

Scalable Extraction of Information from Spatio-Temporal Patterns of Chemoresponsive Liquid Crystals Using Topological Data Analysis

Shengli Jiang¹, Nanqi Bao², Alexander D. Smith¹, Shraddha Byndoor¹, Reid C. Van Lehn¹, Manos Mavrikakis¹, Nicholas L. Abbott² and Victor M. Zavala^{1,*}

¹ Department of Chemical and Biological Engineering, University of Wisconsin-Madison, Madison, WI 53706, USA

² Smith School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, NY 14853, USA

Abstract

Chemoresponsive liquid crystals (LCs) can be engineered to generate information-rich optical responses upon exposure to gas contaminants. We investigate the use of topological descriptors to extract information from these complex responses and with this facilitate sensor design and understand physical phenomena that govern responses. We provide a holistic view of topological descriptors using Minkowski functionals and fractal analysis and show how descriptors can be used for unsupervised (clustering, visualization) and supervised (regression, classification) machine learning (ML) tasks. Specifically, by using high-throughput, experimental data for LC films exposed to diverse contaminants, we show that topological descriptors can be used to effectively detect outliers and predict contaminant concentrations using simple ML models. Notably, these models achieve comparable accuracies to those of powerful convolutional neural networks, but with much lower computational times (from hours to seconds) and using less sophisticated computing hardware. This scalability enables high-resolution, space-time data analysis.

Keywords: liquid crystals; optical response; space-time; topology; gas sensors

Chemoresponsive liquid crystals (LCs) provide a versatile platform for designing affordable and lightweight gas (analyte) sensors¹. Specifically, LCs can be functionalized to change their orientational ordering and optical birefringence when exposed to a target chemical environment². For example, LC sensors can be designed by supporting a micrometer-thick LC film on a chemically functionalized solid surface. The molecules constituting the LC (mesogens) have functional groups that bind to the solid surface to cause a well-defined orientation (e.g., perpendicular to the surface) of the LC¹. The gas analyte diffuses through the LC film after exposure, binds or reacts with the solid surface, and triggers a change in the LC orientation (e.g., homeotropic to planar)³. The LC reorientation alters the polarization of light traveling through the film, resulting in an observable optical response (see **Fig. 1** for an illustration).

The optical response of an LC system comprises a complex space-time evolution of color and brightness patterns that result from the multiple physical processes taking place in the LC film (e.g., diffusion of the analyte through the film and binding/reaction events occurring at the surface). The optical response is influenced by the system design (nature of the supporting surface and of the LC film)⁴; as such, responses encode significant information that can be used to guide sensor design and to gain insights into mechanistic behavior occurring at different spatial and temporal scales.

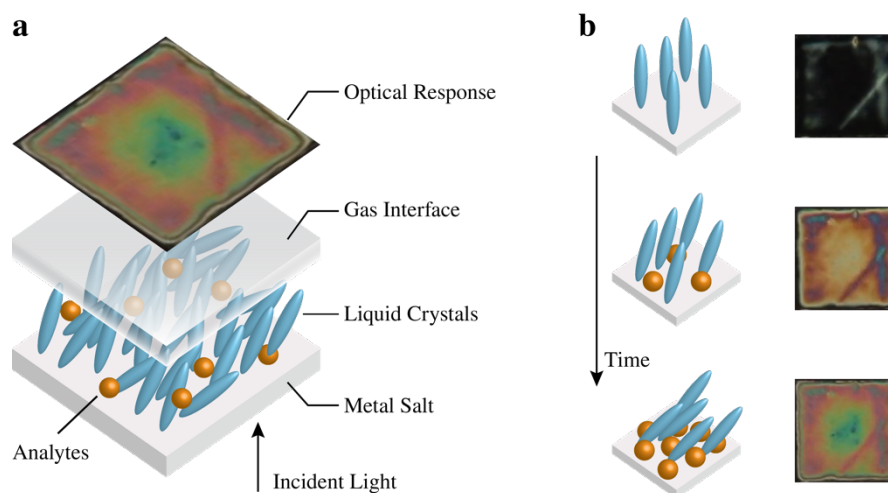


Fig. 1: Optical response of a functionalized LC film to gaseous analytes. (a) Schematic of the LC film design used for detection of gaseous analytes. The LC molecules are initially aligned perpendicularly (homeotropic) to a chemically functionalized surface (e.g., a metal salt-decorated surface); as such, the incident polarized light passing through the LC film from the bottom cannot pass through the top polarizer. The gas analyte diffuses through the LC film and eventually reaches the surface to perturb the binding of the LC to the surface. This perturbation triggers a change in LC orientation (homeotropic to planar) and alters the polarized light passing through the top polarizer, resulting in a complex space-time optical pattern (b).

Machine learning (ML) techniques have recently been used to extract information from LC optical responses to enable prediction of analyte species and concentrations. Cao et. Al. used an off-the-shelf convolutional neural network (CNN), known as AlexNet, to extract features of LCs that were exposed to dimethyl methylphosphonate (DMMP)⁵; the feature information was obtained by applying convolution operations on LC optical micrographs (in the form of grayscale images) and captured spatial brightness patterns that developed in the LC when exposed to DMMP. The authors found that this information correlates strongly with the gas analyte type and this thus provides an approach for detecting DMMP (a simulant of nerve gas agents). Smith et. Al. analyzed response features for this same dataset (in the form of color images) by using a simpler CNN architecture called VGG-16⁶. By analyzing feature information from targeted convolutional layers, the authors found that color patterns of the LC response encode information that correlates strongly to the presence of DMMP. Bao et. Al. recently developed and trained a powerful, three-dimensional (3D) CNN to extract information from space-time responses of LCs (in the form of colored videos) that are exposed to a gaseous mixture containing ozone (O_3) and chlorine (Cl_2)⁷. Their analysis shows that the space-time features extracted using convolution operations accurately predict the concentration of the O_3/Cl_2 mixture and reveal LC response features that are characteristic of O_3 and Cl_2 .

While neural net architectures provide powerful capabilities that can automatically extract information from highly complex space-time LC responses, they also have some important drawbacks. Specifically, due to the “black box” nature of these models, it is difficult to gain insight into the specific features that they are searching for and their physical origin. Attempts have been made to “explain” CNN predictions using saliency maps⁸, but these methods also provide limited interpretability. CNNs (particularly in 3D) also involve many parameters (hundreds of thousands to millions); as a result, they can be computationally expensive to train (particularly for images/videos with high spatial or temporal resolutions) and are prone to overfitting; in addition, CNNs can be sensitive to noise and rotations that commonly appear in real datasets.

Topological data analysis (TDA) has recently emerged as a powerful mathematical framework that provides tools for characterizing data objects. For instance, TDA has been employed to quantify the similarity of pore structures in nanoporous materials, facilitating the identification of materials with comparable pore geometries and the screening of materials similar to top-performing structures⁹. TDA has also gained prominence in brain research, elucidating the dynamic organization of whole-brain activity¹⁰, and uncovering the principles governing transitions in resting-state brain activity¹¹. In medical science, TDA has proven useful for identifying candidate cancer-associated genes and revealing previously undiscovered cancer-related alterations¹². Furthermore, it has been instrumental in uncovering patterns, interactions, and predictors in brain and spinal cord injuries¹³. Among all TDA techniques, topological descriptors are particularly straightforward to implement and have strong connections to physical properties, resulting in their extensive application across numerous fields. The Euler characteristic (EC) is a key topological descriptor that characterizes random fields¹⁴ and encodes statistical information such as means, variances, and characteristic length-scales^{15,16} and has been recently used to characterize images. The fractal dimension (FD) is another topological descriptor that quantifies the scaling factor of a pattern contained within a data object¹⁷. FD has been widely used to analyze and monitor morphological changes on material surfaces¹⁸. Topological descriptors are capable of extracting information from multidimensional data and are tolerant to noise¹⁹; in addition, topological descriptors can provide natural connections with physical phenomena that drives behavior. For instance, the EC has been widely used to characterize topological defects in LCs^{20–27}. These defects have profound effects on the physical properties of the LC^{28,29}; for instance, the total topological charge is correlated to the EC of the surface via the Gauss-Poincaré theorem (sum of all charges on a closed surface is equal to the EC of the surface)³⁰. In this work, we use topology to characterize the optical responses of nematic LC systems, which contain defects as well as many other structural features.

Several studies have characterized the topology of optical responses of different nematic LC systems using the FD. Garbovskiy et al. explored the electro-optical properties of high-resistivity nematic LCs that are sandwiched between ferroelectric polymer films³¹; the authors measured the FD of ferroelectric films to indicate the roughness of the boundaries between adjacent domains. The FD has also been used to study aggregation of inclusions in bulk LCs or at LC interfaces. Lavrentovich et al. studied the dynamics of anisotropic aggregation of spherical colloidal particles dispersed in a nematic LC³²; they used FD to describe the geometry of the aggregated colloidal clusters and understand the aggregation dynamics. Abbott et al. used FD to characterize the aggregation of peptides at interfaces of LCs³³. Goncharuk et al. used the FD to characterize the aggregation, percolation, and phase transition for a nematic LC doped with carbon nanotubes³⁴. Turner et al. used the FD to explore the aggregation and optical properties of a thermotropic nematic LC doped with graphene oxide³⁵. These studies have shown how topology is a common factor that correlates to emergent properties. While the FD has been widely used to characterize nematic LCs, other topological descriptors have rarely been used. For example, Smith et al. recently used the EC to characterize simulated random fields that contain features that are similar to those of chemoresponsive nematic LC films¹⁹. Solis et al. used persistent homology and a derived structural heterogeneity to track the structural changes in an LC nanocomposite and revealed the effect of confined geometry on the nematic-isotropic and isotropic-nematic phase transition³⁶.

In this work, we study the use of different topological descriptors (EC, FD, and lacunarity) to extract information from spatio-temporal optical responses of chemoresponsive nematic LCs. We present a unifying view of these descriptors through the lens of Minkowski functionals (MFs) and fractal analysis, which describe different geometric and topological properties such as shape, convexity, and connectivity that characterize a given topological space¹⁵. This unified view allows us to understand what specific feature information different descriptors are capturing. MFs have been previously proposed for comparing and

evaluating material properties; for example, Slotte et al. used MFs to predict the resistivity and permeability of porous media³⁷. Additionally, MFs have been studied in a variety of other fields such as geological sciences³⁸ and medical imaging³⁹. MFs have been traditionally applied to binary fields (e.g., two-phase materials) because a binary field can be characterized using a simple (scalar) descriptor. However, the optical responses that we analyze here are space-time continuous fields; to characterize these fields, we use a technique known as filtration⁴⁰. The filtration approach enables the extraction of richer topological information because it keeps track of how topological features emerge at different scales. We also show that different color representations reveal different topological features.

Our work aims to demonstrate that the use of topological descriptors provides a computationally scalable and flexible approach for extracting information from complex optical responses of chemoresponsive LCs. We use these capabilities to analyze real, high-throughput experimental data for LC systems that respond to the presence of sulfur dioxide (SO₂) and relative humidity (RH), and for LC systems that respond to gaseous mixtures containing O₃ and Cl₂. Our results reveal that the topological descriptors encode significant and complementary information that can be used for diverse machine learning tasks. Specifically, we demonstrate that the descriptor information can be used for unsupervised learning tasks (such as principal component analysis) to facilitate the identification of experimental outliers and the visualization of response dynamics in low-dimensional spaces. We also show that the topological descriptors can be used in supervised learning tasks to build simple ML models (such as support vector machines) that can achieve comparable prediction accuracies to sophisticated CNNs, but with a computational cost that is 2 to 4 orders of magnitude lower and using less sophisticated computing hardware (CPUs instead of GPUs). This desirable scalability enables the processing of space-time data at high resolutions that might not be accessible to CNNs. All scripts and data needed to reproduce the results are provided as open-source code.

