COSIG
Collection of Open Science Integrity Guides

Anyone can do post-publication peer review.

Anyone can do forensic metascience.

Anyone can sleuth.

However, investigating the integrity of the published scientific literature often requires domain-specific knowledge that not everyone will have. This project is a collection of guides written by publication integrity experts to distribute this domain-specific knowledge so that others can participate in post-publication peer review.

COSIG currently hosts 8 guides and was last updated on 7 February 2025.

Table of contents

COSIG
Collection of Open
Science Integrity Guides

Misidentified and non-verifiable cell lines

Last updated: 5 February 2025

Cell lines

A cell line is a population of cells that can be grown in a laboratory culture indefinitely. Cell lines are an essential tool for biomedical research because they allow biological experiments to be performed in vitro (i.e. outside of a living organism). For instance, researchers developing new cancer drugs will certainly test the effect of their drug candidates on many different cancer cell lines before ever considering testing the drug candidate on live animals, let alone human patients.

Because cell lines can grow indefinitely, one research laboratory or laboratory supplier can take a few cells from their cell line stock and give them to another laboratory for that laboratory to begin culturing their own stock. By far the most popular cell line is HeLa, and the many thousands of HeLa cell stocks used in biomedical research around the world can all be traced back to the original stock of "immortal" cervical cancer cells taken from Henrietta Lacks in 1951.

Misidentified/contaminated cell lines

One complication of relying on cell lines for biomedical research is that stocks will often become cross-contaminated by other cell lines. For instance, HeLa is extremely aggressive as cancer cell lines go and HeLa cells will easily overtake other cell line stocks that are stored nearby. The popular gastric cancer cell line BGC-823 is one such victim of HeLa's aggression; there are no remaining uncontaminated stocks of BGC-823 and thus it is no longer considered an appropriate experimental model for gastric cancer.

Cellosaurus BGC-823 (CVCL_3360)
[Text version]

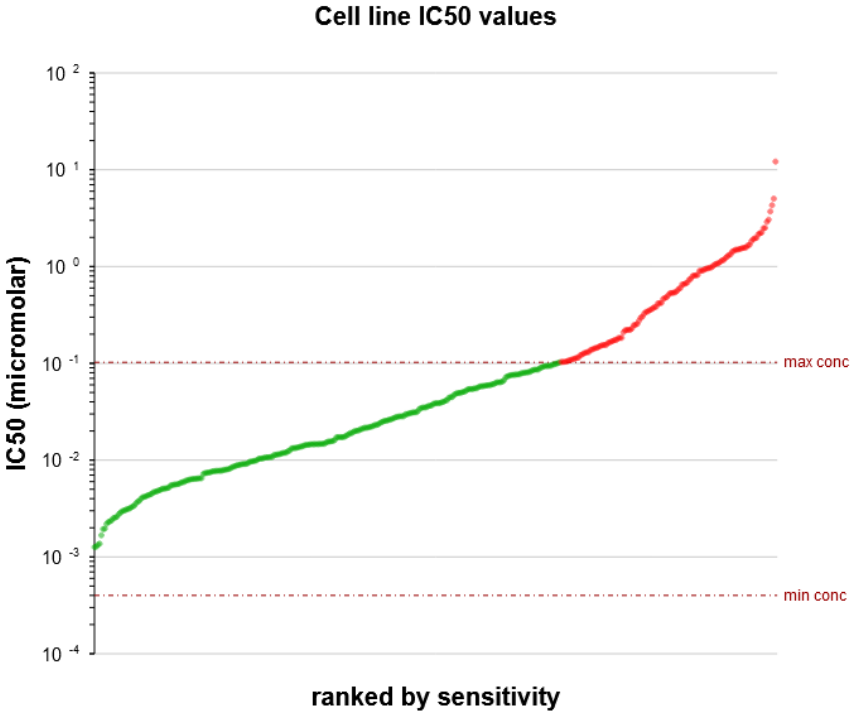
Cell line name	BGC-823
Synonyms	BGC823
Accession	CVCL_3360
Resource Identification Initiative	To cite this cell line use: BGC-823 (RRID:CVCL_3360)
Comments	Problematic cell line: Contaminated. Shown to be a HeLa derivative (CCRID; PubMed=26116706; PubMed=28851942). Originally thought to be a gastric adenocarcinoma. Part of: Cancer Dependency Map project (DepMap) (includes Cancer Cell Line Encyclopedia - CCLE). Registration: International Cell Line Authentication Committee, Register of Misidentified Cell Lines; ICLAC-00570. Population: African American. Transformant: NCBI_TaxID; 333761; Human papillomavirus type 18 (HPV18). Omics: Proteome analysis by 2D-DE. Omics: Transcriptome analysis by microarray. Miscellaneous: Formerly the CCRID database had 2 entries describing this cell line (3111C0001CCC000062, 3131C0001000700011). They were one of the sources for the STR profile of this entry. Misspelling: BSG823; Note=Occasionally. Derived from site: In situ; Uterus, cervix; UBERON=UBERON_0000002.

The Cellosaurus entry for BGC-823 now warns about the cell line being contaminated by HeLa.

The International Cell Line Authentication Committee maintains a register of misidentified cell lines that is currently 593 entries long. A good number of these cell lines were contaminated by cells from a completely different organism, such as human salivary gland cell line CAC2, which is actually made up of unknown cells from a rat.

Cellosaurus is an encyclopedia of thousands of cell lines and will link to studies showing that a cell line is contaminated or otherwise misidentified.

Because different cell lines can have a wide variety of responses to the same treatment, it is essential that researchers avoid misidentified cell lines and only work with authenticated cell lines that actually represent their system of interest. Consider the fact that cancer cell lines can have wildly different sensitivity to the chemotherapy drug paclitaxel.



Number of cell lines screened	395
Maximum IC50 (uM)	12.1
Geometric mean (uM)	0.0482
Minimum IC50 (uM)	0.00126
Min screening concentration (uM)	0.000400
Max screening concentration (uM)	0.102

Half-maximal inhibitory concentration (IC50), in this context, is a measure that describes what concentration of a drug is needed to inhibit the growth of a cell line by 50%. The Genomics of Drug Sensitivity in Cancer (GDSC) database estimated these values for 395 cell lines for the drug paclitaxel. Note that paclitaxel can be more than a thousand times as potent against some cell lines than others, even for cell lines from the same type of cancer.

Cell line verification/authentication

Researchers should always verify that their cell lines stocks are what they believe them to be. There are a number of ways researchers can verify the integrity of their cell line stocks (many of which are detailed by the American Type Culture Collection). The most common authentication technique is short tandem repeat (STR) profiling, which identifies specific molecular signatures in a cell line's genome.

Non-verifiable cell lines

The names of cell lines are usually just a jumble of letters and numbers and thus can often be easily confused. For instance, the name of BGC-823 is very similar to that of MGC-803, another misidentified gastric cancer cell line. One might easily misspell these cell lines as "BGC-803" or "MGC-823". However, it was recently reported by Oste et al. (2024) that many publications will use these misspelled identifiers to refer to another cell line entirely distinct from these existing cell lines. For instance, Zhong et al. (2021) report experiments in cell lines "BGC-803" and "BSG-823" in addition to experiments in the contaminated cell lines BGC-823 and MGC-803.

Hundreds of articles have referred to experiments in these cell lines despite there being no indication that these cell lines actually exist; there are no entries for these cell lines in any cell line indices, they cannot be found in any supplier catalogs, there are no articles describing how these cell lines were established and no one seems to have produced any genetic profiles of these cell lines to confirm their identities. Oste et al. identified eight such "miscellings": BGC-803, BSG-803, BSG-823, GSE-1, HGC-7901, HGC-803, MGC-823 and TIE-3, although there are certainly many more that have not been studied in detail.

Example 1: Contaminated cell lines

Liu et al. (2017) report experiments in the cell lines MGC-803, L02 and SMMC-7721. However, each of these cell lines are contaminated by HeLa are thus are no longer considered suitable models for their respective cancers.

Example 2: Non-verifiable and contaminated cell lines

Yang et al. (2018) mention and provide experimental results in 10 different cell lines, of which 4 are problematic, shown below in bold:

- **SUN-216**: mentioned once in text of paper and again in Figure 1B, both spelled as SUN-216. SNU-216 is an existing cell line. A possible non-verifiable cell line identifier, but not one that has yet been studied in depth.
- **BGC-823**: Contaminated cell line.
- AGS
- **BGC-803**: Mentioned once in text and again in Figure 1B. One of the eight non-verifiable cell line identifiers studied by Oste et al. Likely derived from a typo that confused the cell lines MGC-803 and BGC-823, both contaminated.
- NUGC4
- MKN74
- MKN45
- **SGC-7901**: Contaminated cell line.
- HGC-27
- GES-1

Additional resources

- ATCC Cell Line Authentication Test Recommendations
- Cellosaurus
- International Cell Line Authentication Committee (ICLAC)
- ICLAC-curated reviews on cell line misidentification
- "Misspellings or 'miscellings'—Non-verifiable and unknown cell lines in cancer research publications" (2024)

COSIG
Collection of Open
Science Integrity Guides

Energy-dispersive X-ray spectroscopy

Last updated: 7 February 2025

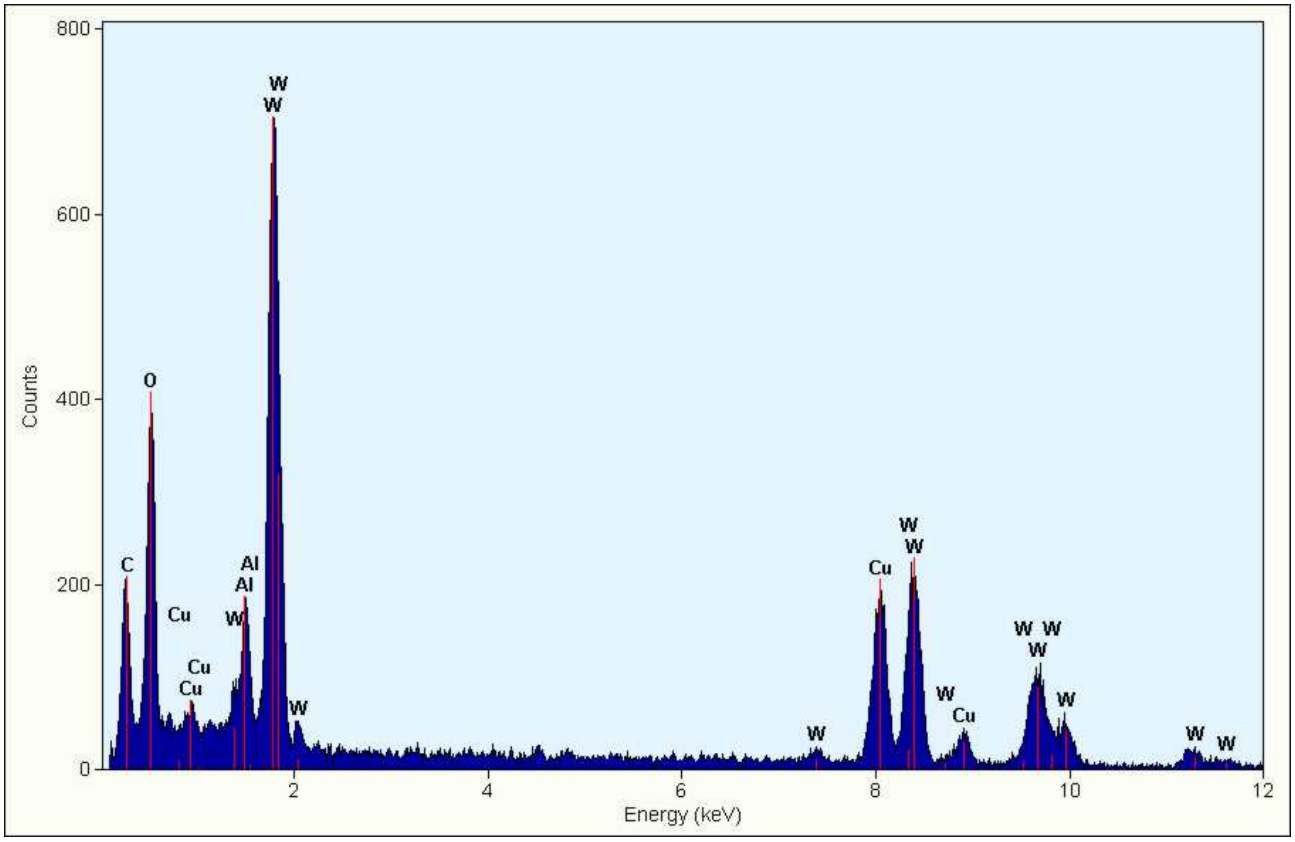
Energy-dispersive X-ray spectroscopy (EDX/EDS/EDAX) is a popular experimental technique used to characterize the elemental composition of a sample.

EDX typically involves bombarding a sample with high-energy electrons, exactly the kind that are used for scanning electron microscopy (SEM) and transmission electron microscopy (TEM). SEM/TEM instruments are often sold with EDX instruments already built in and EDX instruments are often sold as SEM/TEM accessories.

Atoms contain negatively-charged electrons bound to a positively-charged nucleus at well-defined energy levels. When an atom in a sample is exposed to a beam of high-energy electrons, some of these bound electrons will be ejected from the atom and leave their energy level or “shell”, leaving an “electron hole”. This hole can then be filled by another electron dropping from a less tightly-bound shell, which releases the excess energy required to drop to a lower energy level in the form of an X-ray.

The energy levels electrons occupy are well-defined for each element, meaning that each element will emit X-rays at a series of well-defined characteristic energies. An EDX detector will read the energies of emitted X-rays and construct a spectrum, which can then be analyzed quantitively to determine the elemental composition of a sample.

A typical EDX spectrum will show X-ray energy on the x axis (typically expressed in electron volts, eV) and counts or counts per second (cps) on the y axis (with each “count” representing a single detected X-ray emission).



An example EDX spectrum for an aluminum tungsten oxide on a carbon foil supported on a copper grid. Adapted from material prepared by Dr. Frank Krumeich.

Because the characteristic X-ray emission energies of any element are predictable and well-defined, the height and position of peaks they will produce on an EDX spectrum are also predictable and well-defined. For example:

- Carbon will only ever produce a single peak at $\sim$ 277 eV.
- Oxygen will only ever produce a single peak at $\sim$ 525 eV.
- Silicon will only ever produce two characteristic X-rays: one at $\sim$ 1.74 keV and another at $\sim$ 1.84 keV. These peaks are too close to be distinguished by EDX detectors so only appear as one peak at $\sim$ 1.75 keV.

Heavier elements tend to produce more peaks at higher energies.

The characteristic X-ray energies that will be produced by each element can be found in various lookup tables, like this one from Lawrence Berkeley National Laboratory. However, the most convenient and comprehensive resource for determining expected peaks is the software NIST DTSA-II. DTSA-II also allows the user to simulate spectra and quantify elemental abundances from real spectra.

Background signal/continuum/bremsstrahlung and the Duane-Hunt limit

Aside from the peaks produced by the characteristic X-rays of elements in a sample, an EDX spectrum will also feature a background signal (also called a "continuum") caused by bremsstrahlung (German for "braking radiation"). These are X-rays emitted by the incident electrons being redirected by the atomic nuclei in the sample.

