Journal of Materials Science and Chemical Engineering, 2026, 14(8), 12-29
https://www.scirp.org/journal/msce
ISSN Online: 2327-6053
ISSN Print: 2327-6045

Beyond Pretty Clusters: Toward Standardized Validation of Fluorescence Sensor Arrays for Multi-Analyte Discrimination: A Mini Review

Muhammad Jasim

School of Life and Health and Science, Hubei University of Technology, Wuhan, China
Email: jasimgee@hotmail.com

How to cite this paper: Jasim, M. (2026) Beyond Pretty Clusters: Toward Standardized Validation of Fluorescence Sensor Arrays for Multi-Analyte Discrimination: A Mini Review. *Journal of Materials Science and Chemical Engineering*, **14**, 12-29.
https://doi.org/10.4236/msce.2026.148002

Received: July 19, 2026
Accepted: August 8, 2026
Published: August 11, 2026

Abstract

Fluorescence sensor arrays built on lanthanide metal-organic frameworks (Ln-MOFs) have rapidly become a standard approach for discriminating structurally similar analytes. These range from antibiotic congeners and their metabolites to co-occurring metal ions in food and biological matrices. The typical workflow pairs multi-channel fluorescence titration with principal component analysis (PCA) or hierarchical cluster analysis (HCA). A visually separated score plot is then routinely presented as sufficient evidence of discriminatory power. This paper audits that convention across the recent Ln-MOF sensor array literature. We identify five recurring evidentiary gaps. First, circular validation: a model is tested only on the data used to build it. Second, the absence of a quantitative separation metric to replace descriptive language such as "well separated". Third, a concentration-dependence blind spot, where discrimination is demonstrated at a single analyte concentration rather than across the working range. Fourth, the near-total absence of robustness testing under simulated noise or real-world perturbation. Fifth, the conflation of classification claims with quantification claims that are rarely separately validated. Drawing on studies that have already addressed one or more of these gaps individually, we show that closing them requires no new sensing chemistry. Only a different treatment of data most studies already collect is needed. We propose a nine-item minimum reporting checklist and illustrate it against two representative published arrays. We argue for its adoption as a low-cost first step toward comparable, reproducible discrimination claims, drawing an explicit parallel to earlier minimum-reporting movements such as MIQE and STARD.

Keywords

Lanthanide Metal-Organic Framework, Fluorescence Sensor Array, Principal

Component Analysis, Multi-Analyte Discrimination, Validation, Reporting Standards

1. Introduction

Over the past five years, a specific analytical routine has become the default way to claim that a fluorescence sensor can tell chemically similar compounds apart. The material is exposed to a panel of analytes, emission intensities are recorded at two or more wavelengths, the response vectors are projected through principal component analysis (PCA) or hierarchical cluster analysis (HCA), and separated groups in the resulting plot are presented as evidence of discrimination. This workflow has distinguished antibiotic classes [1], structurally related drug candidates and their metabolites [2], natural product families [3], and co-occurring contaminants or metal ions in complex matrices [4] [5]. Lanthanide metal-organic frameworks (Ln-MOFs), with their sharp, spectrally separable 4f-4f emission bands and amenability to dual-lanthanide co-doping, have become a favored platform for this kind of multiplexed sensing [1] [2] [5], and publication in this niche has accelerated sharply since 2023.

Part of the platform's appeal is mechanistic flexibility, which gives researchers design freedom but also freedom in how a response gets explained after the fact. Reported quenching behavior ranges from the inner filter effect alone [6], to the inner filter effect confirmed by density functional theory while photoinduced electron transfer and Förster resonance energy transfer are both ruled out by lifetime measurements [7], to the inner filter effect and photoinduced electron transfer acting together [8] [9], to three mechanisms invoked at once [10]. A bimetallic probe for malachite green showed that a monometallic version of the same material relies on the inner filter effect while its bimetallic counterpart shifts toward resonance energy transfer, improving the detection limit roughly a hundredfold in the process [11]. Two recent reviews confirm this mechanistic diversity is now the norm rather than the exception across the fluorescent MOF literature [12] [13]. The flexibility is a genuine strength, but it means mechanism attribution rests on an unevenly supported evidence base, spectral overlap alone in some studies, lifetime or computational confirmation in others, a parallel inconsistency worth flagging even though it sits outside this paper's scope.

These are not academic distinctions. Fluoroquinolone and nitrofuran residues carry regulatory maximum residue limits and feed antimicrobial resistance selection at sub-therapeutic exposure, which is why resolving structurally related congeners, rather than only flagging one compound, has become a stated design goal in its own right [1] [8] [9]. The same logic drives food-adulteration arrays asked to separate a dye from its reduction product or from co-administered antibiotics in the same matrix [6] [11], and clinical arrays tested against ten-ion metal panels [5] or mixed drug panels [2] in matrices meant to approximate real physiological

samples rather than clean buffer. Because a maximum residue limit either is or is not exceeded, and a metabolite either is or is not distinguishable from its parent compound, a discrimination claim needs to rest on something more defensible than a plot that merely looks convincing.

What has not kept pace is the statistical rigor behind the claim itself. A separated cluster visualizes variance; it is not a validated measure of discriminatory power. It says nothing about whether the separation would survive a blind sample, hold across the working concentration range rather than at one convenient level, resist realistic instrument drift or matrix noise, or exceed a distance large enough to be meaningful rather than merely visible. These are answerable questions, and a handful of studies have already answered some of them, through blind sample accuracy statistics [1], concentration-independent testing across multiple orders of magnitude, or explicit noise robustness checks. But these practices remain scattered rather than shared, and no two papers report the same combination, which makes cross-study comparison of discriminatory power largely impossible.

This paper does not propose a new sensor. It proposes that the field agrees on what counts as evidence. We first characterize the current spread of validation practice across recent Ln-MOF sensor array literature, then identify five specific and recurring gaps, and finally propose a minimum reporting checklist that could be adopted without new experiments in most cases, since the required calculations can usually be performed on data already collected.