Results

We present different case studies using real, high-throughput experimental data of LC optical responses; the experimental procedures and datasets analyzed are reported in previous work^{3,7,41,42}. In a typical experiment, several hundred LC domains supported on a chemically functionalized surface are imaged in parallel. This enables rapid acquisition of LC optical response data. This approach is also amenable to massively parallel data acquisition via simultaneous imaging of tens of thousands to millions of LC domains. Here, we use LC optical response data to highlight how topological information can be used to conduct different unsupervised ML tasks (visualization, clustering) and supervised ML tasks (regression and classification). Our goal is also to demonstrate that these tasks can be conducted in a computationally scalable manner; as such, we provide comparisons with state-of-the-art CNNs. Finally, we demonstrate how these tools can facilitate understanding of underlying phenomena taking place in LC films.

LC Response to SO₂ and H₂O

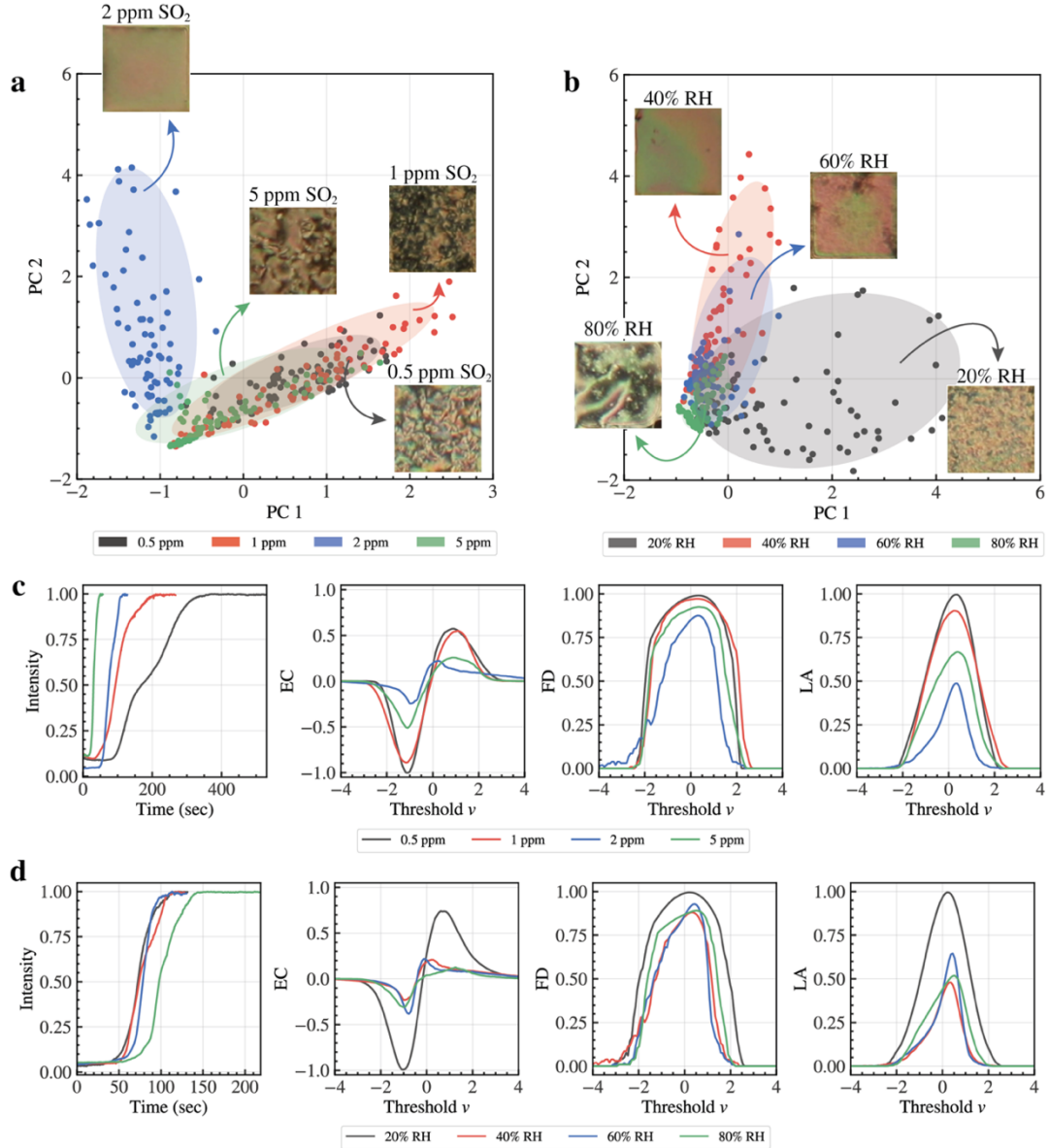
Topological Descriptor Visualization and Analysis. As illustrated in **Fig. 2c-d**, different SO₂ concentrations generate response times that are distinguishable based on large increases in the average image intensity alone; however, changes in RH are not straightforward to distinguish via response times. As a result, the response time cannot accurately reveal changes in gas composition; this motivates our analysis of the spatial topology of endpoint images. **Fig. 2a-b** highlight the differences in spatial topology associated with various SO₂ concentrations and RH; we can see that the endpoint image topology is a more

informative predictor of SO₂ concentrations and RH (compared to response time). In **Fig. 2a-b**, we also show the main principal components of the grayscale EC curves (details on principal component analysis are shown in Section S3.4). A clear cluster separation can be observed between the points representing 2 ppm SO₂ and the points representing the rest of the SO₂ concentrations; the points representing 20% RH are also far apart from the points representing the rest of RHs. We thus expect that concentration prediction is straightforward for 2 ppm SO₂ and 20% RH (due to inherent cluster separation); the overlap of points at other concentrations poses a challenge for ML classification. This highlights how topological information can be used to quickly analyze experimental data and to begin identifying different families of morphologies observed in the responses.

In **Fig. 2c-d**, we show the average grayscale image EC, FD, and LA curves of the LC system exposed to different SO₂ concentrations and RH. The difference in curve shape and peak height and location indicates that the RH has a perceptible effect on the endpoint topology. We also obtained six additional single-channel images to examine the effect of color representations (R, G, B, L*, A*, and B*). Given that each color channel reveals distinct topological features, we proceeded to investigate if combining all the different topologies results in an accurate classification of morphologies (and associate gas analyte concentrations).

Concentration Prediction and Scalability Analysis. We compared the concentration prediction results using AR, BL, EC, FD, and LA as well as the combined descriptors for grayscale images. For each topological descriptor curve, we had a vector of size 100, while the size of the combined vector was 500. To prevent the influence induced by varying vector sizes, we reduced/compressed the feature vector dimension to 100 using principal component analysis (PCA). All subsequent results were derived following PCA dimensionality reduction. **Fig. 3a** and **3c** demonstrate that the combination of topological descriptors on grayscale images outperforms the use of each individual descriptor in terms of classification accuracy and regression RMSE for both SO₂ concentration and RH prediction. Specifically, the SO₂ RMSE for the combined descriptors (1.21 ppm) is 9% lower than the best single descriptor AR result (1.33 ppm); the SO₂ accuracy for the combined descriptor (0.813) is 8% higher than the best single descriptor LA result (0.750). Similarly, the RH RMSE for the combined descriptors (15.7%) is 10% lower than the best single descriptor EC result (17.5%); the RH accuracy for the combined descriptor (0.785) is 8% higher than the best single descriptor EC result (0.729). This indicates that the topological information contained in each descriptor is complementary (non-redundant).

Because each color channel carries different information, we investigated the effect of combining them. For each color channel, we had a vector of topological descriptors of size 500 (including AR, BL, EC, FD, and LA), while the size of the entire combined descriptor vector was 3500. Following the same procedure as for grayscale images, the input vector size to the SVM used in TDA was reduced to 100 using PCA. In **Fig. 3b** and **3d**, we demonstrate that the combination of topological information from color channels outperforms each color channel. Specifically, the SO₂ RMSE for the combined color channels (0.990 ppm) is 11% lower than the best blue channel result (1.114 ppm); the SO₂ accuracy for the combined color channels (0.875) is 7% higher than the best green channel result (0.816). Similarly, the RH RMSE for the combined channels (14.0%) is 10% lower than the best single blue channel result (15.6%); the RH accuracy for the combined channels (0.854) is 4% higher than the best single blue channel result (0.820). This indicates that the topological information contained in the color channels is complementary.



We compared the prediction results using a 2D CNN to a SVM model that uses topological descriptor information. For the input to the SVM, we used a size 100 vector from PCA containing all topological descriptors and color channels. A parity plot is used to compare the regression results (**Fig. 3f-g**). The dots represent the average predicted values, while the error bars represent the standard deviations. We found that both the CNN and TDA approaches have average prediction values that are close to the diagonal (representing a perfect prediction). Specifically, the RMSE of CNN for SO₂ is 0.945 ppm, which is 5% lower than the RMSE of TDA (0.990 ppm). Similarly, for RH, the RMSE of CNN is 12.6%, which is 10% lower than the RMSE of TDA (14.0%). In **Fig. 3h-k**, we present the classification results in a confusion matrix, with diagonal entries representing the rate of correct classification and off-diagonal entries representing the rate of misclassification. Specifically, the accuracy of CNN for SO₂ is 0.951, which is 9% higher than the accuracy of TDA (0.875). Similarly, the accuracy of CNN for RH is 0.868, which is 2% higher than the accuracy of TDA (0.854). TDA is comparable to CNN methods in both regression and classification. However, we note that the number of parameters in CNN was close to 17,000, which is 170 times the number of parameters used in the SVM model (100). As such, the TDA approach uses a much simpler learning model. Additionally, CNN requires rotational data augmentation to ensure that the prediction is rotation-invariant. The topological descriptors, on the other hand, are globally invariant, which means they are not affected by rotation, bending, or translation. It is also important to highlight that the CNN extracts feature information using repetitive convolution operations (at each iteration in the training procedure), while the TDA approach extracts information using filtration operations that are only conducted once. All of these observations have implications on computational scalability.

To compare computational scalability, we compared the training time required for TDA and CNN (**Fig. 3e**). The average time required for the TDA approach was 14.04 seconds, while the average time required to train a CNN was 154 seconds. The differences observed in CNN training time between SO₂ and RH is due to early stopping⁴³, which is a technique for stopping training when the validation loss converges. The TDA training time includes the calculation of all the topological descriptors (13.28 seconds) and the SVM training time (0.76 seconds). Considering seven color channels, the total number of images processed for the computation of the topological descriptors was 3,528. The average computational time to obtain the topological descriptors from a single image is less than 4 milliseconds, indicating that the computation of all topological descriptors is fast (we also note that these computations only need to be performed once). This contrasts with the CNN approach, in which convolutions are performed repetitively with different operators that the model is learning. Overall, we see that TDA provides a significant reduction in computing time over CNN; moreover, TDA only requires a CPU to conduct computations, while CNN requires the use of a GPU (more sophisticated architecture). Details on computational hardware used for obtaining these results can be found in the SI.

