing labeling bias during annotation, as the identities of given functional groups varied, forcing the annotator to rely on implicit chemical knowledge beyond what is strictly observable in the IR spectrum alone. By merging these groups, we aimed to assess the model’s response to the collapse of potentially overlapping functional groups and to reduce ambiguity in the quality of predictions as compared to our 12FG model.

For example, a terminal, non-symmetrical alkene typically shows a very sharp C–H stretching band near $3095\text{--}3010\text{ cm}^{-1}$ and a moderately sharp C=C stretching band near $1660\text{--}1600\text{ cm}^{-1}$, and both peaks would be labeled as the alkene class. Hence, multi-peak patterns like this would inevitably introduce detection challenges to YOLO models. To mitigate this phenomenon, the 13FG model represented alkenes, alkynes, and amines as position-dependent labels when their annotated regions occupied distinct spectral domains.

It is worth mentioning that aromatic compounds and carboxylic acids also exhibit multiple absorption bands. However, both groups achieved comparatively high F1-scores in the 10FG and 12FG models; see Section S4. Hence, from a data-driven perspective, these categories were retained as single classes in the 13FG model.

Table S4 Training, validation and test box counts for the model vIR-OLO-13FG.

index	name	train	validation	test
0	Aromatics	8100	136	178
1	Alcohols	2502	56	63
2	Amines_L	3936	87	90
3	Amines_R	966	20	21
4	carbonyl subgroups	2958	49	44
5	Alkenes_L	1200	25	23
6	Alkenes_R	2130	44	53
7	Carb. acids	1542	31	40
8	Phenols	828	13	22
9	Nitriles	786	22	18
10	Amides_R	1266	33	28
11	Alkynes_L	330	3	10
12	Alkynes_R	492	7	13
	Total	27036	526	603

S3 YOLO-based Model Architecture and Training

All models were trained using the Ultralytics package available on GitHub. The model yolo11l.pt was chosen because it was the most recent model with compatibility available through the python-ultralytics package optimized by Compute Canada. The

model is composed of 25.3 million parameters, divided into three main submodules called the backbone, neck, and head (Table S5).

Table S5 High-level architecture of the YOLO-based detection model.

Submodule	Blocks
Backbone	Conv, C2f block Conv, C2f block Conv, C2f block Conv, C2f block SPPF
Neck	Top-down FPN: Upsample, Concat (P4), C2f block Upsample, Concat (P3), C2f block Bottom-up PAN: Downsample, Concat, C2f block Downsample, Concat, C2f block
Head	Decoupled detection head Bounding box regression branch Classification branch

In a YOLO-based architecture, the backbone network progressively downsamples the input image while extracting increasingly abstract semantic features. This process is implemented through repeated pairs of convolutional layers and “C2f” blocks, where each C2f block performs partial channel splitting followed by multi-path feature processing and concatenation. The backbone ends with a Spatial Pyramid Pooling-Fast (SPPF) module that aggregates multi-scale contextual information at the deepest level. The neck then performs bidirectional feature fusion: a top-down feature pyramid network (FPN) propagates high-level semantic information into higher-resolution feature maps, while a bottom-up path aggregation network (PAN) reintegrates fine-grained spatial details into deeper layers. This bidirectional fusion produces three enriched feature maps (P3, P4, and P5), balancing localization precision and semantic discrimination. Finally, the detection head operates on these multi-scale feature maps using decoupled branches for bounding box regression and classification, enabling efficient single-stage, multi-scale object detection.

S4 Models performance

The input size of the IR spectral images was provided at the same resolution for all models. The native resolution of each image was 715×353 pixels, and it was subsequently resized to square inputs of 640×640 pixels. The model was trained with a batch size of 64 on a single compute node using an NVIDIA A100 SXM4 GPU with 40 GB of memory.

Optimization followed the Ultralytics default setting `optimizer=auto`, which selects the optimization algorithm based on training duration. For longer training runs, this resolves to stochastic gradient descent (SGD) with momentum and Nesterov acceleration, while for shorter runs it may resolve to AdamW. The learning rate was initialized at the framework-defined value lr_0 and decayed during training toward a final fraction $lr_f \cdot lr_0$ using Ultralytics’ built-in scheduling strategy, which includes a warmup phase for both weights and bias parameters. When SGD is selected, momentum is applied explicitly;

for Adam-family optimizers, the first moment coefficient (β_1) serves an analogous role.

The training objective comprised the standard YOLO multi-component loss, including bounding box regression loss, classification loss, and distribution focal loss (dfl) for improved localization precision. These components are combined using Ultralytics' default loss weights. Early stopping was controlled via a patience parameter, halting training when validation metrics failed to improve over a fixed number of epochs. Model checkpoints, training logs, and evaluation metrics were automatically recorded throughout training for subsequent analysis.

Figure S1 presents a radial plot summarizing the localization-based F1 scores across functional groups for three models: vIR-OL0-10FG (Fig. S1A), vIR-OL0-12FG (Fig. S1B), and vIR-OL0-13FG (Fig. S1C).

vIR-OL0-12FG: The model was trained on 10,560 samples using a batch size of 64 with early stopping enabled, terminating training at epoch 94. Training losses decreased smoothly and plateaued, indicating convergence, with near-final values reached by approximately epoch 60 (Fig. S2). In contrast, validation losses exhibited substantial instability during the early training phase (first 40 epochs), particularly in the classification (cls_loss) component, which showed repeated spikes exceeding values of 25. Box and distribution focal losses also displayed significant oscillations (Fig. S3). After approximately epoch 60, validation losses stabilized and converged toward their training counterparts, with final values of 1.561 (box loss), 0.880 (classification loss), and 1.379 (distribution focal loss). This early instability suggests an initial mismatch between training and validation distributions or transitional overfitting, which resolved during later stages of training, resulting in stable convergence with a minimal generalization gap. Evaluation on 220 test images/annotation pairs yielded a micro-averaged precision of 0.828 and recall of 0.862, corresponding to an incidence-based F1 score of 0.845. Macro-averaged precision, recall, and F1 were 0.793, 0.864, and 0.821, respectively, with an exact match ratio (EMR) of 0.682 (Table S9). These metrics assess the presence of functional groups on a per-image basis without incorporating localization accuracy. Table S6 provides details on the incidence-based metrics across functional groups for model 12FG.

To evaluate localization performance, bounding box predictions were assessed using an IoU threshold of 0.7, consistent with the training configuration. Across all twelve functional groups, the model achieved a precision of 0.646, recall of 0.668, and a localization F1 score of 0.657, with a mean IoU of 0.805 across matched detections (Table S10). Despite the high mean IoU, indicating accurate spatial alignment for detected objects, the moderate F1 score reflects challenges in recall for certain functional groups, particularly those with limited representation in the training dataset or more complex spectral features. Table S11 provides a detailed summary of localization metrics across functional groups for the 12FG model.

vIR-OL0-10FG: Trained on the same number of data points, batch size and early stopping as model 12FG. Training terminated at epoch 100. Training losses decreased smoothly and plateaued, with some signs of overtraining after epoch 60 when compared to

the validation curves (Fig. S2).

