International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 2, Number 3, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v2n3.08
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Research on Visualization Techniques Based on High-Dimensional Attribute Data of Chemical Materials

Xiaokun Tian*

College of Computer Science and Engineering, Sichuan University of Science & Engineering, Yibin, China

*Corresponding Author: Xiaokun Tian

ABSTRACT

Machine learning is currently used to analyze and predict materials in fully automated applications in the materials domain. However, human interpretation and involvement are limited due to the opaque nature of these algorithms. This study explores for the first time the use of semi-human and semi-automated analysis techniques in materials data research. The analytics approach combines machine learning techniques with data visualization and analysis so that the analysis of material properties is dominated by human intelligence, with machine learning techniques as a complementary method. This helps experts understand the associations between material properties from a broader perspective, enabling them to generalize the concept of micro-properties to macro-properties.

KEYWORDS

High-dimensional data; Dimensionality reduction; Visual analytics; Machine learning

1. INTRODUCTION

The field of materials has seen a further application of computer technology due to the development of big data technology [1]. Recently, research combining machine learning and deep learning with materials has yielded some promising results; the content primarily focuses on the prediction [3], classification [4], and search of material performance [5]. Some scholars have also used the classification model to investigate the relationship between the material's microstructure and macroscopic features [6]. The following are some of the drawbacks of the machine learning applications now available in the materials field:

(1) Inadequate interpretability of machine learning techniques: It might be challenging to understand how a training model processes and produces outputs, even when it performs exceptionally well in prediction or classification on the dataset.

(2) Model Training takes a lot of effort, and their effectiveness is limited to particular issues; they are not sufficiently generalizable for useful study.

(3) The training process lacks the human input and intellectual guidance of subject experts, instead relying on automated tweaking and trial-and-error.

These elements can be better complemented by visualization techniques, which is a crucial part of artificial intelligence. Data visualization techniques can be used to analyze data distribution, comparison, composition, and relationships [7, 8]. Chemical material data is challenging to examine using conventional approaches due to its high dimensionality and large volume.

2. RELATED WORK

Scholars have recently conducted high-dimensional data visualization research in the fields of geographic environment [9], ecology [10], and mass spectrometry imaging analysis [11]. Currently, the application of this technology to chemical materials is less studied. The majority of the research uses interactive techniques in addition to the more generalized scatterplot, dimensionality reduction, parallel coordinates, and other visualization techniques to accomplish step-by-step visual analysis; nonetheless, there are certain drawbacks: single-view visualization, essentially depending on the brush selection of local selection; inadequate scalability, making it challenging to raise or decrease the data dimensions and real-time graphic; absence of multi-view integrated analysis; and so on. To assist with the analysis of chemical material properties, it would be helpful that current techniques be expanded upon and improved whenever feasible, in addition to being applied to chemical material data. There are five primary groups into which high-dimensional data visualization techniques fall.

2.1. Axis-based Methods

Radar plots, parallel coordinate plots, and scatterplot matrices are the three primary types of axis-based visualization techniques. These techniques map the coordinate axes onto the data dimensions and are very intuitive. In essence, a scatterplot matrix is a collection of scatterplots. To compare the relationship between variables, any pair of variables can be plotted on a scatterplot within a right-angled coordinate system. However, there are $n(n-1)/2$ combinations of n dimensional data, and as the number of dimensions increases [12], the complexity of the analysis increases correspondingly.

Depending on the number of data dimensions, parallel coordinate charts encode data properties with a matching number of axes. A material's properties are represented by each coordinate axis, and all of the attributes are recorded by the folded line created by joining the data points on each axis. This allows material properties to be shown without any information being lost. Regarding materials with n dimensions, there are $n!$ possible arrangements for the axes [13], and visual analysis techniques based on parallel coordinate graphs typically demand attention to this ordering. The choice of the number and order of the dimensions to be given is an inevitable issue when employing parallel axes since the information presented might be considerably different if the axes are in a different sequence. A common limitation for users is that they can only analyze the axes from among their neighbors, that is, the correlation between the three variables they represent through the three neighboring axes. However, choosing the three variables at random necessitates an ordering of the axes, which is not well studied at the moment.

Multiple radial axes are used at the same origin to encode multidimensional attributes in radar charts. Each piece of data is plotted as a closed graph through the attribute points on the axes, allowing for the visualization of multidimensional data. However, when the amount of data is large, multiple graphs will cause serious occlusion and make it difficult to filter the data interactively [14]. Additionally, a graph with an excessive number of radial axes will be extremely cluttered and challenging to use with chemical materials when the amount of data is small but the dimensionality is high.