2. Auditing Current Practice

This was a narrative mini review. Searches were executed in January 2026 in PubMed, Embase, Scopus, Web of Science, and Google Scholar for records published between 2020 and 2026 using combinations of the terms lanthanide MOF, Ln-MOF, fluorescent sensor array, ratiometric sensor array, PCA discrimination, and HCA antibiotic, and supplemented with citation chaining from two recent field-wide reviews [14] [15]. Records were included if they reported an Ln-MOF-based fluorescence sensor system evaluated against two or more analytes using PCA, HCA, or an equivalent multivariate discrimination method, and were excluded if they described a single-analyte sensor with no discrimination claim, a non-lanthanide or non-MOF material, a conference abstract without an accessible full text, or a study duplicated across databases. Title and abstract screening against these criteria was performed by the author, followed by full-text assessment of the retained records; ambiguous or borderline cases (for example, studies reporting differential quenching across analytes without an explicit multivariate discrimination step) were resolved by re-reading the methods and results sections against the inclusion criteria above. This approach was chosen to capture the most active recent portion of a fast-growing subfield rather than to provide an exhaustive, PRISMA-style census, and no formal risk-of-bias tool was applied. The 27 sources discussed in this paper, drawn from that search, illustrate a pattern we believe is representative of current practice, not a complete accounting of every Ln-MOF

sensor array paper published in this window; a full record of search strings, per-database hit counts, and screening decisions is provided in **Appendix** alongside the study-level checklist extraction described in Section 4.

Looking across the Ln-MOF sensor array literature published in the last five years, validation effort is not absent, but it is unevenly distributed and rarely cumulative. Most papers stop at the same point: a PCA score plot or an HCA dendrogram in which the analyte groups do not visually overlap, described in the text as well separated or clearly discriminated. The strength of that separation is almost never expressed as a number.

A smaller group of studies goes further and tests the model against samples it was not trained on. Xie *et al.* built a ratiometric Eu/Tb-MOF array that classified 25 antibiotics from eight structural classes using the emission ratio at 545 and 616 nm, then challenged the model with 48 blind samples and reported 98 percent correct identification [1]. A related array built from four Ln-MOFs and applied to 26 natural product ingredients was similarly tested against blind samples and reported 100 percent accuracy [3]. These two studies are useful reference points precisely because blind testing is still the exception rather than an expected step. Most papers in the same window report only the training-set clustering that produced the model in the first place, which is a weaker claim than it is usually presented as, since a model will almost always separate the data it was built from.

A separate axis of rigor concerns how the classification accuracy itself is reported. Wang *et al.* paired a bimetallic EuxTb1-x-MOF array with PCA and HCA to identify prostate cancer drugs and their mixtures, and supported the assignment with density functional theory calculations of the underlying interaction, which strengthens the mechanistic case even where the statistical case is left mostly visual [2]. A more recent pH-regulated lanthanide array aimed at ten metal ions in biological fluids used PCA together with an Extra Trees classifier and reported a confusion matrix for the ten-way classification, along with separate testing on binary and ternary mixtures [5]. A confusion matrix is a meaningfully different kind of evidence than a score plot, since it quantifies exactly which analytes are confused with which, at what rate, rather than asserting that no confusion is visible.

Real-matrix and field-relevant testing follows its own inconsistent pattern. A dual-emissive single-component MOF array was shown to distinguish 12 antibiotics and was also described as having robust quantitative performance for individual compounds [16]. That phrasing blends two distinct claims, classification of identity and quantification of concentration, without separating the evidence for each. A Tb-MOF array built for nitrofuran antibiotics reported discrimination of both individual compounds and their binary and ternary mixtures, together with a portable paper-based readout validated in real food samples [17]. This is a strong showing on field applicability, but it does not report a numerical separation metric for the mixture discrimination claim. A 3D-printed smartphone platform aimed at two structurally similar fluoroquinolones combined a self-calibrating lanthanide sensor with PCA to achieve discrimination and reported point-of-care oper-

ation in environmental water [18]. Again, no quantitative index of how far apart the two clusters actually sit is reported. A related water-stable Tb-MOF was reported to discriminate three separate classes of antibiotics as well as D_2O from H_2O using a self-calibrating ratiometric response [19]. The discrimination is described as fast, accurate, and highly sensitive, but as with several examples above, those adjectives are not accompanied by a computed distance or accuracy figure that would let a reader judge the claim independently of the authors' own description.

Two broad reviews published within the last several months, one surveying optical sensing arrays across material classes [14] and one focused on porous-framework arrays and their AI-assisted analysis [15], confirm that this is now a large and fast-moving subfield rather than a marginal one. Neither review, however, audits the internal consistency of how discrimination is validated across the primary studies it covers.

Two further cases sharpen the picture. Zhang *et al.* constructed six isostructural lanthanide MOFs and used PCA to distinguish Fe^{3+}, two chromium oxyanion species, and ceftriaxone sodium in water, alongside separate ratiometric detection of uric acid in simulated urine. The authors describe the work as the first MOF sensor to use PCA for differentiating water contaminants [4]. The claim of novelty is plausible, and the recovery data for uric acid are solid, 95.08 to 102.07 percent, with an RSD below 2.63 percent. But the PCA-based discrimination of the four water contaminants is again reported through score-plot separation alone, with no accuracy statistic attached to the classification itself. A more recent hydrogel-based kanamycin sensor takes the opposite approach to sophistication: it layers four detection modes, single-emission screening, self-calibrated ratiometric sensing, full-spectrum machine learning analysis, and smartphone readout, into one hierarchical workflow [20]. This is a meaningfully more rigorous architecture than most array papers attempt, since each mode is positioned to cross-check the others rather than standing alone, and it points toward where the field is heading. But it is also, at the time of writing, closer to an isolated example of best practice than a reflection of what is typically expected of a new sensor array submission.

Taken together, these additional cases reinforce rather than complicate the pattern established above. Rigor is increasing at the frontier of the field, largely through machine learning and multimodal cross-checking. But the median paper still relies on visual cluster separation as its primary or sole evidence of discriminatory power. The inconsistency is not a matter of a few outlier studies falling short. It is that no shared minimum standard exists for what a validated discrimination claim should contain, which makes the frontier and the median difficult to compare on equal terms. The following section breaks this inconsistency down into five specific, recurring gaps.