Validation losses exhibited nearly identical instability during the early training phase (first 40 epochs), particularly in the classification component. Box and distribution focal losses also displayed significant oscillations (Fig. S3). After epoch 60, validation losses stabilized and converged toward their training counterparts, with final values of 1.528 (box loss), 0.860 (classification loss), and 1.360 (distribution focal loss). This behavior closely mirrors that observed in the 12FG model, suggesting a common training dynamic likely associated with dataset composition or class imbalance effects. Evaluation on the test dataset yielded micro-averaged precision, recall and F1 scores of 0.840. Macro-averaged precision, recall, and F1 were 0.819, 0.830, and 0.821, respectively, with an EMR of 0.718 (Table S9). Per-class F1 scores ranged from 0.69 (amides) to 0.94 (aromatics), with the merged *carbonyl subgroups* class achieving an F1 score of 0.80. Table S7 provides a detailed summary of incidence-based metrics across functional groups for the 10FG model.

Localization performance was also evaluated using an IoU threshold of 0.7. Across all ten functional groups, the model achieved a precision of 0.654, a recall of 0.652, and a localization F1 score of 0.653, with a mean IoU of 0.810 across matched detections (Table S10). Despite the high mean IoU, indicating accurate spatial alignment of detected objects, the moderate F1 score reflects challenges in recall across certain functional groups, particularly those with fewer training examples or more complex spectral characteristics. Table S12 details localization metrics across functional groups for the 10FG model.

vIR-OL0-13FG: Followed the same training protocols as the models 12FG and 10FG. Training terminated at epoch 79. Training losses decreased smoothly and plateaued, with near-final values reached by approximately epoch 60 (Fig. S2), indicating stable convergence.

Validation losses exhibited the same substantial instability during the early training phase with an additional peak at around epoch 35, and stabilizing after epoch 40. Box and distribution focal losses also displayed significant oscillations (Fig. S3). After epoch 60, validation losses stabilized and converged toward their training counterparts, with final values of 1.537 (box loss), 0.822 (classification loss), and 1.367 (distribution focal loss).

Evaluation on the test dataset yielded a micro-averaged precision of 0.857 and recall of 0.853, corresponding to an incidence-based F1 score of 0.855. Macro-averaged precision, recall, and F1 were 0.861, 0.826, and 0.837, respectively, with an EMR of 0.673. Per-class F1 scores ranged from 0.69 (amides-R) to 0.94 (aromatics), with the *carbonyl subgroups* class achieving an F1 score of 0.80. Table S8 provides a detailed summary of incidence-based metrics across functional groups for the 13FG model.

Localization performance was also evaluated using an IoU threshold of 0.7. Across all thirteen functional groups, the model achieved a precision of 0.663, a recall of 0.673, and a localization F1 score of 0.668, with a mean IoU of 0.806 across matched detections (Table S10). Despite the high mean IoU, which indicates accurate spatial alignment of detected objects, the moderate F1 score reflects challenges with recall across certain functional groups. The introduction of position-dependent labels

resulted in heterogeneous performance: *carbonyl subgroups* remained broadly consistent with the 10FG model, *alkynes-L* improved relative to the unsplit alkyne class, whereas *alkynes-R* remained lower. Splitting amines produced asymmetric results, with stronger localization for *amines-R* than for *amines-L*, whereas subdividing alkenes led to reduced F1 scores relative to the original labels. Table S13 provides a detailed summary of localization-based metrics across functional groups for the 13FG model.

S5 vIR-OLO software specifications

S5.1 Information workflow

The vIR-OLO software is organized into two main components: a graphical user interface (GUI) and a filesystem-based project workspace. The GUI provides an interactive environment for image visualization and annotation, replacing more traditional command-line-based workflows (e.g., Jupyter notebooks or Google Colab). This design choice reflects the inherently visual nature of IR spectral data and object detection tasks, which benefit from direct graphical interaction.

To support scalability and compatibility with standard object detection pipelines, the software organizes user-provided data into a structured project directory (Fig. S4;A). This workspace contains subfolders for images, labels, and models, following conventions commonly used in YOLO-based workflows. Selected images are stored in `./images/train`, while the corresponding validation directory remains empty at this stage, as dataset splitting is performed during postprocessing using auxiliary scripts provided in the repository. Upon project initialization, configuration files such as `dataset.yaml` and `config.json` are generated at the root of the workspace.

When a project is loaded, vIR-OLO automatically imports images into the GUI canvas for visualization and annotation (Fig. S4;B). User-generated annotations are stored using the YOLO format, where each bounding box is represented by normalized center coordinates (x, y) and dimensions (width, height), relative to the image size. These annotations are saved as `.txt` files with filenames matching the corresponding image, and are stored in `./labels/train` (Fig. S4;C). The software continuously synchronizes annotations with the displayed images by matching filenames.

After annotation is complete, model training is performed via helper scripts provided in the repository (full GUI integration is planned for future releases). The resulting trained models can then be reloaded into the application to assist with further annotation through automated predictions (Fig. S4;D).

S5.2 Modular architecture and design

The vIR-OLO software adopts a modular, event-driven architecture centered around a main application controller (App). This design enables coordinated interaction between multiple interdependent components while maintaining a consistent shared application state. In contrast to notebook-based environments such as Jupyter or Google Colab, where execution is linear and state is distributed across cells, this architecture supports real-time, bidirectional communication between modules and ensures precise

synchronization between user actions and backend processes.

As illustrated in Fig. S5, the App Module acts as the central orchestrator, receiving input from external sources (I/O) and coordinating operations across all other modules. The UI Module provides the primary interaction layer, exposing both high-level widgets and a dedicated canvas for image visualization and annotation. User actions performed on the canvas (e.g., drawing or modifying bounding boxes) are propagated to the Box Manager Module, which maintains the collection of annotations and ensures their consistency.

Image data is handled by the Image Loader Module, which supplies images to the UI for rendering and manages transformations between image and display coordinates. The File Manager Module is responsible for persistence, handling the storage and retrieval of images, labels, and configuration files within the project workspace. This module ensures that all annotation updates are reflected in the corresponding label files.

Model-related functionality is encapsulated within the model downloader and model prediction module. The model downloader retrieves pretrained or user-defined models, which are then used by the prediction module to perform inference on loaded images. The resulting predictions are returned to the UI and integrated into the annotation workflow via the Box Manager Module.

The arrowheads in Fig. S5 indicate dependency relationships between modules, highlighting how information flows through the system under the coordination of the App. This modular decomposition enables extensibility, as individual components (e.g., prediction models or data handling routines) can be modified or replaced without affecting the overall system behavior.

S5.3 The Ultralytics pipeline

We used the **YOLO11l** object detection model implemented in the **Ultralytics** framework. The pretrained model weights were obtained from the official **Ultralytics YOLO11 documentation**, and training was conducted using a custom Python wrapper adapted from the official usage examples provided in the **Ultralytics GitHub repository**. This wrapper specifies the path to the pretrained checkpoint and the training hyperparameters. For all experiments, we used an input image size of **640 pixels**, a **batch size of 64**, and an **early stopping patience of 50 epochs**. Default data augmentation from Ultralytics was handled internally by the built-in training pipeline rather than being explicitly coded in our wrapper. The default augmentations applied during training included **HSV-based color variation** (hue, saturation, and brightness adjustments), **translation**, **scale-based resizing**, **horizontal flipping**, and **mosaic augmentation**, in which multiple training images are combined into a single composite image. Most of the detection and training pipeline, including these augmentation procedures, was therefore managed directly by the Ultralytics framework.

Table S6 Functional group presence-based classification performance for model vIR-OL0-12FG. Supp.: Support; Pred.: Predicted positives; TP: True positives; FP: False positives; FN: False negatives; Prec.: Precision; Rec.: Recall.

