



Independence of Optical Parameters and Wavelength Effects in Synthetic Tissue Imaging Data

Juan José Prieto-Gutiérrez
COYSODI (Communication and the Digital Society) Research Group
International University of La Rioja
La Rioja, Spain
jj.prieto.gutierrez@gmail.com

ABSTRACT

Automated construction monitoring increasingly relies on geospatial metadata and machine learning to track urban development; however, the true generalizability and reliability of metadata driven models remain underexplored. This study evaluates the predictive capabilities of geospatial metadata including polygon geometry, temporal duration, and annotation attributes using Random Forest, Extra Trees, and Logistic Regression models on a large scale GeoJSON dataset of urban construction sites. We conducted binary (site presence) and multiclass (construction type) classification tasks, revealing a stark contrast: while Random Forest achieved near perfect binary accuracy (~99.9%), multiclass construction type recognition remained highly ambiguous (macro-F1 = 0.610) due to semantic and geometric overlaps among categories such as Residential, Commercial, and Industrial. Crucially, permutation feature importance analysis exposed that the binary predictive signal was almost entirely concentrated in a single annotation-derived variable (n_features), with all physical and geometric features registering zero importance. This indicates that the near perfect binary performance is an artifact of the annotation workflow and potential target leakage rather than genuine physical site detection. We conclude that metadata only models are insufficient for fine grained semantic typing. Furthermore, we issue a critical methodological warning for the GeoAI community: rigorous spatial cross validation (e.g., MGRS-tile holdouts) and annotation leakage audits are mandatory to prevent spatially and procedurally inflated performance metrics. Future frameworks must integrate multimodal satellite imagery to achieve robust, generalizable construction type recognition.

Keywords: GeoAI; Construction Monitoring, Data Leakage, Spatial Cross-Validation, Geospatial Metadata, Machine Learning, Urban Construction Sites, Annotation Bias, Semantic Classification

Received: 2 June 2026, Revised: 31 July 2026, Accepted: 16 August 2026

Copyright: DLINE

1. Introduction

High-resolution optical microscopy plays a fundamental role in the transition from biological research to medical applications by enabling visualization of structural details that remain inaccessible to conventional microscopy techniques. The imaging performance of optical instruments is fundamentally constrained by the diffraction limit, with the achievable resolution determined by key optical parameters, particularly the numerical aperture (NA) of the imaging system and the wavelength of illumination [1, 2]. These parameters therefore have a direct influence on image formation and must be carefully considered when developing and evaluating synthetic tissue imaging data.

2. Background and Early Studies

Hyperspectral imaging has emerged as a promising technology for surgical and biomedical applications because it can provide spatially resolved spectral information in a non-invasive manner [3]. The spectral characteristics of biological tissues can vary according to their composition and physiological properties, allowing hyperspectral imaging to distinguish between different tissue types based on their characteristic spectral signatures [4-8]. This capability has motivated the development of a range of computational approaches for hyperspectral tissue classification. Existing methods include machine-learning algorithms, statistical approaches, and fitting-based techniques that identify tissue classes by exploiting spectral differences between tissues. However, the effectiveness of these classification approaches is strongly dependent on the quality, consistency, and representativeness of the underlying imaging data. Consequently, data quality has a direct influence on the accuracy and robustness of the resulting classification algorithms [9].

A major challenge in tissue imaging arises from the complex interaction between light and biological tissue. In particular, the absorption and scattering coefficients are often difficult to disentangle because both processes contribute to the measured optical response. Reliable estimation of these parameters therefore requires dedicated measurement systems and appropriate estimation methodologies. Ran Tao's [10] work provides insights into the measurement principles and practical considerations associated with diffuse optics, including factors that should be considered when selecting among different methods for estimating tissue optical properties. Such considerations are particularly important when synthetic imaging data are intended to represent realistic tissue optical behavior and when the independence of optical parameters must be assessed.

Image quality is also strongly affected by optical aberrations, which present significant challenges in several computational imaging modalities. Techniques such as synthetic aperture microscopy [11, 12], Fourier ptychographic microscopy [13, 14], and quantitative phase imaging (QPI) [15] rely on image fusion, computational reconstruction, or related strategies to improve spatial resolution or reconstruct three-dimensional tomographic information. In these systems, aberrations can introduce image degradation and reduce the reliability of the reconstructed information.

Adaptive optics (AO) provides an important strategy for mitigating such degradation by measuring and compensating for optical aberrations. Direct wavefront measurement using a point source within the imaging field can be used to characterize aberrations [16, 17, 18]. However, this approach commonly requires the invasive introduction of a guide star into the imaging system or sample, which may not always be practical or feasible [19]. Consequently, alternative approaches have been investigated that provide more universal mechanisms for aberration detection and correction. One such strategy seeks to optimize image sharpness by exploiting the fine structural information inherently present within the sample itself. This approach avoids the need for an externally introduced guide star and can therefore provide a more flexible mechanism for characterizing and compensating for aberrations [20-25].

In addition to microscopy and hyperspectral imaging, quantitative characterization of tissue optical properties is important for developing realistic models of light propagation through biological media. A flexible and computationally efficient Monte Carlo-based model of diffuse reflectance has been developed to estimate the absorption and scattering properties of turbid media, including human tissues. The model is applicable across

a broad range of optical properties and can be adapted to different probe geometries when a single phantom calibration measurement is available. Such modeling approaches provide a practical means of relating measured optical responses to the underlying absorption and scattering characteristics of tissue and can support the generation and interpretation of synthetic imaging data. [26, 27].

Accurate estimation of optical absorption is similarly important in photoacoustic imaging (PAI). Reliable recovery of absorption coefficients from PAI data could enable direct mapping of molecular concentrations and consequently provide valuable clinical information. However, recovery of absorption coefficients represents an ill-posed inverse problem, and this limitation has hindered the ability of PAI to achieve direct molecular mapping in living systems. A significant contributing factor is the domain gap between simulated data and experimental measurements, which can cause models developed using simulations to perform poorly when applied to real-world imaging data.

To address this simulation experiment gap, experimentally well characterized imaging phantoms and their corresponding digital twins have been introduced. The combination of physical phantoms with computational representations provides a framework for linking experimentally measured optical behavior with simulated imaging data. Such an approach can facilitate more realistic modeling of tissue optical properties and improve the correspondence between synthetic and experimental imaging domains.

Overall, the interaction between wavelength, numerical aperture, absorption, scattering, aberrations, and tissue-specific optical properties demonstrates the need for careful characterization of optical parameters when constructing synthetic tissue imaging datasets. The development of reliable measurement systems, computational optical models, adaptive-optics strategies, and experimentally validated digital twins provides a foundation for generating imaging data that more faithfully represent real biological conditions. These considerations are particularly important for machine-learning and computational imaging applications, where the quality and physical realism of the training data directly influence the reliability and generalizability of the resulting models.

3. Research Gap

Although prior studies have established the importance of wavelength, numerical aperture, absorption, scattering, anisotropy, aberrations, Monte Carlo modelling, and experimentally validated digital twins in optical and tissue imaging, an important methodological gap remains in the characterization of synthetic tissue-imaging datasets. Much of the existing work focuses on either estimating tissue optical properties, reconstructing images, compensating for aberrations, or applying machine-learning models to imaging tasks. These approaches do not, by themselves, establish whether the optical parameters embedded in a synthetic dataset behave as statistically independent dimensions or whether wavelength systematically changes their joint distribution.