3. Five Recurring Gaps

Figure 1 maps these gaps onto the workflow itself, showing where each one

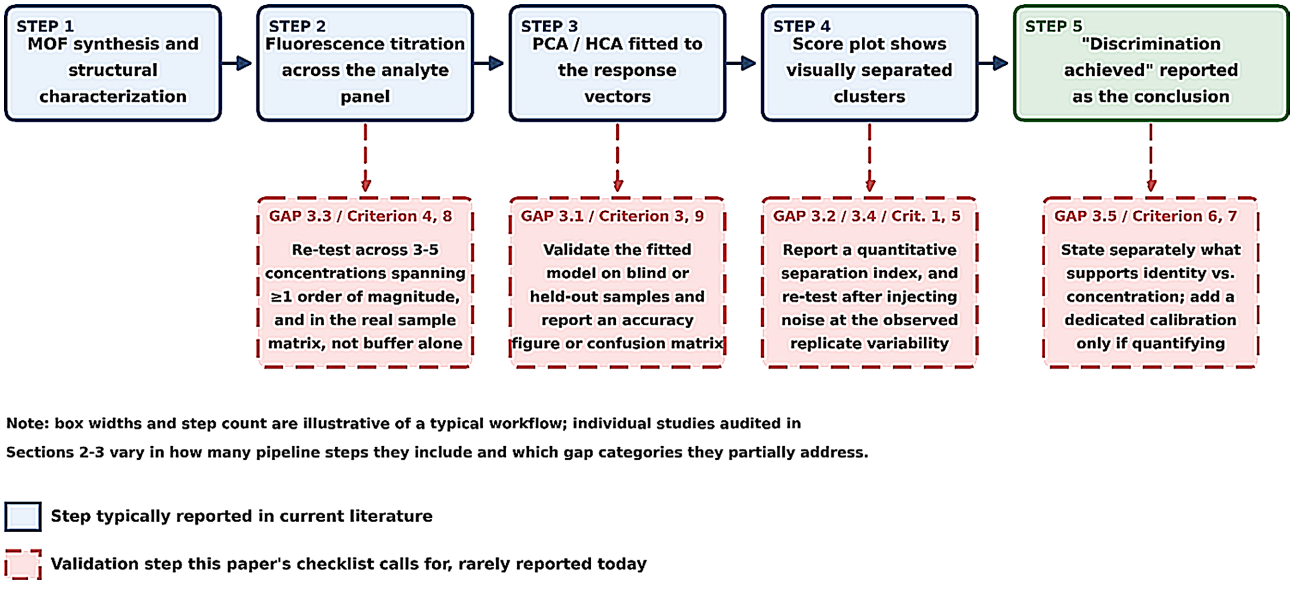

Figure 1. The standard sensor array workflow, and where its validation evidence usually stops.

typically opens up between an accepted step (solid boxes) and the validation evidence the checklist in Section 4 asks for (dashed boxes). The five subsections below discuss each gap in turn.

3.1. Circular Validation

The most basic problem is also the most common. A model is built from a set of fluorescence response vectors, PCA or HCA is applied to that same set, and the resulting separation is presented as proof that the model discriminates. This is closer to a description of the training data than a test of the model's predictive value. Almost any set of chemically distinct analytes measured across several emission channels will produce some visual separation once projected into two or three principal components. This is simply because PCA is designed to maximize apparent variance along its leading axes. Figure 2 illustrates this with a synthetic three-analyte dataset. A nearest-centroid classifier was fitted to the PC1-PC2 training scores, assigning each new point to whichever training-class centroid it sits closest to; this explicit decision rule, not PCA itself, is what produces a class assignment and an accuracy figure. The same fitted model appears perfectly separated when only the training points are shown (Figure 2(A)), but misclassifies one of eight blind samples, 7/8 correct (87.5%), once independent points are introduced and passed through that decision rule (Figure 2(B)). This is a distinction that a training-set-only score plot cannot reveal by construction.

A recent triple-channel Eu-MOF array built for baijiu discrimination illustrates the pattern clearly. Distinct liquor samples formed separated clusters in score-plot space, but the model was not subsequently challenged with an independent batch of unlabeled samples to confirm that the separation generalized beyond the calibration set [21]. A coumarin-embedded Eu-MOF nanosensor designed to discriminate six tetracycline analogs reported 100 percent discrimination accuracy

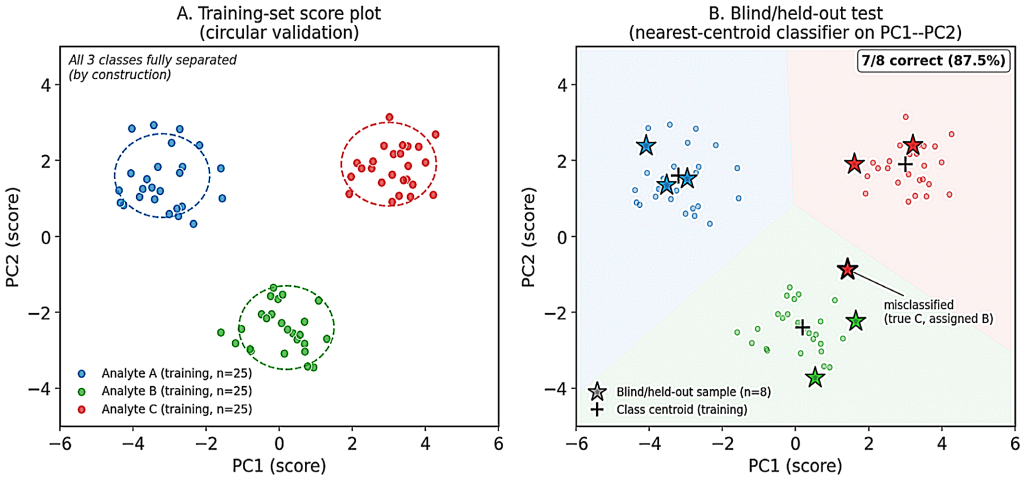

Figure 2. Circular validation versus held-out validation of the same three-analyte model, classified with a nearest-centroid decision rule fitted to the PC1-PC2 training scores (synthetic illustrative data, not drawn from any cited study). Panel B shows 8 blind/held-out samples (3 true A, 3 true C, 2 true B); one true-C sample is assigned to B, giving 7/8 (87.5%) correct.

through PCA [22]. That figure reads as definitive, but based on the reported methodology, it again describes separation within the calibration data rather than performance on an independently held-out set. The studies that do test held-out or blind samples, such as the 48-sample antibiotic panel [1], and the 26-compound natural product array [3], are informative precisely because they show what happens when the model meets data it did not shape. That so few papers take this step is not a minor omission. It means that for the majority of published arrays, the reported accuracy of discrimination is, strictly speaking, unknown.

