Supplementary Materials

APPENDIX I. METHODS

Using rectangular shapes, primary colors in bold blocks, and a strict grid structure, Mondrian is remembered for achieving a universal and harmonious visual language with his abstraction framework [1]. Harnessing these spatial features possesses an enormous potential for providing an intuitive representation of intricate scientific data. While this approach enhances comprehension, its implementation demands multiple stages of data processing, particularly for complex biological networks. In this section, we described the detailed development process of our visualization tool, how we leveraged language modeling to generate pathway embeddings, and subsequently address the essential data processing steps required for creating the visualizations.

A. Language Model Embeddings for Pathways

*1) **Selection of Language Models**:* We utilized a diverse set of language models to generate pathway embeddings which captures intricate neighborhood and interaction relationships. The models utilized include all-MiniLM-L6-v2 (MiniLM) [2] and all-mpnet-base-v2 (mpnet) [3], whose embeddings were produced using the Sentence Transformer library [4]. Additionally, LLM2Vec-Mistral-7B-Instruct-v2-mntp (Mistral-7B) [5] and LLM2Vec-Meta-Llama-3-8B-Instruct-mntp (Llama-3-8B) [6] were used, utilizing the llm2vec framework, which transforms large, decoder-only language models into bidirectional encoders. This transformation involves enabling bidirectional attention, masked next token prediction, and unsupervised contrastive learning, effectively creating rich, context-aware text embeddings [7].

TABLE I
COMPARISON OF DIFFERENT LANGUAGE MODELS

Model	# Params.	Vocab. Size	Embed. Dim.
MiniLM [2]	22M	30,522	384
mpnet [3]	109M	30,527	768
Mistral-7B [5]	7B	32,000	4096
Llama-3-8B [6]	8B	128,256	4096

In contrast, the Sentence Transformers (a.k.a. SBERT) adapts the pre-trained BERT architecture into siamese and triplet network structures to derive semantically meaningful embeddings efficiently, greatly reducing computational demands while preserving high accuracy in semantic similarity tasks [4]. The specific attributes of these models, such as the number of parameters, vocabulary size, and embedding dimensions are detailed in the Table I to contextualize their use in our analysis.

*2) **Prompt Generation Techniques**:* We have generated the prompts to create the pathway embeddings in two phases. In the first phase, we used the Meta-Llama-3.1-8B-Instruct model to summarize the descriptions of the pathways, enhancing clarity and precision. Initially, the model removes alphanumeric identifiers, URLs, and external references to focus on the essence of the content. Then, for extensive descriptions, it distills the text to emphasize key functions and the pathway's biological significance. From brief or one-liner descriptions, we synthesized concise summaries grounded in recent scientific literature, avoiding speculative content. The output encapsulates the summarized pathway description in a coherent paragraph not exceeding 300 words.

In the second phase, we generate structured prompts across four distinct categories, each targeting a specific aspect of pathway data, designed to be fed into language models listed in Table I for subsequent embedding generation.

Type 1. Gene Symbol. Lists up to 100 gene symbols sorted by their significance in the pathway.

Type 2. Gene Description. Augments gene symbols from Type 1 with it's descriptions.

Type 3. Pathway Name. Simplifies the prompt using the pathway name only.

Type 4. Pathway Description Summary. Utilizes concise summaries generated in the first phase.

These prompts are crafted to guide the language models in generating high-quality embeddings that accurately reflect the complex dynamics of genetic and pathway interactions. Examples of the prompt generation techniques for both of the phases can be found in the separated prompt generation supplementary file.

*3) **Efficacy Evaluation of Embeddings**:* We employed the pathway embeddings along with their two-dimensional projections (UMAP and t-SNE) in a multi-class classification setting to evaluate the effectiveness of the embeddings. A standard Support Vector Machine (SVM) from Scikit-learn library [8] was applied without any hyper-parameter tuning to assess the inherent data patterns and structural properties.