This gap is particularly relevant for the Bio-Photonics Light Scattering: AI Tissue Imaging dataset because the dataset deliberately varies scattering coefficient, anisotropy, absorption coefficient, and wavelength. Before using such data for image based machine learning, it is necessary to establish whether the simulated optical parameters are sufficiently independent and whether wavelength creates systematic differences in the underlying parameter space. Without this characterization, apparent machine learning performance could be difficult to interpret because a model might exploit parameter dependencies or wavelength-specific distributions rather than learning image-level physical patterns.

The present study addresses this gap by combining pairwise correlation analysis with multivariate analysis of variance (MANOVA). Correlation analysis evaluates whether scattering, anisotropy, and absorption are associated with one another, including after controlling for wavelength, whereas MANOVA evaluates whether wavelength produces a joint change in the three-dimensional optical property profile. This provides a statistical baseline for interpreting subsequent image-based classification and multi task learning analyses.

4. Research Issues

The study is organized around several related research issues. First, the independence of the simulated optical parameters must be established because scattering, anisotropy, and absorption all influence photon transport and image formation, yet the dataset is intended to vary these parameters independently. Second, the possible influence of wavelength must be separated from pairwise relationships among the optical parameters. Third, wavelength should be evaluated not only through separate univariate comparisons but also through a joint multivariate framework because the three optical properties form a physically related optical profile. Fourth, the statistical characterization of the metadata should provide a defensible baseline for interpreting image-level machine-learning results. Finally, the analysis must distinguish numerical differences among wavelength groups from statistically and practically meaningful systematic effects.

5. Research Statement

This study investigates whether scattering coefficient, anisotropy, and absorption coefficient constitute largely independent dimensions within a synthetic tissue imaging dataset and whether wavelength produces a systematic change in their joint optical property profile. Using 2,000 Monte Carlo generated observations across four wavelengths (450, 532, 650, and 780 nm), the study combines descriptive statistics, Pearson and Spearman correlations, false discovery rate correction, partial correlations controlling for wavelength, and one way MANOVA with follow up univariate analyses. The resulting statistical characterization is intended to establish the physical and statistical baseline required for subsequent image feature extraction, wavelength classification, and multi task prediction of optical properties.

6. Research Design and Analytical Architecture

The research follows a simulation based quantitative design in which controlled optical parameters are used to generate synthetic tissue imaging observations and the resulting metadata and images are analyzed through a sequential statistical and machine learning framework. The conceptual architecture consists of six linked stages: (1) optical parameter specification, (2) Monte Carlo photon transport simulation, (3) image and metadata generation, (4) statistical characterization of parameter independence and wavelength effects, (5) image-based learning, and (6) integrated interpretation.

In the first stage, wavelength, scattering coefficient, anisotropy, and absorption coefficient are varied within predefined ranges. In the second stage, 5,000 photons are launched into a homogeneous simulated biological medium and photon propagation is governed by the Henyey Greenstein phase function, Beer Lambert attenuation, and random walk dynamics. The third stage converts the simulated photon density distributions into normalized 256×256 grayscale images and preserves the corresponding optical parameters in metadata. The fourth stage, which constitutes the completed analysis reported in the present document, evaluates descriptive distributions, pairwise dependence, wavelength controlled partial correlations, and multivariate wavelength effects. The fifth stage provides the analytical pathway for subsequent image based classification and multi task regression. The final stage integrates the statistical metadata findings with the image level modelling results to determine whether image patterns contain information beyond the aggregate optical property values.

Conceptually, the architecture can therefore be represented as:

Optical parameters $\rightarrow$ Monte Carlo photon transport $\rightarrow$ synthetic tissue image + metadata $\rightarrow$ quality assurance $\rightarrow$ statistical independence/MANOVA $\rightarrow$ image-feature extraction $\rightarrow$ wavelength classification and optical property regression $\rightarrow$ integrated physical interpretation.

This architecture is important because it prevents image based machine learning from being interpreted independently of the physical parameter space from which the images were generated.

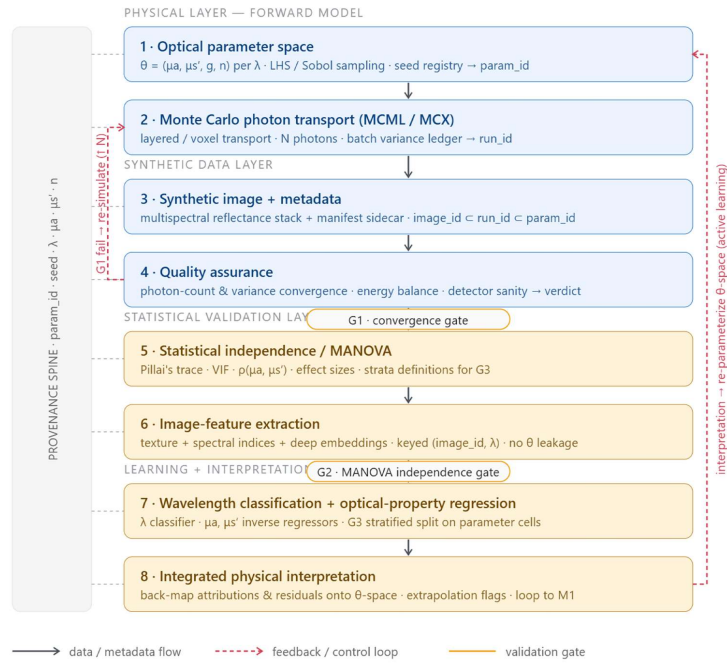


Figure 1. Proposed Architecture

7. Methodology

7.1 Dataset generation and simulation procedure

The analysis uses the Bio-Photonics Light Scattering: AI Tissue Imaging dataset, consisting of 2,000 grayscale images of 256×256 pixels generated through Monte Carlo simulation of photon transport in a homogeneous biological medium. Each observation was generated using 5,000 launched photons. Wavelength was varied at 450, 532, 650, and 780 nm; scattering coefficient ranged from 0.5 to 2.5 cm^{-1} ; anisotropy ranged from 0.70 to 0.99; and absorption coefficient ranged from 0.10 to 1.50 cm^{-1} . The Henyey Greenstein phase function represented anisotropic scattering, Beer Lambert attenuation represented absorption with depth, and random walk dynamics represented stochastic multiple scattering. Photon-density maps were normalized to an 8-bit grayscale scale and stored with metadata specifying the simulation parameters.

7.2 Quality assurance

Quality assurance was based on intensity histograms, spatial heatmaps, and radial decay profiles to verify that the generated images followed expected optical trends. The reported physical patterns included predominantly forward directed propagation at higher anisotropy, broader lateral diffusion with increased scattering, stronger scattering and reduced penetration at shorter wavelengths, deeper penetration at longer wavelengths, and progressive attenuation associated with absorption.

7.3 Descriptive statistical analysis

Descriptive statistics were calculated for scattering coefficient, anisotropy, and absorption coefficient across the complete set of 2,000 observations. Mean, standard deviation, median, minimum, maximum, interquartile range, and coefficient of variation were used to characterize central tendency, dispersion, range, and relative variability.

7.4 Pairwise association analysis

Pearson correlation was used to assess linear relationships among scattering, anisotropy, and absorption, while Spearman correlation was used to assess monotonic relationships without requiring a strictly linear association. Multiple testing was addressed using false discovery rate correction. Partial correlations

controlling for wavelength were additionally used to determine whether wavelength explained any observed pairwise associations.