This background signal will take on a broad, hill-like distribution that tapers off at higher energies when using a low-energy electron beam (such as on an SEM instrument with an accelerating voltage on the order of 10 kV). When using a higher-energy electron beam (such as on a TEM instrument with an accelerating voltage on the order of 100 kV), this distribution will be much more spread out. Over the range of energies typically shown in EDX spectra (usually around 0 to 20 keV), the bremsstrahlung signal on a TEM EDX spectra will appear as low, constant-intensity background noise.

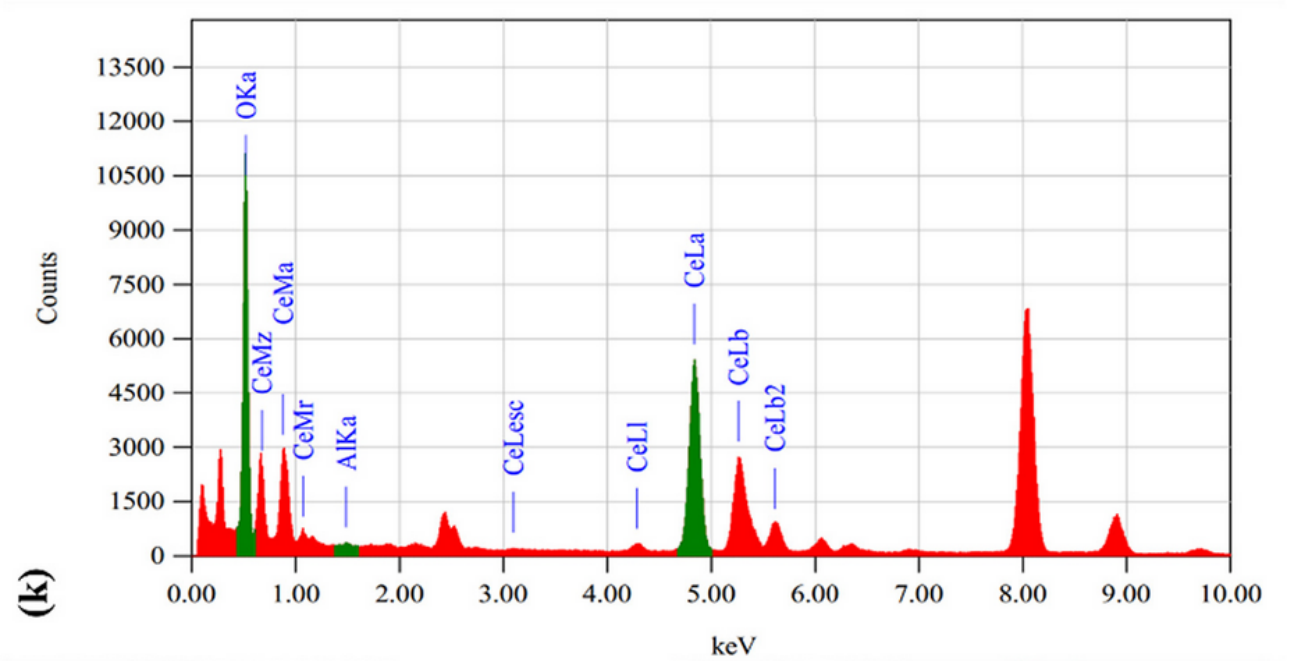
An EDX spectrum will only ever have a nonzero signal up to the energy of the incident electron beam. For example, if EDX is performed with an electron beam with an accelerating voltage of 10 kV, no elemental peaks or bremsstrahlung will be observed at energies higher than 10 keV. This is known as the Duane-Hunt limit.

Elements not detectable by EDX

Hydrogen (atomic number 1) and helium (atomic number 2) do not produce characteristic X-rays and are thus not detectable by EDX. Lithium (atomic number 3) and beryllium (atomic number 4) produce characteristic X-rays at such a low energies ($\sim$ 54 and $\sim$ 108 keV, respectively) that most EDX detectors will fail to capture them.

Escape peaks

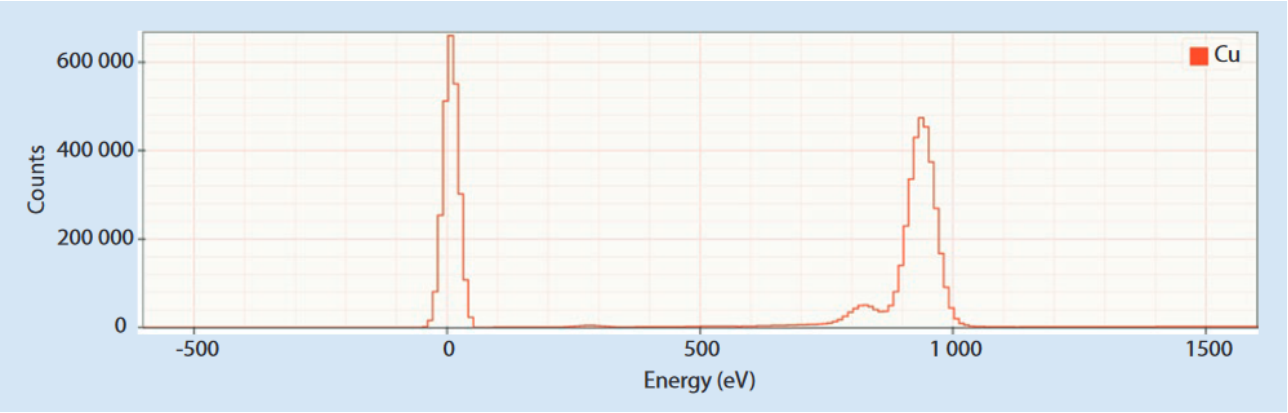
Some elements will produce silicon "escape" peaks. These peaks occur with silicon-based detectors (used in the vast majority of EDX instruments) and will appear exactly one Si $K\alpha$ emission energy ($\sim$ 1.7 keV) below prominent peaks produced by an element. Some models of EDX instrument will remove escape peaks from the spectrum automatically.



An example EDX spectrum for a material containing cerium. A faint cerium escape peak is labeled at $\sim$ 3.1 keV. Adapted from Figure 3K of Shah et al. (2023).

Zero strobe peak

Many EDX detectors will insert an artificial peak centered at 0 eV. This peak, called the “zero strobe”, “zero strobe peak” or “strobe peak”, is used for instrument calibration is often removed before further analysis. If it is left included, it can give the appearance of detection of X-rays with negative energies. The appearance of a zero strobe peak in a published EDX spectrum is usually no cause for concern.



An example EDX spectrum for elemental copper showing a zero strobe peak centered around 0 eV. Adapted from Figure 17.7 of Scanning Electron Microscopy and X-ray Microanalysis.

Sample preparation

The optimal samples for EDX and the most suitable for quantitative analysis will have a flat, smooth, polished surface. Spectra from fibers, particles and rough surfaces can be collected, but only with a signficant reduction in accuracy.

If a sample is not conductive, a negative charge from the incident electron beam can build up on the sample surface, reducing the effective beam energy. This harms both EDX analysis and SEM visualization. To counteract this, samples are often coated with a conductive material. These materials will often appear as additional peaks in EDX spectra. Coatings are most commonly composed of carbon, gold, platinum, palladium, silver, iridium or chromium.

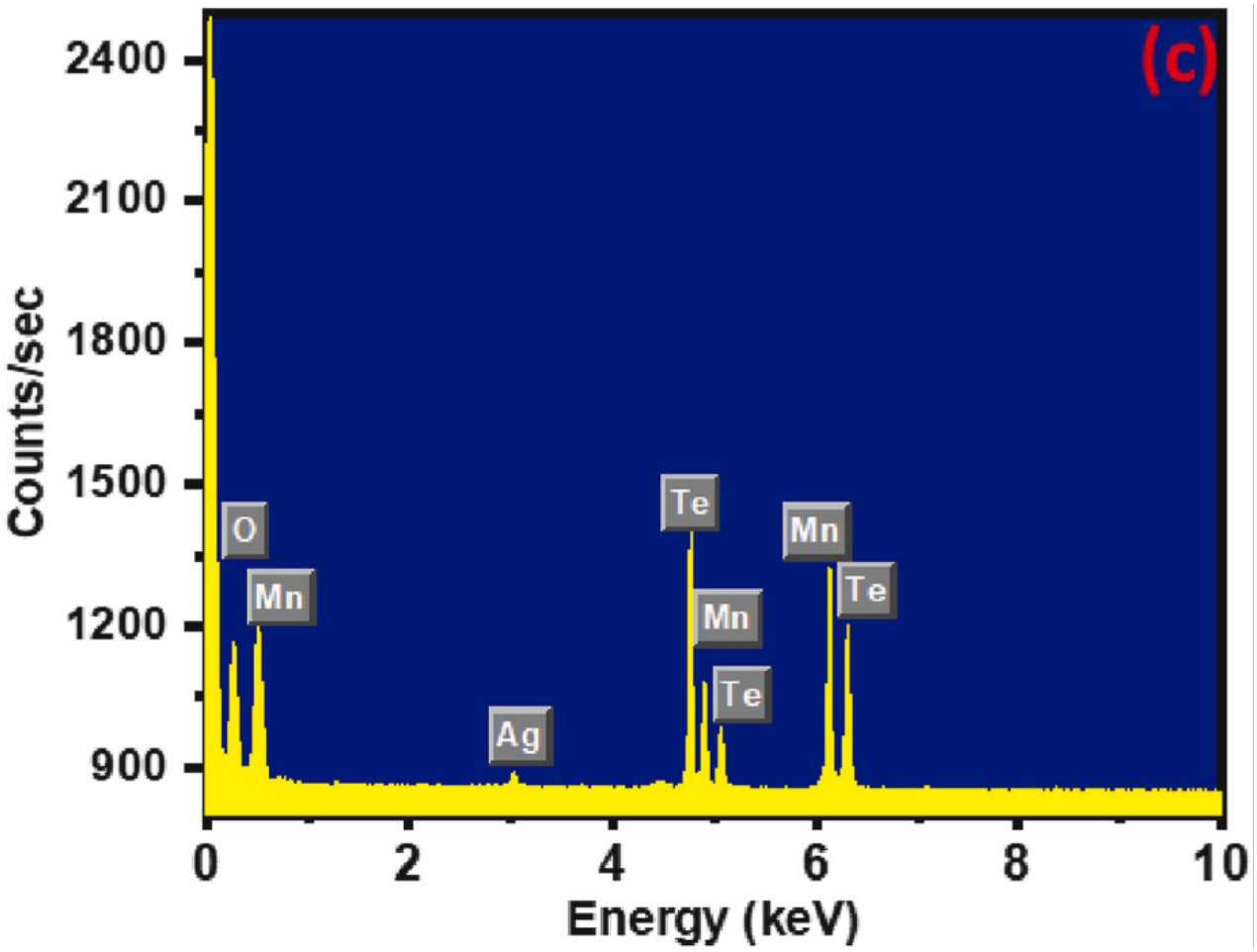
Mounting grids

Samples are sometimes mounted to a metallic grid and/or a thin foil to facilitate handling (this is more common for EDX conducted with a TEM instrument). The materials used in these grids will often appear in EDX spectra. Grids and foils are most commonly composed of gold, copper, carbon or aluminum.

Example 1: Problematic EDX spectrum

Hussain et al. (2024) claim to use EDX to characterize the elemental composition of an AgMnTe composite. However, the EDX spectrum they provide in Figure 2C has several issues:

1. Tellurium has no peaks between 6 keV and 7 keV.
2. Manganese has no peak at $\sim$ 6.2 keV.
3. Manganese has no peak at $\sim$ 5 keV.
4. The background signal is unusually tall and flat, which is not consistent with EDX performed on an SEM instrument.

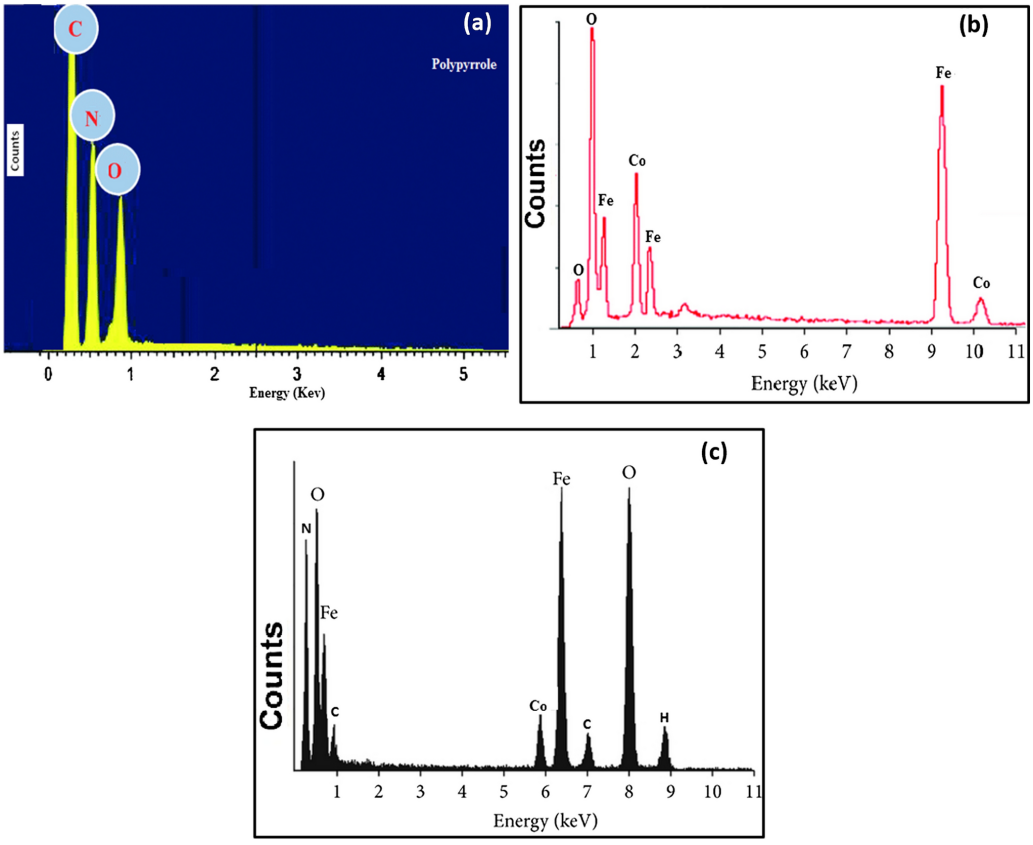


A problematic EDX spectrum with several nonsensical peak labels. Adapted from Figure 2C of Hussain et al. (2024).

Example 2: Problematic EDX spectrum

Alwadai et al. (2022) claim to use EDX to characterize the elemental composition of polypyrrole, cobalt ferrite, and a poylpyrrole/cobal ferrite composite. However, the spectra they provide have numerous issues:

1. Oxygen's sole peak should be at 524.9 eV, but appears in 3A at $\sim$ 0.9 keV.
2. Oxygen has two peaks in 3B. Oxygen has one peak.
3. Iron has no peaks between 1 and 4 keV. 3B shows two peaks for iron in this range.
4. Iron has no peaks > 8 keV, unlike what is shown in 3B.
5. Cobalt has no peak at $\sim$ 2 keV and no peak at $\sim$ 10 keV, unlike what is shown in 3B.
6. In increasing order, the lowest energy peaks in 3C should be carbon, nitrogen, oxygen, iron. The order shown in 3C is nitrogen, oxygen, iron, carbon.
7. Cobalt has no peaks at $\sim$ 5.9 keV. The nearest expected peak is an escape peak at $\sim$ 5.2 keV.
8. Carbon does not have a peak at $\sim$ 7 keV.
9. Oxygen does not have a peak at $\sim$ 8 keV.
10. Hydrogen does not have any peaks, let alone one at $\sim$ 9 keV.