Class	Supp.	Pred.	TP	FP	FN	Prec.	Rec.	F1
aromatics	79	84	77	7	2	0.92	0.97	0.94
alkyne	15	16	14	2	1	0.88	0.93	0.90
carb. acids	20	23	19	4	1	0.83	0.95	0.88
alcohols	63	57	52	5	11	0.91	0.83	0.87
nitriles	18	19	16	3	2	0.84	0.89	0.86
amines	44	49	40	9	4	0.82	0.91	0.86
esters	19	19	16	3	3	0.84	0.84	0.84
alkene	53	44	38	6	15	0.86	0.72	0.78
amides	27	30	22	8	5	0.73	0.81	0.77
aldehydes	11	18	11	7	0	0.61	1.00	0.76
phenol	22	22	16	6	6	0.73	0.73	0.73
ketones	14	20	11	9	3	0.55	0.79	0.65

Table S7 Functional group presence-based classification performance for model vIR-OL0-10FG. Supp.: Support; Pred.: Predicted positives; TP: True positives; FP: False positives; FN: False negatives; Prec.: Precision; Rec.: Recall.

Class	Supp.	Pred.	TP	FP	FN	Prec.	Rec.	F1
aromatics	79	81	75	6	4	0.93	0.95	0.94
carb. acids	20	23	19	4	1	0.83	0.95	0.88
alcohols	63	56	52	4	11	0.93	0.83	0.87
nitriles	18	21	17	4	1	0.81	0.94	0.87
amines	44	50	39	11	5	0.78	0.89	0.83
alkenes	53	41	38	3	15	0.93	0.72	0.81
carbonyl subgroups	41	44	34	10	7	0.77	0.83	0.80
alkynes	15	13	11	2	4	0.85	0.73	0.79
phenols	22	22	16	6	6	0.73	0.73	0.73
amides	27	31	20	11	7	0.65	0.74	0.69

S6 Repeated random-split analysis without explicit external augmentation

To estimate the sensitivity of model performance to dataset partitioning, we performed an additional repeated-split analysis independent of the augmented validation benchmark described in the main text. For each functional group configuration (10FG, 12FG, and 13FG), the available IR spectra were randomly partitioned eight times into training, validation, and test sets using a 70/15/15 split, yielding 1540 training, 330 validation, and 330 test spectra per partition from a pool of 2200 annotated spectra. Each partition was used to train a separate YOLO11l model with the same input size, batch size, and early-stopping settings described above. This protocol was not intended to replace the augmented benchmark, but rather to quantify split-level variability under a more conservative, unaugmented setting.

S6.1 Training stability across random splits

Training stability was assessed using the validation mAP50-95(B) metric recorded by Ultralytics at each epoch for bounding-box detection metrics. The mAP50-95 score is the mean average precision computed over multiple localization criteria, averaging AP across IoU thresholds from 0.50 to 0.95. Compared with individual loss terms such as box loss or classification loss, this metric provides a validation-set measure of detection performance that

Table S8 Functional group presence-based classification performance for model vIR-OL0-13FG. Supp.: Support; Pred.: Predicted positives; TP: True positives; FP: False positives; FN: False negatives; Prec.: Precision; Rec.: Recall.

Class	Supp.	Pred.	TP	FP	FN	Prec.	Rec.	F1
Aromatics	79	85	77	8	2	0.91	0.97	0.94
Carb. acids	20	22	19	3	1	0.86	0.95	0.90
Nitriles	18	18	16	2	2	0.89	0.89	0.89
Alkenes_L	23	20	19	1	4	0.95	0.83	0.88
Alcohols	63	57	53	4	10	0.93	0.84	0.88
Phenols	22	17	17	0	5	1.00	0.77	0.87
Amines_L	55	60	50	10	5	0.83	0.91	0.87
Amines_R	21	23	18	5	3	0.78	0.86	0.82
Alkynes_R	12	10	9	1	3	0.90	0.75	0.82
Alkenes_R	52	42	38	4	14	0.90	0.73	0.81
Carbonyl subgroups	41	46	35	11	6	0.76	0.85	0.80
Alkynes_L	10	7	6	1	4	0.86	0.60	0.71
Amides_R	27	34	21	13	6	0.62	0.78	0.69

Table S9 Overall incidence-based classification performance comparison between models on the test subdataset.

Model	Micro	Prec.	Rec.	F1
vIR-OL0-10FG		0.8403	0.8403	0.8403
vIR-OL0-12FG		0.8279	0.8623	0.8448
vIR-OL0-13FG		0.8571	0.8533	0.8552
Model	Macro	Prec.	Rec.	F1
vIR-OL0-10FG		0.8188	0.8303	0.8208
vIR-OL0-12FG		0.7930	0.8640	0.8212
vIR-OL0-13FG		0.8611	0.8255	0.8373
Model	EMR			
vIR-OL0-10FG		0.7182		
vIR-OL0-12FG		0.6818		
vIR-OL0-13FG		0.6727		

jointly reflects classification confidence and bounding-box localization quality. Loss curves remain useful for diagnosing optimization behavior, but mAP50-95 more directly captures whether the trained detector is improving on held-out data.

Across the eight random splits, all three model configurations showed a consistent increase in validation mAP50-95 during training, followed by stabilization (Fig. S6). The shaded bands represent the standard deviation across runs at each epoch and therefore summarize the variability introduced by the random partitioning procedure.

S6.2 Held-out test performance

Final model performance was evaluated on the held-out test split for each random partition. Predictions were generated using a confidence threshold of 0.10, and localization-aware true positives were defined using one-to-one class-wise matching with an IoU threshold of 0.70. For each model configuration, summary metrics were first computed within each split and then reported as mean $\pm$ standard deviation across the eight runs (Table S14).

The three configurations achieved closely comparable held-out performance, with micro-F1 scores spanning only 0.674–0.688 and

Table S10 Overall localization-based performance comparison between models on the test set. Prec.: Precision; Rec.: Recall; mIoU: mean Intersection over Union.

Model	Prec.	Rec.	F1	mIoU
vIR-OL0-10FG	0.6539	0.6517	0.6528	0.8096
vIR-OL0-12FG	0.6458	0.6683	0.6569	0.8046
vIR-OL0-13FG	0.6634	0.6733	0.6683	0.8064

Table S11 Localization-based performance by functional group class for model vIR-OL0-12FG. TP: True positives; FP: False positives; FN: False negatives; Prec.: Precision; Rec.: Recall.

Class	TP	FP	FN	Precision	Recall	F1	mIoU
esters	16	3	3	0.842	0.842	0.842	0.807
carb. acids	35	10	5	0.778	0.875	0.824	0.814
alcohols	49	14	14	0.778	0.778	0.778	0.808
phenol	15	7	7	0.682	0.682	0.682	0.799
aldehydes	10	9	1	0.526	0.909	0.667	0.824
aromatics	116	64	62	0.644	0.652	0.648	0.806
alkene	45	19	31	0.703	0.592	0.643	0.795
amines	53	41	29	0.564	0.646	0.602	0.803
nitriles	11	8	7	0.579	0.611	0.595	0.792
alkyne	12	7	11	0.632	0.522	0.571	0.792
amides	33	27	24	0.550	0.579	0.564	0.804
ketones	8	12	6	0.400	0.571	0.471	0.804

macro-F1 scores 0.670–0.681; the 13FG model showed the largest split-level variability, consistent with the main-text interpretation that finer subdivision of chemically overlapping classes introduces additional ambiguity. The error profile is asymmetric but moderate: macro recall (0.776–0.795) exceeds macro precision (0.593–0.601), and the models emit on average 1.34–1.35 predicted boxes for every ground-truth annotation. The unaugmented detectors therefore recover a high proportion of annotated bands at the cost of a modest excess of unmatched predictions. The consistently high mean IoU values (0.907–0.913) indicate that, when detections are successfully matched, the predicted regions are almost exactly aligned with the annotated boxes. The dominant source of error in this repeated-split analysis is therefore not poor spatial overlap among matched detections, but class-specific false positives together with missed detections in the least represented classes. A complete per-class breakdown of counts and metrics is given in Table S18.