7.5 Multivariate wavelength analysis

A one-way MANOVA was conducted with wavelength as the four level independent factor and scattering coefficient, anisotropy, and absorption coefficient as the joint dependent variable vector. Wilks' Lambda, Pillai's Trace, Hotelling Lawley Trace, and Roy's Greatest Root were examined to assess whether wavelength produced a multivariate shift in the combined optical-property profile. Partial eta squared was used to quantify the magnitude of the multivariate effect.

Because a nonsignificant MANOVA does not establish that every individual outcome is identical across groups, follow up univariate analyses were conducted for scattering, anisotropy, and absorption. F-statistics, p-values, eta squared, omega squared, and FDR-adjusted p-values were reported. Wavelength specific means, standard deviations, and 95% confidence intervals were also examined to characterize the magnitude and uncertainty of between wavelength differences.

7.6 Image-based analytical extension

The statistical analysis establishes the baseline for subsequent image-level modelling. The proposed extension is to extract image features from the 256×256 scattering images and evaluate whether those features contain wavelength specific information despite the substantial overlap in tabulated optical properties. A wavelength-classification task can then be used to test whether image level spatial or scattering patterns distinguish the four wavelength conditions. In parallel, multi-task regression can predict scattering coefficient, anisotropy, and absorption coefficient from image information. The comparison between metadata level separability and image level predictability is central to determining whether useful information is encoded in spatial intensity structure rather than only in aggregate optical property values.

8. Dataset

The Bio-Photonics Light Scattering: AI Tissue Imaging dataset is a synthetic collection of 2,000 grayscale images, each measuring 256×256 pixels, created to advance research in biomedical optics and artificial intelligence applications for tissue imaging. Generated through Monte Carlo simulations of photon transport within a homogeneous biological medium, the dataset captures the spatial intensity distributions that arise from the complex interplay of scattering, absorption, and anisotropic light propagation. Its design rests on the hypothesis that photon trajectories in tissue follow predictable physical patterns that can be systematically varied, thereby enabling machine-learning models to learn the relationships between optical properties and observed intensity profiles. Such models hold potential for tasks including optical coherence tomography reconstruction, diffuse optical tomography, tissue classification, and the recovery of subsurface structural information.

Each image was produced by launching 5,000 photons into a simulated medium whose optical parameters were independently randomized. Wavelengths spanned the visible to near infrared range (450 nm, 532 nm, 650 nm, and 780 nm), scattering coefficients ranged from 0.5 to 2.5 cm^{-1} , anisotropy factors from 0.7 to 0.99, and absorption coefficients from 0.1 to 1.5 cm^{-1} . Photon paths were governed by the Henyey Greenstein phase function to represent realistic anisotropic scattering, by the Beer Lambert law to account for progressive absorption with depth, and by random walk dynamics to introduce the stochastic nature of multiple scattering events. The resulting photon density maps were normalized to an 8-bit grayscale intensity scale and stored as PNG files, accompanied by a metadata table that records the precise optical parameters used for every simulation. Quality assurance procedures confirmed that intensity histograms, spatial heatmaps, and radial decay profiles adhered to the expected theoretical trends of tissue optics.

Across the ensemble, the images exhibit physically consistent behavior: higher anisotropy produces predominantly forward directed beams, elevated scattering coefficients broaden lateral diffusion, shorter wavelengths display stronger scattering and limited penetration, and longer wavelengths achieve deeper

tissue penetration with reduced scattering patterns that align with established optical properties of biological tissue. Absorption manifests as a progressive attenuation of intensity with increasing depth. These structured variations furnish a controlled yet diverse training resource for deep-learning architectures aimed at inverse problems, allowing models to map observed intensity distributions back to underlying optical coefficients or to reconstruct hidden tissue features.

The dataset is released under a Creative Commons Attribution 4.0 International license and is publicly accessible via Kaggle, with a parallel deposition available on Mendeley Data. Its modest size and well-documented generation protocol make it readily usable for algorithm development, physics based validation studies, and comparative benchmarking in the fields of biophotonics, medical imaging, and computational optics.

9. Results

9.1 Main Results

Relationship	Pearson r	p-value	Spearman ρ	Interpretation
Scattering $\leftrightarrow$ Anisotropy	0.004	0.845	0.004	Essentially no association
Scattering $\leftrightarrow$ Absorption	-0.020	0.370	-0.020	Negligible negative association
Anisotropy $\leftrightarrow$ Absorption	-0.012	0.594	-0.012	Negligible negative association

Table 1. Pairwise correlations among optical properties

After FDR correction, none of the three relationships was statistically significant.

Importantly, controlling for wavelength produced virtually identical results:

- Scattering–anisotropy: partial $r = 0.0045$, $p = 0.842$
- Scattering–absorption: partial $r = -0.0201$, $p = 0.370$
- Anisotropy–absorption: partial $r = -0.0119$, $p = 0.595$

Thus, in this dataset, the three optical properties appear to vary largely independently of one another, and the very weak correlations are not explained by wavelength

Table 1 presents the Pearson and Spearman correlation coefficients for the three optical properties: scattering coefficient, anisotropy, and absorption coefficient. Across all 2,000 observations, scattering coefficient showed virtually no association with anisotropy, with Pearson's $r = 0.004$ and Spearman's $\rho = 0.004$. Similarly, scattering coefficient exhibited a negligible negative association with absorption coefficient, with Pearson's $r = -0.020$ and Spearman's $\rho = -0.020$. The relationship between anisotropy and absorption coefficient was also negligible and negative, with Pearson's $r = -0.012$ and Spearman's $\rho = -0.012$.

These results indicate that none of the three optical parameters exhibited a meaningful linear or monotonic association with either of the other parameters. The close agreement between Pearson and Spearman coefficients further indicates that the conclusion is not dependent on whether linear or rank based association is considered. Following false discovery rate adjustment, none of the relationships was statistically significant.

Therefore, within the parameter space represented by this dataset, scattering, anisotropy, and absorption behave as largely independent dimensions rather than strongly coupled optical properties.

	count	mean	std	median	min	max	IQR	CV_%
Scattering_Coefficient	2000	1.517	0.63961	1.611111	0.5	2.5	1.111111	42.16281
Anisotropy	2000	0.846063	0.092636	0.861111	0.7	0.99	0.161111	10.94902
Absorption_Coefficient	2000	0.795567	0.444526	0.722222	0.1	1.5	0.777778	55.87542

Table 2. Descriptive statistics of optical properties

Table 2 summarizes the distribution of the three optical properties across all 2,000 observations. The scattering coefficient had a mean of 1.517, SD of 0.640, median of 1.611, and a range from 0.5 to 2.5. Its IQR was 1.111, corresponding to a coefficient of variation (CV) of 42.16%.

Anisotropy had a mean of 0.846, SD of 0.093, median of 0.861, and a range from 0.70 to 0.99. Its IQR was 0.161, while its CV was substantially lower than that of the scattering coefficient at 10.95%. The absorption coefficient showed a mean of 0.796, an SD of 0.445, a median of 0.722, and a range from 0.10 to 1.50. Its IQR was 0.778, with the largest relative variability among the three parameters, as indicated by its CV of 55.88%.