Pathway annotations were sourced from WikiPathways [9], typically characterized by low-level details. We consolidated these detailed annotations into broader categories—such as classic metabolic pathways, disease pathways, signaling pathways, and regulatory pathways—using ontology data from BioPortal [10]. In instances where pathways corresponded to multiple categories, we assigned a single label randomly to mitigate overlaps.

B. Mondrian Map Generation Algorithm

Essential attributes such as area, color, and central point coordinates are required to represent each entity as a rectangle whereas, the relationships between the entities are illustrated using lines that link the corners of the corresponding rectangles. The definition and significance of these attributes are

subject to variability contingent upon the specific use case and the underlying scientific inquiry. In our case study, area of the rectangles are proportional to fold change of the pathways and the color depends upon fold change and adjusted p-values. For intuitive understanding of the Mondrian Map generation algorithm, we have depicted the key steps in Figure 1. The procedure to create a Mondrian Map is described as follows:

Step 1. Creating Canvas. Initialize an empty canvas featuring a grid system with the central points (x, y) of all entities, depicted in Fig. 1(a).

Step 2. Visualizing Entity Attributes. Utilizing the central points and the designated area, rectangles are generated by approximating the number of rows and columns of grid-blocks surrounding the points and colored according to other attributes (Fig. 1(b)). The approximation of the point and area is deliberately imprecise to maintain adherence to the grid system.

Step 3. Visualizing Entity Relationships. Relationships between entities are established by identifying the two closest vacant corners of their rectangles and drawing red straight lines (Type 1) for vertically or horizontally aligned entities, or L-shaped lines (Type 2) for others (Fig. 1(c)).

Step 4. Mimicking the Art Style. The last step focuses on the art-style, drawing vertical and horizontal gray lines from L-shaped lines (Fig. 1(c)) and each vacant rectangle corners and retaining only the shorter line to enhance visual clarity depicted in Fig. 1(d). These lines are for illustration only and hold no significance in the analysis.

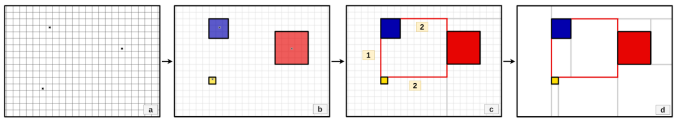

Fig. 1. Implementing Mondrian Map involves 4 key steps: (a) Initiating a grid-based canvas with specified point coordinates, (b) generating rectangles to symbolize entities, (c) illustrating red lines to establish relationships between the entities, and (d) populating the canvas with additional gray lines.

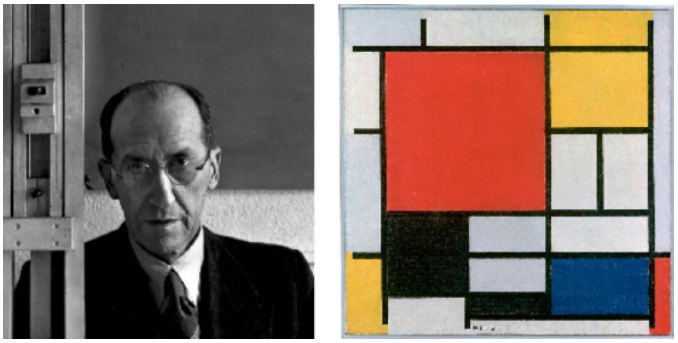

Fig. 2. Mondrian's abstraction. Left: Piet Mondrian (1872–1944), photographed by Arnold Newman in 1942. Right: "Composition with Large Red Plane, Yellow, Black, Grey and Blue" by Mondrian in 1921.

APPENDIX II. RESULTS

A. Data Preparation for Case Study

TABLE II
NUMBER OF UP AND DOWN-REGULATED GENES PER PATIENT PROFILES FROM SECTION II (C) (3).

Profile	# Up-regulated Genes	# Down-regulated Genes
Baseline R1/TP	2409	362
Baseline R2/TP	3808	315
Aggressive R1/TP	745	923
Aggressive R2/TP	985	1101
Non-aggressive R1/TP	1727	3208
Non-aggressive R2/TP	2842	2583

TABLE III
NUMBER OF PAGS AND M-TYPE PAG-TO-PAG RELATIONSHIPS FROM SECTION II (C) (4).

