AISCT® AURORA™ • WHITE PAPER WP-AURORA-001

# When Field Screening Works — and When It Only Looks Like It Does

*Eight vapour screening programmes, 662 laboratory-paired soil samples, one question: does the field reading actually predict what the laboratory will report? Three programmes used a standardised-procedure, parameter-optimised vapour system. Five used conventional bag-headspace PID, run by two different companies with two different sensors.*

## SUMMARY OF FINDINGS

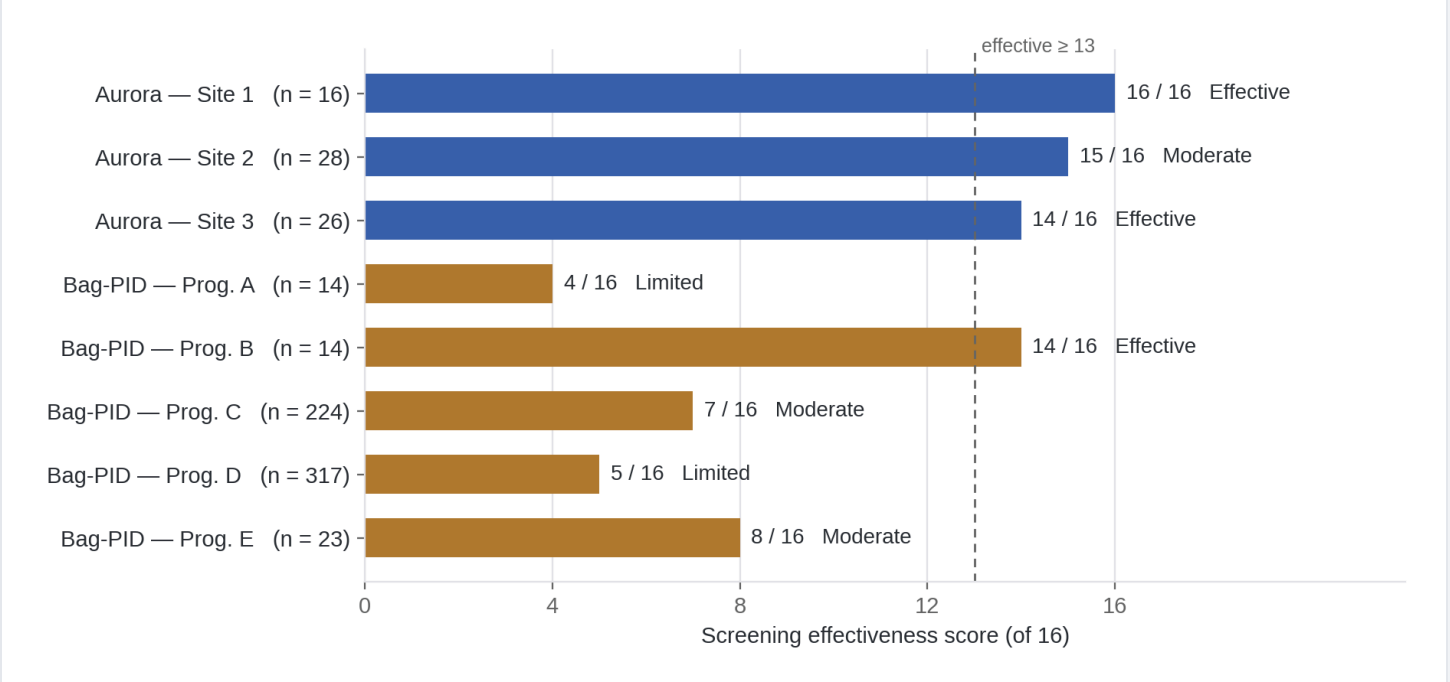
**662**
**Paired samples**
across eight independent programmes