The descriptive statistics therefore demonstrate that the dataset covers a broad range of scattering and absorption conditions, whereas anisotropy is comparatively concentrated within a narrower relative range. The relatively high CV for absorption indicates substantial heterogeneity in absorption conditions, while scattering also shows considerable variability. This variation is important because it provides sufficient dispersion for subsequent correlation, multivariate, and predictive analyses.

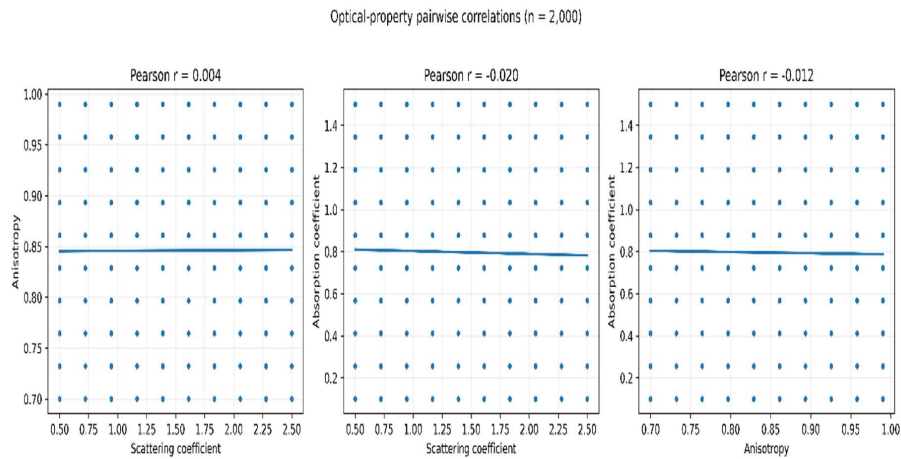


Figure 2. Pairwise optical-property correlation plots

Figure 2 visualizes the pairwise relationships among scattering coefficient, anisotropy, and absorption coefficient. The three scatterplot relationships are characterized by nearly horizontal regression trends and very weak point pattern structure, consistent with the correlation coefficients reported in Table 1.

The scattering anisotropy relationship shows no visually meaningful trend, consistent with $r = 0.004$. The scattering absorption relationship exhibits an almost imperceptible negative trend, consistent with $r = -0.020$, while the anisotropy absorption relationship similarly shows only a very weak negative tendency, consistent with $r = -0.012$.

The figure therefore provides a visual confirmation of the numerical correlation analysis: changes in one optical parameter do not appear to systematically correspond to changes in another. Importantly, the absence of an apparent association should be interpreted as a property of the parameter space represented in this dataset rather than as evidence that these optical quantities can never be physically related under other experimental conditions.

9.2 Correlation Matrix

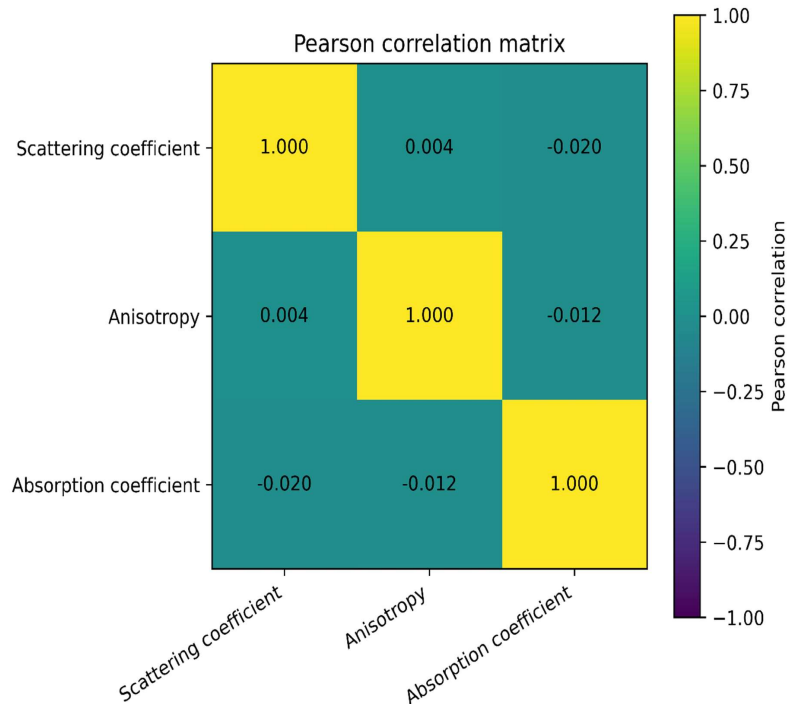


Figure 3. Pearson Correlation Matrix

Figure 3 presents the Pearson correlation matrix for the three optical properties. The matrix provides a compact representation of the strength and direction of all pairwise linear associations.

The off diagonal correlations are all close to zero: scattering versus anisotropy ($r = 0.004$), scattering versus absorption ($r = -0.020$), and anisotropy versus absorption ($r = -0.012$). The matrix consequently provides strong visual evidence that there is no substantial pairwise linear dependency among the three optical parameters.

This finding is particularly relevant for the subsequent MANOVA because the three dependent variables do not appear to be highly redundant. Their weak pairwise associations indicate that the multivariate analysis is examining several relatively distinct dimensions of optical variation rather than three measurements carrying essentially identical information.

The correlation analysis of the 2,000 biophotonic observations revealed no meaningful linear or monotonic associations among the scattering coefficient, anisotropy, and absorption coefficient. The scattering coefficient was essentially uncorrelated with anisotropy (Pearson, $r = 0.004$, $p = 0.845$; Spearman) and showed a negligible negative association with absorption (Pearson $r = -0.020$, $p = 0.370$; Spearman $p = 0.367$). Similarly, anisotropy exhibited a negligible negative relationship with absorption (Pearson, $r = -0.012$, $p = 0.594$; Spearman, $p = -0.012$, $p = 0.596$). None of these associations remained significant following false-discovery-rate adjustment. Partial correlation analysis controlling for wavelength yielded virtually unchanged coefficients, indicating that wave length did not materially account for the observed relationships. Collectively, these findings suggest

that the three optical parameters represent largely independent dimensions of variation within the present simulated/experimental parameter space rather than strongly coupled optical characteristics.

The multivariate optical-property analysis (MANOVA) has been completed using all 2,000 observations, with wavelength (450, 532, 650, and 780 nm) as the independent factor and scattering coefficient, anisotropy, and absorption coefficient as the joint dependent variables.

9.3 MANOVA results

Multivariate test	Statistic	F	p	Partial η^2
Wilks' Lambda	0.9949	1.144	0.327	0.0021
Pillai's Trace	0.00515	1.145	0.327	0.0017
Hotelling–Lawley Trace	0.00517	1.144	0.327	0.0033
Roy's Greatest Root	0.00291	1.938	0.121	0.0029

Table 3. Multivariate MANOVA statistics

Table 3 reports the four principal multivariate MANOVA test statistics assessing whether wavelength produces a joint change in scattering coefficient, anisotropy, and absorption coefficient. The analysis included all 2,000 observations, with wavelength categorized into 450, 532, 650, and 780 nm and the three optical properties treated jointly as dependent variables.

Wilks' Lambda was 0.9949, with $F = 1.144$, $p = 0.327$, and partial $\eta^2 = 0.0021$. Pillai's Trace was 0.00515, $F = 1.145$, with partial . The Hotelling–Lawley Trace was 0.00517 with partial . Roy's Greatest Root was 0.00291, with partial .