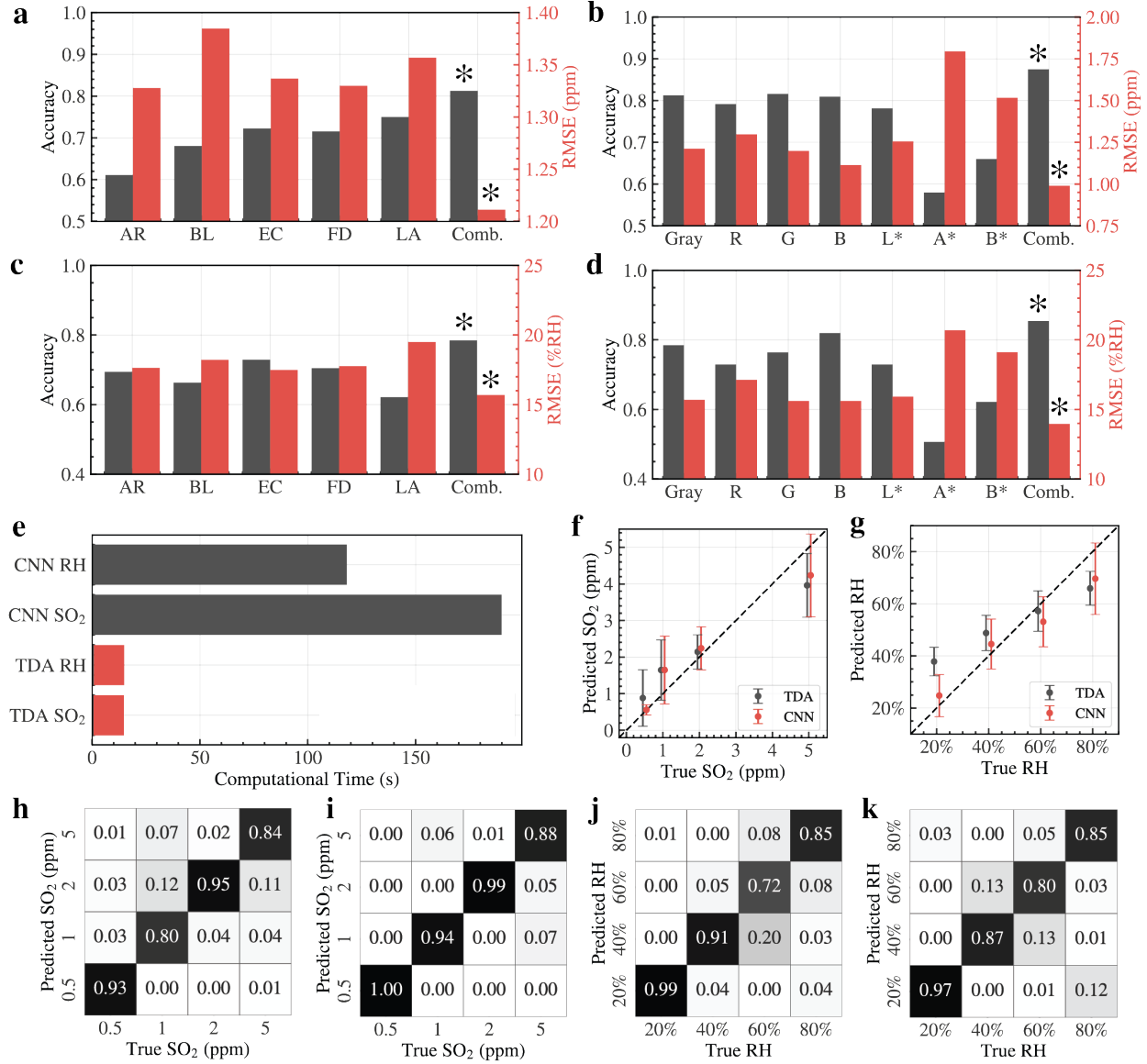


Fig. 3: SO₂/RH concentration prediction results. (a, c) Prediction results for SO₂ and RH based on different topological descriptors of grayscale images. A combination of topological descriptors outperforms the use of single descriptors with higher classification accuracy and lower regression RMSE, showing that the topological information contained in each descriptor is complementary. The asterisk (*) indicates the best-performing descriptor set. (b, d) Prediction results for SO₂ and RH based on different color channels, respectively. The combined color channel descriptors exhibit better prediction accuracy than the single color channel descriptors, indicating that the topological information contained in each color channel is also complementary. (e) Comparison of computation time of TDA and the 2D CNN. The 2D CNN requires 200 times the training time of TDA. Even accounting for the time required to prepare topological data, the total computation time of the CNN is 11 times that of TDA. (f-g) Regression parity plot based on SVM with TDA and 2D CNN for predicting SO₂ concentration and RH, respectively. The error bars represent the standard deviation of the prediction. (h-i) Classification confusion matrix for the prediction of SO₂ concentration based on TDA and 2D CNN, respectively. (j-k) Classification confusion matrix for the prediction of RH based on TDA and 2D CNN, respectively. In general, TDA delivers comparable prediction accuracy than the 2D CNN, but at a lower computing cost, with fewer overfitting problems, and without the need for rotational data augmentation.

LC Response to O₃ and Cl₂

Topological Descriptor Visualization and Analysis. In this study, we investigated four levels of O₃ concentration (i.e., 1.5, 5, 100, and 650 ppm) and 4 levels of Cl₂ concentration (0, 1, 2, and 5 ppm). From our previous study, the concentration of the mixture cannot be accurately predicted if only the endpoint image is considered⁷, therefore we studied the space-time topology of the video. **Fig. 4c** illustrates the differences in the space-time topology of the two LC systems exposed to different concentrations of analytes. The A* channel exhibits a recurring color transition from red to green, indicating a changing tilt angle. **Fig. 4d** depicts the space-time filtration for a grayscale video; the space-time topology varies considerably at different thresholds.

The dataset included 800 responses from LC systems (50 responses per gas mixture concentration). Similar to the SO₂/RH study, we began with a grayscale video and normalized it with a zero mean and unit variance. Direct space-time filtration on the video (**Fig. 4d**) yielded six topological descriptor curves, which are VL, AR, BL, EC, FD, and LA. In **Fig. 4e**, we plotted the first two principal components of space-time EC curves from the videos in **Fig. 4a** (pure O₃) and **Fig. 4b** (O₃/Cl₂ mixture). One can detect a distinct distinction between pure O₃ (black dots) and the O₃/Cl₂ mixture (red dots). The considerable difference between the average EC curves of pure O₃ and O₃/Cl₂ mixture suggests that space-time EC can effectively capture the difference in topology. The presence of outliers implies that response videos from the same batch of experiments still exhibit variations (e.g. caused by imperfect sample preparation, including the presence of scratches and other features on the chemically functionalized surfaces that locally disrupt LC anchoring). **Fig. 4f** demonstrates how different topological descriptors can result in distinct clustering, indicating that they are complementary. Similar to the SO₂/RH case study, we examined the effect of topological descriptors and color channels on the accuracy of concentration prediction; a comparison with the 3D CNN model was also performed.

To confirm that space-time topology information was essential for accurate predictions, we also studied the use of spatial topological descriptors of endpoint images. Response curves (time evolution of the mean brightness), which are often considered in LC studies, were also used to predict the mixture concentrations. For video and image filtration, each input vector to SVM is reduced to a size of 100 using PCA.

Concentration Prediction and Scalability Analysis. We compared the concentration prediction results using VL, AR, BL, EC, FD, and LA as well as the combined descriptors for grayscale videos. For each topological descriptor, we had a vector of size 100, while the size of the combined vector was 600. We reduced the vector size to 100 using PCA. As shown in **Fig. 5a** and **c**, we find that in general, the combination of topological descriptors outperforms each descriptor in terms of classification accuracy and regression RMSE for both O₃ and Cl₂ concentration prediction. Specifically, the O₃ RMSE for the combined descriptors (115 ppm) is 8% lower than the best single descriptor AR result (125 ppm); the O₃ accuracy for the combined descriptor (0.938) is 2% higher than the best single descriptor AR result (0.921). However, the Cl₂ RMSE for the combined descriptors (1.52 ppm) is 3% higher than the best single descriptor LA result (1.48 ppm); the Cl₂ accuracy for the combined descriptor (0.606) is 5% lower than the best single descriptor LA result (0.635). This means that LA (space-time hole size distribution) contains key information for predicting Cl₂ concentration. In general, combining topological descriptors still improves the accuracy of the prediction.

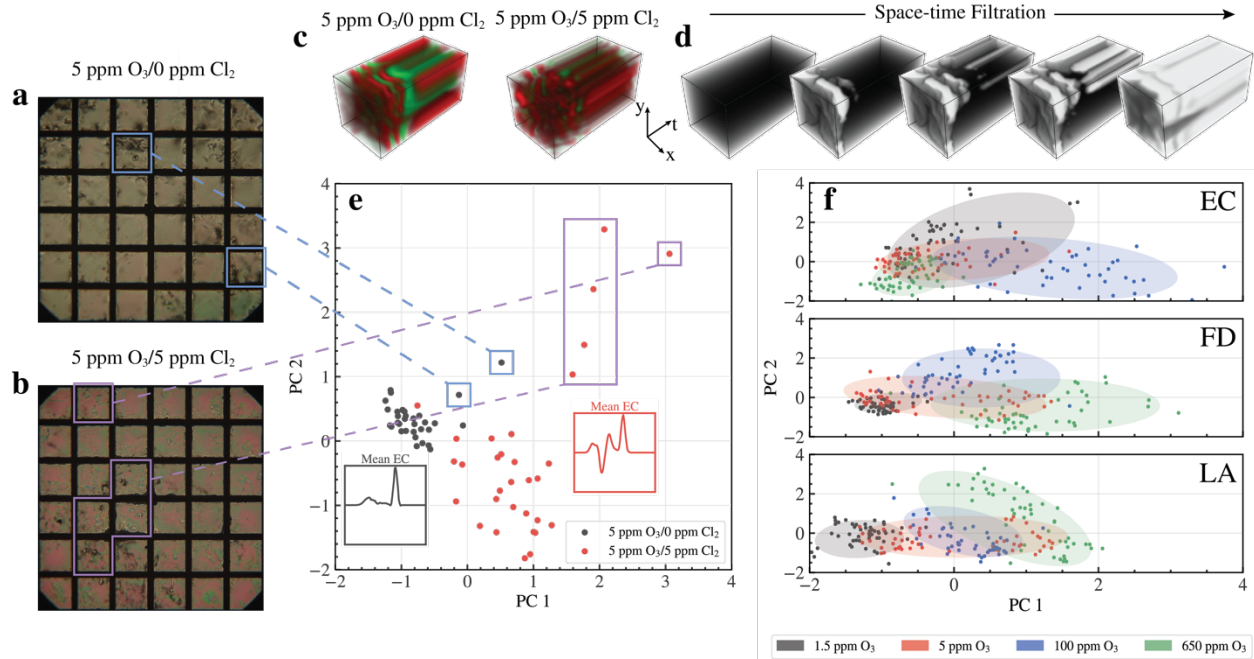


Fig. 4: O_3/Cl_2 mixture dataset. Experimental TEM grids capturing multiple LC responses (endpoint images) when exposed to (a) pure 5 ppm Cl_2 and (b) a mixture of 5 ppm Cl_2 and 5 ppm O_3 . (c) Visualization of the space-time A^* topology when exposed to two concentrations. (d) Space-time filtration of a grayscale video. (e) Principal component analysis of EC curves obtained by space-time filtration of grayscale videos in (a) and (b). Each point represents the main principal components associated with one of the grid videos. A clear separation of black dots (pure O_3) and red dots (O_3/Cl_2 mixtures) can be observed. The significant difference between the average EC curves of pure O_3 and O_3/Cl_2 mixtures indicates that space-time EC can effectively capture the difference in topology caused by different analytes. The blue and purple boxes represent outliers, revealing that response videos from the same batch of experiments exhibit variations. (f) Principal component analysis of EC, FD, and LA curves for mixtures of 1.5, 5, 100, and 650 ppm O_3 and 5 ppm Cl_2 . Different topological descriptor curves generate distinct clustering patterns.