**46%**
**Bag-PID exceedances missed**
58 of 125 true exceedances

**6%**
**Aurora exceedances missed**
1 of 16 true exceedances

**0.91+**
**Aurora R², every site**
0.912–0.923 vs a best bag-PID result of 0.386

## SCREENING EFFECTIVENESS BY PROGRAMME




*Each programme is scored on eight measures — Catch, Restraint, Pass Trust, Flag Trust, Separation, Tracking, Fit and Ranking — two points per measure. Aurora Site 2 scores 15 of 16 but is capped at Moderate because it missed one exceedance; a screen that misses impacts gives false confidence regardless of how well the other measures read.*

All three Aurora programmes correlated to the laboratory at R² 0.912 to 0.923. The best bag-headspace PID programme reached 0.386, and two of the five fell below 0.13 — a reading that carries almost no quantitative information about the soil beneath it.

The bag method was not uniformly poor. It was **inconsistent**, scoring anywhere from 4 to 14 out of 16 with nothing to tell a crew in advance which kind of programme they were on. That unpredictability, not raw detection power, is what this evidence indicts.

# 1. The Question Behind the Comparison

Headspace vapour screening is the most widely used field method in petroleum hydrocarbon site work, and CCME Volume 1 Section 5.5.1 recognises it as a field analytical technique. What the guidance cannot do is tell a practitioner whether the screen in their hand is working. A vapour reading is a proxy for what the laboratory recovers, and whether that proxy holds is an empirical question.

This paper answers it for eight programmes. Three used AISCT® Aurora™, a standardised-procedure, parameter-optimised vapour screening system with integrated data management: sample mass, chamber volume and equilibration temperature are fixed and logged on every sample, and each reading is evaluated against a correlation model built from that programme's own laboratory-confirmed pairs. Five used conventional bag-headspace PID, completed by two different companies with two different sensors.

One detail makes the comparison unusually direct: **one site appears twice** — screened with bag-headspace PID during an earlier investigation and with Aurora during recent works. Section 4 reports that pair on its own.

# 2. Method

## 2.1 Datasets

| PROGRAMME | SCREENING METHOD | SAMPLES | LAB PARAMETERS |
|---|---|---|---|
| **Aurora — Site 1** | AISCT® Aurora real-time vapour | 16 | F1, F2, F3, Total BTEX |
| **Aurora — Site 2** | AISCT® Aurora real-time vapour | 28 | Total BTEX; F1/F2/F3 as reported |
| **Aurora — Site 3** | AISCT® Aurora real-time vapour — delineation | 26 | F1, F2, F3, Total BTEX |
| **Bag-PID — Programme A** | Bag-headspace PID — same site as Site 2 | 14 | F1, F2, F3, Total BTEX |
| **Bag-PID — Programme B** | Bag-headspace PID | 14 | F1, F2, F3, Total BTEX |
| **Bag-PID — Programme C** | Bag-headspace PID | 224 | F1, F2, F3, Total BTEX |
| **Bag-PID — Programme D** | Bag-headspace PID | 317 | F1, F2, F3, Total BTEX |
| **Bag-PID — Programme E** | Bag-headspace PID | 23 | F1, F2, F3 |

## 2.2 Criteria, thresholds and treatment of non-detects

Criteria follow each programme's applicable values — 200 mg/kg F1/VPH and 1,000 mg/kg F2/LEPH and F3/HEPH for seven of the eight, and the site-specific values on Aurora Site 3; the Total BTEX criterion is data-derived per programme. Non-detects are substituted at half the detection limit. Each programme carries its own vapour threshold. A sample counts as exceeding if it exceeds *any* applicable criterion.

## 2.3 Measures

Each programme is evaluated on five classification measures — Catch, Restraint, Pass Trust, Flag Trust and Separation — and three correlation measures: Tracking, Fit and Ranking. Each is rated Limited, Moderate or Effective against a fixed benchmark and scored 0, 1 or 2, for a total out of 16; definitions and bands follow the results table in Section 3.1. Correlation statistics come from each programme's **best-fit laboratory fraction** on **all paired samples**: no outliers excluded, no masks, and every programme credited with the parameter it tracks best. That is the most generous treatment available to the conventional method, applied identically to all eight.