Profile	# PAGs	# PAG-to-PAG relationships
Baseline R1/TP	302	1488
Baseline R2/TP	533	6220
Aggressive R1/TP	117	208
Aggressive R2/TP	140	260
Non-aggressive R1/TP	460	5520
Non-aggressive R2/TP	560	7602

B. Embedding Evaluation

We experimented with four language models using four different prompts to generate embeddings. Our analysis indicates that the t-SNE projection of Llama-3-8B embeddings, derived from the pathway description summaries, provides the most accurate projection. While Llama-3-8B embeddings based on pathway names achieved the highest classification accuracy at 73%, slightly surpassing the 72% accuracy of the description summaries, the latter's richer data content led us to use its 2D coordinates for pathway localization. Figure 4 compares the 2D t-SNE projections across all methods, highlighting distinctive clusters, particularly in the Mistral-7B and Llama-3-8B embeddings. The same goes for UMAP in Fig. 5, however, it is more converged compared to t-SNE projection.

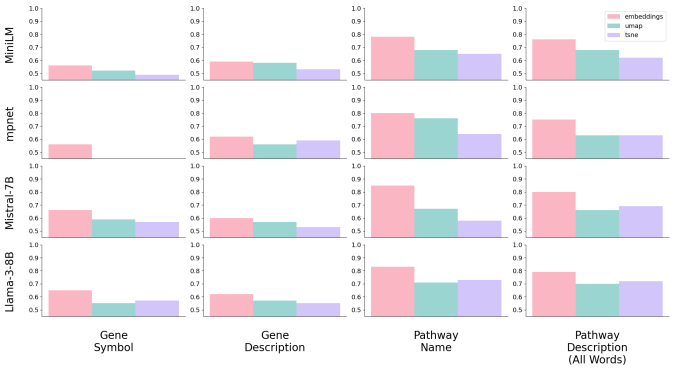

Fig. 3. Classification Accuracy for Embeddings and Projections.

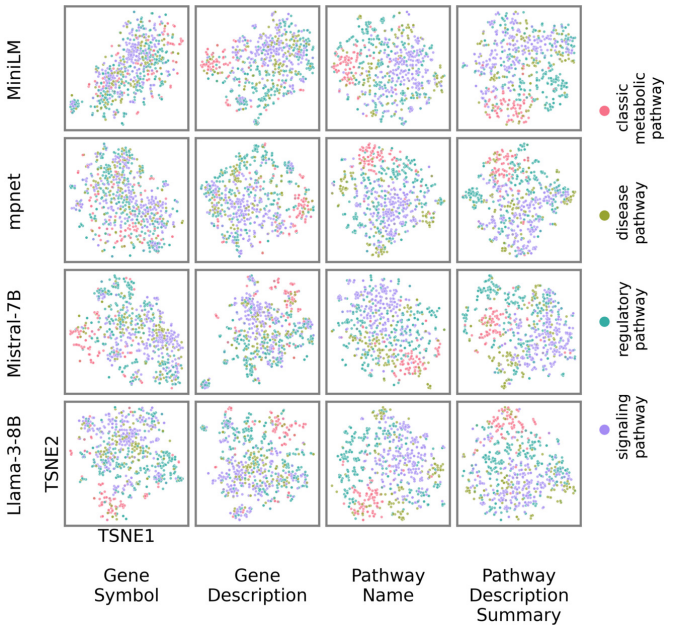

Fig. 4. Comparison of t-SNE projection for the language model embeddings with different prompting strategies.