In **Fig. 5b** and **d**, we demonstrate that the combination of color channels outperforms each color channel. Specifically, the O_3 RMSE for the combined color channels (76.9 ppm) is 18% lower than the best red channel result (94.0 ppm); the O_3 accuracy for the combined color channels (0.986) is 5% higher than the best red channel result (0.941). Similarly, the Cl_2 RMSE for the combined channels (1.15 ppm) is 22% lower than the best single red channel result (1.48 ppm); the Cl_2 accuracy for the combined channels (0.786) is 22% higher than the best single blue channel result (0.646). The significant accuracy improvement using combined color channels shows that the topological information contained in each color channel is complementary.

We compared the prediction results using 3D CNN and SVM with TDA. **Fig. 5f** and **g** show the regression parity plots. Both the CNN and TDA methods have average prediction values that are close to the diagonal. Specifically, the RMSE of CNN for O_3 is 43.2 ppm, which is 44% lower than the RMSE of TDA (76.9 ppm). Similarly, for Cl_2 , the RMSE of CNN is 1.01 ppm, which is 12% lower than the RMSE of TDA (1.15 ppm). The remarkable RMSE difference for O_3 may be due to the lack of sufficient data points between 100 and 650 ppm. However, the coefficient of determination (R^2) of the O_3 concentration regression for CNN is 0.973 and for TDA is 0.971, which is very close.

In **Fig. 5h-k**, we present the classification results in a confusion matrix. Specifically, the accuracy of the CNN for O_3 is 0.991, which is 0.5% higher than the accuracy of TDA (0.986). Similarly, the accuracy of the CNN for Cl_2 is 0.711, which is 10% lower than the accuracy of TDA (0.786). TDA is comparable to the CNN in terms of O_3/Cl_2 concentration prediction. The number of parameters in the CNN is at least 43,000, which is 430 times higher than the number of parameters used in the SVM (total of 100). In addition, we compared the training time required for TDA and CNN in **Fig. 5e**. The average time required for TDA was 274.8 seconds, while the average time required to train a CNN was 9,000 seconds (2.5 hours). The TDA time includes both the computation of all the topological descriptors (271.51 seconds) and the SVM training time (3.29 seconds). Considering seven color channels, the total number of videos processed for the topological descriptor was 5,600, and the size of each video is $150 \times 50 \times 50$. The average time required to extract topological descriptors from a single video is less than 50 milliseconds, indicating again the scalability of the TDA approach. Details of the resources for this case study can be found in the SI. We emphasize that 3D CNNs have limitations in handling large videos (resulting from long experiment duration and high space-time resolutions) without sacrificing resolution or sampling rate, but the extraction of topological descriptors for large videos is feasible and scalable. Our computational results thus suggest that topology can enable processing of space-time data at resolutions that are not accessible to CNNs.

LC State Monitoring. Traditionally, the evolution of the average intensity has been used to visualize the state of the LC system. In this type of analysis, there are a couple of important times of interest, the response time and the steady-state (settling) time. The response time is when the average intensity increases from the initial state, while the steady-state time is when the intensity settles (i.e., response saturates). The steady-state time is important in determining when an experiment can be stopped. For example, in **Fig. S12**, the response time for the LC system exposed to 5 ppm O_3 and 5 ppm Cl_2 is at 8 seconds, while the settling time is at 25 seconds. The increase in intensity is smooth between 8 and 25 seconds. To explore how topological descriptors can be used to visualize the dynamics of the system, we obtained the topological descriptor curve for each snapshot in the video and calculated the first two principal components associated with the curve for each time point. **Fig. S12** shows the temporal evolution of the principal components. The results indicate that, even when the average intensity reaches a steady-state, there are clear topological changes still occurring in the LC film (e.g., EC, FD, and LA have large variations after 25 seconds). As such, if we use only the average intensity to characterize the LC system, important behavior might be neglected. We also observe that, between 8 and 25 seconds, although the increase in average intensity is smooth, all topological descriptors show that the evolution of the topological pattern is not smooth and complex. For example, the second principal component of LA shows that, between 10 and 15 seconds, the variation in the hole size distribution is not monotonic. Examining the time evolution of the topological descriptor curves in **Fig. S13** also reveals that some topological changes continue beyond the steady-state time of the average intensity. For instance, the position and shape of EC curve peaks vary significantly, indicating that the system has not reached a steady-state. Compared to the average intensity, time topology provides a more informative description of the LC state.

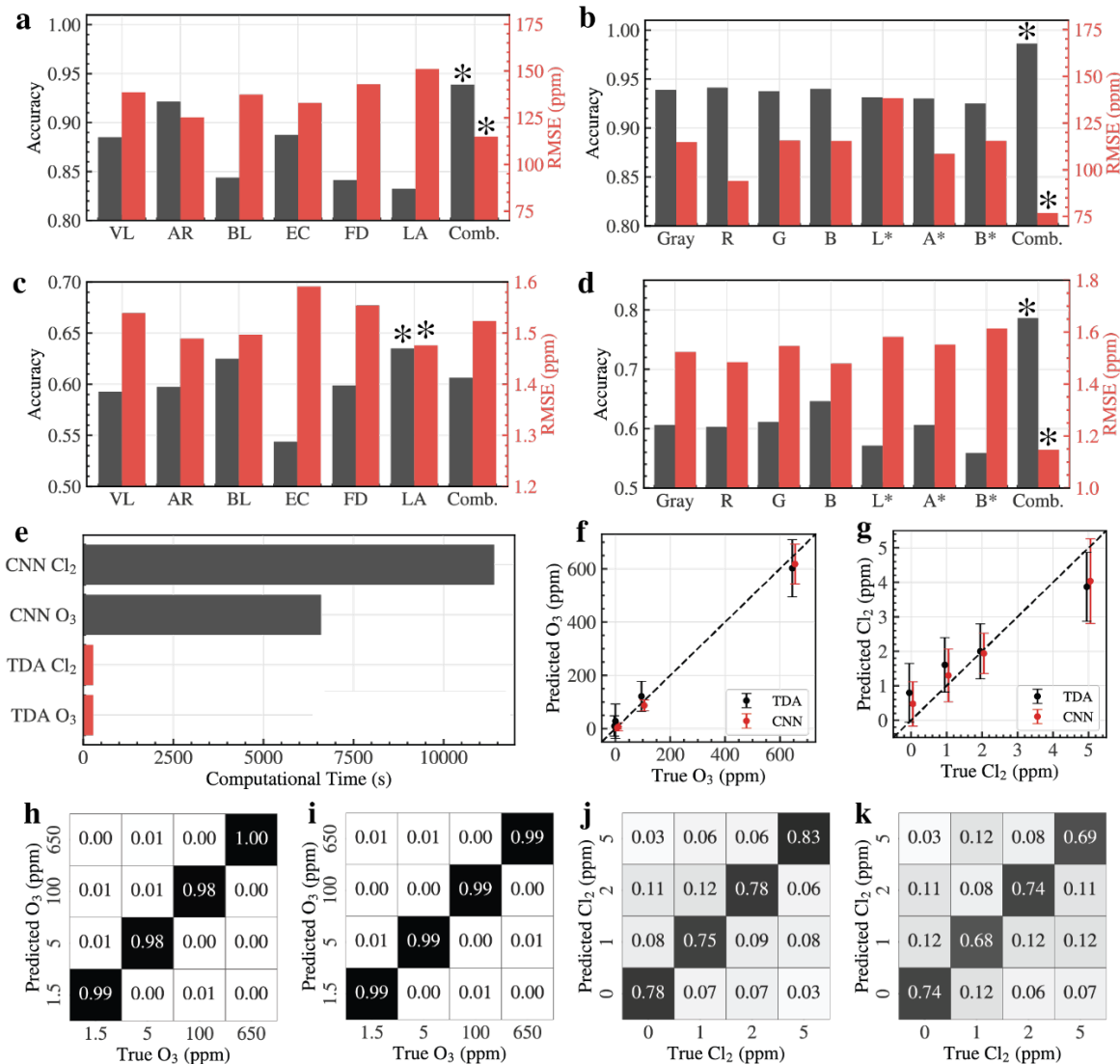


Fig. 5: O₃/Cl₂ concentration prediction results. (a-c) Prediction results for O₃ and Cl₂ based on different topological descriptors of grayscale images, respectively. For O₃ concentration prediction, the combined topological descriptors outperform any single descriptors with higher classification accuracy and lower regression RMSE. For Cl₂ concentration prediction, LA has the highest accuracy, demonstrating that the fractal topology provides important information that can be used to identify Cl₂ concentration. The combination of topological descriptors maintains a relatively high level of prediction precision. The asterisk (*) indicates the best-performing instance. (b)(d) Prediction results for O₃ and Cl₂ based on different color channels, respectively. The combined color channel descriptors provide more accurate predictions than the individual color channel descriptors, demonstrating that the topological information included in each color channel is also complementary. (e) Comparison of computation time of SVM with TDA and 3D CNN. 3D CNN requires more than 2,700 times the training time of SVM with TDA. Even accounting for the time required to prepare topological data, the total computation time of 3D CNN is 33 times that of TDA. (f-g) Regression parity plot based on SVM with TDA and 3D CNN for predicting O₃ and Cl₂ concentrations, respectively. The error bars represent the standard deviation of the prediction. (h-i) Classification confusion matrix for the prediction of O₃ concentration based on SVM with TDA and 3D CNN, respectively. (j-k) Classification confusion matrix for the prediction of Cl₂ concentration based on SVM with TDA and 3D CNN, respectively. Similar to the SO₂/RH case study, SVM with TDA provides comparable or higher prediction accuracy than state-of-the-art 3D CNN at a lower computational cost, with fewer overfitting issues, and without the requirement for rotational data augmentation.

Discussion

We presented topological descriptors for characterizing the space-time optical responses of chemoresponsive LCs exposed to gaseous analytes. We show that topological descriptors provide a scalable and flexible approach for extracting information from complex image and video data and are competitive with state-of-the-art convolutional neural networks. Specifically, we show that topological descriptors enable the construction of simple machine learning models that achieve comparable accuracy to state-of-the-art convolutional neural networks but with a much lower computational cost. This indicates that topological descriptors can be potentially used to create new and powerful machine learning architectures that more effectively capture topological features. Specifically, one could envision the creation of hybrid architectures that extract information using convolution and topological operations. In addition, the evolution of topological descriptors provides a more informative approach to monitoring the state of LC systems than intensity-based timescales (*e.g.*, response time and steady state time). Our future research will focus on studying such architectures and on using molecular dynamics to understand how liquid crystal properties affect observed topologies.

Methods

The optical response data analyzed in this work appears in the form of color images and videos. A grayscale image is represented as a matrix, where each entry contains the pixel light intensity. Images can be considered as a union $X = \bigcup_{i=1}^N A_i$ of compact and convex subsets A_i in the 2D Euclidean space $\mathbb{R}^2$. The restriction of convexity is reasonable because a pixel can be considered as a compact and convex set. In this regard, functionals can be defined to describe the geometric and topological properties of images (which are composed of a finite union of pixels).

A grayscale video is a sequence of images and can be represented as a tensor; here, the tensor can be seen as a union of voxels (3D version of a 2D pixel) in $\mathbb{R}^3$. Color images and videos are the superposition of matrices or tensors, where each matrix or tensor captures a specific color channel. For instance, a color image is typically represented as the superposition of three matrices (red, blue, and green channels).

In the following discussion, we focus on the analysis of images; the analysis of videos is a natural (but important) extension of the discussion. In **Fig. 6**, we provide an overview of information extraction techniques from images and videos using topological descriptors for the characterization of space, time, and space-time topology. Here, we highlight how descriptors such as the Euler Characteristic (EC), fractal dimension (FD), and lacunarity (LA) embed information that can help visualize topological features and can be used to construct predictive models.