Problematic EDX spectra with numerous issues. Adapted from Figure 3 of Alwadai et al. (2022).

Additional resources

- Lecture notes by Dr. Chris Boothroyd
- *Scanning Electron Microscopy and X-ray Microanalysis*
- *“Sample Preparation for Electron Probe Microanalysis — Pushing the Limits"*
- "Performing elemental microanalysis with high accuracy and high precision by scanning electron microscopy/silicon drift detector energy-dispersive X-ray spectrometry (SEM/SDD-EDS)"
- NIST DTSA-II

Thanks to Dr. Nicholas Ritchie for providing feedback on this guide.

COSIG
Collection of Open
Science Integrity Guides

PubPeer Commenting - Best Practices

Last updated: 5 Februrary 2025

PubPeer is a post-publication peer review site where users can comment on any scientific publication. It is presently the foremost forum for post-publication peer review. To date, more than 200,000 scientific publications have been commented upon on PubPeer.

PuPubPeer comments can be posted anonymously, pseudonymously or under your name. Comments should be polite, neutral and should contribute meaningfully to scientific discussion. PubPeer employs a team of moderators that will edit or remove comments that do not meet these guidelines. For more information, read the PubPeer FAQ and Terms of Service.

This guide covers general tips on how to write a high quality, effective PubPeer comment.

Keep comments professional, direct, relevant and substantive

PubPeer is a forum for facilitating scientific discourse. Comments should adopt a polite and neutral tone. Users should write with the goal of engaging readers and authors of a publication in discussion, not with the goal of airing their personal opinions or discussing matters that are not relevant to the publication at hand.

Examples of helpful comments

> *I see several issue with the analysis presenting in this article, which I elaborate upon below.*
>
> *We discussed this article in journal club and found the authors' investigation very thorough. Have the authors considered whether their method can be used for creating graphene-based catalysts?*
>
> *Readers should become aware of a recent preprint by Jeannie Lee's group that used CLAP data to re-inforce the idea that PRC2 is an RNA binding protein. This preprint is entitled "Re-analysis of CLAP data affirms PRC2 as an RNA binding protein" and can be found at: https://www.biorxiv.org/content/10.1101/2024.09.19.613009v1*

Examples of unhelpful comments

> *Great paper!*
>
> *This paper SUCKS! The authors should be ashamed of themselves.*

You are an imbecile extraordinaire. I will not dignify your comment with a response.

Hahahahahaha

This article reminds me of the time I caught the ferry over to Shelbyville. I needed a new heel for my shoe. So I decided to go to Morganville, which is what they called Shelbyville in those days. So I tied an onion to my belt, which was the style at the time. Now, to take the ferry cost a nickel...

Cite your sources and show your work

As stated on the PubPeer FAQ, "the most important rule for commenting is to base your statements on publicly verifiable information". For effective comments, one should always clearly and thoroughly explain their reasoning and link to sources and supporting documents. Note that PubPeer comments can be styled with Markdown, and thus allow for hyperlinking. For instance, a PubPeer comment written as

```
Check out this [pre-print](doi.org/10.48550/arXiv.2107.06751)!
```

will appear on the site as

Check out this pre-print!

Examples of helpful comments

The cell line EC-9706 is contaminated by HeLa and is likely to be a poor model of esophageal cancer. See "Genetic profiling reveals an alarming rate of cross-contamination among human cell lines used in China".

The biological relevant molecular weight of Nrf2 is 95-110 kDa. However, in this paper they are studying an irrelevant protein at 68 kDa using Western immunoblotting. Please see Donna Zhang's article from 2013 for more information: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3503463/

*Figure 7F: Authors source data match data points plotted in the Figure, but statistics appear different. "n = 3 Fire+/+ mice and 4 Fire Δ/Δ mice. <0.3 μm, *P = 0.0417, two-way ANOVA with Sidak's multiple comparisons test" Inputing the authors source data from the website: two-way ANOVA (genotype): F (1,25)=2.905. P=0.1007 (ns) two-way ANOVA (genotype/diameter interaction): F (1,25)=2.905. P=0.4333 (ns) Sidak's multiple comparisons test (even though not valid as two-way ANOVA is ns): <0.3 (+/+ vs. Δ/Δ): t=2.422, DF=25, p=0.1099*

Examples of unhelpful comments

To achieve 16% f>m, 39% f=m, and 45% m>f, the underlying normal distributions of f and m have a difference of d=0.63 [sd units]. [Moderator: you should show your working here.]

Smith et al. covered a similar topic.

Include images and illustrate your observations

As a part of “showing your work”, it is helpful to readers to include pictures and illustrations in your comment. If you are leaving a comment about a particular figure, include that figure in the comment. If your observations about the figure are not immediately apparent, annotate that figure in a software like Microsoft PowerPoint, Adobe Illustrator, GIMP, Inkscape or Microsoft Paint. Note that externally-hosted images are backed up by PubPeer and will remain a part of the comment in perpetuity.

Examples of helpful comments

The SEM image shown in Figure 2B contains many apparent duplications. I have highlighted unexpectedly similar regions with colored boxes in the image below (brightness increased for ease of viewing).

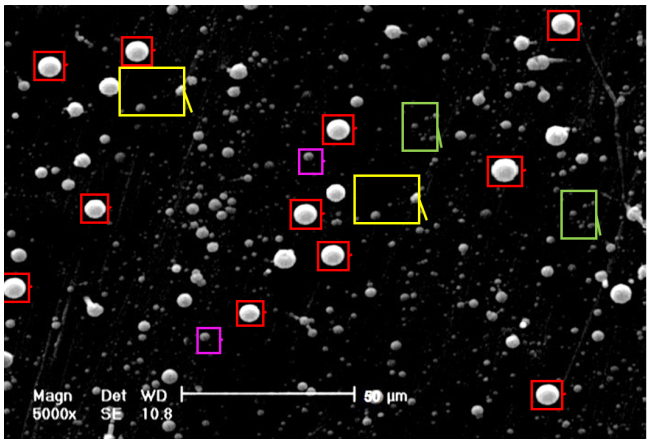
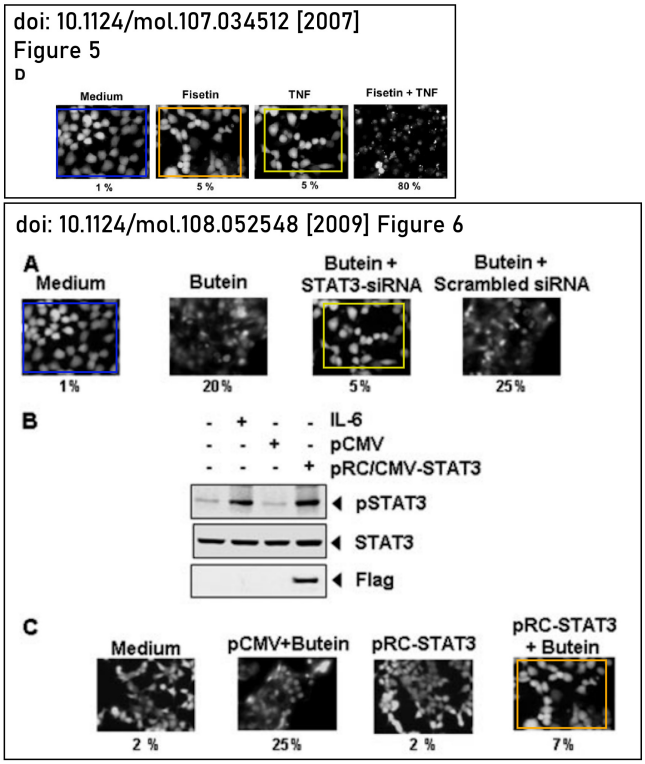


Figure 6: Some of the images previously appeared elsewhere. Identified by Imagetwin.ai.



Examples of unhelpful comments

The nanoparticles in Figure 8 look a bit wonky.

Some of the lanes SDS-PAGE gel shown in Figure 4 look very similar to one another.

Do not speculate on researcher motivations

Even when there is copious evidence that some fabrication, falsification or plagiarism occurred in a publication, it is not helpful to allege research misconduct. From the PubPeer FAQ:

> *Allegations of misconduct are forbidden on PubPeer. They are anyway unnecessary. Your audience on the site is mostly composed of highly intelligent researchers and scientists. They are quite capable of drawing their own conclusions if the facts are clearly presented.*
>
> *You should also avoid personal comments about authors and speculation about researcher actions and motives. This is non-scientific, but also can pose legal problems.*

Beyond this, it is often not possible to know exactly what took place in a publication's preparation and who was responsible for the improprieties observed in PubPeer comments.

Make specific, actionable requests from authors

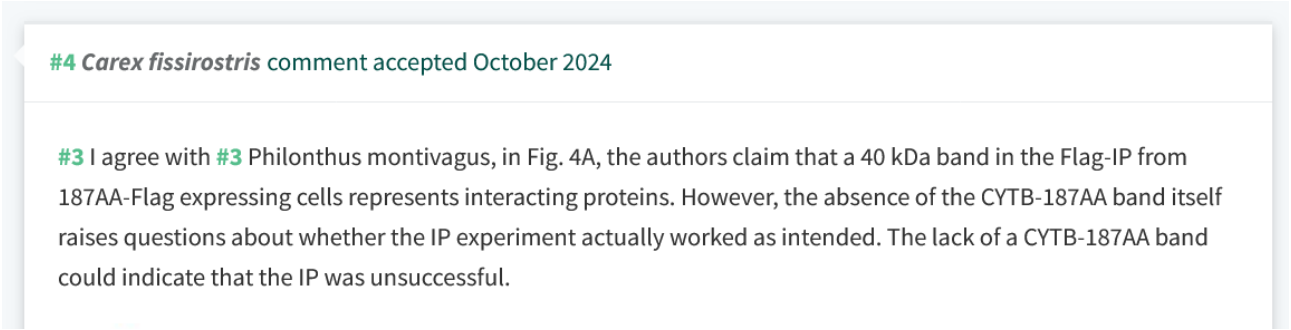
Authors are the best experts on their own publications and PubPeer has the option to provide author emails to loop them into a discussion. Questions about a publication can often be addressed directly by authors, so it is helpful to make clarifying requests where appropriate.

> *Could the authors provide the original scanning electron microscope images shown in Figure 3?*
>
> *Could the authors provide the raw data for the EDX spectrum shown in Figure 4?*
>
> *Could the authors clarify how they estimated crystallite size?*
>
> *Could the authors clarify what "enormous information" means and why it was used instead of "big data"?*

Split multiple observations into multiple comments

PubPeer comments can be referenced from other PubPeer comments on the same article by referring to the comment number preceded by "#".

If you have many observations on different aspects of an article, consider splitting these observations into multiple comments so that other commenters can reference your specific observations one at a time.



This PubPeer comment refers back to a previous PubPeer comment using #. The comment number for this comment can be found in the upper left corner of the comment box, to the left of the commenter's alias.

COSIG
Collection of Open
Science Integrity Guides

Standard deviation versus standard error

Last updated: 6 February 2025

Standard deviation

Standard deviation (SD) is a statistical measure that describes the variability of numerical observations of a variable around the mean of these observations.

When an entire population is measured, the population standard deviation is calculated σ is calculated as

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N}}$$

where x_i is the value of each observation, $\bar{x}$ is the mean of each observation ($\bar{x} = \sum x_i/N$) and N is the number of observations (i.e., the sample size). A lower standard deviation indicates that the data is more closely clustered around its mean, whereas a higher standard deviation indicates that the data is more spread out from its mean.

Note that standard deviation is calculated differently depending on whether you are measuring standard deviation for a population versus just a sample from that population. When only a sample is taken from a population, the sample standard deviation s is calculated as

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{N - 1}}.$$

See this explanation by Dr. Paul Savory for why this correction is made. Most research does not survey a whole population, and thus “standard deviation" usually denotes sample standard deviation s and not population standard deviation σ.

Imagine that we took a sample of ten individuals and measured their height in centimeters, obtaining ten observations: 176, 171, 159, 168, 185, 193, 174, 171, 168 and 189. The mean of this sample would be 175.4 and the sample standard deviation would be approximately 10.6.

Standard error

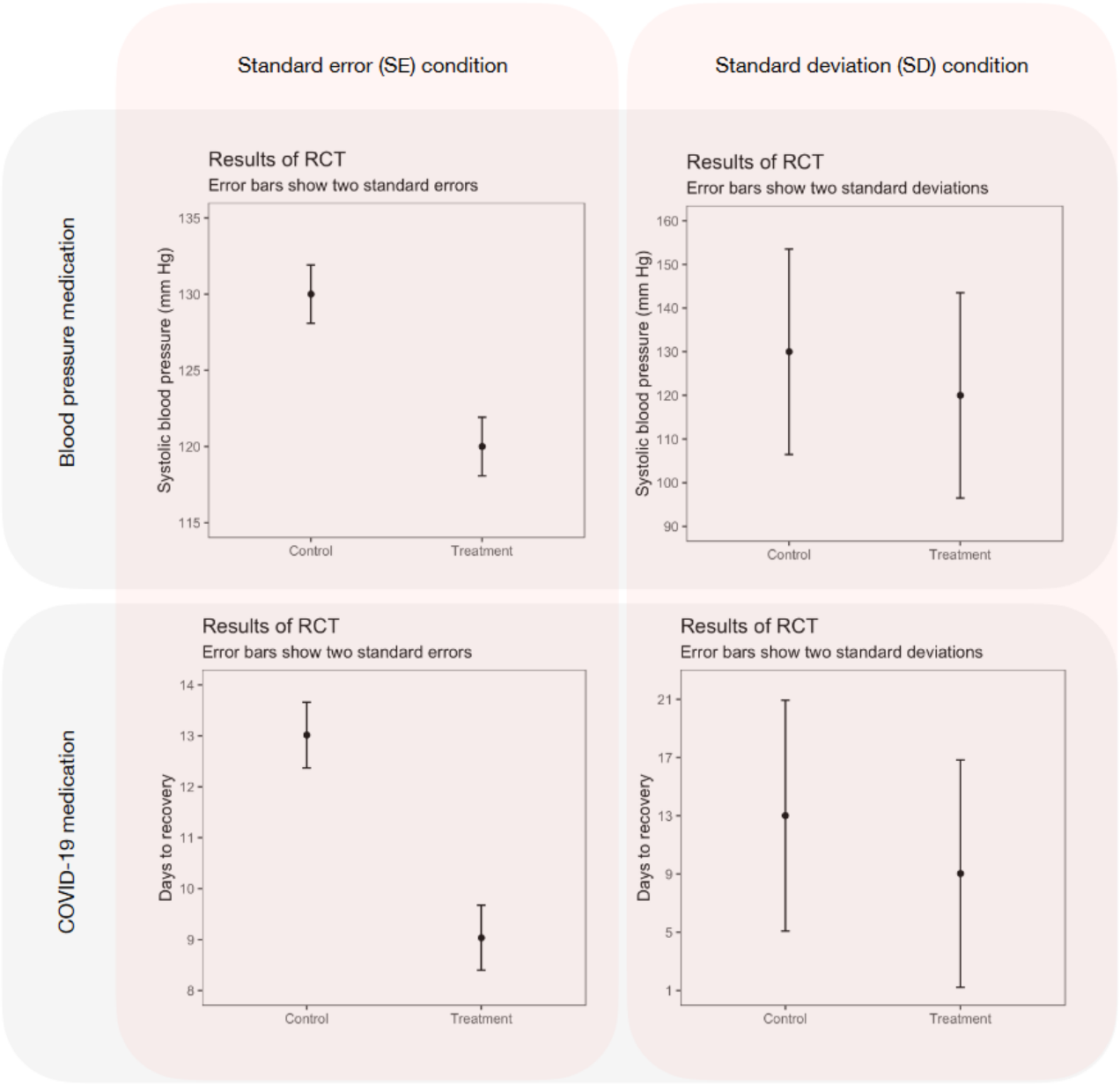
Standard error (SE) is a measure of the variability of an estimate. For instance, for the previous scenario, a different sample of ten individuals would likely yield a slightly different mean height. The standard error SE is calculated as

$$SE = \frac{s}{\sqrt{N}}.$$

Note that as sample size N increases, the standard error decreases. For instance, the mean heights calculated from two different samples of 100 from the same population would likely be closer together than the mean heights calculated from two different samples of only 10. This is why studies with a larger sample size are generally considered more rigorous; their estimates will be more precise. For the previously-listed sample of heights, the standard error of the mean (SEM) would be $10.6/\sqrt{10} \approx 3.3$.