The per-functional-group results further illustrate that performance differences are class dependent (Figs. S7–S9). Nitriles and alcohols were consistently the strongest classes across all three model configurations (F1 0.782–0.800 and 0.783–0.790, respectively), whereas amides, ketones, and the more finely resolved alkyne subclasses were weakest, reaching a minimum of 0.504 ± 0.059 for ketones in 12FG and 0.516 ± 0.074 for alkynes-L in 13FG. This ordering tracks class representation: the labels with the fewest ground-truth boxes per split — alkynes-L (~10), alkynes-R (~14), and esters (~34) — also carry the largest standard deviations, since a change of a few detections moves their F1 substantially. Precision is the weaker component for the chemically coupled carbonyl-adjacent labels (amides 0.450–0.484, ketones 0.428), whereas recall remains high throughout (0.58–0.91). These trends support the interpretation that repeated-split performance is governed by both class representation and intrinsic

Table S12 Localization-based performance by functional group class for model vIR-OL0-10FG. TP: True positives; FP: False positives; FN: False negatives; Prec.: Precision; Rec.: Recall.

Class	TP	FP	FN	Precision	Recall	F1	mIoU
alcohols	50	8	13	0.862	0.794	0.826	0.820
carb. acids	32	11	8	0.744	0.800	0.771	0.810
phenols	16	6	6	0.727	0.727	0.727	0.823
carbonyl subgroups	31	15	13	0.674	0.705	0.689	0.806
aromatics	118	63	60	0.652	0.663	0.657	0.811
alkenes	45	18	31	0.714	0.592	0.647	0.803
alkynes	13	6	10	0.684	0.565	0.619	0.787
amines	47	42	35	0.528	0.573	0.550	0.806
amides	31	28	26	0.525	0.544	0.534	0.806
nitriles	10	11	8	0.476	0.556	0.513	0.809

Table S13 Localization-based performance by functional group class for model vIR-OL0-13FG. TP: True positives; FP: False positives; FN: False negatives; Prec.: Precision; Rec.: Recall.

Class	TP	FP	FN	Precision	Recall	F1	mIoU
Phenols	17	1	5	0.944	0.773	0.850	0.811
Alcohols	52	9	11	0.852	0.825	0.839	0.816
Carb. acids	32	7	8	0.821	0.800	0.810	0.826
Amines_R	17	7	4	0.708	0.810	0.756	0.819
Alkenes_L	15	5	8	0.750	0.652	0.698	0.772
Alkynes_R	8	2	5	0.800	0.615	0.696	0.800
carbonyl subgroups	32	19	12	0.627	0.727	0.674	0.813
Aromatics	118	70	60	0.628	0.663	0.645	0.800
Amines_L	55	40	35	0.579	0.611	0.595	0.800
Nitriles	11	8	7	0.579	0.611	0.595	0.792
Alkynes_L	5	2	5	0.714	0.500	0.588	0.791
Alkenes_R	28	17	25	0.622	0.528	0.571	0.797
Amides_R	16	19	12	0.457	0.571	0.508	0.844

spectral ambiguity.

S6.3 Robustness of localization-based F1 under unaugmented repeated splits

To determine how representative our main training protocol is, we compared the per-functional-group localization-based F1 obtained under two regimes: The first is the main protocol with a single fixed 80/10/10 split in which the training set is augmented (five perturbed copies per spectrum plus original, giving 10,560 training pairs) and 220 unique held-out spectra (10% of the 2,200-spectrum corpus) used as the test set on which F1 is computed (Tables S12–S13). The second is the repeated-split with eight random 70/15/15 partitions and no external augmentation, reporting the mean and standard deviation of the held-out test F1 (Tables S15–S17).

At the model level the two regimes agree closely (micro-F1

Table S14 Held-out test performance across eight random splits. Values are the mean $\pm$ standard deviation across runs. Macro metrics are averaged across functional group classes within each run; micro-F1 is computed from pooled TP, FP, and FN counts within each run. MaP, MaR, MaF1, MiF1, and mIoU denote macro precision, macro recall, macro F1, micro F1, and mean IoU, respectively.

Model	MaP	MaR	MaF1	MiF1	mIoU
10FG	0.601 ± 0.028	0.795 ± 0.018	0.681 ± 0.021	0.680 ± 0.022	0.908 ± 0.003
12FG	0.596 ± 0.030	0.776 ± 0.021	0.670 ± 0.022	0.674 ± 0.022	0.913 ± 0.003
13FG	0.593 ± 0.050	0.787 ± 0.024	0.671 ± 0.036	0.688 ± 0.032	0.907 ± 0.004

0.653 $\rightarrow$ 0.680 $\pm$ 0.022 for 10FG, 0.657 $\rightarrow$ 0.674 $\pm$ 0.022 for 12FG, and 0.668 $\rightarrow$ 0.688 $\pm$ 0.032 for 13FG), indicating that the single augmented split is representative of the broader partition distribution and that the aggregate localization behavior is not an artifact of one favorable split. The small advantage of the repeated-split regime (+0.017 to +0.027) lies within about one standard deviation and should not be over-interpreted. Decomposing it is nonetheless informative: micro recall is substantially higher under repeated splitting (0.800, 0.787, and 0.803, versus 0.652, 0.668, and 0.673 for the augmented benchmark), whereas micro precision is lower (0.592, 0.591, and 0.603, versus 0.654, 0.646, and 0.663). Correspondingly, the ratio of predicted to ground-truth boxes rises from 1.00–1.04 under the augmented benchmark to 1.34–1.35 under repeated splitting. The unaugmented models therefore trade a modest amount of precision for a substantially higher detection rate, arriving at a similar F1 by a different route. The informative divergences are per-class.

Three patterns recur consistently across all three models and are therefore unlikely to be split-specific artifacts. (i) **Abundant, spectrally distinct classes are stable or marginally stronger without augmentation.** Aromatics — the most represented class — scored slightly higher under the repeated-split test (+0.04 to +0.05) with a tight standard deviation (0.025–0.043); alcohols were reproduced to within sampling noise (−0.04, +0.01, −0.06), and amines improved in 10FG, 12FG, and 13FG amines-L (+0.09, +0.06, +0.10). For these classes, augmentation provides no measurable benefit, consistent with their characteristic features already being densely sampled. (ii) **Narrow, sharply localized peaks scored markedly higher under the unaugmented repeated split.** Nitriles improved in every model, and by a wide margin (+0.28 for 10FG, +0.21 for 12FG, and +0.19 for 13FG), far exceeding the corresponding standard deviations. A plausible mechanism is that the augmentation, which enlarges bounding boxes by up to 25% and jitters their centers, biases the detector toward looser boxes; for a narrow absorption such as the nitrile band near $\sim 2200\text{ cm}^{-1}$, such inflation is more likely to fall below the strict $\text{IoU} \geq 0.7$ criterion, so the tight-box unaugmented regime is favored. The evidence from alkynes ($\sim 2100\text{ cm}^{-1}$) is mixed — +0.06 in 12FG but −0.01 in 10FG, −0.04 for alkynes-R and −0.07 for alkynes-L in 13FG, all within one to two standard deviations — so the mechanism remains a hypothesis rather than a demonstrated cause. (iii) **Rare or chemically coupled classes degrade under repeated splitting.** Phenols dropped in all three models (−0.11 in 10FG, −0.07 in 12FG, and −0.23 in 13FG), as did carboxylic acids (−0.05, −0.11, −0.09) and, in 13FG, the position-resolved amines-R (−0.10) and alkynes-L (−0.07). For these labels, the augmented-validation value should be read as an optimistic single draw rather than an expected level of performance, and the contrast most plausibly reflects the combined effect of small per-split sample sizes and the additional training exposure that augmentation provides to underrepresented classes.