# 3. Results

## 3.1 All eight programmes

| PROGRAMME | N | Catch | Restr. | Pass Tr. | Flag Tr. | Separ. | Track. | Fit | Rank. | SCORE | RATING |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Aurora — Site 1** | 16 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.960 | 0.922 | 0.879 | 16/16 | Effective |
| **Aurora — Site 2** | 28 | 0.86 | 1.00 | 0.95 | 1.00 | 0.86 | 0.955 | 0.912 | 0.829 | 15/16 | Moderate * |
| **Aurora — Site 3** | 26 | 1.00 | 0.90 | 1.00 | 0.75 | 0.90 | 0.961 | 0.923 | 0.692 | 14/16 | Effective |
| **Bag-PID — Prog. A** | 14 | 1.00 | 0.18 | 1.00 | 0.25 | 0.18 | −0.338 | 0.114 | −0.341 | 4/16 | Limited |
| **Bag-PID — Prog. B** | 14 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.591 | 0.349 | 0.722 | 14/16 | Effective |
| **Bag-PID — Prog. C** | 224 | 0.51 | 0.89 | 0.897 | 0.50 | 0.40 | 0.450 | 0.202 | 0.279 | 7/16 | Moderate |
| **Bag-PID — Prog. D** | 317 | 0.49 | 0.75 | 0.83 | 0.38 | 0.25 | 0.622 | 0.386 | 0.489 | 5/16 | Limited |
| **Bag-PID — Prog. E** | 23 | 0.80 | 0.89 | 0.94 | 0.67 | 0.69 | 0.315 | 0.099 | 0.546 | 8/16 | Moderate |

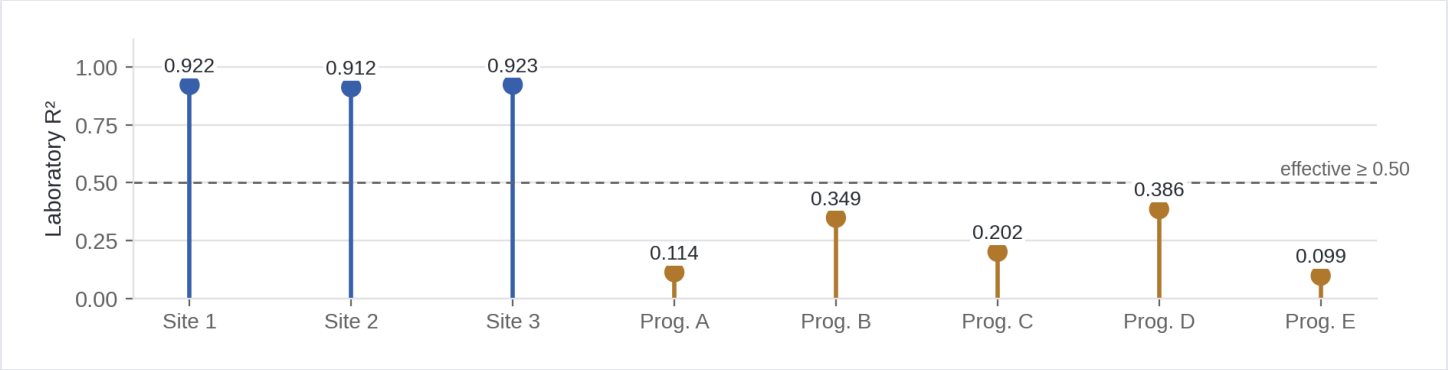
** Scored 15 of 16; capped at Moderate by the sensitivity override described in Section 5. Classification measures are the any-criterion values at each programme's selected threshold.*