None of this is new to chemometrics as a discipline. The principle that a multivariate model must be judged on data it did not see during construction, most rigorously through an independently held-out test set rather than internal resampling alone, has been argued in detail in the chemometrics literature [23], and representative splitting of a data set into calibration and test subsets, rather than arbitrary or convenience splitting, has been standard practice since at least the introduction of the Kennard-Stone algorithm for calibration set selection [24]. What is specific to the Ln-MOF sensor array literature audited here is not the absence of this theory, but its absence in practice: a field that applies PCA and HCA constantly has, with few exceptions, not carried the validation half of that same statistical toolkit into its own workflow. The checklist proposed in Section 4 does not ask the field to adopt a new idea; it asks the field to close the gap between the chemometric methods it already uses and the chemometric validation practice that has accompanied those methods everywhere else they are applied.

3.2. No Quantitative Separation Metric

Terms like well separated and clearly distinguished appear throughout this literature as if they were measurements, when they are descriptions of a plot. Two clusters that look separated at the axis scale chosen for a figure may sit much closer

together in the underlying multidimensional response space, or may only remain separated because outliers were excluded before plotting. This is visible even in mechanistically careful work: a 3d-4f MOF array designed to discriminate phosphate species based on differential quenching patterns reports visually distinct response profiles for each analyte, but the discrimination claim rests on pattern description rather than a computed distance or overlap statistic between the response classes [25]. A separation index, expressed as the ratio of inter-cluster distance to the sum of intra-cluster spreads, or a silhouette coefficient, or a Mahalanobis distance between cluster centroids, converts a visual impression into a number that can be compared across studies and across analyte pairs within the same study. Confusion matrices, used in the ten-metal-ion array [5], serve a related but distinct purpose: they quantify how classification actually performs when tested, rather than how far apart cluster centroids sit in principal component space. Neither type of metric requires new experiments. Both can usually be computed from data a study has already collected, which makes their near-total absence from the majority of array papers a reporting gap rather than a resource constraint.

Because a separation metric is only comparable across studies if it is computed the same way, we specify four choices a study needs to report alongside the number itself. First, preprocessing: whether raw emission intensities, mean-centered values, or autoscaled (mean-centered and unit-variance) channels were used, since autoscaling changes the relative weight of channels with different intensity ranges and therefore changes the resulting separation index. Second, the distance space: whether the metric is computed in the full response-channel space or in the reduced PCA score space, and if the latter, how many components were retained, since a two-component projection can inflate apparent separation relative to the full channel space. Third, treatment of covariance: whether the metric uses a covariance-naive measure such as Euclidean distance between centroids, which implicitly assumes spherical, equally sized clusters, or a covariance-aware measure such as Mahalanobis distance or a silhouette coefficient, which accounts for cluster shape and is preferable whenever intra-class variance differs meaningfully across analytes. Fourth, uncertainty: the metric should be accompanied by a bootstrap or cross-validated confidence interval, or at minimum the number of replicates it is based on, so a reader can judge whether an apparently large separation index reflects a real effect or a small-sample artifact. A study that reports a separation index without specifying these four choices leaves the number as uninterpretable, for comparison purposes, as the descriptive language it was meant to replace.

3.3. Concentration-Dependence Blind Spot

Most discrimination claims are demonstrated at a single, often generously chosen, concentration for each analyte. This leaves an open question that matters a great deal for any claimed real-world application: does the separation hold near the

limit of detection, where signal is weakest and noise proportionally largest, or does it only hold in the comfortable middle of the working range where the fluorescence response is strong and well behaved for every analyte simultaneously? A water-stable, multi-emitting Ln-MOF developed for simultaneous metal ion and pharmaceutical sensing demonstrates the point well: its discrimination capability is showcased at a fixed analyte concentration, and while the sensor's titration-based calibration curves span a wide concentration range for individual analyte quantification, the discrimination claim itself is never re-tested across that same range [26]. It is equally plausible that separation could fail to appear at low concentration because the differential channel response has not yet developed, or that it could collapse at high concentration if multiple analytes drive the sensor toward the same saturated quenching state. Testing across the full working range, at three to five concentration points spanning at least an order of magnitude, is a modest addition to an experiment that has usually already generated titration data for calibration purposes, and would directly address this blind spot without requiring new material synthesis.

3.4. No Robustness-to-Perturbation Testing

A sensor array intended for field or point-of-care use will encounter conditions its calibration data did not include: slightly different instrument gain, ambient temperature drift, matrix variability between sample batches, and simple measurement noise. A dual-mode molecularly imprinted MOF nanozyme platform for chloramphenicol offers a partial counterexample worth noting, since pairing fluorescent and colorimetric readout on the same sample provides an informal cross-check that a single-mode array cannot [27]; even so, the two channels were validated against each other rather than against a deliberately perturbed dataset, which is a weaker test than systematically degrading the input signal and observing whether classification survives. A smartphone-integrated bimetallic Eu-Tb-MOF probe developed for tetracycline antibiotics in milk, beef, and pork reported strong recoveries across three real food matrices and extended the same material to latent fingerprint visualization [28], which is a genuinely useful demonstration of matrix breadth, but it does not report what happens to the classification or quantification accuracy when the smartphone imaging conditions themselves are deliberately varied, a variable that matters a great deal for any camera-based readout intended for uncontrolled field use.

None of the studies reviewed here that report field deployment [17] [18] report what happens to their classification accuracy when synthetic noise is deliberately added to the calibration data before the model is rebuilt. This is a cheap and reproducible check, typically implemented by adding Gaussian noise at a level scaled to the experimentally observed replicate variability and then re-running the classification to see whether separation indices or accuracy statistics degrade gracefully or collapse sharply. Its absence matters because a model that separates clusters perfectly under ideal laboratory conditions provides no guarantee about

how it will behave once deployed.