Reporting of standard deviation versus standard error

When reading a scientific publication, one should take note of whether the authors are reporting variability in their data as standard deviation or standard error. Misidentifying which measure is being reported can give the reader an unrealistic picture of the variability in data and the uncertainty of estimates. There is some evidence that using one measure or the other graphically can lead to this same misinterpretation, even when the reader is aware of which measure is being used (see Zhang et al., 2023).



Within each row, both charts demonstrate the same hypothetical data from a randomized controlled trial (for blood pressure medication, top row, or COVID-19 medication, bottom row). However, the left charts use error bars to display standard error of the mean and the right charts use error bars to display standard deviation. Adapted from Figure S1 of Zhang et al. (2023).

Incorporating variability into a meta-analysis

A meta-analysis is a study that synthesizes the results of independent studies on the same topic to obtain more precise estimates. For instance, one might perform a meta-analysis of randomized controlled trials that all tested the same drug to more precisely determine how effective the drug is at preventing intensive care unit admission or mortality due to COVID-19.

Because larger studies give more precise estimates, they are usually given more weight in meta-analyses so that their findings contribute more to the combined outcome estimate. The most common method for weighting studies for meta-analysis is inverse variance weighting. With inverse variance weighting, each study is weighted by

$$\frac{1}{SE^2}$$

where SE is the standard error of that study's estimate. Thus, studies that yield a smaller standard error have a greater weight. Software for performing meta-analyses like RevMan typical allow the user to enter estimates from included studies alongside the reported standard deviation or standard error in those estimates, automatically performing weighting.

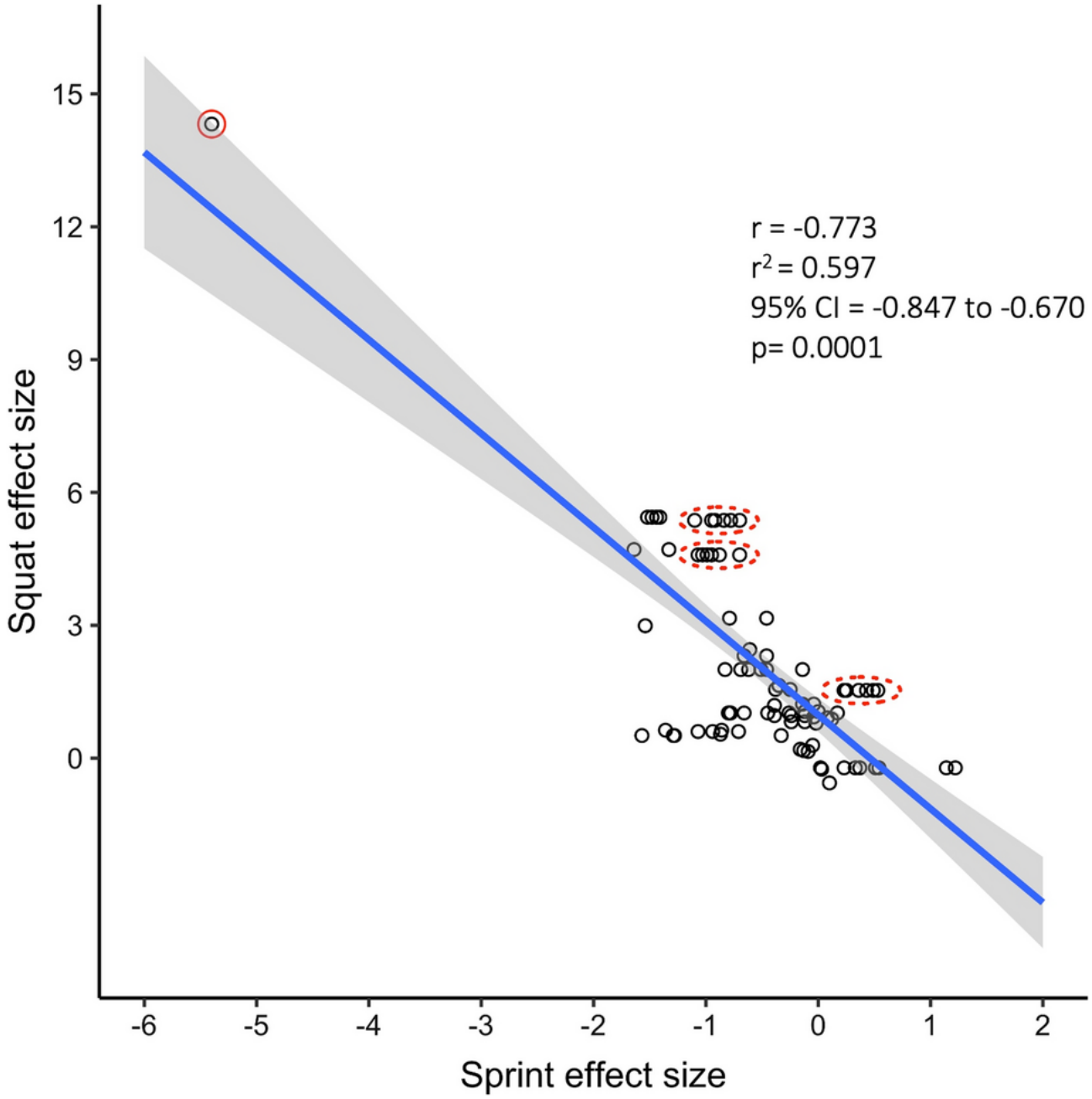
It is critical that those conducting a meta-analysis are cognizant of exactly which measure they are entering into the software. For instance, if the meta-analysis software expects the user to enter means, sample sizes and standard deviations, but the user misreads one study and enters the standard error reported by that study into the software instead of the standard deviation, that study will be weighted much more than it should be in the meta-analysis.

Popular methods for estimating the effect size of studies measuring continuous variables, such as standardized mean difference (SMD) / Hedges' g, will also yield inaccurate results if standard errors are confused for standard deviations.

Confusion of standard errors with standard deviations can often be spotted in a meta-analysis just by looking for outliers. Consider looking closer if one included study appears to have a much larger effect size than the others included in the meta-analysis or if one included study appears to have much smaller standard deviation in outcome than the other included studies.

Example 1: Overestimation of effect size in a meta-analysis due to using standard error instead of standard deviation

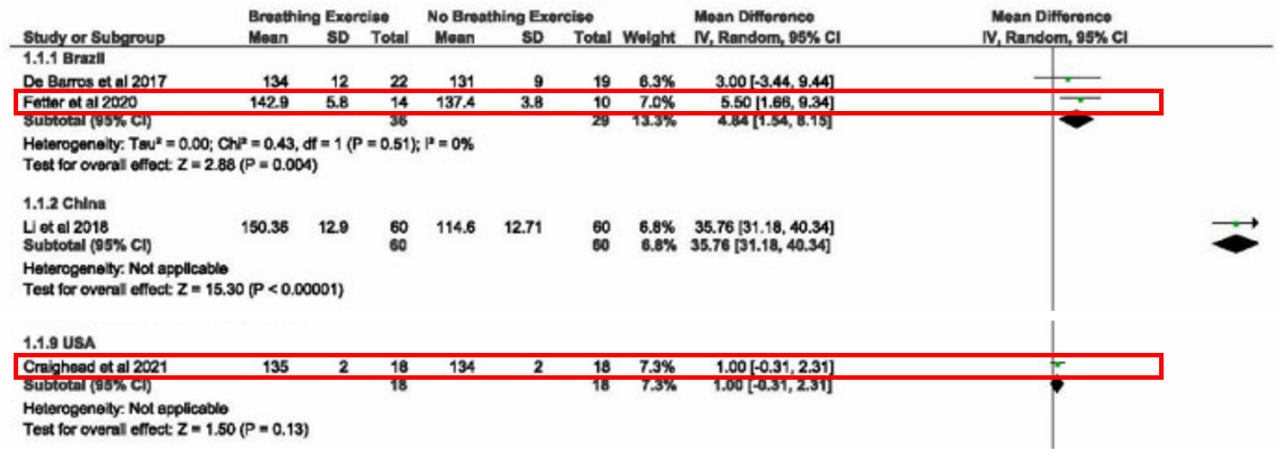
Seitz et al. (2014) performed a meta-analysis to determine whether exercises that increase lower body strength also improve sprinting performance. However, for three of their included studies, they calculated the effect size (in the form of Hedge's g) using standard errors instead of standard deviations. As a result, the effect sizes were over-estimated for these studies, leading to an over-estimation of the overall effect of lower body strengthening on improvement in sprinting performance. The effects of this over-estimation were discussed in detail by Kadlec et al. (2022).



The most severe over-estimation of effect size made by Seitz et al. (2014) was for Wong et al. (2010) (solid red circle in the upper right corner of this plot). Kadlec et al. (2022) corrected the statistical errors made by Seitz et al., finding that Seitz et al. had dramatically overestimated the correlation between lower body strength training and improvement in sprinting performance as a result of these errors. Adapted from Figure 1 of Kadlec et al.

Example 2: Over-weighting studies in a meta-analysis due to using standard error instead of standard deviation

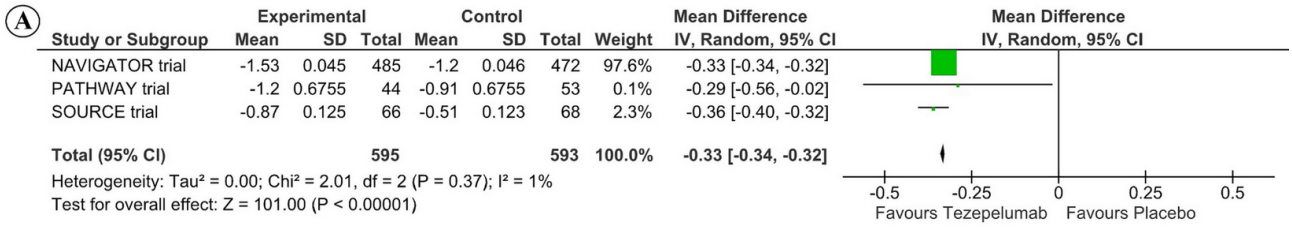
Garg et al. (2024) performed a meta-analysis to assess whether the systolic blood pressure was different for groups performing breathing exercises versus not performing breathing exercises at baseline (i.e. before any intervention was applied; note that there is no logical reason to perform this comparison, as it gives no information on the effectiveness of the intervention, but this meta-analysis has numerous issues that are beyond the scope of this guide). For at least two studies, Fetter et al. (2020) and Craighead et al. (2021), the authors include the standard error as reported by these studies instead of the standard deviation, resulting in these studies being weighted far too heavily.



Garg et al. (2024) overweighted Fetter et al. (2020) and Craighead et al. (2021) (highlighted in their forest plot in red boxes) as a result of confusing standard error for standard deviation. As a result, Fetter et al. and Craighead et al. are each weighted greater than Li et al. (2018) which had three to four times as many participants. Adapted from Figure 2 of Garg et al.

Example 3: Over-weighting studies in a meta-analysis due to using standard error instead of standard deviation

Chagas et al. (2023) performed a meta-analysis to determine whether treatment with the drug tezepelumab reduced asthma patients' scores on the Asthma Control Questionnaire-6 (ACQ-6). For the NAVIGATOR trial, the authors entered the reported standard error as the standard deviation, leading to this trial being weighted to 97.6% for this outcome (meaning that this portion of the meta-analysis was almost entirely based on the results of the NAVIGATOR trial).



Chagas et al. (2023) overweighted the NAVIGATOR trial as a result of confusing standard error for standard deviation. Note that the "standard deviation" reported in the above forest plot is much lower for the NAVIGATOR trial than the PATHWAY trial or the SOURCE trial. Adapted from Figure 2 of Chagas et al.

Additional resources

- "Points of significance: Error bars" (2013)
- "The Standard Error/Standard Deviation Mix-Up: Potential Impacts on Meta-Analyses in Sports Medicine" (2024)
- "With Great Power Comes Great Responsibility: Common Errors in Meta-Analyses and Meta-Regressions in Strength & Conditioning Research" (2022)
- *The Cochrane Handbook for Systematic Reviews of Interventions*
- *The Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy*

COSIG
Collection of Open
Science Integrity Guides

Software for image forensics

Last updated: 7 February 2025

Publication integrity issues often arise from image integrity issues. Performing image forensics is often very useful for taking part in post-publication peer review, especially in fields that frequently use images in scientific articles (e.g., biomedicine).

While many image integrity issues can be spotted by eye, it is often easier and more efficient to use software to automatically spot issues or make issues more visible to readers. This guide is a catalog of tools and software that are commonly used by sleuths.

Here are some factors to consider when using these tools:

- **There can be false positives.** A tool may flag image features that, upon closer manual examination, do not actually indicate any image integrity issues.
- **There can be false negatives.** A tool may not flag image features that are indicative of image integrity issues. For instance, software for detecting image duplications within a figure may miss some duplications that are visible by eye.
- **Tools may not examine all data types.** There can be entire categories of data (including image data) that have been explicitly excluded by the programmers of a tool because they are not yet comfortable with the sensitivity/specificity of their tool for that data type.
- **Analysis may not be reproducible.** Some of these services and software are updated frequently and the current version of a tool may not yield the same results as a previous version.
- **Analysis of the same images in a different format may yield different results.** For example, if one uploads an entire PDF to a duplication detection tool, the tool may yield different results than if individual images are uploaded, even if the individual images and those in the PDF appear identical by eye. Figures in published articles will often be available in multiple resolutions, which can yield differing results. When performing image forensics, it is always preferable to work with the original, full-resolution, uncropped images provided by a study's authors.
- **Tweaking parameters can yield differing results.** Some tools have sensitivity settings that can be changed by the user. Changing these setting may produce different results for the same images.
- **Different tools do different things.** Not every tool described here has the same functionality or use cases as another tool. Another person without access to your tool of choice may not be able to reproduce your analysis.