**WHAT EACH MEASURE MEANS — AND THE BAND EACH NUMBER IS COLOURED AGAINST**

| MEASURE | WHAT IT MEANS | EFFECTIVE | MODERATE | LIMITED |
|---|---|---|---|---|
| **Catch** | Of soil the lab found over criteria, how much the field screen flagged. [sensitivity] | ≥ 0.90 | 0.60–0.89 | < 0.60 |
| **Restraint** | Of soil the lab found clean, how much the field screen correctly left unflagged. [specificity] | ≥ 0.80 | 0.50–0.79 | < 0.50 |
| **Pass Trust** | When the screen passes soil, how often the lab agrees. [negative predictive value] | ≥ 0.90 | 0.60–0.89 | < 0.60 |
| **Flag Trust** | When the screen flags soil, how often the lab agrees. [positive predictive value] | ≥ 0.80 | 0.50–0.79 | < 0.50 |
| **Separation** | One 0–1 score balancing catches against false alarms; higher is better. [Youden's J] | ≥ 0.70 | 0.30–0.69 | < 0.30 |
| **Tracking** | How tightly the field reading follows the lab number in a straight line. [Pearson correlation] | ≥ 0.70 | 0.40–0.69 | < 0.40 |
| **Fit** | How much of the lab result the field reading explains. [coefficient of determination, $r^2$] | ≥ 0.50 | 0.16–0.49 | < 0.16 |
| **Ranking** | Whether a higher field reading means a higher lab result. [Spearman rank correlation] | ≥ 0.70 | 0.40–0.69 | < 0.40 |
| **Score** | Two points per measure rated Effective, one per Moderate, none per Limited. [out of 16] | 13–16 | 7–12 | 0–6 |

## 3.2 Correlation to the laboratory

The classification measures answer whether the screen made the right call on the samples it saw. Correlation is more demanding: whether the reading carries quantitative information about the laboratory result, and so whether a threshold drawn from it means anything on the next sample. This is where the two methods separate most sharply.




*Laboratory $R^2$ on each programme's best-fit fraction, all paired samples, no outlier exclusion. The three Aurora programmes cluster at 91–92% of the variance explained; the bag-PID programmes range from 10% to 39%, with no relationship to data-set size.*

One bag-PID programme returned *negative* correlation even on its best-fit fraction (r = −0.338, ρ = −0.341) — a higher vapour reading was, on balance, associated with a *lower* laboratory concentration. No threshold drawn on that relationship can be defended: it runs the wrong way.

# 4. The Same Site, Screened Both Ways

Cross-programme comparisons always carry the objection that the sites differ. One pair removes it: the same site was screened with bag-headspace PID during an earlier investigation (Programme A, 14 paired samples) and with Aurora during recent works (Site 2, 28 paired samples).


*The same site, two screening methods, at each programme's own threshold. The bag-PID programme scores higher only on Catch and Pass Trust — where flagging almost everything helps — and lower on the other six.*

The bag-PID programme caught the three exceedances in its 14 samples, but it did so by flagging almost everything: 9 false alarms on 11 clean samples, a Restraint of 0.18, and only 5 of its 14 calls matching the laboratory — worse than a coin toss. Its correlation to the laboratory was negative (Pearson r −0.338, Spearman ρ −0.341), meaning a higher vapour reading was, if anything, associated with a *lower* laboratory result. Aurora caught six of the seven exceedances in its 28 samples, raised no false alarms at all, and correlated at r = 0.955.

A screen that flags nearly everything is not conservative. It is uninformative: it hands the decision back to the laboratory while still consuming a crew's day, and it cannot support the targeted delineation or volume reduction that field screening is deployed to deliver.