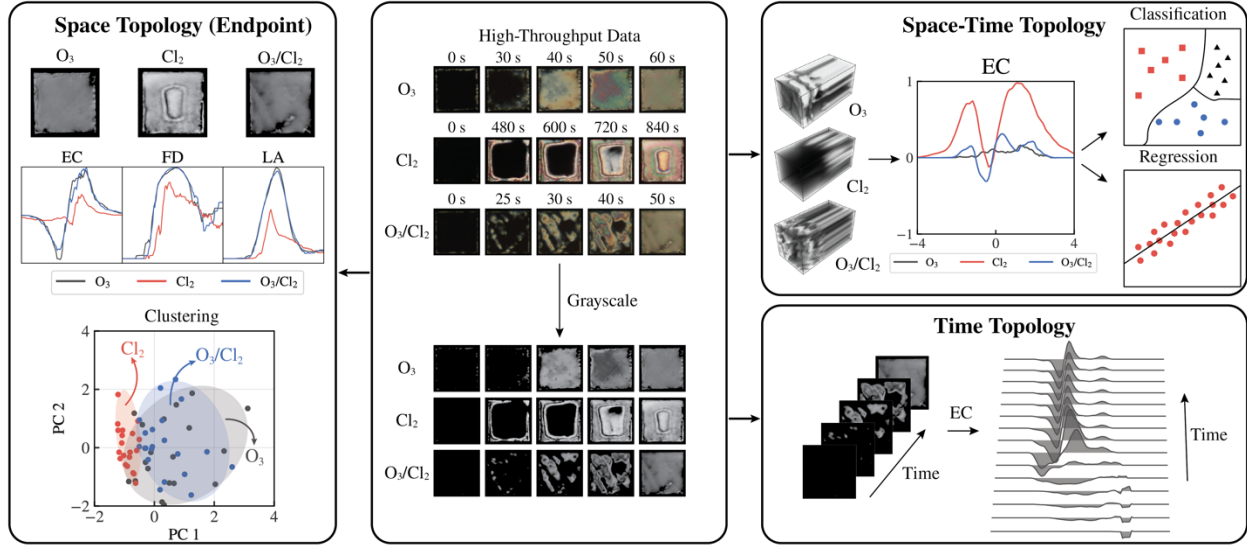


Fig. 6: Overview of information extraction approaches using topological descriptors. We collect high-throughput liquid crystal response data for different analytes (e.g., pure O_3 , pure Cl_2 , and O_3/Cl_2 mixtures). The analytes trigger different space-time responses. We convert these RGB data to different color representations; here, we illustrate a grayscale representation as an example (but others such as LAB are possible). Using the selected color representation, we can extract spatial topology information from endpoint images using a variety of descriptors, such as the Euler characteristic (EC), fractal dimension (FD), and lacunarity (LA). Topological descriptor can then be used to perform unsupervised learning tasks (e.g., principal component analysis) to separate/cluster different responses or to build supervised learning models (e.g., classification and regression) to predict/classify responses. It is also possible to visualize the time evolution of the spatial topology by monitoring the descriptors at each time step; this can be used, for instance, to identify critical times and/or determine when the system has reached a topological steady-state. By treating the video data as a 3D tensor, we can also directly examine the topology of the associated space-time field and we can use the descriptors to conduct unsupervised/supervised learning tasks.

Regular Geometry

Integral geometry provides a set of descriptors, known as Minkowski functionals (MFs), to describe the geometric and topological properties of a *binary lattice* X . The pixels of a binary lattice are either 0 (black) or 1 (white). In the case of a spatial field, the lattice is a binary matrix; in the case of a space-time field, the lattice is a binary tensor. It is important to emphasize that Minkowski functionals can be used to represent spatial and space-time lattices (and even higher dimensional lattices). A binary lattice can be obtained from an image or video (a continuous field) by applying a thresholding/filtration operation, as we describe later.

In general, in a n -dimensional Euclidean space $\mathbb{R}^n$, there are $n + 1$ Minkowski functionals that describe the compact domain of a binary space⁴⁴. These are defined by integrals of curvatures using differential geometry of smooth surfaces⁴⁵. In a 2D case, three functionals M_0, M_1, M_2 are used to characterize X . If ∂X is the regular boundary of X and R is the principal radius of the curvature, the three functionals are:

$$\begin{aligned}
M_0(X) &= \int_X d^2r = A, \\
M_1(X) &= \frac{1}{2\pi} \int_{\partial X} dr = \frac{1}{2\pi} U, \\
M_2(X) &= \frac{1}{2\pi^2} \int_{\partial X} \frac{1}{R} dr = \frac{1}{\pi} \chi,
\end{aligned} \tag{1}$$

An illustration of these definitions and more details on Minkowski functionals can be found in the SI (Section S1.1 and S1.2). In simple terms, M_0 is the covered area A , which we will refer to as AR; M_1 is the boundary length U (up to a factor), which we will refer to as BL; M_2 is the Euler characteristic χ (up to a factor), which we will refer to as EC. The EC is a dimensionless descriptor that captures connectivity and is thus fundamentally different from AR and BL. The EC can be expressed as $\chi = \beta_0 - \beta_1$, where β_0 is the number of connected components and β_1 is the number of holes¹⁶. For a binary spatial lattice, β_0 is the number of connected (white) components and β_1 is the number of connected (black) components. For a binary space-time lattice, the EC is $\chi = \beta_0 - \beta_1 + \beta_2$, where β_2 is the number of voids.

To understand the relevance of Minkowski functionals, it is important to discuss Hadwiger's Theorem⁴⁶. This result states that any additive, motion-invariant, and continuous functional $\mathcal{F}$ of $X \subset \mathbb{R}^n$ can be expressed as a linear combination of $n + 1$ Minkowski functionals M_v :

$$\mathcal{F}(X) = \sum_{v=0}^n c_v M_v(X). \tag{2}$$

Here, the coefficients $c_v \in \mathbb{R}$ are independent of X . This result is key because it highlights that different Minkowski functionals encode unique information (although some functionals might not be particularly informative). Note also that the functional $\mathcal{F}(X)$ can be used to represent an individual MF $M_v(X)$ by setting the coefficients of the other MFs to zero.

The additivity, motion invariance, and continuity properties of the functional $\mathcal{F}$ have key practical consequences, which we summarize as follows:

1. *Additivity*: The functional of the union of domains $X \cup X'$ can be written as $\mathcal{F}(X \cup X') = \mathcal{F}(X) + \mathcal{F}(X') - \mathcal{F}(X \cap X')$. Additivity is important because it implies that a complicated domain X can be partitioned and one can compute the functional separately in each partition. Techniques to compute MFs, such as the marching square algorithm⁴⁴, exploit this property and enable the use of parallel computing. A detailed procedure for the calculation of the MFs for a binary lattice using the marching square algorithm is provided in the SI (Section S1.2). We highlight that these procedures are behind powerful software used for computing topological descriptors.
2. *Motion invariance*: The functional is independent of its position and orientation. If $\mathcal{G}$ is a combination of the translational and rotational group and $g \in \mathcal{G}$, then $\mathcal{F}(gX) = \mathcal{F}(X)$. The motion-invariance property is critical because it states that the functional is invariant to rotations and translations of the object X . This property is not satisfied by other data processing techniques such as convolutions (used in CNNs); as such, CNNs need to deal with translation and rotation using ad-hoc procedures that affect their computational scalability.

3. *Continuity*: If a sequence $X_n \rightarrow X$ converges as $n \rightarrow \infty$, then $\mathcal{F}(X_n) \rightarrow \mathcal{F}(X)$ also converges. This result is important because it indicates that the functional of the true field X can be approximated by an approximating sequence X_n (e.g., obtained from digital imaging). In other words, an approximation X_n of X will yield similar functional values; this has implications in dealing with objects of different spatial or temporal resolutions. Specifically, there is a guarantee that increasing resolution will lead to eventual convergence to the true functional.

Fractal Geometry

Minkowski functionals are defined on regular geometry, whereas many structures found in scientific research and nature are not regular (such as fractal structures). Fractals are specific examples of self-similarity, where a structure or dynamic pattern (sometimes random) is repeated across multiple spatial or temporal scales and have a fine structure on arbitrarily small scales⁴⁷. Well-known fractal structures appear in Brownian motion⁴⁸, polymer networks⁴⁹, aggregate growth phenomena⁵⁰, and porous media⁵¹. Fractal structures have also been studied in LCs⁴⁷. Methods of classical geometry are not suitable for studying fractals; to study fractal geometry, it is necessary to introduce new forms of dimensioning.

The notion of dimension is central to characterizing fractal patterns. Traditional geometry involves the topological dimension (or Lebesgue covering dimension). The first, second, and third topological dimensions are length, area, and volume; however, topological dimensions are not sufficient to describe fractals. For example, from measure theory, the path of a 2D Brownian motion only makes sense as a 2D object (almost space-filling) because it can be arbitrarily large while still having zero area⁵².

The limitations of topological dimensions lead to the emergence of fractal dimensions. A well-known fractal dimension is the Hausdorff dimension, which has the advantage of being defined for any set. A fractal always has a Hausdorff dimension that is greater than or equal to its topological dimension⁵³. The formal definition of the Hausdorff dimension and measure can be found in the SI (Section S1.3).

Unfortunately, the Hausdorff fractal dimension is difficult to compute and thus the so-called box-counting FD is used to approximate it (a detailed discussion is provided in the SI). In each step of the box-counting scheme, the lattice X containing fractals is covered by a grid of squares with side length ϵ , and the number of squares $N(\epsilon)$ intercepted by the object is recorded. One can derive a log-log relation as follows:

$$\log N(\epsilon) \sim D \log \frac{1}{\epsilon} + \log \Lambda, \quad (3)$$

$$D = \lim_{\epsilon \rightarrow 0} \log N(\epsilon) / \log \left(\frac{1}{\epsilon} \right)$$

where D is the fractal dimension (FD) and Λ is the lacunarity (LA). Expressed simply, the LA describes the space occupied by the fractal, while the FD describes the heterogeneity of the fractal⁵⁴. The FD is also commonly referred to as the box-counting dimension or Minkowski dimension. The LA is also commonly referred to as Minkowski content and is computed using box-counting techniques. We refer to the FD and LA as fractal Minkowski descriptors. A fractal with higher FD has more "roughness", while a fractal with higher LA has a more inhomogeneous hole distribution. The exact definitions of these topological descriptors are provided in the SI (Section S1.3 and S1.4).⁵⁵. Lacunarity is the (complementary) counterpart

of the fractal dimension. It is possible that the FD and LA are correlated (do not provide complementary information); however, this is not generally the case. It is also important to highlight FD and LA can be defined for 3D lattices (e.g., arising from space-time objects); we emphasize this because these descriptors have been typically used only to characterize 2D lattices (e.g., arising from spatial objects).

Understanding the scaling properties of Minkowski functionals is important because they can help distinguish fractal and regular geometry. For a 2D lattice containing a regular geometry, we have $M_i(\lambda X) = \lambda^{2-i} M_i(X)$, that is

$$M_0 \sim 1, M_1 \sim \lambda^{-1}, M_2 \sim \lambda^{-2}. \quad (4)$$

For instance, shrinking X by half, in the same area, we get 4 replicates of half M_1 . Based on the additivity of Minkowski functionals, we obtain the new $M'_1 = 2M_1$. For fractals, the scaling is different; the second Hausdorff measure of the fractal in $\mathbb{R}^2$, which is the covered area A under the Lebesgue measure, scales to a non-integer value $M_0 \sim L^{D-2}$, with fractal dimension $D \in (1,2)$ instead of $d = 2$ for the scaling of regular geometry. One can find similar scaling behavior for M_1 and M_2 of fractals⁵⁶. It is important to emphasize that the FD and LA are descriptors that capture fundamentally different features than AR, BL, and EC. However, as we show in our computational results, all these descriptors can provide valuable information when characterizing complex space-time patterns.