Software/tools commonly used for image forensics in post-publication peer review:

- **Adobe Photoshop (subscription-based)**. Photoshop is an image manipulation software that allows users to adjust color levels, adjust brightness and contrast, overlay images and annotate figures among myriad other features. The United States Department of Health and Human Service Office of Research Integrity provides some toolkits for image forensics with Photoshop.
- **GIMP (the GNU Image Manipulation Program) (free to use and open-source)**. GIMP is a image manipulation software with most of the same features and functionality as Photoshop but is free to use (and modify).
- **Imagetwin (subscription-based)**. Imagetwin is a browser-based service that allows users to upload article PDFs and individual images, which it will then compare against a large database of published images to see if parts of any of the uploaded images have been used previously. It also detects within-document image duplication and splicing of certain images (e.g., Western blots). Users can control the sensitivity of detection on the Results page of a scan.
- **Proofig (subscription-based)**. Proofig is a browser-based service that allows users to upload article PDFs and individual images, which it will then compare against a large database of published images to see if parts of any of the uploaded images have been used previously. It also detects within-document image duplication.
- **Sherloq (free to use and open-source)**. Sherloq is a software environment for image forensics that can be installed on Linux and Windows. Users can perform various image transformations that make manipulation more apparent (such as visualizing luminence gradient) as well as inspect image metadata.
- **Forensically (free to use)**. Forensically is a browser-based service that offers several tools for image forensics, such as clone detection and levels adjustment.
- **FotoForensics (free to use)**. FotoForensics is a browser-based service that offers several tools for image forensics, such as error level analysis and metadata inspection.
- **Figcheck (subscription-based, limited free use)**. Figcheck is a browser-based service that detects within-document image duplication. Each user is limited to uploading 10 images a day.
- **Image Duplication Check (Sholto David) (free to use)**. This is a browser-based application that allows the user to upload a PDF and scan for within-document image duplication.
- **Google Lens (free to use)**. Google Lens is an extension of the Google search engine that allows users to upload an image and finding matching and visually-similar images across the web.

COSIG
Collection of Open
Science Integrity Guides

Extracting vector graphics from a PDF

Last updated: 5 February 2025

It is often difficult to tell if two line graphs have shared features from a published figure alone. While the webpage of a scientific article invariable displays these images in raster format (i.e. pixel-based), the PDF version of the article may actually display these images in vector format, meaning that the features of the figure can be more easily magnified and compared.

An easy way to check if a figure in the PDF of a scientific article is in raster format or vector format is to try highlighting the text in the figure with your cursor (e.g. the text that labels the axes of the graph). If the text within the figure can be highlighted, the image is probably in vector format. Consider this article, for which the figures in the PDF are in vector format, and this article, for which the images in the PDF are in raster format.

Figure 1 of PDF version of
10.1007/s12182-020-00439-9
Text in figure *can* be highlighted
Figure is in **vector** format!

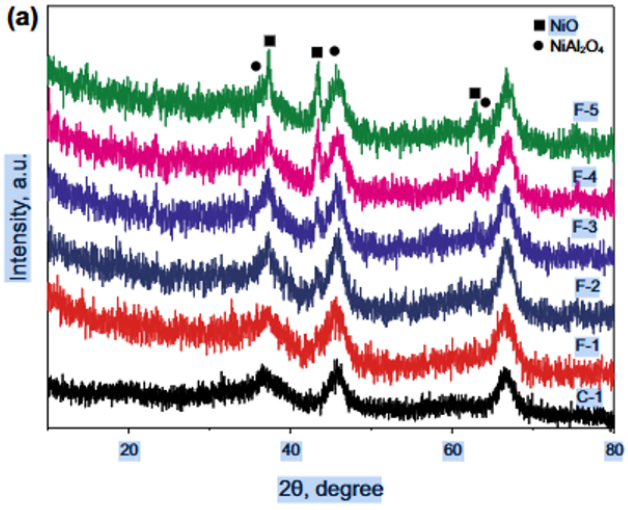
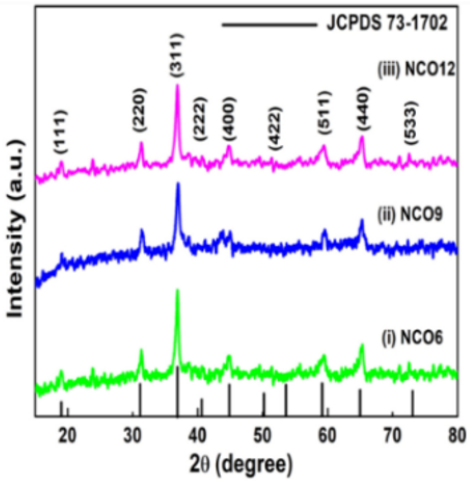


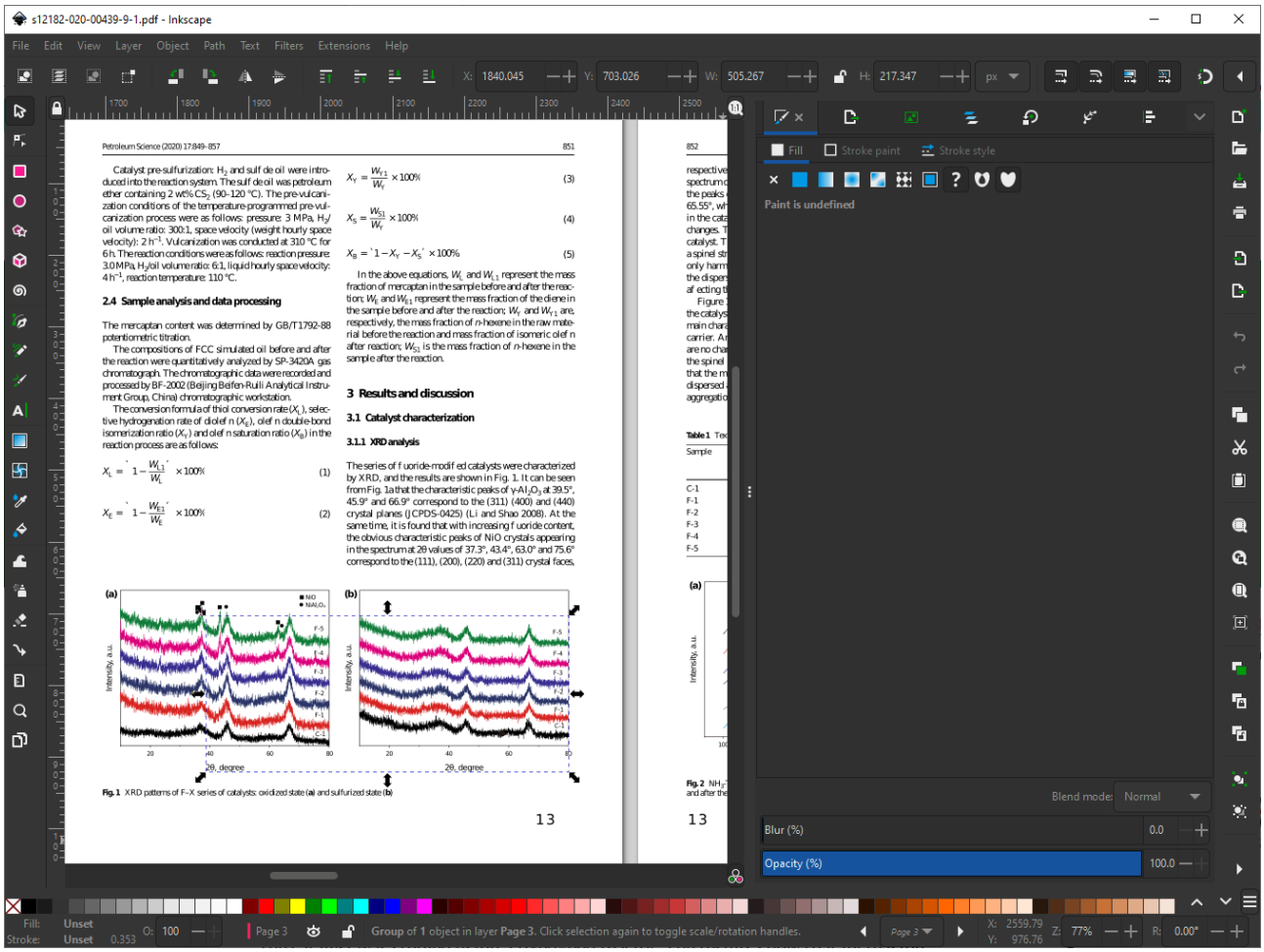
Figure 2 of PDF version of
10.1016/j.est.2023.109227
Text in figure *cannot* be highlighted
Figure is in **raster** format!



The figure on the left is embedded within its PDF in vector format, whereas the image on the right is embedded within its PDF in raster format.

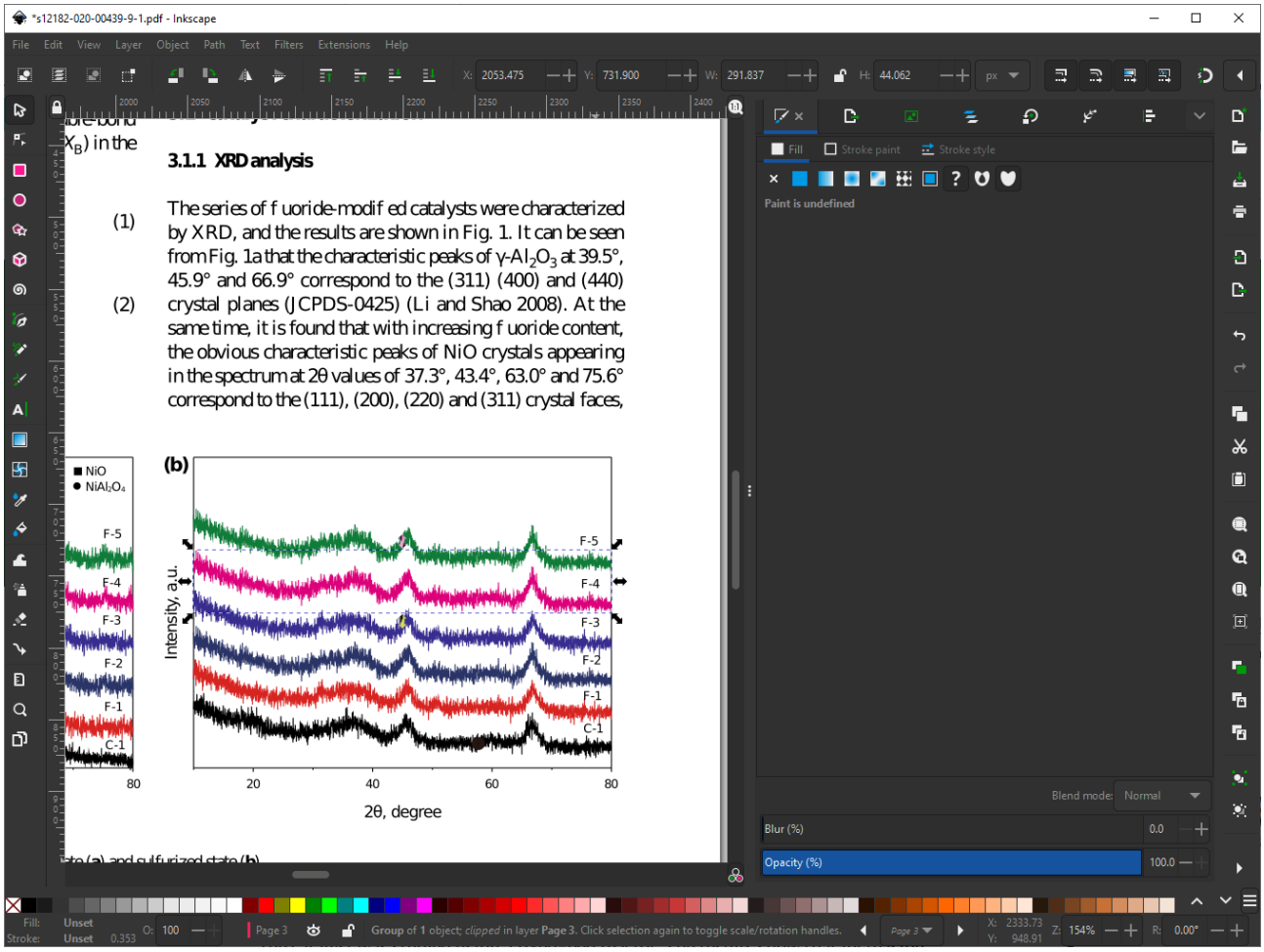
Vector graphics can be isolated in any vector graphics editor like Adobe Illustrator or its free, open-source alternative Inkscape. We'll use Inkscape for this demonstration but the procedure will be more or less the same in other vector graphics editors.

First, open Inkscape, navigate to **File > Open**, and select your PDF of interest. For this demonstration, we'll use Figure 1B in the PDF of this article. After the PDF opens, scroll over to your figure of interest and select your component of interest.



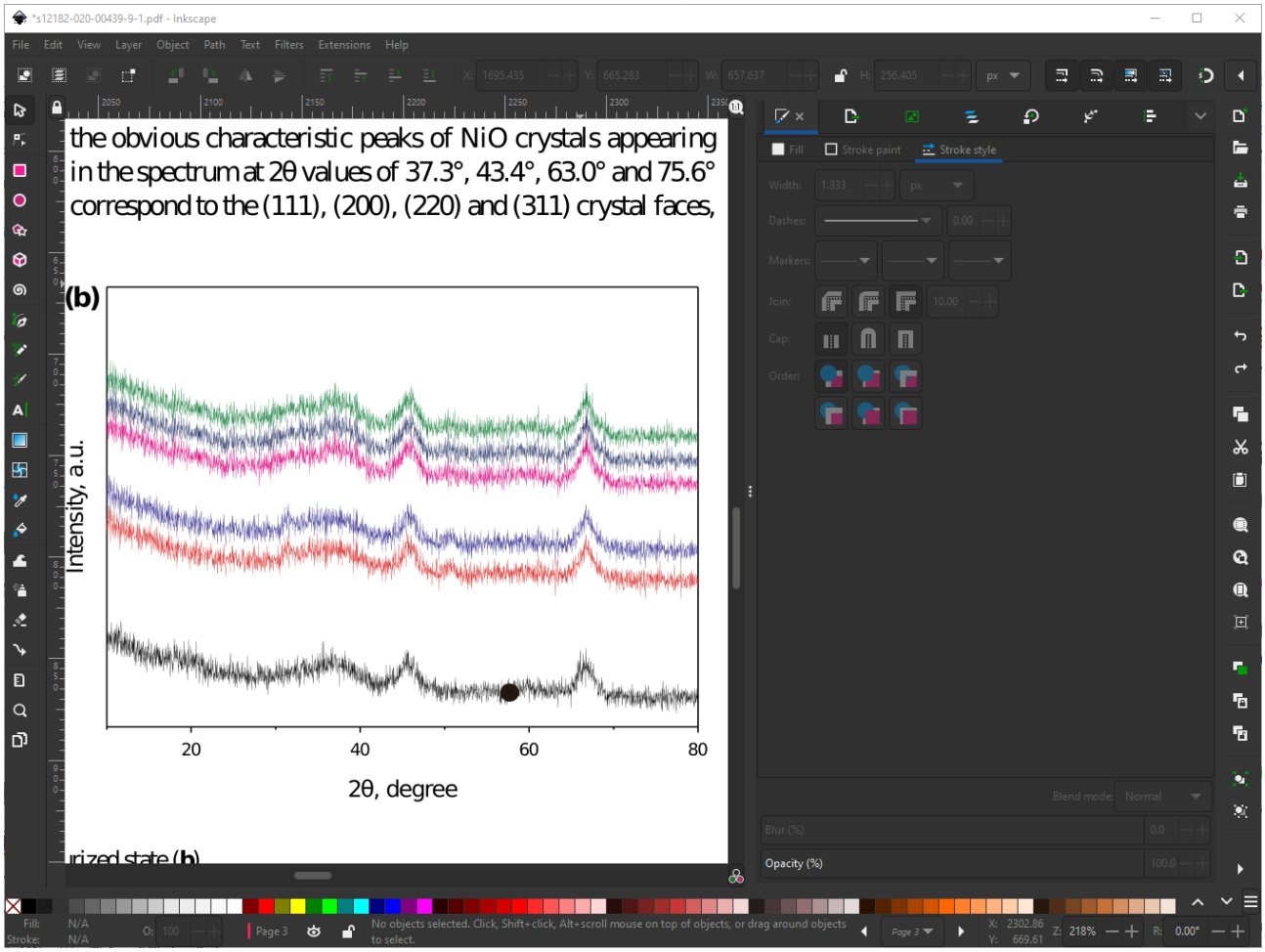
Our PDF opened in Inkscape with the traces in Figure 1B selected.