3.5. Classification and Quantification Treated as One Claim

A pattern that recurs across several of the studies audited here is the blending of two analytically distinct claims into a single sentence: that an array can identify which analyte or combination of analytes is present, and that it can also determine how much of each is present. The single-component antibiotic array described its performance as both discriminatory and quantitatively robust for individual compounds in the same statement [16], without separating the evidence supporting each claim. A device-integrated lanthanide MOF sensor responsive to pH shows a related version of the same issue, where classification of chemical identity and calibration-curve-based quantification are presented as complementary strengths of one system without addressing whether the quantification step remains reliable once multiple analytes are simultaneously present rather than measured one at a time [29]. PCA and HCA are unsupervised methods, not classification tools. PCA compresses a multidimensional response into axes chosen to maximize the total variance captured, and HCA groups samples by similarity in that response space; neither procedure is given class labels to predict, and neither, on its own, assigns a new sample to an analyte category. Predictive classification requires a separately specified decision rule, such as a nearest-centroid assignment, linear discriminant analysis, or another supervised classifier fitted to the PCA or HCA output, and it is that downstream rule, not the projection itself, whose accuracy can meaningfully be reported. A resulting score plot, with or without a decision rule layered on top, says nothing about the concentration of any component within a resolved mixture. Quantifying individual analytes within a mixture that has already been correctly classified requires a separate multivariate calibration step, typically partial least squares regression or an equivalent supervised model trained explicitly for that purpose, and this step is rarely performed even in studies that otherwise demonstrate strong classification performance.

4. A Proposed Minimum Reporting Checklist

The gaps identified above are not evenly weighted. Some, such as reporting a quantitative separation metric, require no new data collection at all, only a different treatment of results already in hand. Others, such as blind sample testing or noise-injection robustness checks, require modest additional experimental or computational effort but nothing approaching the scale of a new synthesis or a new sensing platform. What follows is a proposed minimum set of nine criteria, framed as a checklist rather than a formal standard, that a study reporting a multianalyte discrimination claim for a fluorescence sensor array could reasonably be expected to address, either by satisfying the criterion or by explicitly stating why it does not apply (**Table 1**).

The checklist is deliberately narrow in scope. It does not attempt to specify sensor design, material choice, or analyte selection, since those decisions are legiti-

mately driven by the scientific question a given study is asking. It concerns only the evidentiary basis for a discrimination claim once that claim is made, on the reasoning that a reader should be able to judge how much confidence a cluster separation deserves without needing to reconstruct the underlying statistics from a figure.

To move beyond descriptive language such as rare or occasional, we applied the nine criteria retrospectively to the studies audited in Section 2. Of the 27 primary and review sources cited in this paper, 17 report an explicit multi-analyte discrimination claim and were included in this count; the remainder are cited either for mechanistic context (quenching pathway studies not making a discrimination claim) or as broad field reviews rather than primary discrimination studies. For each of the 17, a criterion was scored as satisfied only where this paper's review explicitly describes that evidence being reported, and scored as explicitly absent only where the review explicitly states the evidence is missing. Where our review of a given study did not address a criterion one way or the other, the case is recorded as not addressed rather than assumed absent, since the absence of a statement is not equivalent to a confirmed gap. This distinction matters: the counts below likely understate true coverage of criteria 2 and 7 in particular, since replicate counts and dedicated mixture-calibration models are reported in methods sections that fell outside the scope of what this review discusses for each paper. A full accounting of those two items would require a dedicated re-read of each paper's methods section, which we flag as a limitation of this table rather than a claim that the field performs better on those two items than it does on the rest. The per-study scoring underlying the aggregate counts in Table 1, together with the specific text passage used as evidence for every satisfied or absent call, is provided in full as **Appendix**, so that any individual cell in Table 1 can be traced back to the study and statement that produced it rather than taken on the authors' summary alone.

None of these nine items is unprecedented. Each has already appeared, individually, in at least one of the studies audited in Section 2: blind testing with an accuracy figure in the antibiotic and natural product arrays [1] [3], a confusion matrix in the ten-metal-ion array [5], multimodal cross-checking in the hydrogel kanamycin sensor [20], and real-matrix testing in the nitrofuran paper-based array [17]. What is missing is not any single technique, but the convention of applying more than one or two of them together and reporting the result plainly enough that a reader unfamiliar with the specific material can judge the strength of the discrimination claim without re-deriving it.

Item 7 deserves a brief note, since it is the criterion most likely to be contested. Not every study needs to quantify mixture components; a sensor intended purely for screening or triage may only need to answer a yes-or-no or which-of-several question, and demanding a full calibration model in that case would be disproportionate. The criterion is phrased conditionally for this reason: it applies only where a quantification claim is actually made. The problem identified in Section

Table 1. Proposed minimum reporting checklist for multi-analyte discrimination claims in fluorescence sensor arrays.

#	Criterion	What it requires, in practice	Gap	Confirmed satisfied	Explicitly absent	Not addressed
1	Quantitative cluster-separation metric	Report a separation index, silhouette coefficient, or Mahalanobis distance between cluster centroids, not just a visual description	3.2	0/17	7/17	10/17
2	Minimum three independent replicates per cluster	State the number of replicates contributing to each analyte's response vector before clustering	3.1, 3.2	0/17	0/17	17/17
3	Blind or held-out sample validation	Test the trained model against samples excluded from model construction and report an accuracy figure	3.1	2/17	2/17	13/17
4	Full concentration-range testing	Repeat the discrimination analysis at three to five concentrations spanning at least one order of magnitude, not one fixed level	3.3	1/17	1/17	15/17
5	Robustness to signal perturbation	Re-run classification after injecting noise scaled to observed replicate variability, or compare across instruments or operators	3.4	0/17	5/17	12/17
6	Separated classification and quantification claims	State explicitly which evidence supports identity assignment and which, if any, supports concentration estimates	3.5	1/17	2/17	14/17
7	Dedicated mixture-quantification model	Where concentration of a mixture component is claimed, report a supervised calibration (e.g., PLSR) distinct from the classification model	3.5	0/17	0/17	17/17
8	Real-matrix validation	Confirm discrimination in the intended sample matrix (food extract, serum, urine), not only in buffer or pure solvent	3.3, 3.4	6/17	0/17	11/17
9	Confusion matrix or per-analyte accuracy	For any blind or held-out test, report which analytes are misclassified as which, and at what rate	3.1, 3.2	1/17 (full matrix); 3/17 (any accuracy figure)	3/17	11/17

Note. The reporting checklist is mapped to the corresponding gap identified in Section 3, with a retrospective count against the 17 primary discrimination-claim studies audited in this paper.