Two caveats bound the interpretation. First, the two regimes differ in three confounded factors simultaneously — the presence or absence of training augmentation, a single versus eight splits, and the training/evaluation proportions (80/10/10 with 10,560 augmented training pairs versus 70/15/15 with 1540 unaug-

mented training spectra) — so observed differences cannot be attributed to augmentation alone. Second, the augmented-benchmark value is a *single* draw without replicates, so the repeated-split standard deviation is our only estimate of how much a single partition can vary; any per-class gap smaller than roughly one to two repeated-split standard deviations is indistinguishable from split-to-split sampling noise. Both regimes are evaluated against the same 2,200-spectrum annotated corpus, so differences in annotation coverage do not contribute to the comparison.

Finally, mean IoU remained high in both regimes for every class — at least 0.772 under the augmented benchmark and at least 0.850 under repeated splitting — confirming that the spatial overlap of matched boxes is robust and that the per-class F1 differences are driven by detection matching (recall and class-specific false positives) rather than by localization precision. Mean IoU is in fact markedly higher under repeated splitting (0.907–0.913 versus 0.805–0.810), which is independently consistent with the box-inflation mechanism proposed above: without the enlarging augmentation, the boxes the detector does match sit tighter around the annotated bands. Taken together, the comparison indicates that a 70/15/15 unaugmented protocol reproduces the main protocol’s overall localization performance and its strongest classes, while augmentation and the larger training fraction are most plausibly beneficial for underrepresented and spectrally overlapping functional groups — precisely the classes for which a single split is least trustworthy.

Table S15 Repeated-split held-out *test* localization-based F1 by functional group for model vIR-OL0–10FG. Values are mean $\pm$ standard deviation across eight random 70/15/15 splits (no external augmentation). These are the values plotted in the corresponding histogram.

Class	F1 (mean)	Std
nitriles	0.794	0.062
alcohols	0.787	0.017
carb. acids	0.724	0.067
carbonyl subgroups	0.712	0.033
aromatics	0.695	0.043
alkenes	0.663	0.035
amines	0.639	0.033
phenols	0.618	0.050
alkynes	0.608	0.095
amides	0.572	0.040

Table S16 Repeated-split held-out *test* localization-based F1 by functional group for model vIR-0L0-12FG. Values are mean $\pm$ standard deviation across eight random 70/15/15 splits (no external augmentation). These are the values plotted in the corresponding histogram.

Class	F1 (mean)	Std
nitriles	0.800	0.078
alcohols	0.790	0.036
esters	0.782	0.042
carb. acids	0.712	0.052
aromatics	0.697	0.025
alkenes	0.661	0.037
amines	0.660	0.045
alkynes	0.635	0.060
aldehydes	0.618	0.095
phenols	0.617	0.058
amides	0.568	0.040
ketones	0.504	0.059

Table S17 Repeated-split held-out *test* localization-based F1 by functional group for model vIR-0L0-13FG. Values are mean $\pm$ standard deviation across eight random 70/15/15 splits (no external augmentation). These are the values plotted in the corresponding histogram.

Class	F1 (mean)	Std
alcohols	0.783	0.033
nitriles	0.782	0.109
alkenes-L	0.730	0.037
carb. acids	0.723	0.107
amines-L	0.697	0.060
aromatics	0.695	0.036
carbonyl subgroups	0.694	0.054
amines-R	0.655	0.053
alkynes-R	0.653	0.076
phenols	0.622	0.063
alkenes-R	0.621	0.058
amides-R	0.548	0.042
alkynes-L	0.516	0.074

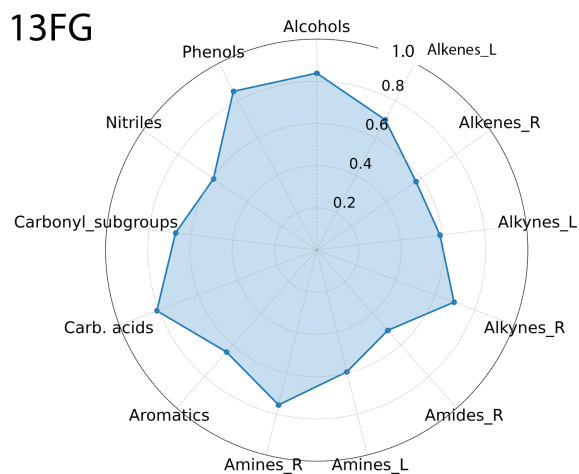
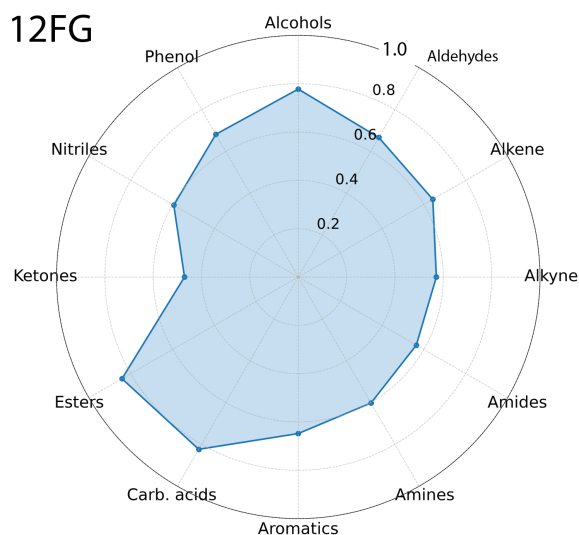
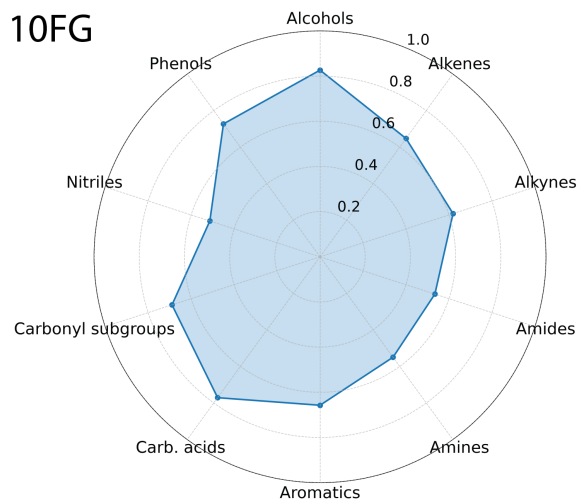


Fig. S1 Comparison of localization-aware F1 scores between YOLO-based models 10FG, 12FG and 13FG.

Table S18 Complete per-functional-group held-out test metrics across eight random 70/15/15 splits (no external augmentation). Count columns (GT, Pred., TP, FP, FN) are means across the eight runs; Precision, Recall, F1, and mIoU are given as mean $\pm$ standard deviation across runs. The systematic excess of predictions over ground-truth boxes is the origin of the low precision discussed in Section S6.3.