2.2. Symbol-based Methods

Symbol-based techniques mostly employ symbols, such as those found on speed limit signs for vehicles, to encode information. When working on a specific subject, visualization experts can use graphical qualities like shape, color, size, opacity, position, and so on to encode high-dimensional information. They can also combine numerous small design symbols into a single symbol to represent different aspects of a material's attributes [15]. Because there are only so many coding channels available, the primary focus of this coding strategy is on how to present multidimensional data in a way that domain experts can easily understand. As a result, the properties that matter most to domain

experts should be encoded into the symbols. The challenge of coding is to better represent the data to be studied and to incorporate domain experts' knowledge to enhance interpretability since the symbols for visualization must be created by the researchers themselves based on the requirements of the data to be analyzed.

2.3. Pixel-based Methods

By encoding the information of all dimensions through dense pixel points—where one pixel represents one dimension—and offering a different main view or subview for each dimension [16], pixel-based visualization techniques are presented as a way to present all the data with extremely high dimensionality. While it is feasible to display the data compactly on the screen, the small size of the data points makes exploration based on visual interaction challenging. This presentation, similar to data stored in a database, makes it difficult to analyze the correlations between multiple dimensions from the view.

2.4. Hierarchy-based Methods

While a large number of dimensions makes it difficult to navigate the data space and causes problems with scalability for visual mapping, a hierarchical organization of dimensions can directly reveal dimensional relationships, reduce dataset complexity, and have an easy-to-expand hierarchical structure. Data dimensions can be interactively evaluated for intra- and inter-data comparisons using tree or other topologies [17], allowing for hierarchical representation. The organization of the hierarchy of dimensions is difficult, the representation takes up a lot of plotting area, and data mapping is time-consuming.

2.5. Animation-based Methods

Animating correlated views in a multi-view analysis is one of the high-dimensional data visualization techniques that make use of animation transformations to improve the perception of point and structure correspondences between multiple correlated graphs. The associated views are updated in a coordinated fashion when a portion of the correlated graph is manipulated [18]. An appropriate animation design must incorporate the knowledge of both visualization professionals and subject experts, just like the symbol-based approach does.

Visualization and visual analysis technology in the field of chemical materials has been studied initially, mainly focusing on (1) Using visualization and dimensional reduction techniques to analyze the similarities between chemical compounds [19]. (2) Realizing digital factories that mimic actual manufacturing processes and direct production by utilizing IoT and AI technology [20]. (3) Use in conjunction with high-throughput computing to visually represent the outcomes of material classification or machine learning predictions to support user interpretation [21]. The visualization of the microscopic features of materials and their macroscopic application properties remains mostly unexplored research area.

3. METHOD

Data on chemical materials commonly involves both numerical and category properties. For instance, the fluoroelastomer's mechanical characteristics and crystalline form. This research uses two widely used data dimensionality reduction algorithms for visualization to give a general overview of the material's features. The data distribution is shown as a scatterplot in two dimensions. For the investigation presented in this work, the dimensionality reduction method based on nonlinear manifolds is more appropriate due to the intricacy of the internal relationship of chemical material data. Both the t-SNE and UMAP techniques will be used in the data dimensionality reduction process.

Years of study have shown that the t-SNE approach generally provides better visualization; nevertheless, because of its square-level time complexity and memory space occupation, it is not suitable for situations involving very large data sizes and high dimensionality. Although UMAP may not preserve the local structure as well as t-SNE, it can handle data with up to 10,000 dimensions because of its almost linear level of time and space complexity.

3.1. The t-sne Dimensionality Reduction Algorithm

t-SNE (t-distributed Stochastic Neighbor Embedding) is a statistical method used for visualizing high-dimensional data by mapping each data point to a location in a two- or three-dimensional map. Unlike linear methods such as PCA (Principal Component Analysis), t-SNE can handle data that cannot be separated by straight lines. It projects high-dimensional data into a low-dimensional space (usually 2D or 3D) while preserving the structure of the data. t-SNE has been used in various domains, including genomics, natural language processing, music analysis, and biomedical signal processing.