Extracting Topological Information using Filtration

Minkowski functionals and fractal properties discussed previously are defined for binary lattices. Grayscale images, however, are continuous fields; in other words, pixels that can take continuous values between the minimum of 0 (black) and the maximum of 1 (white). A lattice representation of an image can be obtained by defining a threshold ν between 0 and 1 and setting all pixels with lower intensity level than the threshold to 0 (and all other pixels to 1). Once the image has been binarized, we can compute the corresponding Minkowski functionals. However, this approach requires pre-selecting a threshold value; to avoid this limitation, we apply a filtration procedure. Here, the idea is to propose a sequence of threshold values that is used to binarize the grayscale image at different intensity levels, and then compute Minkowski functionals at each level. This allows us to keep track of how topological features emerge at different scales.

To explain how the filtration process extracts rich topological information, note that a 2D image $\mathbf{X}$ can be seen as a 3D topography that can be decomposed into multiple binary images X (also known as isoheight planes)⁴⁴. The filtration process is also known as a “watershed” in image processing. Importantly, this approach can be applied to high-dimensional objects such as videos; in such a case, the 3D video $\mathbf{X}$ can be seen as a 4D topography that is decomposed into multiple binary videos X . As shown in **Fig. 7**, we can obtain the topological descriptors (AR, BL, EC, FD, and LA) of each binary image and summarize this information in topological descriptor curves.

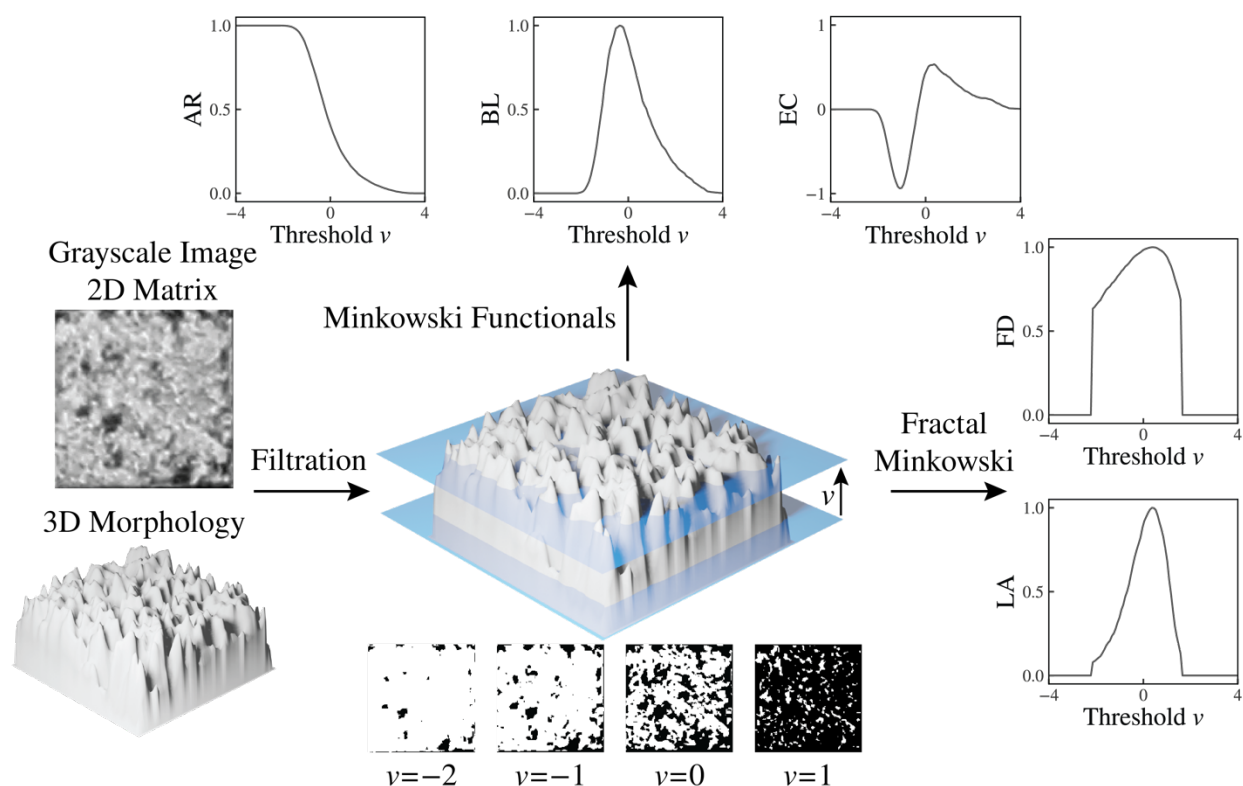


Fig. 7: Filtration process of a grayscale image. The LC optical response is captured as a grayscale image; this image can be represented as a 2D matrix, where the pixel intensity is the value of each matrix entry. The matrix representation reveals the topological features of the grayscale image (which can be visualized as a 3D morphology); this reveals peaks (large values) and valleys (small values) of intensity. To facilitate the next step of the filtration process, the matrix is normalized to have a zero mean and unit variance. Filtration decomposes the matrix into a series of binary matrices with certain thresholds (i.e. matrix values greater than the threshold are set to 1, otherwise they are set to 0). Topological descriptors collected at each threshold ν are summarized in descriptor curves that capture various aspects of the topology. The filtration process can also be applied to higher-dimensional objects, such as videos, using space-time filtration.

Topological descriptor curves are effective tools for summarizing and visualizing image topology. In **Fig. 8**, we illustrate how EC curves can be used to highlight differences in the topology of LC optical responses to various analytes. Each image represents the endpoint optical response of the sensor, which corresponds to the optical response after a fixed time after exposure to the analyte and after the optical response stops evolving in time (i.e., average intensity is at steady state) (also see **Fig. 6**). Specifically, different analyte types and concentrations will result in different LC surface orientations and corresponding morphologies. For example, a mixture of 2 ppm SO_2 and 20% relative humidity (RH) produces a rough response with small domains (**Fig. 8d**), whereas a mixture of 2 ppm SO_2 and 80% RH produces a smooth response with a large domain (**Fig. 8e**). The differences in topology are revealed by the EC curves. In **Fig. 8g**, we see that the EC peak at 20% RH is higher than that at 80% RH. Other topological descriptor curves can also capture different types of variations, as shown in SI (Section S2.1 and S2.2). In **Fig. 8**, the LC responses to pure O_3 and the O_3/Cl_2 mixture⁷, pure Cl_2 ³, SO_2/RH ⁴², and DMMP⁴¹ are obtained from previous studies.

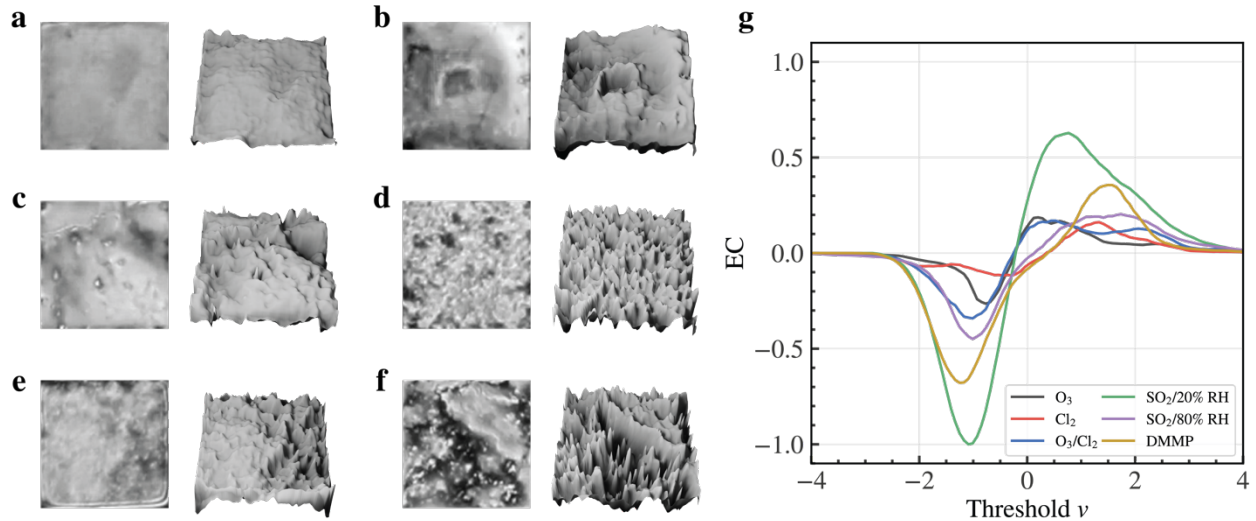


Fig. 8: EC curves of LC optical responses to various gas analytes. Endpoint LC optical responses to (a) 5 ppm O_3 ⁷, (b) 5 ppm Cl_2 ³, (c) a mixture of 5 ppm O_3 and 5 ppm Cl_2 ⁷ (d) a mixture of 2 ppm SO_2 and 20% RH⁴², (e) a mixture of 2 ppm SO_2 and 80% RH⁴² and (f) 10 ppm DMMP⁴¹ and their corresponding 3D morphologies. (g) Normalized EC curves for different analytes; the EC curve shapes reveal differences in the topology of the optical responses.

Extracting Topological Information from Color Images

The optical response of LCs is frequently analyzed using grayscale images. As shown in **Fig. 9a**, however, we can capture the response as a color image and decompose the image into red, green, and blue (RGB) channels and analyze the topology of each channel. The color image in **Fig. 9a** was obtained after the average intensity had reached a steady state. Similar to the RGB color space, the LAB color space is also widely used in color decomposition, where the L^* channel describes the brightness, the A^* channel describes the red/green level, and the B^* channel describes the blue/yellow level. **Fig. 9a** shows how different color spaces highlight different topological features of a response. In this example, the repeated transition from red to green in the A^* image is particularly noticeable, indicating different orders of interference fringes (consistent with a continuous LC anchoring transition)⁷. This information can thus be used to characterize spatial patterns of LC orientations that are not revealed in other color channels. There is no definitive advantage or disadvantage of using RGB or LAB; different color channels, in general, carry distinct information. This is illustrated in **Fig. 9b**, which shows that the EC highlights topological differences between the color channels. The presence of complementary information from the color channels illustrates that they can all be useful in characterizing responses. This has been previously observed in the analysis of LC responses using CNNs⁷ (which conduct convolution operations in different color channels to extract different types of pattern and texture information). These observations are important because, as is done in CNNs via convolution, we can extract information from different color channels using topological analysis. As we will show, this provides a flexible and general approach to constructing models that correlate topological information to emergent properties of interest and that are competitive with CNNs. Specifically, we will see that these models are computationally more scalable than CNNs, can achieve similar accuracies, and require more modest computational hardware.