Next, repeatedly use **Right click > Ungroup (Shift + Ctrl + G)** on the selected objects until individual objects (in this case, individual line traces) can be selected.



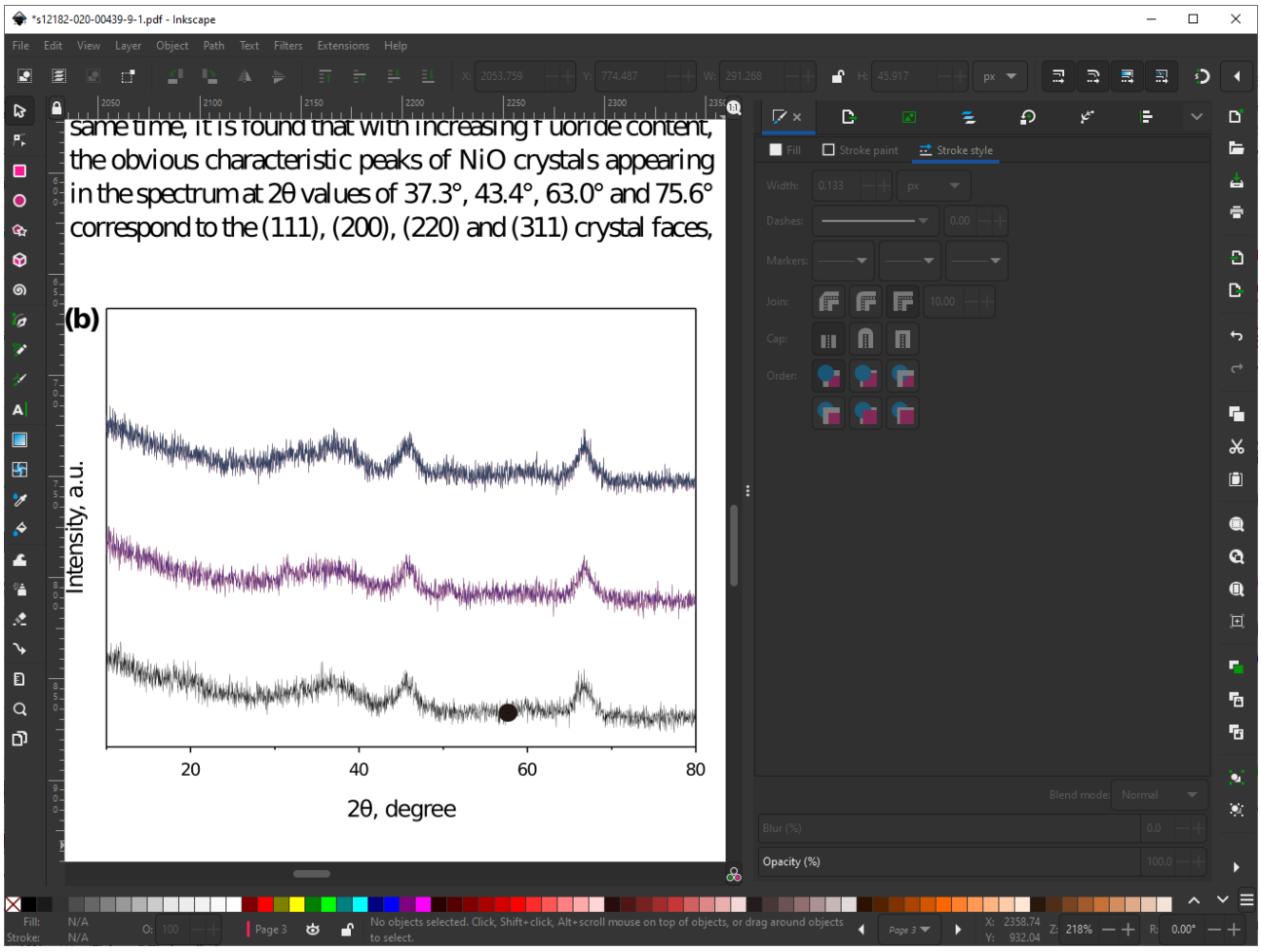
After repeatedly ungrouping objects, we can now select individual line traces.

When viewing line traces, it is helpful, but not necessary, to thin out the traces to enable comparison. To do so, navigate to **Objects > Fill and Stroke (Shift + Ctrl + F)**, under which the **Stroke Style** menu allows you to set line thickness. Afterwards, moving the traces to be closer to one another reveals that the F-2 (navy), F-4 (pink) and F-5 (green) traces are identical and the F-3 (blue) and F-1 (red) traces are identical.



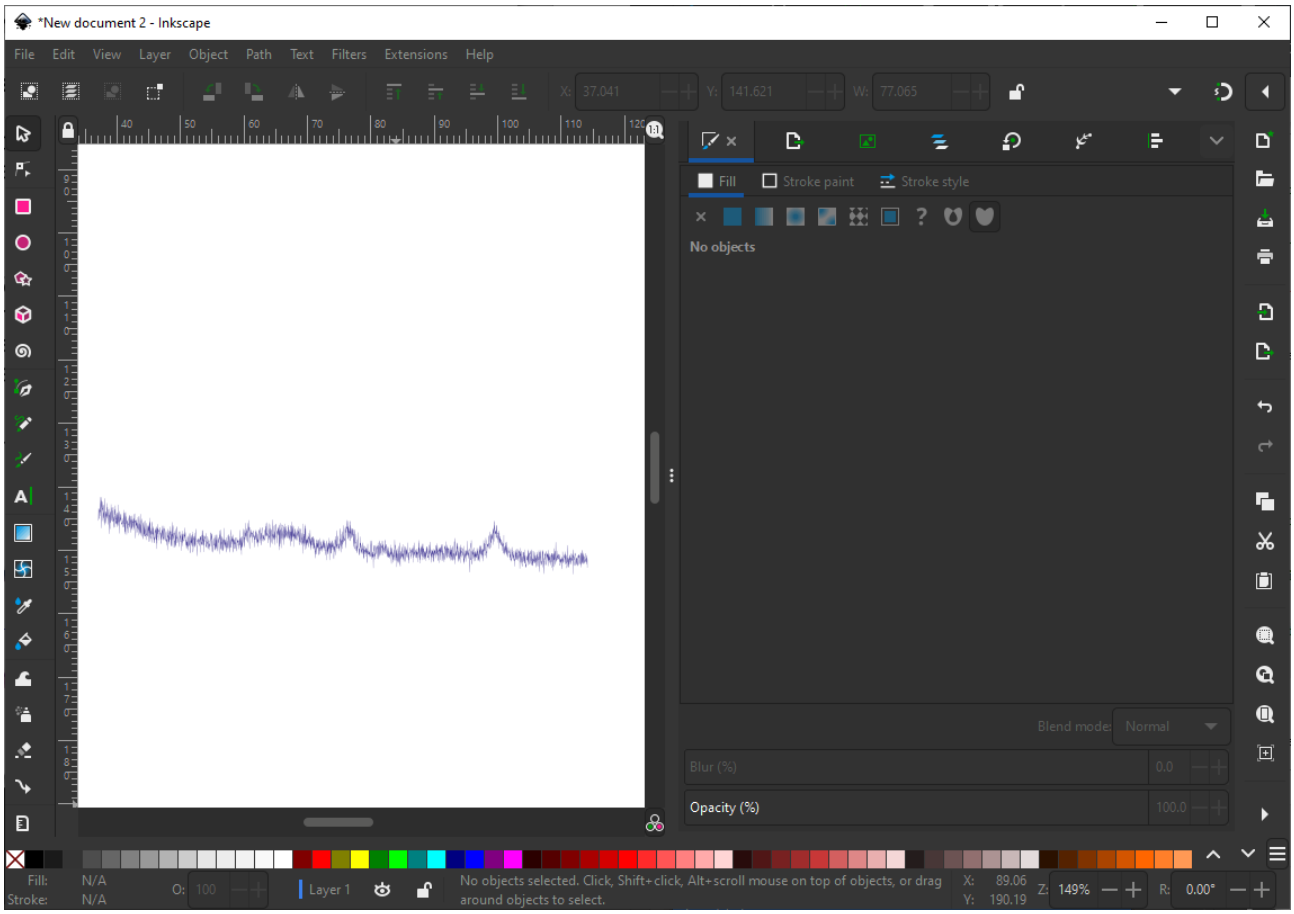
Moving around and thinning out traces in Inkscape allows for easier comparison.

We can also overlap the traces to make it crystal clear that these are the same data.



Now that traces are individual objects, we can easily overlap them to show that certain traces are identical.

Objects can also be copy-pasted into a new document.



An individual trace copied into a new Inkscape document.

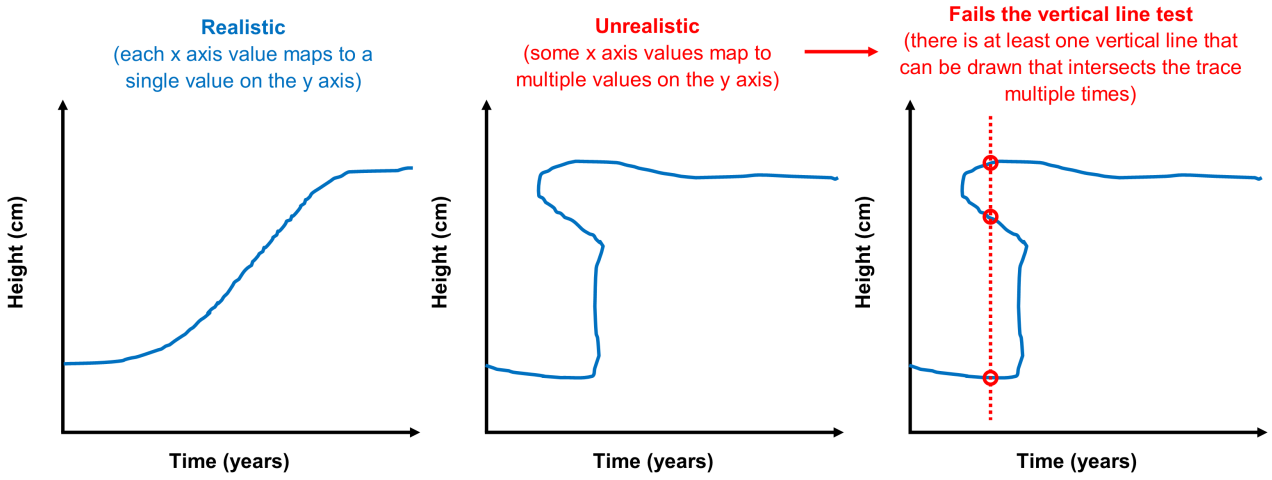
COSIG
Collection of Open
Science Integrity Guides

The vertical line test

Last updated: 7 February 2025

For many types of data plotted on a x/y plane, it is expected that each x-axis value maps uniquely to a y-axis value. For instance, if you plotted your height over time, you would expect that each x-axis value (a point in time) corresponds to a single y-axis value (your height at that time). If not, this plot would imply that there were some points in your life where you were two different heights at the same time.

This property defines a mathematical function, for which there is one unique output (y) for every unique input (x). The vertical line test is a simple method for gauging if a trace/curve is a function or not: given a trace/curve on an x/y plane, can you draw a vertical line that intersects the trace/curve multiple times? If there is at least one such vertical line, the trace/curve is not a mathematical function and may not make sense for the type of data it is describing.



If you plotted your height over time, you would expect each point in time (i.e., each point on the x axis) to correspond to a single height (i.e., a single point on the y axis). The graph on the left matches this expectation and is realistic for this kind of data. The graph in the middle does not match this expectation; there are some points in time that correspond to multiple heights. The graph on the right shows this trace failing the the vertical line test. This is just one of many vertical lines that could be drawn that show that this trace is not a function and thus does not realistically describe data representing height over time.

Data types that should always pass the vertical line test

Some data types, when plotted, will invariably be functions and thus should always pass the vertical line test. For instance, taking the absorption spectrum of a material should yield one value for absorption for every wavelength. Data types that are always expected to pass the vertical line test include, but are not limited to:

- any light absorption spectrum, including ultraviolet-visible (UV-VIS) absorption spectra, infrared (IR) absorption spectra, microwave absorption spectra, X-ray absorption spectra (XAS), etc. (*Note that essentially anything that gets called a "spectrum" should pass the vertical line test.*)
- nuclear magnetic resonance (NMR) spectra
- electron spin resonance (ESR)/electron paramagnetic resonance (ESR) spectra
- Fourier-transform infrared (FTIR) spectra
- Raman spectra
- X-ray diffraction (XRD) patterns/diffractograms
- energy-dispersive X-ray (EDX/EDS/EDAX) spectra
- X-ray photoelectron spectra (XPS)
- photoluminescence (PL) spectra
- Mass spectra (MS/mass spec)
- differential thermal analysis (DTA) curves
- differential scanning calorimetry (DSC) curves
- thermogravimetric analysis (TGA)curves
- photocurrent response/transient photocurrent (TPC) curves
- electroencephalograms (EEG)
- patch-clamp recordings

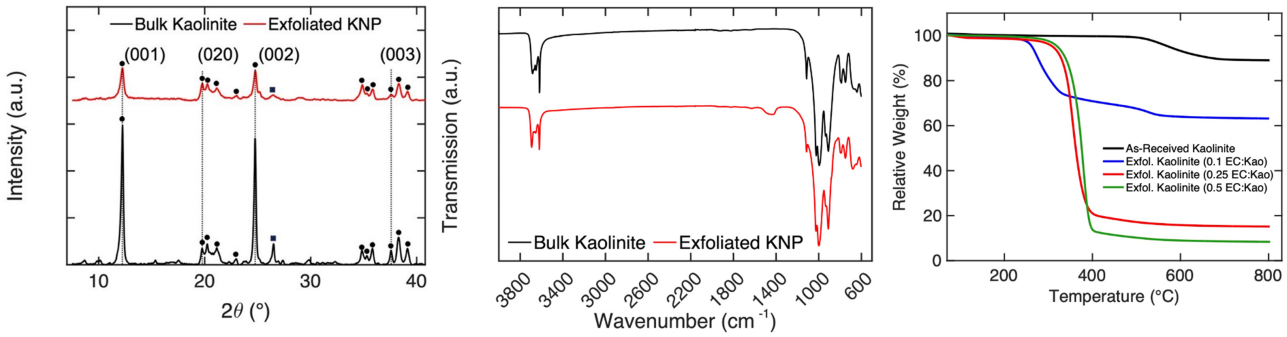
Data types that may not pass the vertical line test

For some data types, it is entirely expected that some x axis values will map to multiple y axis values. Data types that may or may not pass the vertical line test include, but are not limited to:

- cyclic voltammetry (CV) curves
- traces of an object moving in two dimensions, such as a mouse solving a Morris water navigation task
- Pressure-volume (PV) diagrams
- diagrams of mathematical graphs/networks

Example 1: Passes the vertical line test

Thomas et al. (2024) use X-ray diffraction (XRD), Fourier-transform infrared spectroscopy (FTIR) and thermogravimetric analysis (TGA), among other techniques, to characterize kaolinite and kaolinite nanoplatelets. Each of these techniques yield data with one value on the y axis for each value on the x axis. All the plots corresponding to these techniques provided by Thomas et al. pass the vertical line test, as expected.