Denote by $X \in R^{N\times P}$ the original data set, and $Y \in R^{N\times 2}$ the dataset mapped from the original data to the two-dimensional space. For each observation value $i \in [N]$, use x_i and y_i to denote the corresponding data in X and Y, i.e., the original data of a single entry and its mapping in two-dimensional space, respectively. The distance function $Distance_{i,j}$ is used to represent the distance metric of R^P between the data values X_i and X_j. The definition of the distance metric varies by algorithm. The t-SNE algorithm defines the conditional probability for two data points i and j as:

$$p_{j|i} = \frac{e^{-\|x_i - x_j\|^2/2\sigma_i^2}}{\sum_{k\neq i} e^{-\|x_i - x_k\|^2/2\sigma_i^2}} \tag{1}$$

Where σ_i is found by dichotomizing the following equation:

$$Perplexity = 2^{-\sum_j p_{j|i}\log_2 p_{j|i}} \tag{2}$$

The user-defined parameter $Perplexity$ monotonically increases with σ_i, i.e., the larger the $Perplexity$, the more uniform the probability distribution. Define the symmetric probability as:

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n} \tag{3}$$

This probability does not depend on the decision variable Y but is derived from X. Then define the probability:

$$q_{ij} = \frac{(1+\|y_i - y_j\|^2)^{-1}}{\sum_{k\neq l}(1+\|y_k - y_i\|^2)^{-1}} \tag{4}$$

This probability is a function of the decision variable Y. Define the loss function as：

$$\sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{5}$$

In the dimensionality reduction computation, Y is first initialized with a stochastic initialization of the initial state using a multivariate normal distribution $N(0,10^{-4}I)$, where I denotes a two-dimensional unit matrix. Then several iterations of computation are performed, applying momentum gradient descent, with the momentum term set to 0.5 for the first 250 iterations and 0.8 for the subsequent 750 iterations, with the learning rate initially set to 100 and adaptive updating.

Overall, for each data, i, t-SNE defines a relative distance metric $p_{i|j}$ and tries to find a low embedding where the relative distance $q_{i|j}$ in the low-dimensional space matches the high-dimensional distance, whereby the preservation of high-dimensional local structure to the low-dimensional local structure is realized. However, the global true structure is often not realistically represented because the loss function has less of an impact on the high-dimensionally farther away and low-dimensionally nearer data. This means that in the two-dimensional plane, clusters with small overall differences may be farther away than those with larger differences.

3.2. The UMAP Dimensionality Reduction Algorithm

UMAP, which stands for Uniform Manifold Approximation and Projection, is a dimensionality reduction technique that is particularly effective for visualizing high-dimensional data. Unlike other popular methods like t-SNE, UMAP maintains much of the global data structure, making it useful not only for visualization but also for general non-linear dimension reduction. The algorithm operates under three main assumptions:

(1) The data is uniformly distributed on a Riemannian manifold.

(2) The Riemannian metric is locally constant, or can be approximated as such.

(3) The manifold is locally connected.

From these assumptions, UMAP constructs a high-dimensional graph to represent the data and then optimizes a low-dimensional graph to be as structurally similar as possible. This process involves a concept known as a "fuzzy simplicial complex" to model the manifold with a fuzzy topological structure. The goal is to find a low-dimensional projection of the data that preserves this topological structure as closely as possible. The algorithm is derived geometrically and the steps for the construction of the graph are as follows:

For each data value, find the k closest neighboring points for a given distance metric such as Euclidean distance, and compute the minimum positive distance ρ_i from the current data point to the neighbor i. Establish the equation:

$$\log_2(k) = \sum_{j=1}^{k} \exp\left(\frac{-\max\{0, Distance_{i,i} - \rho_i\}}{\sigma_i}\right) \tag{6}$$

A scale parameter σ_i is computed for each data point i. σ_i is used to normalize the distance between each data value and its neighboring points, which is used to maintain a relatively high-dimensional proximity. Define the weight function as:

$$w(X_i, X_j) = \exp\left(\frac{-\max(0, Distance_{i,j} - \rho_i)}{\sigma_i}\right) \tag{7}$$

Accordingly, a weighted graph G is defined, whose vertices are data values, and for each of its edges (i, j) denotes that X_i is the nearest neighbor of X_j, and vice versa. For a particular edge (i, j), its symmetric weight is denoted as:

$$\overline{w}_{i,j} = w(X_i, X_j) + w(X_j, X_i) - w(X_i, X_j) \cdot w(X_j, X_i) \tag{8}$$