Class	GT	Pred.	TP	FP	FN	Precision	Recall	F1	mIoU
vIR-OL0-10FG									
aromatics	253.9	347.5	208.9	138.6	45.0	0.604 $\pm$ 0.059	0.823 $\pm$ 0.021	0.695 $\pm$ 0.043	0.883 $\pm$ 0.005
alcohols	81.5	89.8	67.4	22.4	14.1	0.754 $\pm$ 0.027	0.826 $\pm$ 0.032	0.787 $\pm$ 0.017	0.960 $\pm$ 0.005
amines	110.4	162.9	87.1	75.8	23.3	0.539 $\pm$ 0.042	0.788 $\pm$ 0.033	0.639 $\pm$ 0.033	0.902 $\pm$ 0.004
carbonyl subgroups	95.9	125.4	78.8	46.6	17.1	0.630 $\pm$ 0.054	0.821 $\pm$ 0.028	0.712 $\pm$ 0.033	0.930 $\pm$ 0.006
alkenes	102.8	139.9	80.3	59.6	22.5	0.576 $\pm$ 0.048	0.782 $\pm$ 0.022	0.663 $\pm$ 0.035	0.884 $\pm$ 0.004
carb. acids	50.9	63.8	41.5	22.3	9.4	0.657 $\pm$ 0.102	0.814 $\pm$ 0.033	0.724 $\pm$ 0.067	0.936 $\pm$ 0.008
phenols	26.3	37.0	19.4	17.6	6.9	0.535 $\pm$ 0.073	0.741 $\pm$ 0.061	0.618 $\pm$ 0.050	0.944 $\pm$ 0.010
nitriles	27.8	36.3	25.5	10.8	2.3	0.704 $\pm$ 0.070	0.914 $\pm$ 0.059	0.794 $\pm$ 0.062	0.859 $\pm$ 0.013
amides	82.8	122.9	58.6	64.3	24.1	0.484 $\pm$ 0.052	0.710 $\pm$ 0.070	0.572 $\pm$ 0.040	0.919 $\pm$ 0.008
alkynes	23.9	33.5	17.5	16.0	6.4	0.529 $\pm$ 0.121	0.734 $\pm$ 0.069	0.608 $\pm$ 0.095	0.866 $\pm$ 0.008
vIR-OL0-12FG									
aromatics	253.9	343.9	207.6	136.3	46.3	0.609 $\pm$ 0.046	0.818 $\pm$ 0.021	0.697 $\pm$ 0.025	0.882 $\pm$ 0.005
alcohols	81.5	92.9	68.9	24.0	12.6	0.742 $\pm$ 0.040	0.846 $\pm$ 0.043	0.790 $\pm$ 0.036	0.959 $\pm$ 0.004
amines	110.4	152.0	86.1	65.9	24.3	0.575 $\pm$ 0.066	0.781 $\pm$ 0.033	0.660 $\pm$ 0.045	0.904 $\pm$ 0.005
esters	34.4	37.4	28.1	9.3	6.3	0.752 $\pm$ 0.039	0.816 $\pm$ 0.054	0.782 $\pm$ 0.042	0.935 $\pm$ 0.009
alkenes	102.8	136.0	78.8	57.3	24.0	0.583 $\pm$ 0.058	0.768 $\pm$ 0.047	0.661 $\pm$ 0.037	0.880 $\pm$ 0.008
carb. acids	50.9	62.0	40.4	21.6	10.5	0.651 $\pm$ 0.063	0.789 $\pm$ 0.059	0.712 $\pm$ 0.052	0.936 $\pm$ 0.009
ketones	35.4	51.5	21.9	29.6	13.5	0.428 $\pm$ 0.073	0.627 $\pm$ 0.080	0.504 $\pm$ 0.059	0.926 $\pm$ 0.013
phenols	26.3	37.3	19.6	17.6	6.6	0.529 $\pm$ 0.063	0.744 $\pm$ 0.064	0.617 $\pm$ 0.058	0.949 $\pm$ 0.010
nitriles	27.8	34.6	24.9	9.8	2.9	0.727 $\pm$ 0.098	0.896 $\pm$ 0.053	0.800 $\pm$ 0.078	0.859 $\pm$ 0.014
amides	82.8	127.3	59.5	67.8	23.3	0.472 $\pm$ 0.044	0.718 $\pm$ 0.062	0.568 $\pm$ 0.040	0.919 $\pm$ 0.006
aldehydes	26.1	39.9	20.5	19.4	5.6	0.518 $\pm$ 0.109	0.779 $\pm$ 0.049	0.618 $\pm$ 0.095	0.930 $\pm$ 0.007
alkynes	23.9	30.6	17.4	13.3	6.5	0.566 $\pm$ 0.067	0.732 $\pm$ 0.103	0.635 $\pm$ 0.060	0.876 $\pm$ 0.020
vIR-OL0-13FG									
aromatics	253.9	338.5	206.4	132.1	47.5	0.612 $\pm$ 0.037	0.810 $\pm$ 0.071	0.695 $\pm$ 0.036	0.883 $\pm$ 0.006
alcohols	81.5	93.0	68.3	24.8	13.3	0.737 $\pm$ 0.049	0.838 $\pm$ 0.034	0.783 $\pm$ 0.033	0.960 $\pm$ 0.006
amines-L	121.3	160.6	97.4	63.3	23.9	0.620 $\pm$ 0.086	0.805 $\pm$ 0.048	0.697 $\pm$ 0.060	0.894 $\pm$ 0.004
amines-R	29.6	40.0	22.6	17.4	7.0	0.585 $\pm$ 0.100	0.765 $\pm$ 0.076	0.655 $\pm$ 0.053	0.944 $\pm$ 0.006
carbonyl subgroups	95.9	131.4	78.8	52.6	17.1	0.605 $\pm$ 0.077	0.822 $\pm$ 0.027	0.694 $\pm$ 0.054	0.929 $\pm$ 0.006
alkenes-L	36.1	46.1	30.0	16.1	6.1	0.657 $\pm$ 0.082	0.834 $\pm$ 0.057	0.730 $\pm$ 0.037	0.870 $\pm$ 0.013
alkenes-R	66.6	95.3	49.9	45.4	16.8	0.533 $\pm$ 0.079	0.751 $\pm$ 0.054	0.621 $\pm$ 0.058	0.886 $\pm$ 0.007
carb. acids	50.9	63.8	41.3	22.5	9.6	0.666 $\pm$ 0.155	0.806 $\pm$ 0.056	0.723 $\pm$ 0.107	0.935 $\pm$ 0.006
phenols	26.3	37.6	19.6	18.0	6.6	0.533 $\pm$ 0.077	0.752 $\pm$ 0.054	0.622 $\pm$ 0.063	0.949 $\pm$ 0.012
nitriles	27.8	37.8	25.4	12.4	2.4	0.694 $\pm$ 0.140	0.912 $\pm$ 0.050	0.782 $\pm$ 0.109	0.857 $\pm$ 0.012
amides-R	42.3	67.1	30.0	37.1	12.3	0.450 $\pm$ 0.050	0.708 $\pm$ 0.054	0.548 $\pm$ 0.042	0.933 $\pm$ 0.006
alkynes-L	9.6	11.8	5.5	6.3	4.1	0.478 $\pm$ 0.100	0.576 $\pm$ 0.092	0.516 $\pm$ 0.074	0.903 $\pm$ 0.027
alkynes-R	14.3	23.1	12.3	10.9	2.0	0.532 $\pm$ 0.075	0.851 $\pm$ 0.086	0.653 $\pm$ 0.076	0.850 $\pm$ 0.028