Plots from Thomas et al. (2024) that pass the vertical line test (adapted from Figures 2B, S8 and S2).

Example 2: Does not pass the vertical line test

Mandizadeh et al. (2014) use FTIR to characterize barium hexaferrite nanostructures. However, the FTIR spectrum they show backtracks on itself several times, failing the vertical line test.

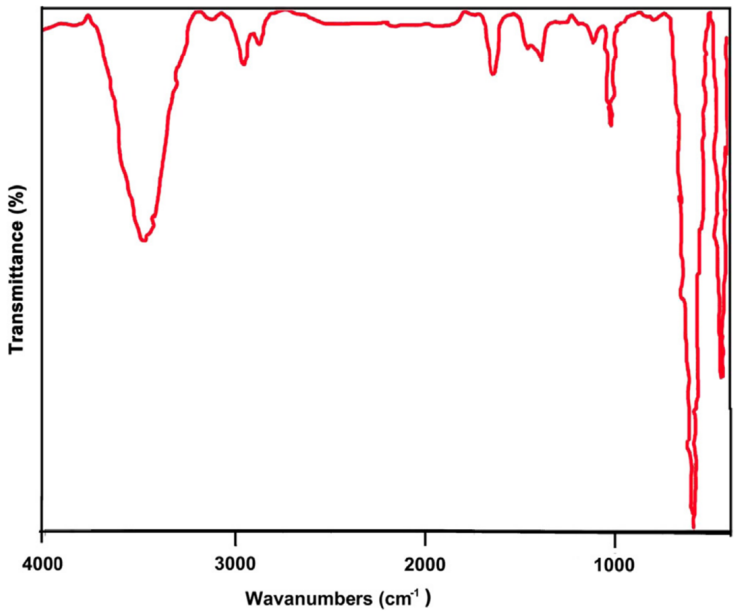
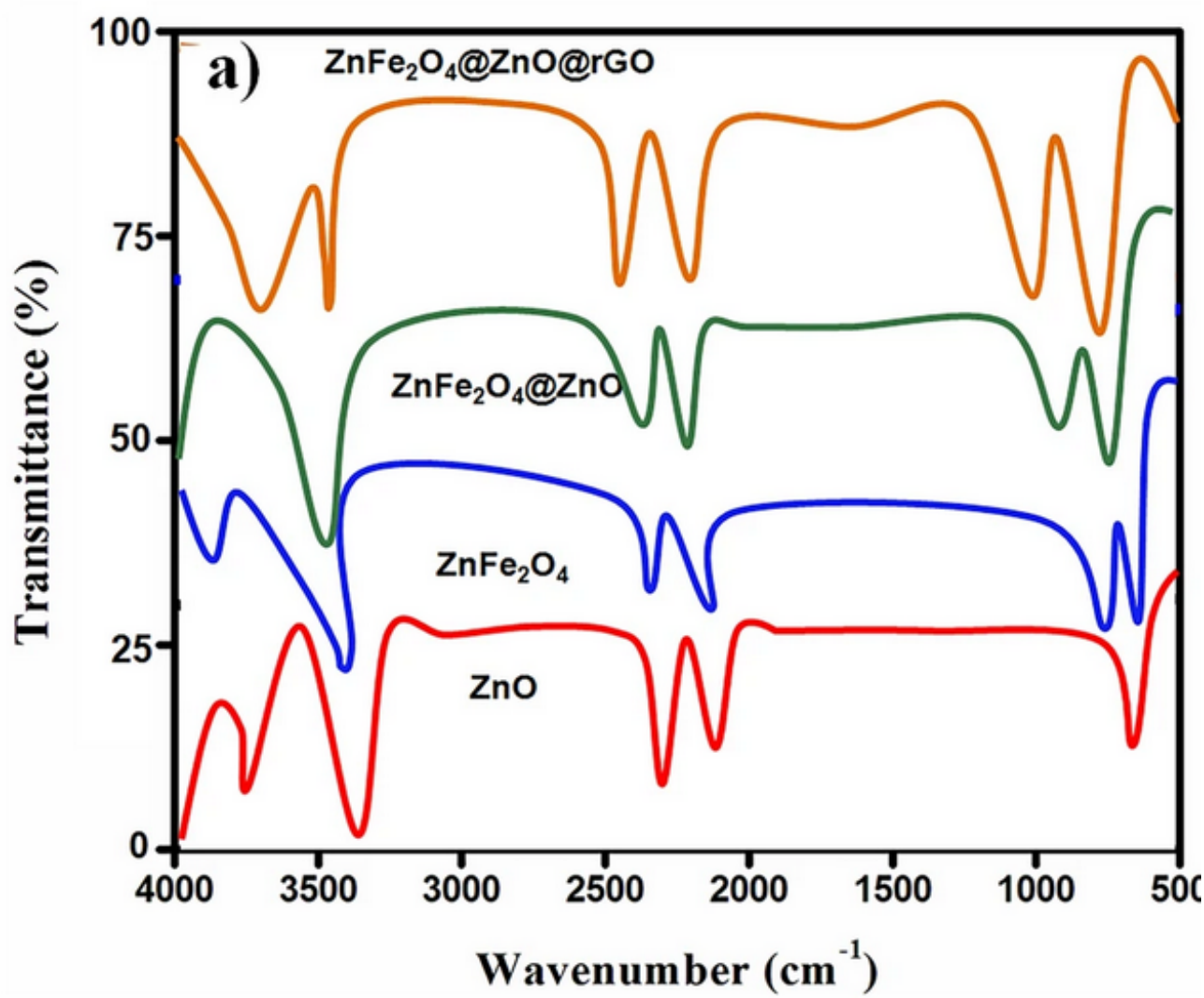


Fig. 1. FT-IR spectra of barium hexaferrite nanostructures.

An FTIR spectrum from Figure 1 of Mandizadeh et al. (2014) that fails the vertical line test.

Example 3: Does not pass the vertical line test

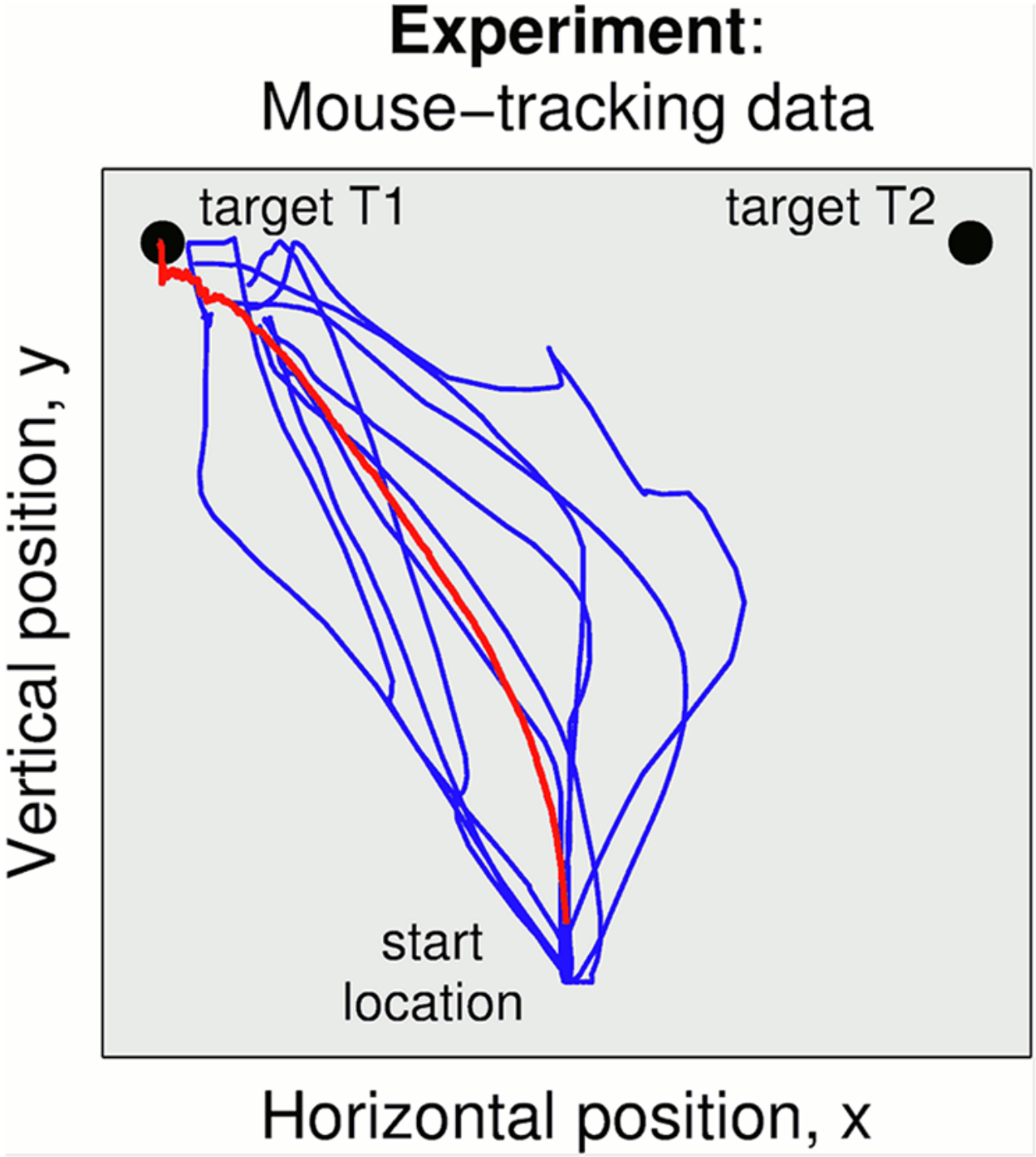
Suguna et al. (2024) use FTIR to characterize nanocomposite photocatalysts. However, the FTIR spectra they show appear to be hand-drawn and some fail the vertical line test.



FTIR spectra from Figure 3A Suguna et al. (2024). The $ZnFe_2O_4$ spectrum (in blue) fails the vertical line test.

Example 4: Vertical line test does not apply

Lepora and Pezzulo (2015) describe tracking mice in 2D space. They include several mouse trajectories in figures which follow the path of a mouse with a line. These trajectories are not expected to be functions and thus the vertical line test does not apply.



Mouse trajectory traces in 2D space, for which the vertical line test does not apply. Adapted from Figure 2 of Lepora and Pezzulo (2015).

COSIG
Collection of Open
Science Integrity Guides

X-ray diffraction patterns - Scherrer's equation

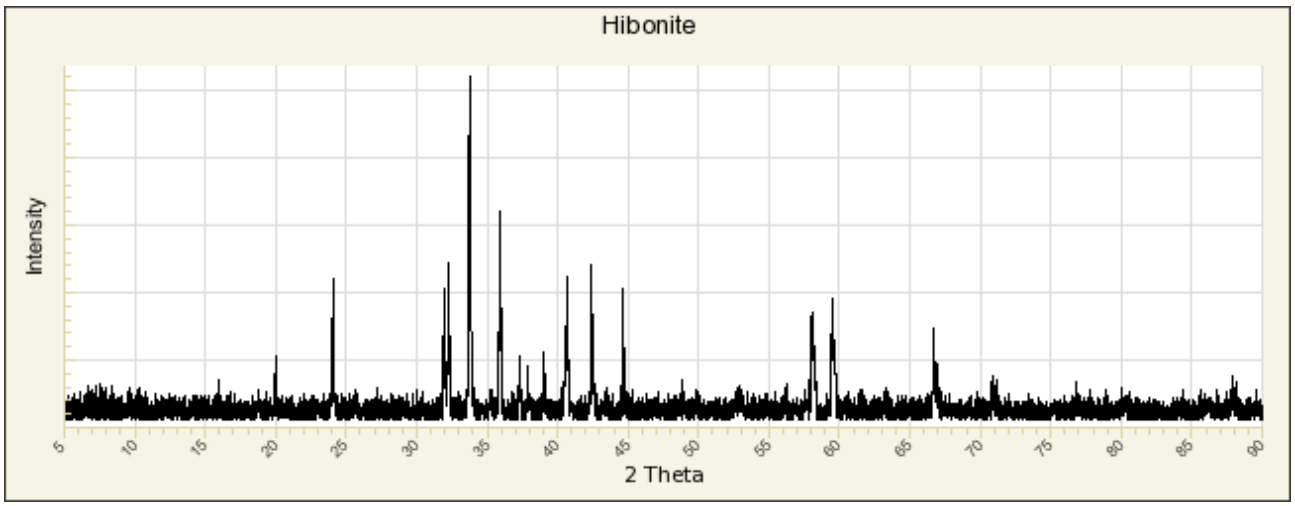
Last updated: 6 February 2025

X-ray diffraction patterns/diffractograms

X-ray diffraction (XRD) is a popular experimental technique used to characterize the nanoscale structure of materials like crystals, glasses and even liquids. XRD involves bombarding a sample with high-energy X-rays at a range of incident angles.