Optimize the graph layout after constructing the high-dimensional spatial graph sequentially: first, initialize Y using the spectral embedding, iterate over the edges in graph G, and apply gradient descent on each edge (i, j) of data point i: Apply an attraction to data point I in the low-dimensional space:

$$F_{i,j} = y_i + \alpha \cdot \frac{-2ab\left\|y_i - y_j\right\|_2^{2(b-1)}}{1 + a(\left\|y_i - y_j\right\|_2^2)^b} \overline{w}_{i,j}(y_i - y_j) \tag{9}$$

So that it is close to the data point j. The hyperparameters a and b are obtained by fitting the function:

$$\left(1 + a\left(\left\|y_i - y_j\right\|_2^2\right)^b\right)^{-1} \tag{10}$$

Fitting the (10) function to a non-normalized weight function to compute the values of a and b:

$$\exp(-\max\left\{0, Distance_{i,j} - \rho_i\right\}) \tag{11}$$

The calculation above yields a smooth approximation. The hyperparameter α denotes the learning rate. Then a repulsion force is added to data point i to push it away from data point k, which is a non-adjacent vertex chosen by random sampling, i.e., from an edge (i,k) that is not in G. The repulsion force is denoted:

$$F_{i,k} = y_i + \alpha \cdot \frac{b}{\left(\varepsilon + \left\|y_i - y_k\right\|_2^2\right)\left(1 + a\left(\left\|y_i - y_k\right\|_2^2\right)^b\right)}\left(1 - \overline{w}_{i,k}\right)\left(y_i - y_k\right) \tag{12}$$

Where ε takes the value of a very small constant. First, the algorithm builds a weighted graph of nearest neighbors, in which the weights stand for probability distributions. This can be thought of as a stochastic gradient descent on individual data during the optimization phase. Because there is no randomness aspect in its initialization process, it is more stable than t-SNE.

4. VISUALIZATION

The visual analysis system serves researchers in the field of chemical materials by plotting the material data into a three-part view, namely, low-dimensional mapping, parallel coordinate plot, and scatter plot matrix, and realizing the interactive functions as an operation panel and binding to graphical elements, respectively. Since the low-dimensional mapping plot deals with data from the entire dataset, it is used as the first view, and two-dimensionality reduction visualization algorithms, t-SNE, and UMAP, are provided as options. Analysts can clearly observe the trends of the data clusters on this view and can box in some of the data points to be used as input data for parallel coordinate plots and scatter plot matrices.

The visualization results of 2D mapping of chemical material properties using UMAP are shown in Figure 1. Each vertex represents a material that is color coded according to a certain class of material properties. The distance between vertices on the graph represents the similarity of various materials, and this dimensionality reduction approach can show more global information in two dimensions. A global overview of material properties is provided by the view, which uses visual channels including axes, distances, and colors. This view allows the user to perform an initial screening of the material and select data of interest for further visual analysis and research.

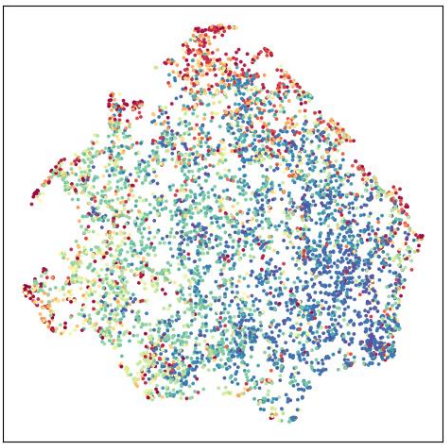

Figure 1. UMAP two-dimensional mapping graph

Following the selection of the data to be studied, the data is processed and sorted by the back-end server. The data is then assigned axes in triads with contribution and variable relevance as the basis for sorting in the parallel coordinate chart, and different data is coded with different colored folds. This prevents occlusion during detailed study because the analyst can selectively select the data of interest, reducing the visibility of the remaining data to highlight the current data. The scatterplot can be plotted in a Cartesian coordinate system, allowing the analyst to gradually explore the high-dimensional material data through the interaction of multiple views. Figure 2 illustrates the coordination of the two views. Two axes can also be selected as the two dimensions of the scatterplot matrix in order to specifically study the relationship between two variables.