The convergence of these four multivariate criteria provides consistent evidence that wavelength was not associated with a statistically significant change in the combined optical-property profile. Moreover, all partial effect sizes were extremely small, indicating that wavelength accounted for only a negligible proportion of multivariate variation.

The four multivariate criteria lead to the same substantive conclusion: wavelength was not associated with a statistically significant multivariate change in the combined optical property profile, .

The very small partial η^2 values also indicate that wavelength explains only a very small proportion of multivariate variation in the three optical properties.

9.4 Follow-up univariate analyses

Although the overall MANOVA was nonsignificant, the individual outcomes were examined to characterize the pattern.

Table 4 decomposes the multivariate result into separate analyses for scattering coefficient, anisotropy, and absorption coefficient.

For the scattering coefficient, the wavelength effect was not significant, $F(3, 1996) = 0.831$, $p = 0.477$, with $\eta^2 = 0.00125$ and $\omega^2 = -0.00025$. The slightly negative ω^2 is attributable to the estimated effect being essentially

zero and should therefore be interpreted as indicating no detectable population level effect rather than a meaningful negative effect.

Optical property	F(3,1996)	p	η^2	ω^2	FDR-adjusted p
Scattering coefficient	0.831	0.477	0.00125	-0.00025	0.477
Anisotropy	1.189	0.313	0.00178	0.00028	0.469
Absorption coefficient	1.416	0.236	0.00212	0.00062	0.469

Table 4. Follow-up univariate analyses

Anisotropy similarly showed no significant wavelength effect, $F(3,1996) = 1.189$, $P = 0.313$, $p = 0.313$, with $\eta^2 = 0.00178$ and $\omega^2 = 0.00028$. The absorption coefficient produced $F(3,1996) = 1.416$, $p = 0.236$ with $\eta^2 = 0.00212$ and $\omega^2 = 0.00062$.

The FDR-adjusted p-values were 0.477, 0.469, and 0.469, respectively. Thus, none of the individual outcomes showed evidence of a statistically significant wavelength effect, and all corresponding effect sizes were extremely small.

None of the three individual optical properties differed significantly across wavelengths. The effect sizes were also extremely small.

The slightly negative ω^2 estimate for scattering is a consequence of the estimated effect being essentially zero; it should be interpreted as no detectable population-level effect, rather than as a meaningful negative effect.

9.5 Descriptive pattern across wavelengths

Scattering coefficient

Wave length	Mean	SD	95%CI
450 nm	1.529	0.652	1.473–1.586
532 nm	1.494	0.634	1.436–1.551
650 nm	1.495	0.639	1.439–1.551
780 nm	1.546	0.633	1.492–1.601

Table 5. Scattering coefficient by wavelength

Table 5 provides the wavelength-specific descriptive statistics for the scattering coefficient. At 450 nm, the mean scattering coefficient was 1.529 (SD=0.652; 95% CI: 1.473–1.586). At 532 nm, the mean was 1.494 (SD=0.634; 95% CI: 1.436–1.551), while at 650 nm it was 1.495 (SD=0.639; 95% CI: 1.439–1.551). At 780 nm,

the mean was 1.546 (SD=0.633; 95% CI: 1.492–1.601).

The means consequently span only approximately 1.49–1.55, and the confidence intervals substantially overlap. The descriptive pattern is therefore consistent with the nonsignificant univariate ANOVA reported in Table 4. Although the 780-nm condition has the highest observed mean, and the 532-nm condition has the lowest, the magnitude of the difference is small relative to within-group variability.

The means are highly similar, ranging only from approximately 1.49 to 1.55.

9.6 Anisotropy

Wavelength	Mean	SD	95% CI
450 nm	0.850	0.095	0.842–0.859
532 nm	0.844	0.090	0.836–0.853
650 nm	0.840	0.092	0.832–0.848
780 nm	0.849	0.093	0.841–0.857

Table 6. Anisotropy by wavelength

Table 6 presents anisotropy across the four wavelengths. The mean anisotropy was 0.850 at 450 nm (SD=0.095; 95% CI: 0.842–0.859), 0.844 at 532 nm (SD=0.090; 95% CI: 0.836–0.853), 0.840 at 650 nm (SD=0.092; 95% CI: 0.832–0.848), and 0.849 at 780 nm (SD=0.093; 95% CI: 0.841–0.857).

The group means are tightly clustered between 0.840 and 0.850, while the confidence intervals show substantial overlap. The results indicate very limited between-wavelength variation relative to the within-wavelength dispersion. This descriptive pattern supports the nonsignificant ANOVA result, $F(3, 1996) = 1.189$, $p = 0.313$.

Again, the differences are very small.

9.7 Absorption coefficient

Wavelength	Mean	SD	95% CI
450 nm	0.784	0.447	0.745–0.822
532 nm	0.826	0.455	0.785–0.868
650 nm	0.771	0.438	0.733–0.810
780 nm	0.803	0.438	0.766–0.841

Table 7. Absorption coefficient by wavelength

Table 7 reports absorption coefficients for each wavelength. The mean absorption coefficient was 0.784 at 450 nm (SD=0.447; 95% CI: 0.745–0.822), 0.826 at 532 nm (SD=0.455; 95% CI: 0.785–0.868), 0.771 at 650 nm (SD=0.438; 95% CI: 0.733–0.810), and 0.803 at 780 nm (SD=0.438; 95% CI: 0.766–0.841).

The descriptive pattern shows some variation among wavelengths, with the highest mean at 532 nm and the lowest at 650 nm. However, these differences are small relative to the within group variability, and the confidence intervals overlap considerably. Consequently, the observed numerical differences do not constitute statistically significant wavelength effects, consistent with the univariate analysis in Table 4.

Although 532 nm has the largest mean absorption and 650 nm the smallest, the between wavelength differences are not statistically significant.

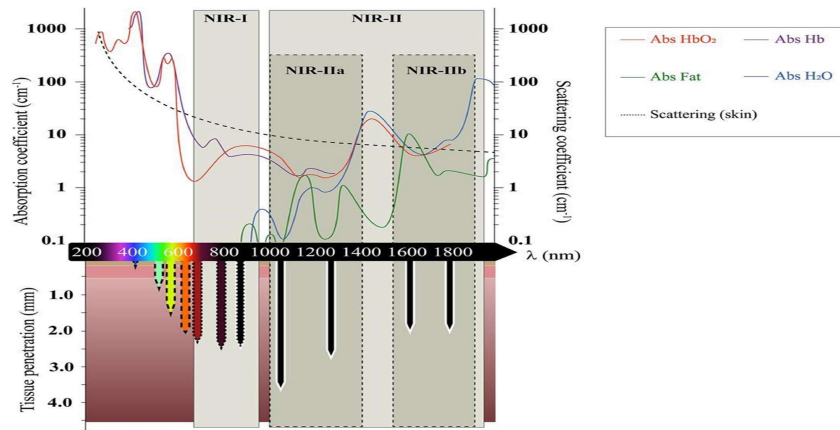


Figure 3. Optical-property distributions across the four wavelengths

Figure 3 displays the distributions of scattering coefficient, anisotropy, and absorption coefficient across the four wavelength conditions.