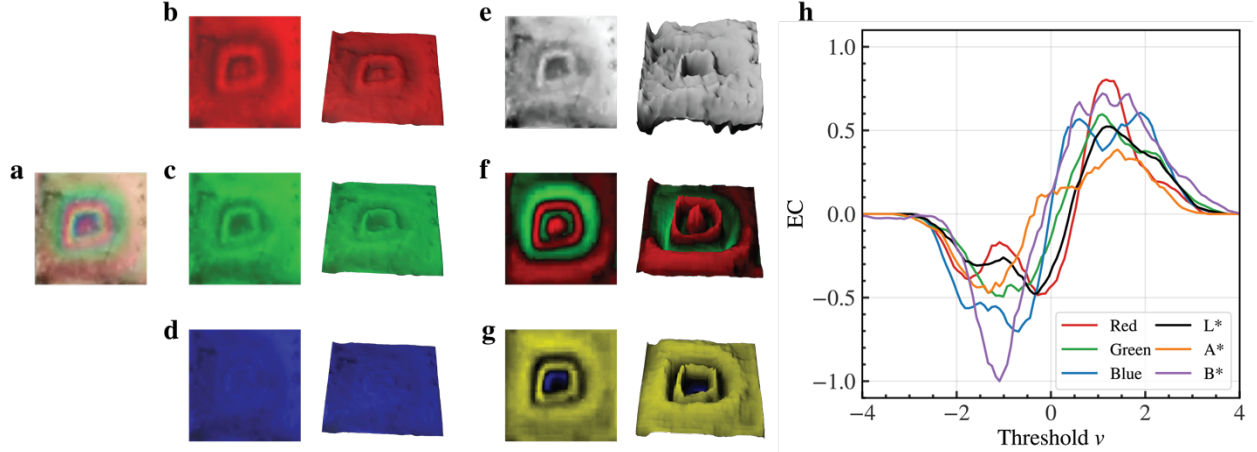


Fig. 9: Extracting topological information color images. (a) Optical response of LC exposed to Cl_2 at a concentration of 5 ppm. The RGB image is typically decomposed into (b) red, (c) green, and (d) blue channels. The image can also be decomposed into (e) an L^* channel that specifies brightness, (f) an A^* channel specifies the red or green level of the color, and (g) a B^* channel specifies the blue or yellow level of the color. The corresponding 3D morphology is next to each single-channel image. The A^* channel reveals a repeated red-to-green transition, indicating different orders of interference fringes (consistent with a continuous LC anchoring transition). The other channels do not highlight this topology clearly. (h) Normalized EC curves for different color channels; differences in EC curves show that different color channels encode different topological information.

Extracting Topological Information from Space-Time Responses

Topological descriptors can be used to characterize objects in high dimensions and in a scalable manner. This provides a powerful approach to visualizing, quantifying, and summarizing complex space-time responses. In **Fig. 10**, we visualize the characterization of a space-time topology of a LC optical response using different descriptors. In **Fig. 10b**, the repeated transition from green to red in the space-time topology of the A^* channel reveals the continuous evolution of the tilt angle of the LCs (indicated by different orders of interference fringes).

Similar to the 2D case (images), the lattice representation of a video is obtained by defining a filtration threshold ν and assigning all pixels with intensity levels below the threshold to 0 and all other pixels to 1. Once the video has been binarized, the Minkowski functionals and fractal Minkowski descriptors can be computed. The 3D Minkowski functionals have an additional component, volume (VL), and the EC now also captures void spaces in the space-time field. In **Fig. 10d**, we illustrate six different topological descriptor curves used to characterize a space-time topology. We can see how the EC reveals the complex dynamics of the tilt angles.

The extraction of topological information for space-time fields can be combined with different color representations to provide a scalable and flexible framework to conduct diverse unsupervised and supervised ML tasks. For instance, as we show in the next section, we can construct ML models that achieve similar accuracies to those obtained with powerful 3D CNNs. We also highlight that it is possible to characterize the evolution of 2D spatial topologies (as opposed to characterizing the topology of the entire 3D space-time field); this is illustrated in one of our case studies. This approach is expected to lose information, as it assumes that spatial topologies are not correlated in time (which might not be the case); however, this approach can provide interesting insights on how topological features appear and disappear in time, can help identify critical times at which features appear or disappear, and can help identify when

the system topology has settled (achieved a topological steady-state). All these capabilities can be useful in facilitating the visualization of complex optical responses and in gaining insight.

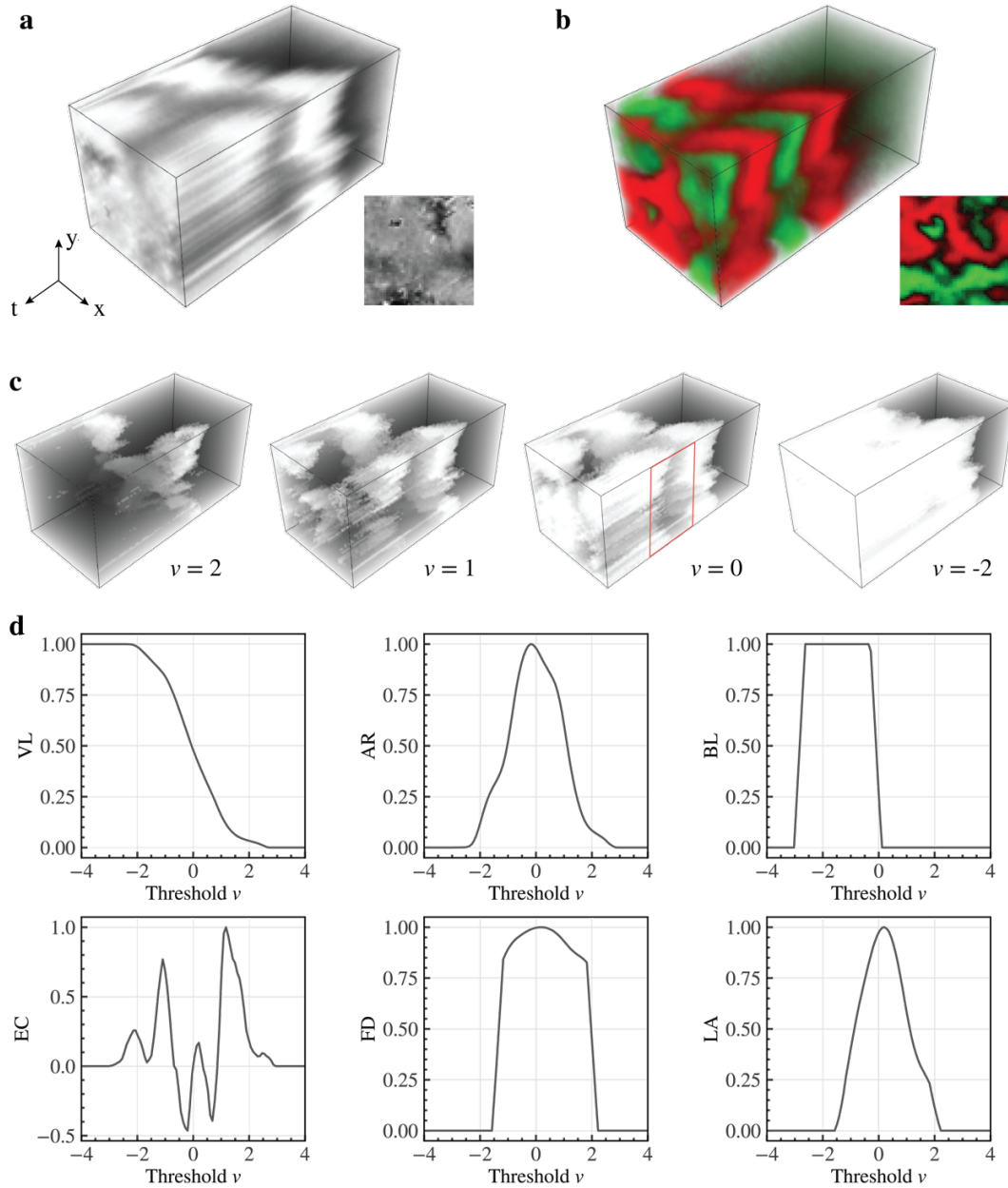


Fig. 10: Space-time topology and filtration. (a) Space-time topology of a grayscale video (1.5 ppm O_3 and 2 ppm Cl_2). The video object has three dimensions, x, y, and time. The duration of the video is 75 seconds. The dark volume at the beginning of the response indicates that no optical response has been observed. (b) Space-time topology of the same video in A* color space. The recurring transition from green to red is indicative of the continuous tilt angle change of liquid crystals. The topological patterns of the A* video are notably different from the pattern of the grayscale video, demonstrating that the different color channels carry distinct information. (c) Space-time filtration of a grayscale video, with pixel values normalized with zero mean and unit variance. We show different space-time topologies at four thresholds. For example, when the threshold value is 0, a discontinuity of intensity over time can be observed (shown in a red box). (d) The space-time filtration results in four Minkowski functional curves and two fractal Minkowski curves. We have an additional curve to describe the occupied volume compared to the 2D case.

Dataset: LC Response to SO₂ and H₂O

High-throughput video data was used to analyze the LC response to different concentrations of SO₂ and relative humidity (RH) in a mixed gaseous environment⁴². LC samples are made of 0.005 mol% 4'-cyanobiphenyl-4-carboxylic acid (CBCA) in 4-cyano-4'-pentylbiphenyl (5CB) and are hosted within the slots of transmission electron microscopy (TEM) grids supported on the TiO₂-decorated surfaces. We split the TEM grid into multiple squares; each square represents a chemoresponsive LC system. For each square, we capture an image of the endpoint response (134×134) which can be represented as a 2D matrix; in other words, here we only analyze the spatial topology of the response at the end of the responses. Each grid square generates a 3-channel (RGB) object. We investigated 4 levels of SO₂ concentration (i.e., 0.5, 1, 2, and 5 ppm) with 40% RH, and 4 levels of RH (i.e., 20%, 40%, 60%, and 80%) with 2 ppm SO₂. The dataset studied contains 504 LC responses (72 responses for every gaseous mixture concentration). We begin by analyzing the (single channel) grayscale images; after standardizing the image with a zero mean and unit variance, we performed 2D filtration and obtained five topological descriptor curves (AR, BL, EC, FD, and LA). Each of the curves is a vector with 100 entries (100 different threshold values were used).

Dataset: LC Response to O₃ and Cl₂

Similar to the experimental setup for the SO₂/RH study, video was used to record the LC responses to various concentrations of ozone (O₃) and chlorine (Cl₂) gas mixtures⁷. LC samples were prepared and embedded in a TEM grid supported on a Mn(ClO₄)₂ decorated surface (**Fig. 4a,b**). The analysis conducted in this case study focused on characterizing the space-time topology of the responses. Each response snapshot is a matrix of dimension 50×50 ; we took 150 snapshots and stacked them chronologically to create a space-time data object (a tensor) of dimension $150 \times 50 \times 50$. A snapshot is taken every 0.5 seconds and the total collected video has a length of 75 seconds. Each square in the TEM grid generates a 3-channel 3D tensor because of the RGB data format.

Concentration Prediction of SO₂ and H₂O

The topological descriptor curves obtained from the responses were fed into a support vector machine (SVM) with a radial basis function (RBF) kernel, which is used to regress and classify the SO₂ concentrations and RH. We refer to this approach as topological data analysis (TDA). Predictions of SO₂ concentration and RH were performed separately. We performed 5-fold cross-validation (CV) to tune the hyperparameters of the SVM. The dataset is split into 5 components; training occurs 5 times, each time leaving out a different part of the dataset for testing. A subset of each training set is chosen as the validation set to tune the hyperparameters of the SVM. We average the error on 5 test sets; for each fold, we use stratified random splitting to divide the 288 responses into 184 training samples (64%), 46 validation samples (16%), and 58 test samples (20%). For the detailed implementation of the SVM and of CV and hyperparameter tuning, please see SI (Section S3). To determine prediction performance, we evaluated regression with root mean square error (RMSE) and classification with accuracy scores.