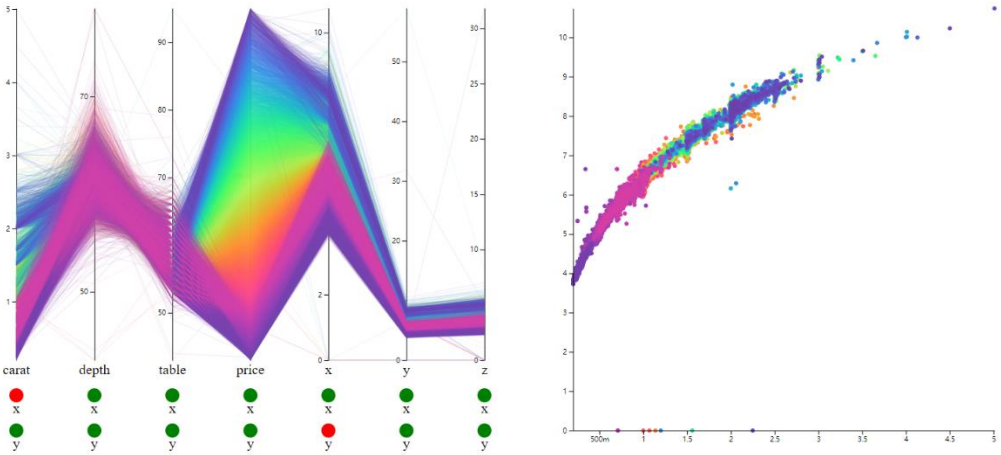

Figure 2. Schematic of the coordination of parallel coordinates with the scatter plot matrix

To provide more flexibility for researchers when evaluating data on chemical materials. An analysis of correlations between several variables can be done by the user arbitrarily combining tuples of three coordinates in a parallel coordinate plot, which can be manually ordered by the user via the visualization system. To create a scatter plot, the user can also choose any two variables to use as the x and y axes. The study of the correlation between two variables is facilitated by the scatterplot matrix.

5. SUMMARY

Currently, machine learning is used in fully automated applications of AI in the materials domain to analyze and forecast materials. However, because these algorithms are opaque, human interpretation and involvement is limited. This is the first investigation of semi-artificial and semi-automated analysis technology in material data research. The analysis method that combines machine learning technology with data visualization and analysis makes the analysis of material properties dominated

by human intelligence and machine learning technology as an auxiliary method. It can help specialists understand the connection between the material's attributes from a wider perspective, allowing them to extrapolate the idea of micro-properties to macro-performance.

This study examines the state of high-dimensional data visualization techniques currently in use in the analysis of chemical materials. Two nonlinear dimensionality reduction-based algorithms are used to map the high dimensional material data to a 2D screen, giving rise to a global view. Next, machine learning techniques and visual interaction are used to design a multi-view coordinated visual analysis system for chemical materials. With the use of artificial intelligence and visualization technologies, the system gives the conventional chemical industry the ability to effectively encode high-dimensional properties of chemical materials into visual channels, giving subject matter experts a useful tool for analysis.

ACKNOWLEDGEMENTS

This research was supported by the Graduate Innovation Fund project of Sichuan University of Science & Engineering, Project Number: Y2022181.

REFERENCES

[1] Morgan, D., & Jacobs, R. (2020). Opportunities and challenges for machine learning in materials science. Annual Review of Materials Research, 50, 71-103.

[2] Esposito, A., Lappa, M., Pagliara, R., Spada, G. (2022). A Mixed Radiative-Convective Technique for the Calibration of Heat Flux Sensors in Hypersonic Flow. FDMP-Fluid Dynamics & Materials Processing, 18(2), 189–203. https://doi.org/10.1146/annurev-matsci-070218-010015

[3] Zhao, Y., Al-Fahdi, M., Hu, M., Siriwardane, E. M., Song, Y., Nasiri, A., & Hu, J. (2021). High-throughput discovery of novel cubic crystal materials using deep generative neural networks. Advanced Science, 8(20), 2100566. https://doi.org/10.1002/advs.202100566

[4] Khushaba, R. N., & Hill, A. J. (2022). Radar-based materials classification using deep wavelet scattering transform: A comparison of centimeter vs. millimeter wave units. IEEE Robotics and Automation Letters, 7(2), 2016-2022. https://doi.org/10.1109/LRA.2022.3143200

[5] Patel, D., Shah, D., Modi, P., & Roy, M. (2022). Artificial intelligence powered material search engine. Materials Today: Proceedings, 57, 11-13. https://doi.org/10.1016/j.matpr.2022.01.120

[6] Acosta, C. M., Ogoshi, E., Souza, J. A., & Dalpian, G. M. (2022). Machine learning study of the magnetic ordering in 2D materials. ACS Applied Materials & Interfaces, 14(7), 9418-9432.