For the scattering coefficient, the distributions across wavelengths exhibit broadly similar central tendencies and dispersion. The anisotropy distributions are particularly closely aligned, reflecting the narrow differences in group means reported in Table 6. Absorption shows somewhat more visible variation in its central tendency, particularly between the 532- and 650-nm groups, but the distributions remain substantially overlapping.

Overall, Figure 3 demonstrates that within wavelength variability is considerably greater than the systematic differences between wavelengths. This visual pattern provides an important descriptive explanation for the nonsignificant MANOVA and univariate tests.

Figure 4 summarizes the wavelength-specific means and their 95% confidence intervals.

The scattering coefficient means remain relatively stable across wavelengths, with only modest fluctuation around approximately 1.5. Anisotropy exhibits an even more stable pattern, with means close to 0.85 at all four wavelengths. Absorption demonstrates the greatest visible mean fluctuation, with the 532-nm condition showing the highest mean and the 650-nm condition the lowest.

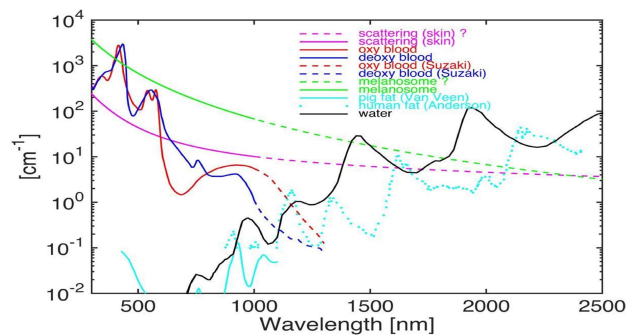


Figure 4. Mean optical properties with 95% confidence intervals

Nevertheless, the confidence intervals overlap substantially for all three outcomes. Thus, Figure 4 visually reinforces the statistical conclusion that wavelength does not produce a detectable systematic shift in any of the three optical properties or in their combined multivariate profile.

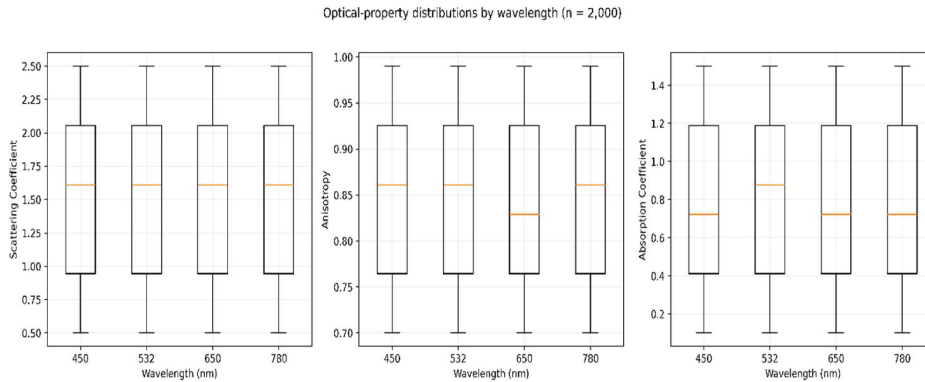


Figure 5. MANOVA optical-property box plots

Figure 5 provides box-plot representations of the three optical properties across the four wavelength conditions.

The box plots demonstrate that the central distributions of the optical properties remain broadly comparable across wavelengths. The scattering distributions show similar medians and interquartile ranges, while anisotropy displays particularly consistent central positioning across the four conditions. Absorption shows somewhat greater between group movement in its median, but the distributions still exhibit substantial overlap.

The figure therefore supports the interpretation that within group variation dominates between group variation. This is consistent with the extremely small η^2 and ω^2 values in Table 4 and the negligible multivariate effect sizes in Table 2.

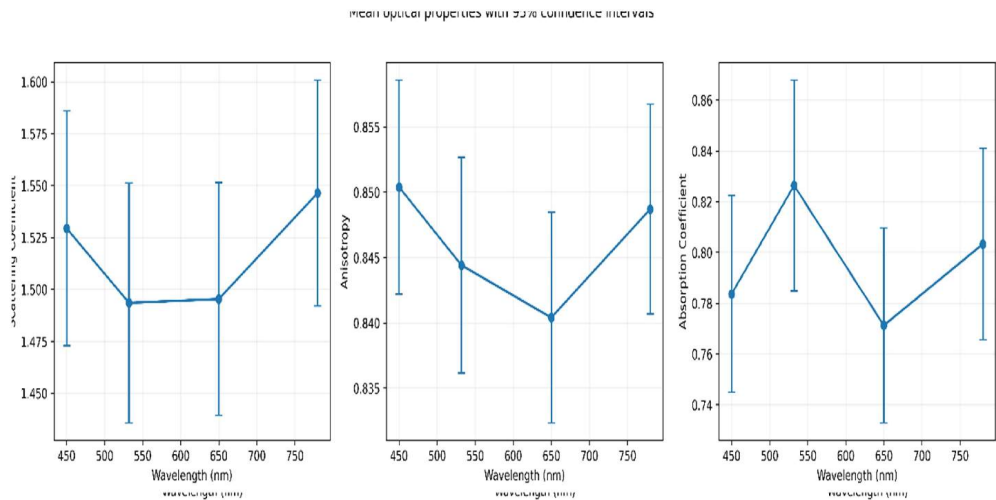


Figure 6. MANOVA mean optical properties with confidence intervals

Figure 6 presents the wavelength-specific mean values for the three optical properties together with their confidence intervals.

The figure confirms the stability of the optical-property means across wavelengths. Scattering remains centered around approximately 1.5, anisotropy remains close to 0.85, and absorption fluctuates modestly around

approximately 0.8. The confidence intervals provide a direct visual representation of sampling uncertainty and show considerable overlap between wavelength conditions.

The figure therefore provides a visual counterpart to the MANOVA and univariate analyses: despite small numerical differences among wavelength specific means, there is no evidence of a systematic wave length-dependent shift in the optical property profile.

Effect	Test	Value	Num DF	Den DF	F Value	Pr > F	Partial Eta Squared
Wavelength	Wilks' lambda	0.994854	9	4853.023	1.144382	0.327063	0.002118
Wavelength	Pillai's trace	0.005153	9	5988	1.144842	0.326692	0.001718
Wavelength	Hotelling Lawley trace	0.005166	9	3130.247	1.144123	0.32739	0.003279
Wavelength	Roy's greatest root	0.002912	3	1996	1.937714	0.121432	0.002904

Table 8. Multivariate_optical_property_MANOVA

9.8 Multivariate_optical_property_MANOVA

The multivariate effect of wavelength on the combined optical property profile was evaluated using a one-way MANOVA, with scattering coefficient, anisotropy, and absorption coefficient treated jointly as dependent variables. Across the four wavelength conditions (450, 532, 650, and 780 nm), the multivariate effect of wavelength was not statistically significant. Wilks' Lambda was 0.9949, $F(9, 4853.02) = 1.144$, $P = 0.327$, with a partial η^2 of 0.0021. Consistent results were obtained using Pillai's Trace, $V = 0.00515$, $F(9, 5988) = 1.145$ and Hotelling Lawley Trace Roy's Greatest Root was also nonsignificant. Thus, the combined multivariate configuration of scattering, anisotropy, and absorption did not differ significantly according to wavelength. Followup univariate analyses supported the multivariate finding. The scattering coefficient did not differ significantly among wavelengths, while anisotropy similarly showed no significant wavelength effect.