To compare the effectiveness of using topological information to predict gas concentrations, we compared performance against CNNs⁷. The CNN accepts as input a 3-channel matrix (RGB image) with a dimension of 134×134 and conducts repetitive convolution operations to extract feature information. We evaluated the performance of the CNN using the same five-fold CV for the SVM, with the same training, validation, and test samples. The hyperparameters we tuned for the CNN included the learning rate, the batch size, and the number of filters. The detailed structure and training procedure of the CNN can be seen in SI

(Section S3.2). While the architecture of the CNN is predetermined, the parameter values for the convolutional layers and fully-connected layers are learned from training and validation data. The CNN is trained and validated using a set of input tensors drawn from the training set and their corresponding output labels. The CNN will then attempt to predict the concentration from extracted features and compare it to the true concentration. The difference between predicted and true values is referred to as the error. The CNN will then adjust its parameters to reduce the magnitude of the error (by minimizing a loss/error function). The loss function for classification is cross entropy, while the loss function for regression is mean squared error. More details on the CNN architecture used can be found in the SI (Section S3.2).

Concentration Prediction of O₃ and Cl₂

The same SVM setup was used to both regress and classify the O₃ and Cl₂ concentrations. Predictions of O₃ and Cl₂ concentrations were performed separately. The same 5-fold cross-validation is used to tune the hyperparameters. Each fold has 512 training samples (64%), 128 validation samples (16%), and 160 test samples (20%). The detailed implementation of CV and hyperparameter tuning can be seen in SI. The 3D CNN accepts as input a 3-channel tensor (RGB image) with a dimension of 150 × 50 × 50. We followed the same procedure for training and tuning hyperparameters as the 2D CNN used in the SO₂/RH study.

Data availability

The data for this work is available at https://github.com/zavalab/ML/tree/master/LC_TDA

Code availability

The source code of this study is publicly available at https://github.com/zavalab/ML/tree/master/LC_TDA

Acknowledgments

We acknowledge funding from the U.S. National Science Foundation (NSF) under BIGDATA grant IIS-1837812 and IIS-1837821.

References

1. Shah, R. R. & Abbott, N. L. Principles for Measurement of Chemical Exposure Based on Recognition-Driven Anchoring Transitions in Liquid Crystals. *Science* (1979) **293**, 1296–1299 (2001).
2. Carlton, R. J. *et al.* Chemical and biological sensing using liquid crystals. *Liq Cryst Rev* **1**, 29–51 (2013).
3. Szilvási, T. *et al.* Redox-Triggered Orientational Responses of Liquid Crystals to Chlorine Gas. *Angewandte Chemie International Edition* **57**, 9665–9669 (2018).
4. Hunter, J. T. & Abbott, N. L. Dynamics of the chemo-optical response of supported films of nematic liquid crystals. *Sens Actuators B Chem* **183**, 71–80 (2013).
5. Cao, Y., Yu, H., Abbott, N. L. & Zavala, V. M. Machine Learning Algorithms for Liquid Crystal-Based Sensors. *ACS Sens* **3**, 2237–2245 (2018).

6. Smith, A. D., Abbott, N. & Zavala, V. M. Convolutional Network Analysis of Optical Micrographs for Liquid Crystal Sensors. *Journal of Physical Chemistry C* **124**, 15152–15161 (2020).
7. Bao, N. *et al.* Sensing Gas Mixtures by Analyzing the Spatiotemporal Optical Responses of Liquid Crystals Using 3D Convolutional Neural Networks. *ACS Sens* **7**, 2545–2555 (2022).
8. Adebayo, J. *et al.* Sanity checks for saliency maps. in *Advances in Neural Information Processing Systems* vols 2018-December (2018).
9. Lee, Y. *et al.* Quantifying similarity of pore-geometry in nanoporous materials. *Nat Commun* **8**, (2017).
10. Saggar, M. *et al.* Towards a new approach to reveal dynamical organization of the brain using topological data analysis. *Nat Commun* **9**, (2018).
11. Saggar, M., Shine, J. M., Liégeois, R., Dosenbach, N. U. F. & Fair, D. Precision dynamical mapping using topological data analysis reveals a hub-like transition state at rest. *Nat Commun* **13**, (2022).
12. Rabadán, R. *et al.* Identification of relevant genetic alterations in cancer using topological data analysis. *Nat Commun* **11**, (2020).
13. Nielson, J. L. *et al.* Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nat Commun* **6**, (2015).
14. di Bernardino, E., Estrade, A. & León, J. R. A test of Gaussianity based on the euler characteristic of excursion sets. *Electron J Stat* **11**, (2017).
15. Pranav, P. *et al.* Topology and geometry of Gaussian random fields I: On Betti numbers, Euler characteristic, and Minkowski functionals. *Mon Not R Astron Soc* **485**, (2019).
16. Adler, R. J., Bobrowski, O., Borman, M. S., Subag, E. & Weinberger, S. Persistent homology for random fields and complexes. in *Borrowing strength: theory powering applications--a Festschrift for Lawrence D. Brown* vol. 6 124–144 (Institute of Mathematical Statistics, 2010).
17. Theiler, J. Estimating fractal dimension. *Journal of the Optical Society of America A* **7**, (1990).
18. Carriazo, J. G., Molina, R. & Moreno, S. Fractal dimension and energetic heterogeneity of gold-modified Al-Fe-Ce pilc's. *Appl Surf Sci* **255**, (2008).
19. Smith, A. & Zavala, V. M. The Euler characteristic: A general topological descriptor for complex data. *Comput Chem Eng* **154**, (2021).
20. Kiselev, A. D., Vovk, R. G., Egorov, R. I. & Chigrinov, V. G. Polarization-resolved angular patterns of nematic liquid crystal cells: Topological events driven by incident light polarization. *Phys Rev A* **78**, (2008).
21. Pairam, E. *et al.* Stable nematic droplets with handles. *Proc Natl Acad Sci U S A* **110**, (2013).
22. Tsung, J. W., Wang, Y. Z., Yao, S. K. & Chao, S. Y. Crystal-like topological defect arrays in nematic liquid crystal. *Appl Phys Lett* **119**, (2021).
23. Wang, Y. S., Yang, G. H., Tian, L. J. & Duan, Y. S. Stability of disclinations in nematic liquid crystals. *Commun Theor Phys* **45**, (2006).
24. Das, S. K. & Rey, A. D. Texture formation under phase ordering and phase separation in polymer-liquid crystal mixtures. *Journal of Chemical Physics* **121**, (2004).
25. Fernández-Nieves, A. *et al.* Novel defect structures in nematic liquid crystal shells. *Phys Rev Lett* **99**, (2007).
26. Muševič, I. Nematic liquid-crystal colloids. *Materials* **11**, 24 (2018).
27. Liu, Q., Senyuk, B., Tasinkevych, M. & Smalyukh, I. I. Nematic liquid crystal boojums with handles on colloidal handlebodies. *Proc Natl Acad Sci U S A* **110**, 9231–9236 (2013).
28. Kleman, M. & Lavrentovich, O. D. Topological point defects in nematic liquid crystals. *Philosophical Magazine* **86**, (2006).
29. Wang, X., Miller, D. S., Bukusoglu, E., de Pablo, J. J. & Abbott, N. L. Topological defects in liquid crystals as templates for molecular self-assembly. *Nat Mater* **15**, 106–112 (2016).

30. Lavrentovich, O. D. Topological defects in dispersed words and worlds around liquid crystals, or liquid crystal drops. *Liq Cryst* **24**, 117–126 (1998).
31. Garbovskiy, Y. A., Evans, D. R., Banerjee, P. P. & Glushchenko, A. v. Static and dynamic electro-optical properties of liquid crystals mediated by ferroelectric polymer films. *RSC Adv* **8**, (2018).
32. Pishnyak, O. P., Shiyanovskii, S. v. & Lavrentovich, O. D. Aggregation of colloidal particles in a non-equilibrium backflow induced by electrically-driven reorientation of the nematic liquid crystal. *J Mol Liq* **164**, (2011).
33. Wang, X. *et al.* Interfacial ordering of thermotropic liquid crystals triggered by the secondary structures of oligopeptides. *Chemical Communications* **51**, (2015).
34. Goncharuk, A. I., Lebovka, N. I., Lisetski, L. N. & Minenko, S. S. Aggregation, percolation and phase transitions in nematic liquid crystal EBBA doped with carbon nanotubes. *J Phys D Appl Phys* **42**, (2009).
35. Al-Zangana, S., Iliut, M., Turner, M., Vijayaraghavan, A. & Dierking, I. Properties of a Thermotropic Nematic Liquid Crystal Doped with Graphene Oxide. *Adv Opt Mater* **4**, (2016).
36. Membrillo Solis, I. *et al.* Tracking the time evolution of soft matter systems via topological structural heterogeneity. *Communications Materials* **2022 3:1** **3**, 1–11 (2022).
37. Slotte, P. A., Berg, C. F. & Khanamiri, H. H. Predicting Resistivity and Permeability of Porous Media Using Minkowski Functionals. *Transp Porous Media* **131**, (2020).
38. Vogel, H. J., Weller, U. & Schlüter, S. Quantification of soil structure based on Minkowski functions. *Comput Geosci* **36**, 1236–1245 (2010).
39. Monetti, R. *et al.* Assessment of the human trabecular bone structure using Minkowski Functionals. in *Medical Imaging 2009: Biomedical Applications in Molecular, Structural, and Functional Imaging* vol. 7262 (2009).
40. Bubenik, P. Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research* **16**, (2015).
41. Szilvási, T. *et al.* The role of anions in adsorbate-induced anchoring transitions of liquid crystals on surfaces with discrete cation binding sites. *Soft Matter* **14**, 797–805 (2018).
42. Bao, N. *et al.* Ordering Transitions of Liquid Crystals Triggered by Metal Oxide-catalyzed Reactions of Sulfur Oxide Species. *J Am Chem Soc* **144**, 16378–16388 (2022).
43. Prechelt, L. Early stopping - But when? *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **7700 LECTU**, 53–67 (2012).
44. Mantz, H., Jacobs, K. & Mecke, K. Utilizing Minkowski functionals for image analysis: A marching square algorithm. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P12015 (2008).
45. Mecke, K. R. Integral Geometry In Statistical Physics. *Int J Mod Phys B* **12**, 861–899 (1998).
46. Hadwiger, H. *Vorlesungen Über Inhalt, Oberfläche und Isoperimetrie*. (Springer-Verlag, 1957).
47. Hashemi, S. M. *et al.* Fractal nematic colloids. *Nat Commun* **8**, 1–9 (2017).
48. Mandelbrot, B. B. & van Ness, J. W. Fractional Brownian Motions, Fractional Noises and Applications. *SIAM Review* **10**, 422–437 (1968).
49. Stanley, H. E. Application of fractal concepts to polymer statistics and to anomalous transport in randomly porous media. *J Stat Phys* **36**, 843–860 (1984).
50. Martin, J. E. & Schaefer, D. W. Dynamics of fractal colloidal aggregates. *Phys Rev Lett* **53**, 2457–2460 (1984).
51. Yu, B. & Li, J. Some fractal characters of porous media. *Fractals* **9**, 365–372 (2001).
52. Lévy, M. P. Le Mouvement Brownien Plan. *American Journal of Mathematics* **62**, 487–550 (1940).
53. Hurewicz, W. & Wallman, H. Dimension Theory (PMS-4), Volume 4. in *Dimension Theory (PMS-4), Volume 4* (Princeton university press, 2015).

54. Mandelbrot, B. B. Measures of Fractal Lacunarity: Minkowski Content and Alternatives. in *Fractal Geometry and Stochastics* 15–42 (Springer, 1995).
55. Ivanovici, M., Richard, N. & Decean, H. Fractal dimension and lacunarity of psoriatic lesions- A colour approach. in *Proceedings of the 2nd WSEAS International Conference on Biomedical Electronics and Biomedical Informatics, BEBI '09* (2009).
56. Mecke, K. R. Additivity, Convexity, and Beyond: Applications of Minkowski Functionals in Statistical Physics. in *Statistical Physics and Spatial Statistics: The art of analyzing and modeling spatial structures and pattern formation* 111–184 (Springer, 2000).