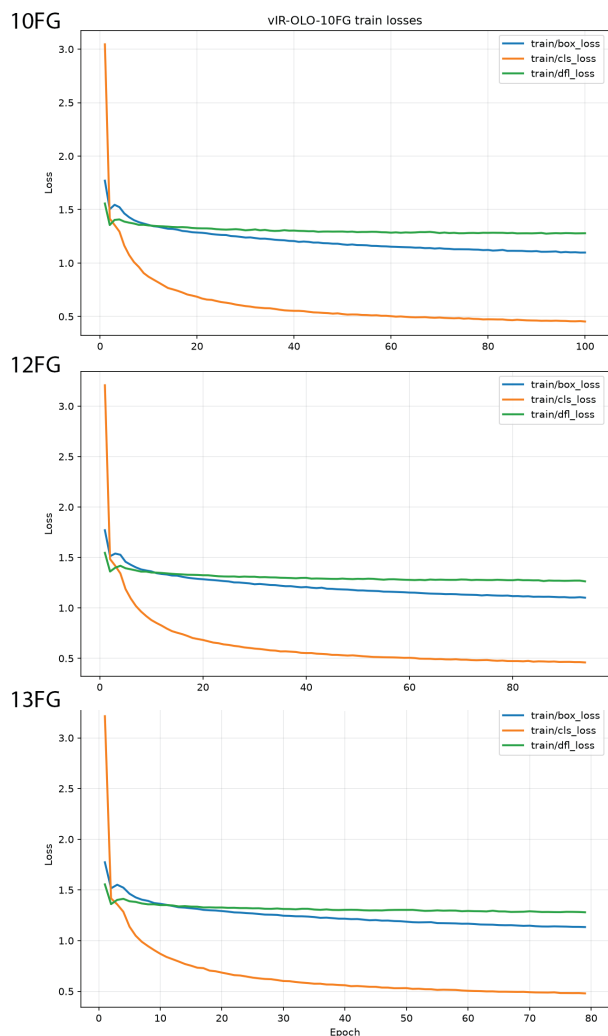


Fig. S2 Training losses for the three models. Each plot shows the three loss components optimized by the YOLO Ultralytics framework: box loss, classification loss, and DFL loss.

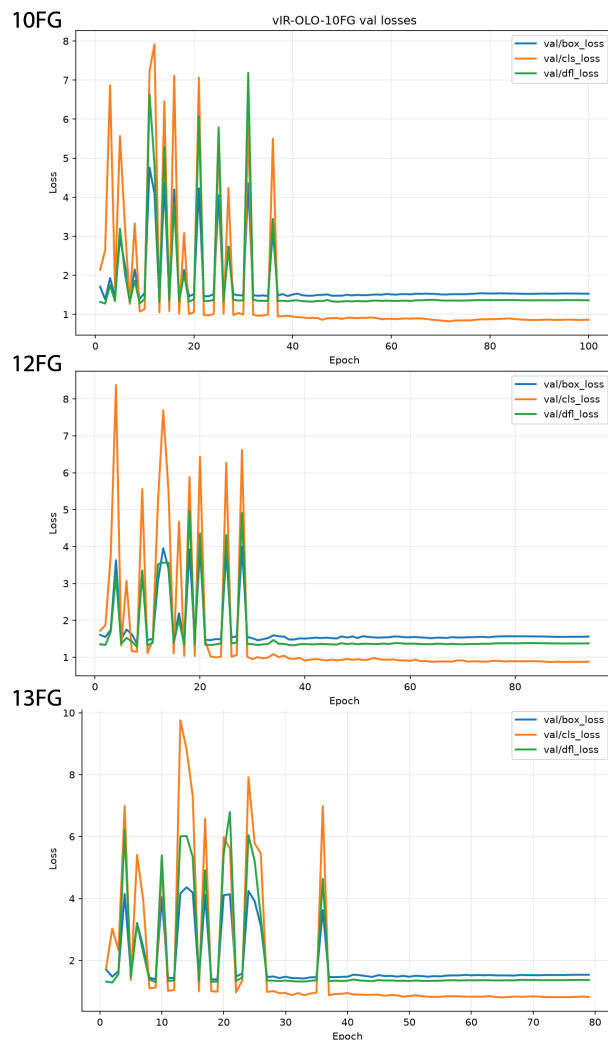


Fig. S3 Validation losses for the three models. Each plot shows the three loss components evaluated on the validation set by the YOLO Ultralytics framework: box loss, classification loss, and DFL loss. All three models exhibit high variance in the early epochs before stabilizing.

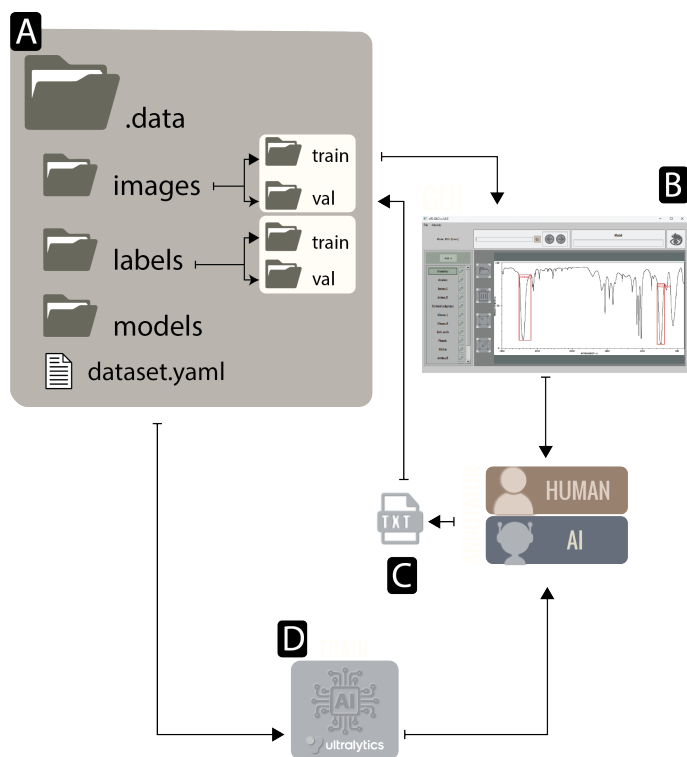


Fig. S4 Flow of information in the program vIR-OLO. A. data file structure. B. GUI interface. C. An annotator can be a human or a model embedded within the software. D. The data is fed back into a YOLO-based model to retrain and update the AI annotator, completing the program's life cycle.

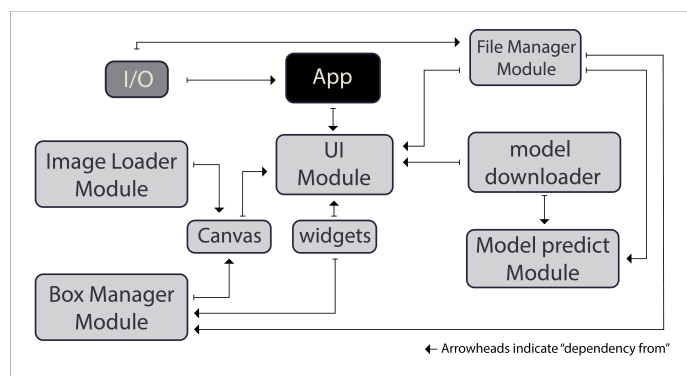


Fig. S5 Modular architecture of the vIR-OLO software. The system is centered around a main application controller (App), which orchestrates interactions among the user interface (UI Module), data-handling components (Image Loader and File Manager Modules), annotation management (Box Manager Module), and model-related components (model downloader and model prediction Module).