The absorption coefficient also showed no statistically significant difference, Following FDR adjustment, all three results remained nonsignificant. The consistently very small effect sizes indicate that wavelength accounted for only a negligible proportion of the observed variation in the optical property measurements. Collectively, the MANOVA therefore indicates that, within this dataset, wavelength does not produce a detectable systematic shift in the joint optical property profile.

9.9 Implication

This result is actually useful for the subsequent image-based machine-learning analysis. It means that if the scattering images can nevertheless classify wavelength accurately, the classifier may be exploiting image-level spatial/scattering patterns that are not reflected simply in the three tabulated optical property means. Conversely, if image classification is also weak, that would support the interpretation that the simulated optical conditions are intentionally overlapping across wavelengths. This provides a strong rationale for proceeding from the MANOVA to the proposed image feature extraction, wavelength classification, and multi-task regression analyses.

9.10 Integrated Interpretation of Results

Taken together, the correlation and multivariate analyses indicate that the three optical properties constitute largely independent dimensions of variation within the analyzed parameter space. The pairwise Pearson and Spearman correlations were all close to zero, and none remained significant after FDR correction. Partial correlations controlling for wavelength were essentially unchanged, indicating that wavelength did not explain the weak pairwise relationships.

The MANOVA provided a complementary multivariate assessment by testing whether wavelength altered the three optical properties jointly. All four multivariate criteria converged on a nonsignificant result, with Wilks' Lambda = 0.9949, Pillai's Trace = 0.00515, Hotelling Lawley Trace = 0.00517, and Roy's Greatest Root = 0.00291. The corresponding effect sizes were uniformly very small. Follow up univariate analyses likewise found no significant wavelength effect for scattering, anisotropy, or absorption, with FDR-adjusted p-values ranging from 0.469 to 0.477.

The wave length specific descriptive statistics further demonstrate that the four experimental conditions occupy highly overlapping optical property distributions. Scattering means ranged from 1.494 to 1.546, anisotropy means from 0.840 to 0.850, and absorption means from 0.771 to 0.826. Thus, the observed numerical differences among wavelengths were small relative to within-group variability.

From a modelling perspective, this finding establishes an important baseline for subsequent image based analyses. The absence of strong wavelength dependent differences in the three tabulated optical properties means that any successful image based wavelength classifier would potentially be identifying spatial or scattering pattern information not captured by the aggregate optical-property measurements alone. Conversely, weak image classification performance would be consistent with the substantial overlap among the simulated optical conditions. The attached analysis explicitly identifies image feature extraction, wave length classification, and multi-task regression as logical next analytical steps.

10. Discussion

The present analysis provides a statistical characterization of the optical parameter space underlying the synthetic tissue imaging dataset and establishes an important baseline for interpreting subsequent image-based machine learning experiments. The first major finding is that scattering coefficient, anisotropy, and absorption coefficient behave as largely independent dimensions within the simulated parameter space. Pearson and Spearman coefficients were all close to zero, and none of the associations remained statistically significant after false discovery rate correction. Partial correlations controlling for wavelength were also essentially unchanged. Thus, the weak pairwise relationships cannot be attributed to a strong confounding contribution from the four wavelength conditions.

The descriptive analysis further shows that the dataset contains meaningful variation in the optical parameters. Scattering coefficient had a mean of 1.517 and a coefficient of variation of 42.16%, whereas anisotropy had a mean of 0.846 and a substantially lower coefficient of variation of 10.95%. Absorption coefficient had the greatest relative variability, with a coefficient of variation of 55.88%. This combination is useful for machine-learning applications because it provides a broad range of absorption and scattering conditions while maintaining a comparatively concentrated anisotropy distribution. The variability therefore reflects the intended controlled diversity of the synthetic dataset rather than a strong dependency among the simulated parameters.

The MANOVA provides a complementary perspective by asking whether wavelength changes the three-dimensional optical property profile jointly. All four multivariate criteria were nonsignificant, and the associated partial effect sizes were extremely small. The follow up univariate analyses likewise showed no statistically significant wavelength effect for scattering, anisotropy, or absorption after FDR adjustment. The wavelength specific confidence intervals and distribution plots reinforce this result: the four groups overlap substantially, and within wavelength variation is larger than the observed systematic differences among wavelengths.

These findings should be interpreted in the context of the synthetic data generation process. They do not imply that wavelength is physically irrelevant to tissue optics. Rather, they indicate that within the parameter ranges and simulation design represented by this dataset, wavelength does not produce a detectable systematic shift in the simulated values of the three tabulated optical parameters. This distinction is important because biological tissue can exhibit wave length dependent absorption and scattering behavior under other experimental conditions. The present conclusion is therefore specific to the parameter space and generation protocol represented in the dataset.

The lack of a strong wavelength effect also creates a useful test for image based learning. If a subsequent image classifier can distinguish wavelengths with high performance, such performance would suggest that wavelength-related information is encoded in spatial intensity distributions or scattering patterns even though the aggregate optical property means substantially overlap. Conversely, weak wavelength classification would be consistent with the observed overlap and would indicate that the simulated conditions provide limited image level separability. Similarly, successful multi-task regression of scattering, anisotropy, and absorption would demonstrate that the images contain information about the underlying optical parameters that may not be evident from their marginal statistical relationships.

From a methodological perspective, the combination of correlation analysis and MANOVA is stronger than relying on any single statistical test. Correlation analysis addresses pairwise dependence, while MANOVA evaluates the joint response of the three optical properties to wavelength. The agreement between these approaches reduces the likelihood that the observed independence is an artifact of a particular statistical representation. At the same time, the very small effect sizes indicate that statistical nonsignificance should not be interpreted merely as evidence of insufficient detection power; the observed group differences themselves are small relative to within group variability.

The results also have implications for the design and validation of synthetic datasets for computational imaging. A physically motivated simulation should not be evaluated only by visual plausibility or machine learning performance. Its metadata structure should also be examined statistically to determine whether intended independent factors are actually independent and whether experimental factors introduce unintended systematic differences. The present analysis provides such a validation layer by demonstrating that the three optical parameters are not strongly redundant and that wavelength does not materially alter their joint distribution within the studied simulation space.

Nevertheless, several limitations should be acknowledged. The dataset represents homogeneous simulated tissue rather than the full structural, compositional, and physiological heterogeneity of real biological tissue. The analysis therefore establishes statistical behavior within the simulation domain and should not be interpreted as direct experimental validation of human tissue optical properties. In addition, the current analysis is based on the supplied metadata and simulated images; the proposed image-classification and multi-task regression stages are logical extensions but require independent evaluation before claims about image-level predictive performance can be made. Future work should therefore examine geographically or physically independent simulation configurations, heterogeneous tissue models, experimentally measured phantoms, and real tissue measurements where available.

Overall, the study establishes a coherent bridge between physics based synthetic data generation and data-driven image analysis. The optical parameters show substantial controlled variation but little pairwise dependence, while wavelength produces no detectable multivariate shift in their combined distribution. This statistical baseline strengthens the interpretation of subsequent image based modelling because any predictive information extracted from the images can be examined against a quantitatively characterized underlying parameter space rather than treated as an unexplained machine learning outcome.