## 4.1 Where the false alarms came from

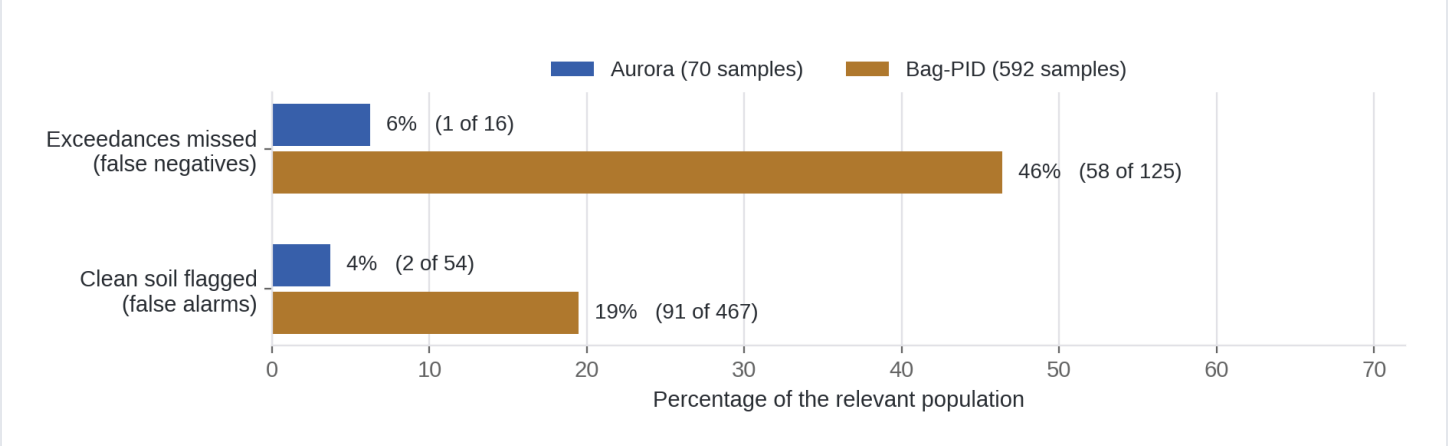
The figures above are on the any-criterion basis used throughout this paper: a flagged sample counts as a false alarm only if it exceeds no criterion at all. Parameter by parameter the picture is harsher on both methods — and the gap is still stark.

| PARAMETER | BAG-PID RESTRAINT | BAG-PID CLEAN SAMPLES FLAGGED | AURORA RESTRAINT | AURORA CLEAN SAMPLES FLAGGED |
|---|---|---|---|---|
| **F1 / VPH** | 0.14 | 12 of 14 flagged | 0.79 | 6 of 28 flagged |
| **F2 / LEPH** | 0.11 | 8 of 9 flagged | — | no F2 lab pairs |
| **F3 / HEPH** | 0.11 | 8 of 9 flagged | — | no F3 lab pairs |
| **Total BTEX** | 0.18 | 9 of 11 flagged | 1.00 | 0 of 21 flagged |

*Per-parameter false alarms at each programme's own threshold. Aurora's six F1 flags all exceeded another criterion, which is why its any-criterion Restraint is 1.00; the bag-PID programme's did not. Counts run over the samples with laboratory data for that parameter.*

## 5. Pooled Error Profile

Aggregating the any-criterion outcomes across all programmes gives the two methods' error profiles at scale: 70 paired samples for Aurora, 592 for bag-headspace PID.




*False negatives are counted against the samples that truly exceed criteria; false positives against the samples that are truly clean.*

Against the USEPA field screening acceptance limits used throughout the AISCT® programme — no more than 5% false negatives and no more than 20% false positives, expressed against the total sample count — Aurora returns 1 false negative in 70 samples (1.4%) and 2 false positives (2.9%), inside both limits. The pooled bag-PID programmes return 58 false negatives in 592 samples (9.8%), outside the limit, and 91 false positives (15.4%), inside it.

### THE SENSITIVITY OVERRIDE

Where a programme misses more than 10% of true exceedances, its overall rating is capped at Moderate no matter how strong the other seven measures are. A screen that misses impacts does not merely underperform — it produces confident clearance of contaminated soil, which is the one failure mode a screening programme cannot absorb.