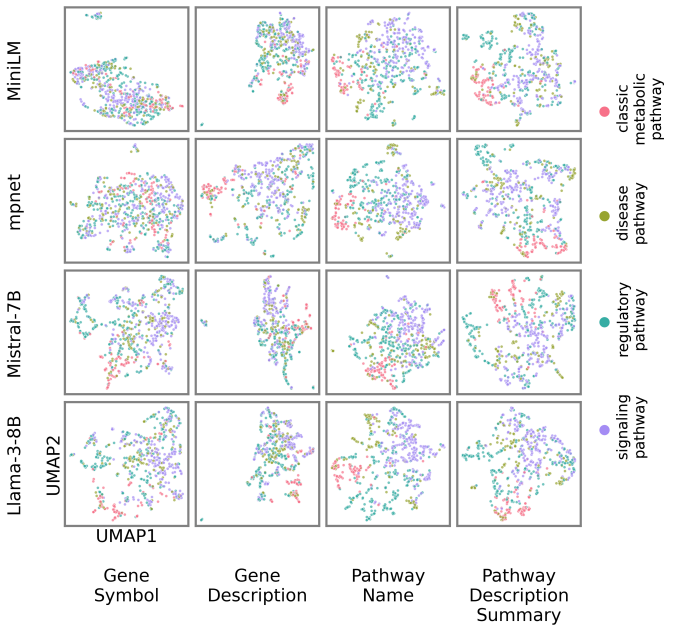

Fig. 5. Comparison of UMAP projection for the language model embeddings with different prompting strategies.

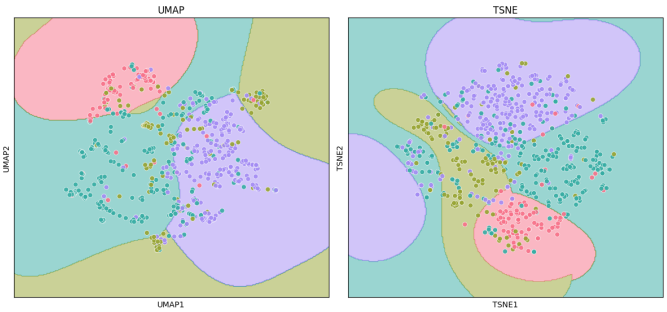

Fig. 6. model: Llama-3-8B, type: Pathway Name (embedding=0.83, umap=0.71, tsne=0.73)

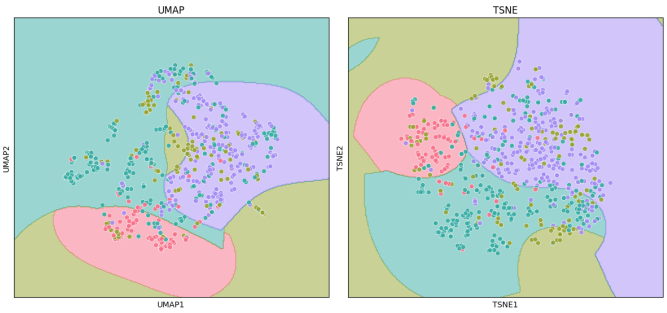

Fig. 7. model: Llama-3-8B, type: Pathway Description Summary (embedding=0.79, umap=0.7, tsne=0.72)

References

[1] C. Blotkamp, *Mondrian: The art of destruction*. Reaktion Books, 2001.

[2] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5776–5788, 2020.

[3] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "Mpnet: Masked and permuted pre-training for language understanding," *Advances in neural information processing systems*, vol. 33, pp. 16 857–16 867, 2020.

[4] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[5] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[6] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[7] P. BehnamGhader, V. Adlakha, M. Mosbach, D. Bahdanau, N. Chapados, and S. Reddy, "Llm2vec: Large language models are secretly powerful text encoders," *arXiv preprint arXiv:2404.05961*, 2024.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[9] A. Agrawal, H. Balcı, K. Hanspers, S. L. Coort, M. Martens, D. N. Slenter, F. Ehrhart, D. Digles, A. Waagmeester, I. Wassink *et al.*, "Wikipathways 2024: next generation pathway database," *Nucleic acids research*, vol. 52, no. D1, pp. D679–D689, 2024.

[10] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen, "Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications," *Nucleic acids research*, vol. 39, no. suppl_2, pp. W541–W545, 2011.