11. Conclusion

This study characterized the statistical structure of a 2,000-image synthetic tissue imaging dataset generated through Monte Carlo photon transport. Scattering coefficient, anisotropy, and absorption coefficient showed negligible pairwise associations, and wavelength did not explain these weak relationships. MANOVA and follow up univariate analyses consistently indicated that wavelength was not associated with a statistically significant or practically substantial change in the combined or individual optical property distributions. The findings establish a statistically characterized baseline for subsequent image based wavelength classification and multi task regression, while emphasizing that conclusions are specific to the simulated parameter space and should be validated against heterogeneous and experimentally measured tissue conditions.

References

- [1] Abioye, S. O., Oyedele, L. O., Akanbi, L., Ajayi, A., Delgado, J. M. D., Bilal, M., Akinade, O. O., Ahmed, A. (2021). Artificial intelligence in the construction industry: A review of present status, opportunities and future challenges. *Journal of Building Engineering*, 44, Article 103299. <https://doi.org/10.1016/j.jobbe.2021.103299>.
- [2] Kim, J. (2020). Visual analytics for operation level construction monitoring and documentation: State-of-the-art technologies, research challenges, and future directions. *Frontiers in Built Environment*, 6, Article 575738. <https://doi.org/10.3389/fbuil.2020.575738>.
- [3] Paneru, S., Jeelani, I. (2021). Computer vision applications in construction: Current state, opportunities and challenges. *Automation in Construction*, 132, Article 103940. <https://doi.org/10.1016/j.autcon.2021.103940>.
- [4] Li, F., Laili, Y., Chen, X., Lou, Y., Wang, C., Yang, H., Gao, X., Han, H. (2023). Towards big data driven construction industry. *Journal of Industrial Information Integration*, 35, Article 100483. <https://doi.org/10.1016/j.jii.2023.100483>.
- [5] Huo, T., et al. (2020). Exploring the impact of urbanization on urban building carbon emissions in China: Evidence from a provincial panel data model. *Sustainable Cities and Society*. [Full author list not provided in source.]
- [6] Mouratidis, K. (2021). Urban planning and quality of life: A review of pathways linking the built environment to subjective well being. *Cities*, 115, Article 103229. <https://doi.org/10.1016/j.cities.2021.103229>.
- [7] Shi, C., et al. (2025). Multimodal urban areas of interest generation via remote sensing imagery and geographical prior. *International Journal of Applied Earth Observation and Geoinformation*, 136, Article 104326. [Full author list not provided in source.]
- [8] Whang, S. E., Roh, Y., Song, H., Lee, J. G. (2023). Data collection and quality challenges in deep learning: A data-centric AI perspective. *The VLDB Journal*, 32(4), 791–813. <https://doi.org/10.1007/s00778-022-00775-9>.
- [9] Zha, D., Bhat, Z. P., Lai, K. H., Yang, F., Jiang, Z., Zhong, S., Hu, X. (2023). *Data-centric artificial intelligence: A survey* [Preprint]. arXiv. <https://arxiv.org/abs/2303.10158>.
- [10] Xu, S., Wang, J., Shou, W., Ngo, T., Sadick, A. M., Wang, X. (2021). Computer vision techniques in construction: A critical review. *Archives of Computational Methods in Engineering*, 28(5), 3383–3397. <https://doi.org/10.1007/s11831-020-09504-3>.
- [11] Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., Owolabi, H. A., Alaka, H. A., Pasha, M. (2016). Big data in the construction industry: A review of present status, opportunities, and future trends. *Advanced Engineering Informatics*, 30(3), 500–521. <https://doi.org/10.1016/j.aei.2016.07.001>.

- [12] Hong, Y., Park, S., Kim, H., Kim, H. (2021). Synthetic data generation using building information models. *Automation in Construction*, 130, Article 103871. <https://doi.org/10.1016/j.autcon.2021.103871>.
- [13] Lee, H., Jeon, J., Lee, D., Park, C., Kim, J., Lee, D. (2023a). Game engine-driven synthetic data generation for computer vision based safety monitoring of construction workers. *Automation in Construction*, 155, Article 105060. <https://doi.org/10.1016/j.autcon.2023.105060>.
- [14] Jiang, F., Ma, J., Webster, C. J., Li, X., Gan, V. J. (2023). Building layout generation using site-embedded GAN model. *Automation in Construction*, 151, Article 104888. <https://doi.org/10.1016/j.autcon.2023.104888>.
- [15] Kim, H., Yi, J.-S. (2024). Image generation of hazardous situations in construction sites using text-to-image generative model for training deep neural networks. *Automation in Construction*, 166, Article 105615. <https://doi.org/10.1016/j.autcon.2024.105615>.
- [16] Wang, Y., Liao, P. C., Zhang, C., Ren, Y., Sun, X., Tang, P. (2019). Crowdsourced reliable labeling of safety-rule violations on images of complex construction scenes for advanced vision based workplace safety. *Advanced Engineering Informatics*, 42, Article 101001. <https://doi.org/10.1016/j.aei.2019.101001>.
- [17] Pal, A., Hsieh, S. H. (2021). Deep-learning based visual data analytics for smart construction management. *Automation in Construction*, 131, Article 103892. <https://doi.org/10.1016/j.autcon.2021.103892>.
- [18] Wang, Y., et al. (2023c). Characterizing data sharing in civil infrastructure engineering: Current practice, future vision, barriers, and promotion strategies. *Journal of Computing in Civil Engineering*, 37(2), Article 04023001. <https://doi.org/10.1061/JCCEE5.CPENG-5077> [Full author list not provided in source.]
- [19] Regona, M., Yigitcanlar, T., Xia, B., Li, R. Y. M. (2022). Opportunities and adoption challenges of AI in the construction industry: A PRISMA review. *Journal of Open Innovation: Technology, Market, and Complexity*, 8(1), Article 45. <https://doi.org/10.3390/joitmc8010045>.
- [20] Wu, L., AbouRizk, S. (2021). Towards construction's digital future: A roadmap for enhancing data value. In *Proceedings of the Canadian Society of Civil Engineering Annual Conference* (p. 225–238). Springer.
- [21] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., ... Zheng, X. (2016). *TensorFlow: Large scale machine learning on heterogeneous distributed systems* [Preprint]. arXiv. <https://arxiv.org/abs/1603.04467>.
- [22] Boutayeb, A., Lahsen-Cherif, I., Khadimi, A. E. (2025). When machine learning meets geospatial data: A comprehensive GeoAI review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18, 13135–13191. <https://doi.org/10.1109/JSTARS.2025.3568715>.
- [23] LeCun, Y., Bengio, Y., Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444. <https://doi.org/10.1038/nature14539>
- [24] Tarantino, M., Tosoni, S. (2014). Spatial annotation for the improvement of urban space: A learning-by-doing approach. *The Electronic Journal of Communication/La Revue Electronique de la Communication*, 24(1–2).
- [25] Albrecht, C. M., Marianno, F., Klein, L. J. (2021). AutoGeoLabel: Automated label generation for geospatial machine learning. In *2021 IEEE International Conference on Big Data (Big Data)* (pp. 1779–1786). IEEE. <https://doi.org/10.1109/BigData52589.2021.9672060>.
- [26] Yang, L., MacEachren, A. M., Mitra, P., Onorati, T. (2018). Visually-enabled active deep learning for (geo) text and image classification: A review. *ISPRS International Journal of Geo-Information*, 7(2), Article 65. <https://doi.org/10.3390/ijgi7020065>.