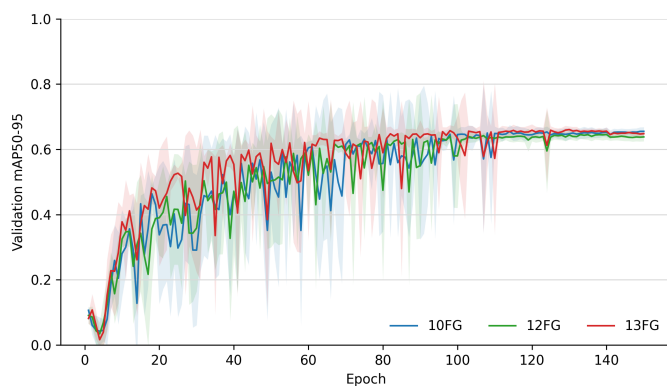


Fig. S6 Training stability across eight random train/validation/test splits for the 10FG, 12FG, and 13FG YOLO models. Lines show the mean validation mAP50-95 across runs, and shaded bands represent one standard deviation.

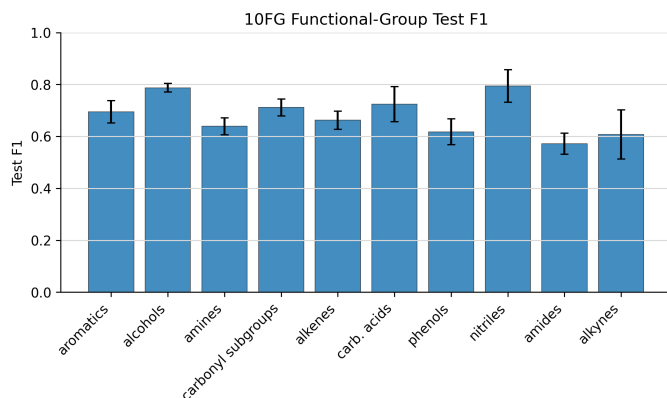


Fig. S7 Held-out test F1 by functional group for the 10FG model. Bars show the mean across eight random splits and error bars show one standard deviation.

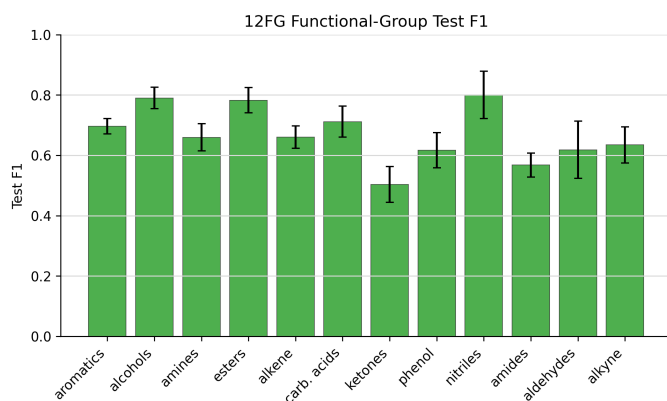


Fig. S8 Held-out test F1 by functional group for the 12FG model. Bars show the mean across eight random splits and error bars show one standard deviation.

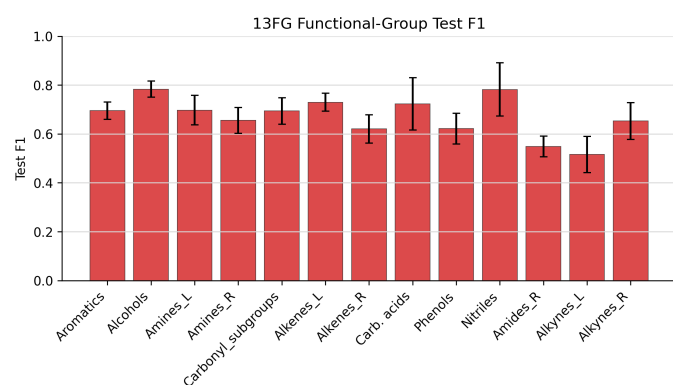


Fig. S9 Held-out test F1 by functional group for the 13FG model. Bars show the mean across eight random splits and error bars show one standard deviation.

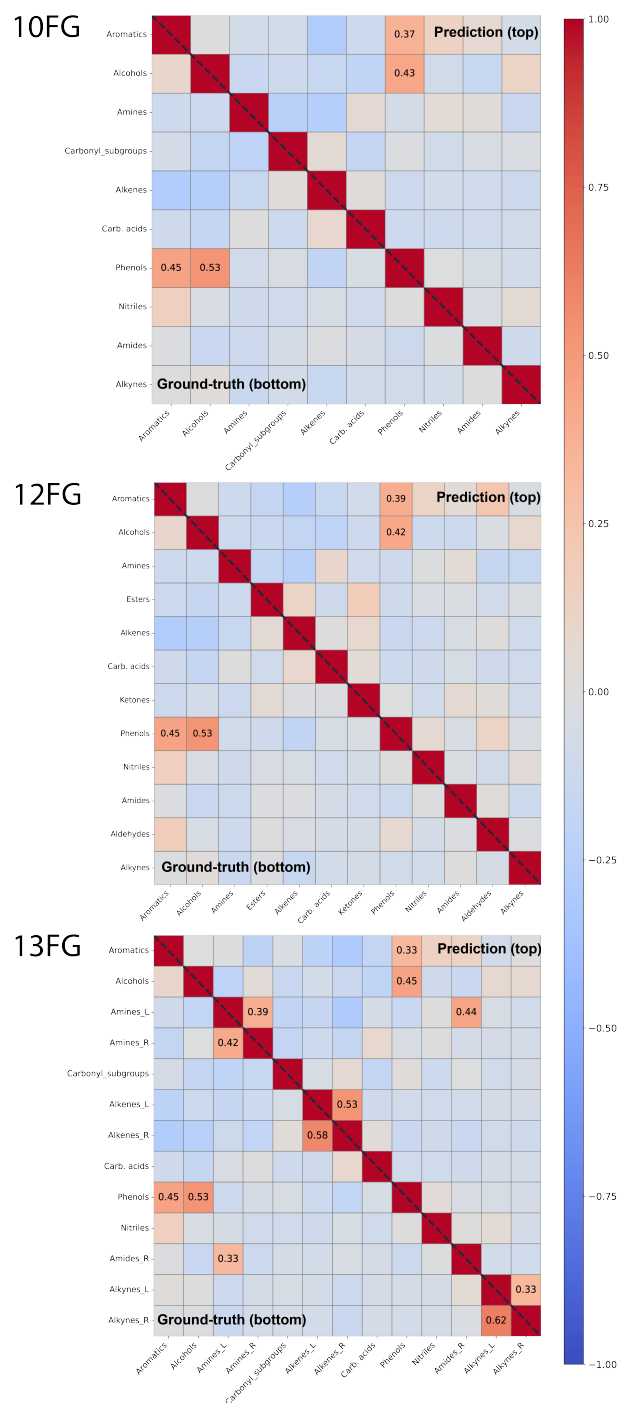


Fig. S10 Heatmaps of the Matthews Correlation Coefficients (MCCs) between functional groups across the three model label schemes.