## 6. Limitations

- The Aurora data set is 70 paired samples across three programmes against 592 across five for bag-PID. The comparison is unbalanced, and the Aurora figures carry wider uncertainty.
- The two methods report in different units on different scales, so each programme carries its own vapour threshold. Classification results are therefore threshold-dependent; correlation results are not.
- Programmes were completed at different times by different personnel. Aurora enforces a fixed protocol, which is itself part of what is compared: methods as deployed, not sensors in isolation.
- Total BTEX criteria are data-derived per programme rather than set by a single regulatory value. On Aurora Site 3 the fraction criteria are the site's own applicable values (210 mg/kg F1, 150 mg/kg F2, 1,300 mg/kg F3) rather than the values used on the other seven programmes.
- Aurora Site 3's own field log records two data-quality notes: a small number of readings were taken with a saturated moisture filter, and several carried an instrument bump-test-due status. None of the affected readings has a paired laboratory result, so none enters the statistics reported here.
- F3/HEPH never exceeded its criterion anywhere in the Aurora Site 3 data set, so no F3 Catch value could be computed for that programme.

# 7. The Two Results That Complicate the Story

Two findings run against a clean narrative, and both are more instructive than the headline.

### 7.1 Aurora missed an exceedance

At Site 2, Aurora missed 1 of 7 Total BTEX exceedances at the selected threshold. Catch is 0.86, and the override caps the programme at Moderate despite full marks — 14 of 14 points — on every other measure. The mechanism that made the miss visible is the point: it is detectable *because* the programme was evaluated against a laboratory-paired subset with a confusion matrix, during the work. A programme that never computes its own error profile does not have fewer misses — it has misses no one has counted.

### 7.2 One bag-PID programme scored Effective

Programme B scored 14 of 16 — perfect classification on 14 samples containing 3 exceedances. Its correlation was weaker (r 0.591, $R^2$ 0.349, ρ 0.722): the reading did not track the laboratory number, and the result rests on a small sample with wide separation between the impacted and clean populations. This is the pattern that makes conventional screening feel reliable. On a small, strongly contrasted data set a weak proxy still sorts the obvious cases — until it meets a programme like C or D, where samples are numerous, concentrations sit nearer the criteria, and half the exceedances are missed.

> **THE PATTERN ACROSS THE FIVE BAG-PID PROGRAMMES**
>
> Scores of 4, 14, 7, 5 and 8 out of 16. The problem is not that the bag method always fails. It is that its performance is unpredictable from one programme to the next, and nothing in the method tells the crew which kind of programme they are on. Aurora scored 16, 15 and 14 — and the one capped rating came with a named, quantified reason.

# 8. Conclusions

- On its three programmes — 70 of the 662 paired samples — the standardised-procedure vapour system correlated to the laboratory at $R^2$ 0.912, 0.922 and 0.923 — every programme more than double the best bag-PID result of 0.386, and on the same generous, no-exclusion basis.
- Pooled, the bag-PID programmes missed 46% of true exceedances. Aurora missed one in 16, across 70 samples, with two false alarms.
- On the one site screened by both methods, the bag-PID programme returned negative correlation and flagged 9 of 11 clean samples; Aurora returned r = 0.955 and no false alarms.
- Bag-PID performance was not uniformly poor — it was *inconsistent*, ranging from 4 to 14 out of 16 with no way for a crew to know in advance which they were getting. The three Aurora programmes scored 14, 15 and 16.
- The decisive difference is not how responsive the sensor is to vapour. It is that the process is standardised, the result is correlated to laboratory data, and the error profile is computed while the crew is still on site.

**662**
**paired samples, eight programmes**

**THE BOTTOM LINE**

A field screen is only worth deploying if someone can say how often it is wrong. On this evidence the conventional method could not — and where it was measured, it was wrong far more often than its users would have assumed.