[7] Fang, Y., Zhang, Q., Yang, H., Zhuang, X., Deng, S., Zhang, W., ... & Chen, H. (2022, June). Molecular contrastive learning with chemical element knowledge graph. In Proceedings of the AAAI conference on artificial intelligence (Vol. 36, No. 4, pp. 3968-3976). https://doi.org/10.1609/aaai.v36i4.20313

[8] Holzinger, A., Dehmer, M., Emmert-Streib, F., Cucchiara, R., Augenstein, I., Del Ser, J., ... & Díaz-Rodríguez, N. (2022). Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Information Fusion, 79, 263-278. https://doi.org/10.1016/j.inffus.2021.10.007

[9] Liu, J., Wan, G., Liu, W., Li, C., Peng, S., & Xie, Z. (2023). High-dimensional spatiotemporal visual analysis of the air quality in China. Scientific Reports, 13(1), 5462. https://doi.org/10.1038/s41598-023-31645-1

[10] Mitku, A. A., Zewotir, T., North, D., & Naidoo, R. N. (2020). Exploratory data analysis of adverse birth outcomes and exposure to oxides of nitrogen using interactive parallel coordinates plot technique. Scientific reports, 10(1), 7363. https://doi.org/10.1038/s41598-020-64471-w

[11] Schwarz, C., Buchholz, R., Jawad, M., Hoesker, V., Terwesten-Solé, C., Karst, U., ... & Faber, C. (2022). Fingerprints of element concentrations in infective endocarditis obtained by mass spectrometric imaging and t-Distributed Stochastic neighbor embedding. ACS infectious diseases, 8(2), 360-372. https://doi.org/10.1021/acsinfecdis.1c00485

[12] Li, Y., Han, H., Shan, S., & Chen, X. (2023). Disc: Learning from noisy labels via dynamic instance-specific selection and correction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 24070-24079). https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.02305

[13] Heinrich, J., & Weiskopf, D. (2013). State of the Art of Parallel Coordinates. Eurographics (state of the art reports), 95-116. http://dx.doi.org/10.2312/conf/EG2013/stars/095-116

[14] Duan, R., Tong, J., Sutton, A. J., Asch, D. A., Chu, H., Schmid, C. H., & Chen, Y. (2023). Origami plot: a novel multivariate data visualization tool that improves radar chart. Journal of clinical epidemiology, 156, 85-94. https://doi.org/10.1016/j.jclinepi.2023.02.020

[15] Du, M., & Yuan, X. (2021). A survey of competitive sports data visualization and visual analysis. Journal of Visualization, 24, 47-67. https://doi.org/10.1007/s12650-020-00687-2

[16] Kammer, D., Keck, M., Gründer, T., Maasch, A., Thom, T., Kleinsteuber, M., & Groh, R. (2020). Glyphboard: Visual exploration of high-dimensional data combining glyphs with dimensionality reduction. IEEE transactions on visualization and computer graphics, 26(4), 1661-1671. https://doi.org/10.1109/TVCG.2020.2969060

[17] Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The science of visual data communication: What works. Psychological Science in the public interest, 22(3), 110-161. https://doi.org/10.1177/15291006211051956

[18] Yao, L., Bezerianos, A., Vuillemot, R., & Isenberg, P. (2022). Visualization in motion: A research agenda and two evaluations. IEEE Transactions on Visualization and Computer Graphics, 28(10), 3546-3562. https://doi.org/10.1109/TVCG.2022.3184993

[19] Vodka, O., Zadorozhniy, I., & Lavshhenko, R. (2020, September). Application algorithms of nonlinear dimensionality reduction to material database visualization. In 2020 IEEE 15th International Conference on Computer Sciences and Information Technologies (CSIT) (Vol. 1, pp. 100-104). IEEE. https://doi.org/10.1109/CSIT49958.2020.9321965

[20] Örs, E., Schmidt, R., Mighani, M., & Shalaby, M. (2020, June). A conceptual framework for AI-based operational digital twin in chemical process engineering. In 2020 IEEE international conference on engineering, technology and innovation (ICE/ITMC) (pp. 1-8). IEEE. https://doi.org/10.1109/ICE/ITMC49519.2020.9198575

[21] Ludwig, A. (2019). Discovery of new materials using combinatorial synthesis and high-throughput characterization of thin-film materials libraries combined with computational methods. npj Computational Materials, 5(1), 70. https://doi.org/10.1038/s41524-019-0205-0