3.5 is not that quantification is attempted and falls short, but that the word is used without the corresponding evidence being present at all.

Item 2 deserves a similar note. The floor of three independent replicates is not proposed as a formal power calculation, but as the minimum number of measurements from which any variance or outlier can be assessed at all; below three replicates, a response vector's spread cannot be estimated, only assumed. This floor is consistent with long-standing convention in single-laboratory method validation, where replicate measurement in triplicate or more is treated as a baseline requirement for estimating precision [30]. Studies working near the limit of detection, or with material that is expensive or difficult to synthesize in bulk, may reasonably need more replicates rather than fewer to characterize response varia-

bility adequately. There is proposed here as a minimum below which a discrimination claim should be treated cautiously, not as a target that on its own guarantees sufficiency.

Two of the nine items, numbers 1 and 6, are proposed here as the practical starting point. Both can typically be satisfied through re-analysis of data a study has already collected, cost almost nothing in additional experimental time, and would on their own resolve the two gaps that recur most consistently across the literature audited in Section 2. A reviewer or editor who begins by asking for these two alone would already shift the baseline for what a discrimination claim is expected to demonstrate.

5. A Worked Illustration

To show that the checklist in Section 4 is operational rather than aspirational, it is useful to apply it directly to two of the papers already introduced in Section 2, chosen because between them they satisfy more of the nine criteria than most single papers in this literature, which makes them a fair test of what the checklist rewards and what it still catches.

The ratiometric Eu/Tb-MOF antibiotic array of Xie *et al.* [1] is the strongest available example against criterion 3. After building a classification model from the F545/F616 response of 25 antibiotics across eight structural classes, the authors challenged that model with 48 blind samples and reported 98 percent correct identification. This is a direct, textbook satisfaction of the blind-testing requirement described in Section 3.1: the accuracy figure describes performance on data the model did not see during construction, not on the training set itself. The same study also reports testing in the presence of interfering substances and quantitative recovery for antibiotic mixtures in real water samples, which speaks partially to criterion 8. What the paper does not report is a numerical cluster-separation index for the PCA space itself, criterion 1, or a robustness check under deliberately injected signal noise, criterion 5, so a checklist-based read of this study would credit it clearly on blind validation while still flagging two open items.

The pH-regulated lanthanide array aimed at ten metal ions in biological fluids [5] offers a complementary strength. Its PCA and Extra Trees classification are backed by an explicit confusion matrix for the full ten-ion discrimination task, which satisfies criterion 9 far more precisely than a score plot alone: a confusion matrix shows exactly which ion pairs are occasionally misclassified and at what rate, rather than only asserting that the ten clusters look separated. The same study tested classification at three separate concentrations, 100, 25, and 10 micromolar, for each ion, which goes some way toward criterion 4, and extended testing to binary and ternary mixtures, which addresses part of the concern in Section 3.5 by treating mixture discrimination as its own tested condition rather than an extrapolation from single-analyte results. What remains unclear from the available reporting is whether the ion identification claim and any associated concentration values are backed by separate calibration evidence, or whether, as in several stud-

ies audited in Section 3.5, the two claims are presented together without being separately validated.

Read side by side, these two studies show that no single paper in this literature currently satisfies the full checklist, but that most of it is already achievable using techniques already in active use somewhere in the field. Neither the blind-sample accuracy statistic in the antibiotic array nor the confusion matrix in the metal ion array required a new sensing platform to produce; both are analytical choices made after the same kind of fluorescence titration data that nearly every study in this space already collects. That is the more general point the two examples are meant to illustrate: the checklist is not asking the field to do harder chemistry, it is asking the field to stop treating a subset of already-available statistical practice as optional.

6. Outlook

Analytical and clinical research have been through this kind of correction before. Quantitative PCR faced a similar problem in the early 2000s, when a proliferating literature reported cycle threshold values and fold-change claims without a shared minimum standard for what counted as an adequately validated result. The response was not a new instrument or a new chemistry, but a published checklist, MIQE, specifying the minimum information an experiment needed to report before its quantification claims could be taken at face value [31]. A comparable movement took hold in diagnostic accuracy research through the STARD statement, which set out the essential items a study needed to report before a diagnostic test's accuracy could be evaluated on equal terms across publications [32] and more broadly across biological and biomedical fields through the MIBBI initiative, which coordinated a family of minimum-information checklists spanning multiple experimental domains under the shared premise that a result is only as comparable as its reporting is complete [33]. A related precedent exists for a different sensor class: IUPAC has issued terminology and evaluation recommendations for electrochemical electronic tongue arrays. No comparable proposal yet exists for the fluorescence-based, PCA/HCA-driven discrimination claims that now dominate the Ln-MOF literature, which is the gap this paper addresses.

Adoption of each of the above guidelines was gradual and uneven, but the checklists gave editors and reviewers a concrete reference point to point to, rather than relying on individual judgment about what sufficient validation meant, and citation of the standard became, in itself, a signal that a study had been held to a defined bar. The parallel to fluorescence sensor arrays is not exact, since these earlier standards addressed measurement technologies used across thousands of laboratories, while Ln-MOF sensor arrays remain a smaller and more heterogeneous field. But the underlying problem, a fast-growing body of work converging on one analytical workflow without converging on what evidence that workflow requires, is close enough that the same style of solution is worth attempting early, before the reporting habits described in Sections 2 and 3 become any more entrenched.

The checklist proposed in Section 4 is deliberately modest in what it asks. Two of its nine items, a quantitative separation metric and a clear separation of classification from quantification claims, require no new experiments in the large majority of cases, only a different treatment of data a study has already collected. A further two or three items, blind sample testing and noise-robustness checks in particular, require additional effort but not additional material synthesis, and both have already been demonstrated as low-cost additions in the studies audited here [1] [3] [5]. The remaining items, concentration-range testing and real-matrix validation, are frequently already partially present in the titration and recovery data most studies generate for other purposes, and mainly require that the same data be re-examined through the lens of the discrimination claim rather than only the calibration claim.

Two audiences can act on this without waiting for a formal standard to be published. Reviewers can begin by requesting item 1 and item 6 as a routine condition of acceptance for any manuscript that claims multi-analyte discrimination, since both are inexpensive to satisfy and directly address the two gaps that recur most often in the literature surveyed here. Editors can support this by treating a quantitative separation metric as an expected component of a discrimination claim in reviewer guidelines, in the same way that a limit of detection is already an expected component of a sensitivity claim. Neither step requires community-wide consensus before it can start; either could be applied to the next submission a given journal receives.