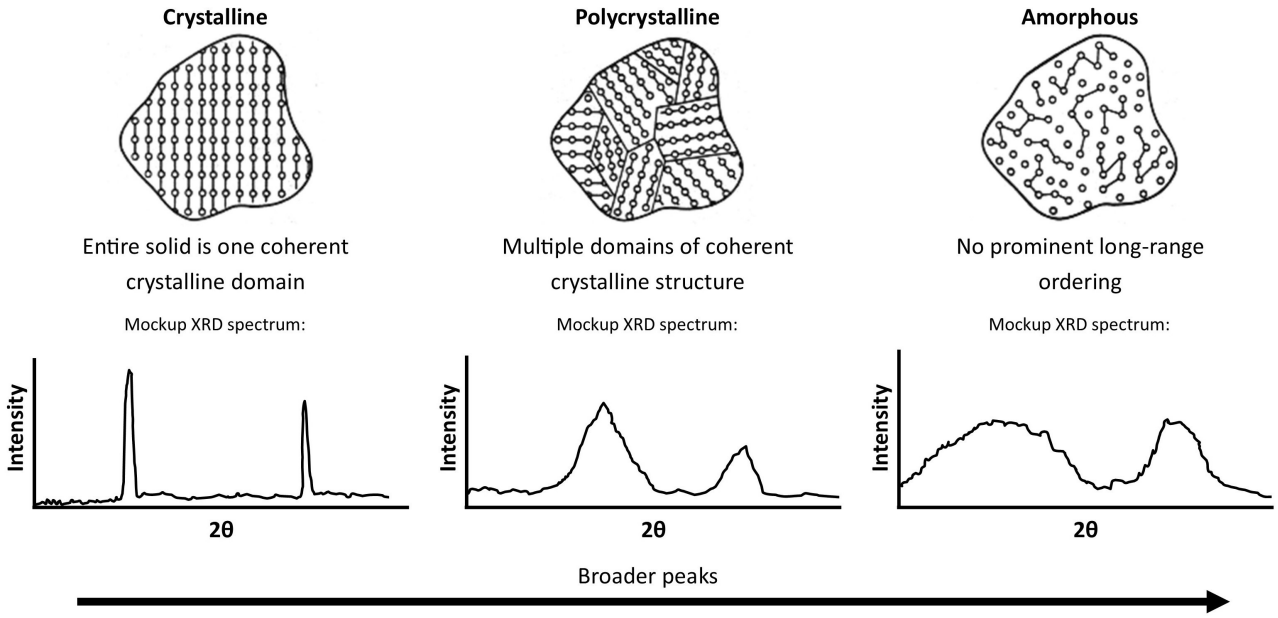
For a material with some amount of repetitive atomic structure, the reflected X-rays will have higher intensity at specific angles ("peaks") corresponding to the distance between successive layers of atoms.

A typical graph of an XRD pattern will have intensity of signal on the y axis and the Bragg angle θ (expressed as 2θ by convention) on the x axis. Peaks may be labeled with the Miller indices of the corresponding crystal plane.



An example XRD pattern for the mineral Hibonite.

Solids may be described as amorphous (having no discernable long-range ordering of atoms), polycrystalline (having long-range ordering, but only within small domains) or crystalline (being made up of one large crystalline domain). Generally, materials with a more ordered structure will yield XRD patterns with sharper peaks. Thus, crystalline materials will tend to yield XRD patterns with very sharp, needle-like peaks, polycrystalline materials will tend to yield XRD patterns with somewhat broader, mountain-like peaks and amorphous materials will tend to yield XRD patterns with very broad, hill-like peaks.



Materials with large crystalline domains will tend have very sharp peaks, whereas a material with no crystalline domains will tend to have very broad peaks. Note that these XRD patterns are hand-drawn to demonstrate peak broadening and do not correspond to any real material. Images adapted from lecture notes of Dr. Alan Doolittle.

Crystallites, grains and particles

Crystallite size describes the typical size of crystalline domains that scatter X-rays coherently (i.e., regions of the solid where the crystal lattice is all oriented the same way). Grain size has a similar definition, also referring to the typical size of regions of the solid where the crystal is oriented in the same direction. However, grain size is usually estimated from high resolution transmission electron microscopy images and crystallite size is usually estimated from XRD patterns. A grain can be made up of a single crystallite or multiple crystallites. 'Grain' and 'crystallite' are often used interchangeably.

$$\text{crystallite size} \leq \text{grain size} \leq \text{particle size}$$

Scherrer's equation

Scherrer's equation (sometimes called the Scherrer relation or the Debye-Scherrer equation, though not without controversy) is a formula for estimating the mean crystallite size within a material based on the width of peaks in its XRD pattern. For a mean crystallite size of D, Scherrer's equation is

$$D = \frac{K\lambda}{\beta \cos\theta}$$

where K is a dimensionless shape factor (usually taken as 0.89, 0.94 or 1.0, depending on the shape of the crystal), λ is the wavelength of the X-rays used (usually 1.5406 Å), β is the

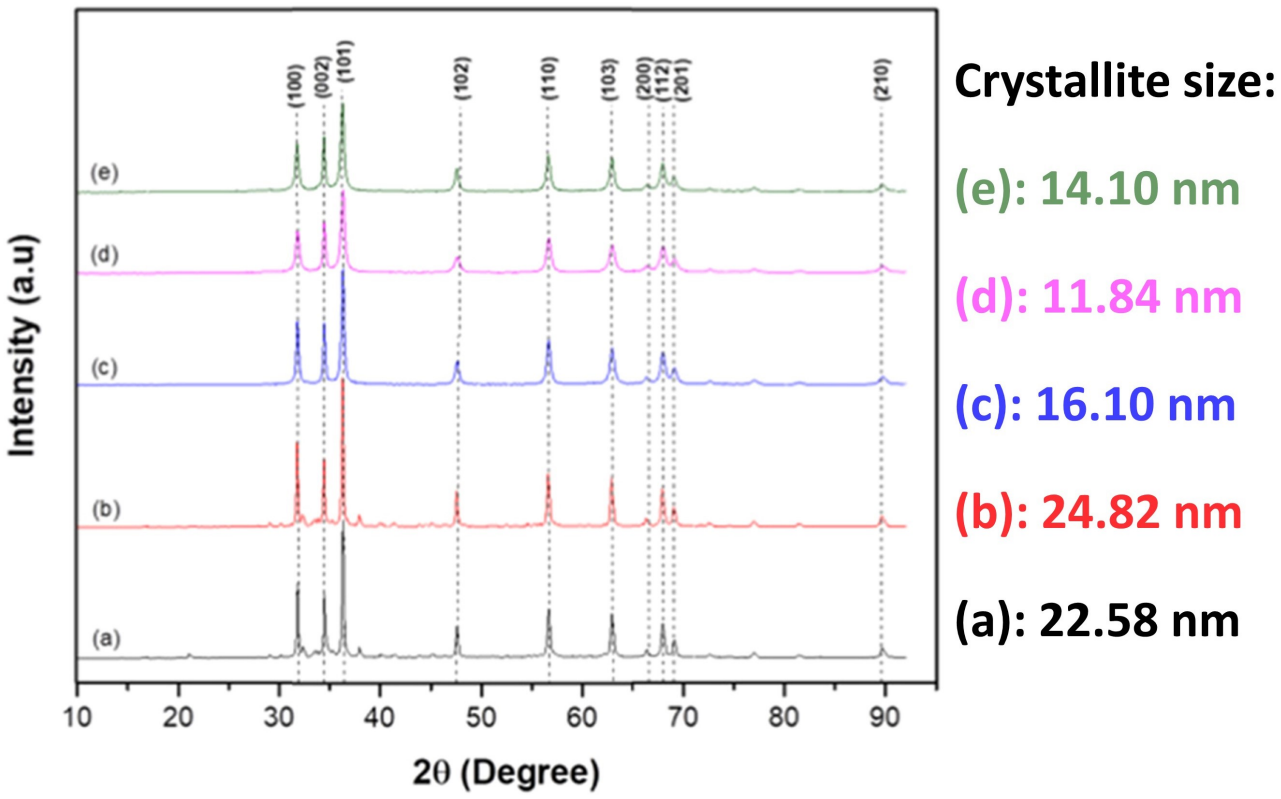
full width at half-maximum of the peak in being used (in radians) and θ is the Bragg angle (in radians). Usually D is estimated using only the tallest peak in the pattern. Full width at half maximum (FWHM) describes the width of the peak at the point halfway between the bottom of the peak and the top of the peak and is usually measured by fitting peaks with a Gaussian curve in software.

For assessing the integrity of crystallite size estimates made in an article using Scherrer's equation, there are three important details to note:

1. D is inversely proportional to β. In other words, as peaks become wider, crystallite sizes decrease; as peaks become narrower, crystallite sizes increase.
2. Scherrer's equation is used for estimating crystallite size, *not* particle size.
3. Crystallite size must always be smaller than or equal to particle size.

Example 1: No detectable integrity issues concerning Scherrer's equation

In Mustapha et al. (2019), the authors use Scherrer's equation to estimate crystallite size of ZnO nanoparticles synthesized at varying pH. The XRD patterns shown in Figure 2 are consistent with the the calculated crystallite sizes shown in Table 1. Observe that as peaks become wider from pattern (a) to pattern (e), crystallite size decreases. There is nothing unexpected here.



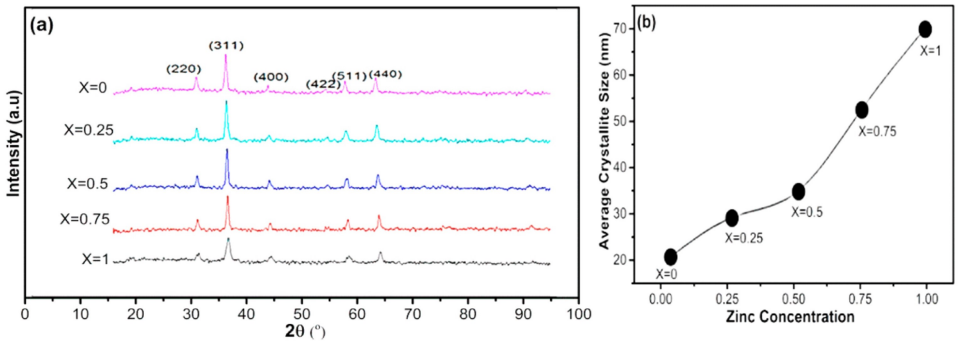
Adapted from Figure 1 and Table 1 of Mustapha et al. (2019).

Example 2: Unexpected results from Scherrer's equation, inconsistency between particle sizes and crystallite sizes

In Khan et al. (2019), the authors use Scherrer's equation to estimate crystallite sizes for Ni- and Zn-based nanoferrites. The authors state:

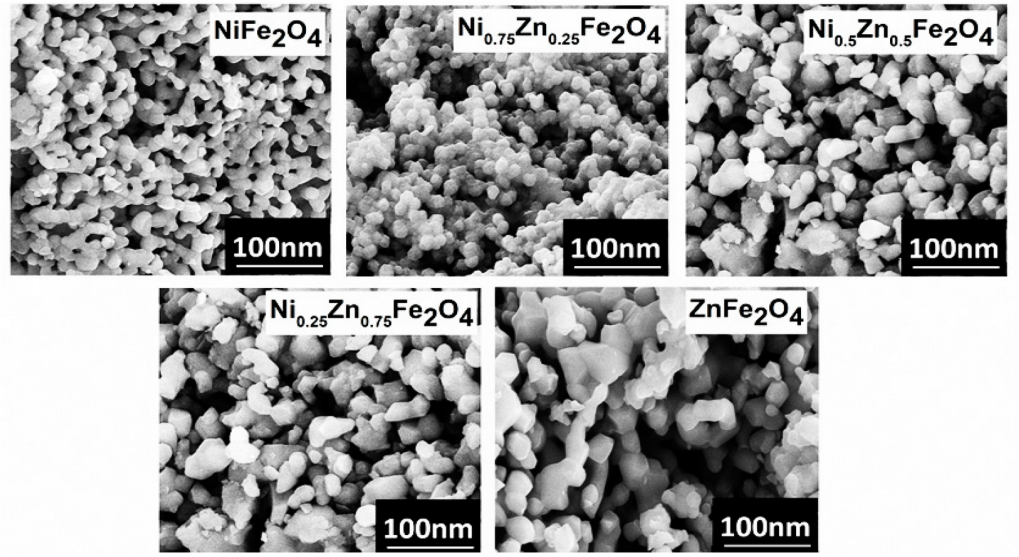
> *The Scherrer formula is used to evaluate the particle size using the extreme intense peak (311). The experimental results demonstrate that precipitated particles' size was in the range of 20–60 nm.*

Recall that Scherrer's equation is used to estimate crystallite size, not particle size. Moreover, the authors' claimed crystallite sizes shown in Figure 2B and Table 1 imply that the (311) peak should be about three times as wide for X = 0 than X = 1. However, in Figure 2A, no such peak broadening is present in the XRD patterns shown. If anything, the (311) peak is wider for X = 1 than X = 0.



Adapted from Figure 2A and Figure 2B of Khan et al. (2019).

Finally, some of the claimed crystallite sizes are larger than the nanoparticles shown for the same materials in the SEM images in Figure 1.



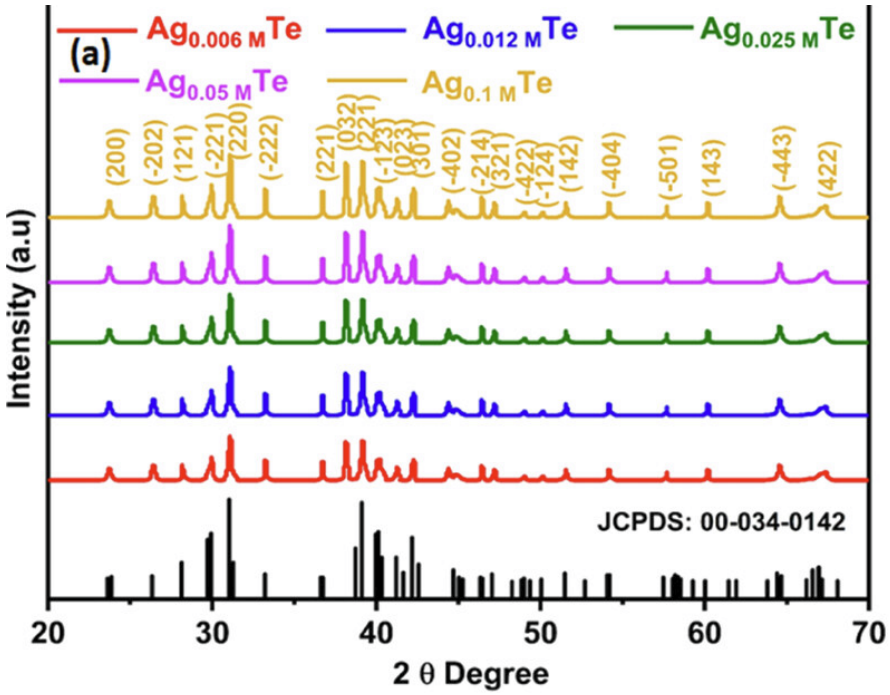
Adapted from Figure 1 of Khan et al. (2019).

Example 3: Unexpected results from Scherrer's equation

In Abdullah et al. (2023), the authors use Scherrer's equation to estimate crystallite sizes of AgTe nanostructures. They state:

> *Effect of Ag0.006Te, Ag0.012Te, Ag0.025Te, Ag0.05Te, and Ag0.1Te nanostructure was determined with crystallite size found in the range of 85 nm, 73 nm and 39 nm, and 67 nm and 88 nm, respectively, measured with Debye Scherer.*

If the crystallite size of Ag0.025Te is twice as small as its counterparts, one would expect the peaks in the XRD pattern for Ag0.025Te to be twice as wide. However, this is not observed in the XRD patterns shown in Figure 1A. In fact, the patterns shown for all materials in Figure 1A are identical except for vertical scaling.



Adapted from Figure 1A of Abdullah et al. (2023).

Example 4: confusion of crystallite size with particle size

In Upadhyay et al. (2016), the authors use Scherrer's equation to estimate crystallite sizes of magnetite nanoparticles. However, throughout the article the authors refer to crystallite size and particle size interchangeably, several times claiming that particle size can be determined by Scherrer's equation. For instace, the caption of Table 1 reads:

> *Table 1. Particle size, lattice parameter and strain in the sample calculated from X-ray data. Size (S) represent particle/crystallite size calculated using Scherer formula while Size (WH) represent particle/crystallite size calculated from Williamson–Hall method.*