The broader aim of this checklist is not to slow the field down. Ln-MOF sensor arrays have moved, within a few years, from single-analyte turn-on and turn-off probes to genuinely multiplexed platforms capable of resolving structurally similar antibiotics, dyes, and metabolites in real food and biological matrices, a capability that conventional chromatographic methods cannot match on cost, speed, or portability. That progress deserves an evidentiary standard that keeps pace with it. The gap this paper has described is not a shortage of good chemistry. It is a shortage of shared agreement about what, beyond a separated cluster on a plot, counts as proof that the chemistry works.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Xie, R., Yang, P., Liu, J., Zou, X., Tan, Y., Wang, X., *et al.* (2021) Lanthanide-Functionalized Metal-Organic Frameworks Based Ratiometric Fluorescent Sensor Array for Identification and Determination of Antibiotics. *Talanta*, **231**, Article ID: 122366. https://doi.org/10.1016/j.talanta.2021.122366

[2] Wang, X., Gopalsamy, K., Clavier, G., Maurin, G., Ding, B., Tissot, A., *et al.* (2024) Lanthanide MOF-Based Luminescent Sensor Arrays for the Detection of Castration-Resistant Prostate Cancer Curing Drugs and Biomarkers. *Chemical Science*, **15**, 6488-6499. https://doi.org/10.1039/d3sc06899d

[3] Yin, K., Wu, S., Zheng, H., Gao, L., Liu, J., Yang, C., *et al.* (2021) Lanthanide Metal-Organic Framework-Based Fluorescent Sensor Arrays to Discriminate and Quantify Ingredients of Natural Medicine. *Langmuir*, **37**, 5321-5328. https://doi.org/10.1021/acs.langmuir.1c00412

[4] Zhang, W., Sun, J., Li, X., Wang, S., Zhang, W., Gong, Y., *et al.* (2025) Lanthanide MOF-Based Luminescent Sensor Array for Detection and Identification of Contaminants in Water and Biomarkers. *Talanta*, **281**, Article ID: 126853. https://doi.org/10.1016/j.talanta.2024.126853

[5] Shi, X., Wu, Y., Hu, G. and Li, Y. (2025) Machine Learning-Assisted Lanthanide Fluorescence Sensor Array Based on Ph Regulation for Pattern Recognition of Metal Ions in Biological Fluids. *The Journal of Physical Chemistry C*, **129**, 12437-12449. https://doi.org/10.1021/acs.jpcc.5c03409

[6] Yue, X., Fu, L., Li, Y., Xu, S., Lin, X. and Bai, Y. (2023) Lanthanide Bimetallic MOF-Based Fluorescent Sensor for Sensitive and Visual Detection of Sulfamerazine and Malachite. *Food Chemistry*, **410**, Article ID: 135390. https://doi.org/10.1016/j.foodchem.2023.135390

[7] Gao, G., Wang, L., Wang, X., Zhang, X., Lai, X., Yang, Q., *et al.* (2026) A Dual-Emission Ln-MOF as Ratiometric Fluorescence Sensor for Detecting NF and OTC with High Sensitivity and Selectivity. *Spectrochimica Acta Part A*: *Molecular and Biomolecular Spectroscopy*, **345**, Article ID: 126831. https://doi.org/10.1016/j.saa.2025.126831

[8] Chen, S.Y., Ye, J.T., Wang, L.Y., Lin, Y., Gong, Y.R., Xiao, M.D., *et al.* (2026) Lanthanide Metal-Organic Framework Fluorescent Sensor for the Detection of Furazolidone in Tap Water and Animal Feed. *Inorganic Chemistry*, **65**, 5214-5221. https://doi.org/10.1021/acs.inorgchem.6c00129

[9] Yue, X., Wu, C., Zhou, Z., Fu, L. and Bai, Y. (2022) Fluorescent Sensing of Ciprofloxacin and Chloramphenicol in Milk Samples via Inner Filter Effect and Photoinduced Electron Transfer Based on Nanosized Rod-Shaped Eu-MOF. *Foods*, **11**, Article 3138. https://doi.org/10.3390/foods11193138

[10] Zhang, H., Ma, L., Ma, F.L., Wang, X.F., Wang, L.L. and Wang, D.Z. (2025) Design of "Turn-Off" Luminescent 3D Ln-MOFs for Sensitive Detection of Fe^{3+} and $Cr_2O_7^{2-}$. *Journal of Molecular Structure*, **1335**, Article ID: 141979. https://doi.org/10.1016/j.molstruc.2025.141979

[11] Zhou, Y., Jiang, Y., Chen, X., Long, H., Zhang, M., Tang, Z., *et al.* (2024) Enhanced Sensitivity and Accuracy of Tb^{3+}-Functionalized Zirconium-Based Bimetallic MOF for Visual Detection of Malachite Green in Fish. *Foods*, **13**, Article 2855. https://doi.org/10.3390/foods13172855

[12] Sharma, P. and Siddiqui, K.A. (2025) Metal-Organic Frameworks as Fluorescent and Colorimetric Sensors for Antibiotic Tracing. *Discover Chemistry*, **2**, Article No. 199. https://doi.org/10.1007/s44371-025-00282-0

[13] Zhang, Y., Li, C., Jiang, M., Liu, Y. and Sun, Z. (2025) Advancements and Prospects of Metal-Organic Framework-Based Fluorescent Sensors. *Biosensors*, **15**, Article 709. https://doi.org/10.3390/bios15110709

[14] Chen, X. and Wang, Q. (2026) Recent Advances of Optical Sensing Arrays (2019-2025): From Design of Optical Elements to Multidimensional Analysis for Applications. *Analytical Chemistry*, **98**, 2633-2679. https://doi.org/10.1021/acs.analchem.5c05597

[15] Zhou, X., Xie, C., Ye, L. and Ding, B. (2026) Smart Porous-Framework Sensor Arrays: Design Principles, AI-Driven Performance and Multiscenario Precision Detection.

Coordination Chemistry Reviews, **549**, Article ID: 217306.
https://doi.org/10.1016/j.ccr.2025.217306

[16] Chi, J., Qin, Y., Song, Y. and Feng, L. (2025) Single-Component Dual-Emissive MOF Sensor Array for Fingerprint Identification of Multiplex Antibiotics. *Analytical Chemistry*, **97**, 18299-18307. https://doi.org/10.1021/acs.analchem.5c03414

[17] Xia, Y.F., Yu, X.S., Wen, S.T., Yuan, H.Q., Xu, C.C., Yang, L., *et al.* (2025) A Facile and Stable Single-System Fluorescent Sensor Array for the Quantification and Discrimination of Nitrofuran Antibiotics. *Food Chemistry*, **494**, Article ID: 146136.
https://doi.org/10.1016/j.foodchem.2025.146136

[18] Wang, S., Xu, J., Yue, F., Zhang, L., Jia, L. and Zhao, T. (2025) Lanthanide-Empowered 3D-Printed Smartphone Interfacing Device Realizes Point-of-Care Discrimination of Multiple Fluoroquinolone Antibiotics in Aquatic Environments. *Sensors and Actuators B*: *Chemical*, **443**, Article ID: 138258.
https://doi.org/10.1016/j.snb.2025.138258

[19] Lei, M., Ge, F., Gao, X., Shi, Z. and Zheng, H. (2021) A Water-Stable Tb-MOF as a Rapid, Accurate, and Highly Sensitive Ratiometric Luminescent Sensor for the Discriminative Sensing of Antibiotics and D_2O in H_2O. *Inorganic Chemistry*, **60**, 10513-10521. https://doi.org/10.1021/acs.inorgchem.1c01145

[20] Lian, X., Chen, F., Li, Y., Zhang, R., Niu, H. and Zhang, Z. (2026) A Smartphone-Read, Machine Learning-Enhanced Lanthanide Hydrogel Sensor for Multimodal and On-Site Detection of Kanamycin in Food Safety. *Analytical Chemistry*, **98**, 3239-3251. https://doi.org/10.1021/acs.analchem.5c07000

[21] Leng, Y., Zhang, J., Xu, T., Ma, Y., Luo, H., Huo, D., *et al.* (2026) A Triple-Channel Fluorescent Sensor Array Based on R6G@Eu-MOF for Baijiu Discrimination. *Microchemical Journal*, **226**, Article ID: 118455.
https://doi.org/10.1016/j.microc.2026.118455

[22] Yuan, P.X., Wang, Y.P., Du, F., Yang, L.P. and Wang, L.L. (2025) Ratiometric Fluorescence Sensing and Discrimination of Tetracycline Analogs by Using Coumarin-Embedded Eu-MOF Nanosensor. *Talanta*, **281**, Article ID: 126914.
https://doi.org/10.1016/j.talanta.2024.126914

[23] Esbensen, K.H. and Geladi, P. (2010) Principles of Proper Validation: Use and Abuse of Re-Sampling for Validation. *Journal of Chemometrics*, **24**, 168-187.
https://doi.org/10.1002/cem.1310

[24] Kennard, R.W. and Stone, L.A. (1969) Computer Aided Design of Experiments. *Technometrics*, **11**, 137-148. https://doi.org/10.1080/00401706.1969.10490666

[25] Du, S.Z., Sun, Z., Han, L., Qing, M., Luo, H.Q. and Li, N.B. (2020) Two 3d-4f Metal-Organic Frameworks as Fluorescent Sensor Array for the Discrimination of Phosphates Based on Different Response Patterns. *Sensors and Actuators B*: *Chemical*, **324**, Article ID: 128757. https://doi.org/10.1016/j.snb.2020.128757

[26] Gao, Y., Zhu, Y., Wang, Y. and Bi, J. (2024) Water-Stable Ln-MOF as a Multi-Emitting Luminescent Sensor for the Detection of Metal Ions and Pharmaceuticals. *Spectrochimica Acta Part A*: *Molecular and Biomolecular Spectroscopy*, **323**, Article ID: 124915. https://doi.org/10.1016/j.saa.2024.124915

[27] He, X.Y., Wang, Y., Xue, Q., Qian, W.F., Li, G.L. and Li, Q. (2025) Molecularly Imprinted MOF Nanozymes: Demonstration of Smartphone-Integrated Dual-Mode Platform for Ratiometric Fluorescent/Colorimetric Detection of Chloramphenicol. *Food Chemistry*: *X*, **26**, Article ID: 102322.
https://doi.org/10.1016/j.fochx.2025.102322

[28] Sun, Y.X., Ji, B.T., Chen, J.H., Liu, L.P., Gao, L.L., Deng, Z.P., *et al.* (2025) A

Smartphone-Integrated Bimetallic Ratiometric Fluorescent Probe for Specific Visual Detection of Tetracycline Antibiotics in Food Samples and Latent Fingerprinting. *Food Chemistry*, **464**, Article ID: 141782. https://doi.org/10.1016/j.foodchem.2024.141782

[29] Zhang, W., Zhao, Y., Sun, J., Peng, D., Li, X., Lv, Y., *et al.* (2024) Fluorescent Sensors Based on Lanthanide-Based Metal-Organic Frameworks via Devices and Ph Response. *Inorganic Chemistry*, **63**, 15527-15536. https://doi.org/10.1021/acs.inorgchem.4c02795

[30] Thompson, M., Ellison, S.L.R. and Wood, R. (2002) Harmonized Guidelines for Single-Laboratory Validation of Methods of Analysis (IUPAC Technical Report). *Pure and Applied Chemistry*, **74**, 835-855. https://doi.org/10.1351/pac200274050835

[31] Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., *et al.* (2009) The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. *Clinical Chemistry*, **55**, 611-622. https://doi.org/10.1373/clinchem.2008.112797

[32] Bossuyt, P.M., Reitsma, J.B., Bruns, D.E., Gatsonis, C.A., Glasziou, P.P., Irwig, L., *et al.* (2015) STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. *BMJ*, **35**, h5527. https://doi.org/10.1136/bmj.h5527

[33] Taylor, C.F., Field, D., Sansone, S.A., Aerts, J., Apweiler, R., Ashburner, M., *et al.* (2008) Promoting Coherent Minimum Reporting Guidelines for Biological and Biomedical Investigations: The MIBBI Project. *Nature Biotechnology*, **26**, 889-896. https://doi.org/10.1038/nbt.1411

Appendix

https://zenodo.org/records/21854123