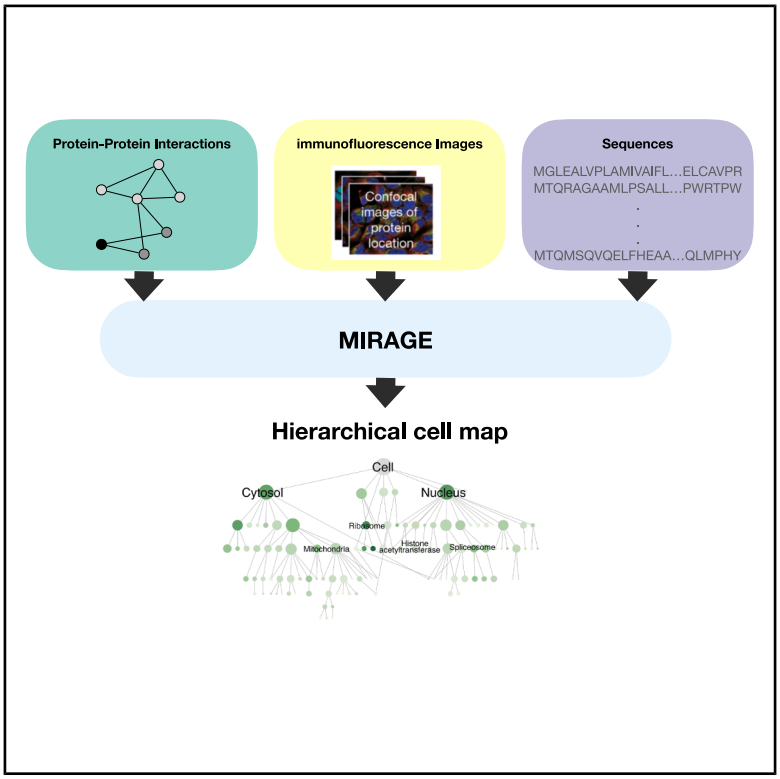
Methods

Cell Systems

An adversarial scheme for integrating multi-modal data on protein function

Graphical abstract



Authors

Rami Nasser, Leah V. Schaffer, Trey Ideker, Roded Sharan

Correspondence

roded@tauex.tau.ac.il

In brief

Nasser et al. present MIRAGE, a multi-modal generative model that integrates protein sequence, interaction, and localization data into a single representation. MIRAGE representations yield state-of-the-art performance in protein function prediction and module detection. MIRAGE is applied to construct a hierarchical map of subcellular organization in HEK293T cells.

Highlights

- MIRAGE integrates protein multi-modal data into a single representation
- Adversarial training enables learning from unaligned data with missing modalities
- MIRAGE outperforms existing methods in protein function prediction and module detection
- Protein representations yield hierarchical maps of subcellular organization

Nasser et al., 2025, Cell Systems *16*, 101444
November 19, 2025
https://doi.org/10.1016/j.cels.2025.101444

CellPress

Please cite this article in press as: Nasser et al., An adversarial scheme for integrating multi-modal data on protein function, Cell Systems (2025), https://doi.org/10.1016/j.cels.2025.101444

Cell Systems

CellPress

Methods

An adversarial scheme for integrating multi-modal data on protein function

Rami Nasser,[1] Leah V. Schaffer,[2] Trey Ideker,[2,3,4] and Roded Sharan[1,5,*]
[1]Blavatnik School of Computer Science and AI, Tel Aviv University, Tel Aviv 69978, Israel
[2]Department of Medicine, University of California, San Diego, La Jolla, CA, USA
[3]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA
[4]Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA
[5]Lead contact
*Correspondence: roded@tauex.tau.ac.il
https://doi.org/10.1016/j.cels.2025.101444

SUMMARY

To begin deciphering the hierarchical structure of the cell, we need to integrate multiple types of data of different scales on subcellular organization. To this end, we developed MIRAGE, a multi-modal generative model for integrating protein sequence, protein-protein interaction, and protein localization data. Our adversarial approach successfully learns a joint embedding space that captures the complex relationships among these diverse modalities and allows us to generate missing modalities. We evaluate our model's performance against existing methods, obtaining superior performance in protein function prediction and protein complex detection. We apply MIRAGE to construct a hierarchical map of subcellular organization in HEK293T cells, recovering known protein assemblies across multiple scales.

INTRODUCTION

Sparsity and incomplete data present significant challenges in bioinformatics, hindering the analysis and interpretation of large biological datasets.[1] These issues necessitate the development of specialized computational algorithms and data imputation methods to accurately predict missing values and extract meaningful insights from incomplete biological information. A single protein encompasses multiple biological dimensions: its amino acid sequence reveals insights into its structure and molecular function, its protein-protein interactions (PPIs) reflect the biological processes it takes place in and provide information on its subcellular organization, and its localization images illustrate its spatial distribution within cellular compartments. This multifaceted nature of proteins provides various sources of information for learning meaningful representations by integrating different biological modalities. Ideally, a unified joint embedding space would allow for an integrated representation of proteins by aligning these diverse modalities.

However, acquiring comprehensive datasets that include all these modalities is often impractical due to the high costs and complexities associated with experimental data collection about the interactions and localization of a protein. Thus, state-of-the-art integration methods consider only a subset of these modalities. For instance, some approaches focus on integrating localization images with PPI information,[2] while others aim to connect sequence data with localization images.[3] Similar approaches have been successful in other fields, such as combining images with text or audio in computer vision.[4,5] Nevertheless, these methods typically operate on limited pairs of modalities, resulting in embeddings that are restricted to the specific combinations used during training. A recent study[6] attempted to address this limitation by using images as a central point to connect with other types of data. While this is a step forward, it depends on comprehensive image data, which are not available as of yet in the protein world.

Our work addresses this gap by proposing the MIRAGE (multi-modal integrative representation using adversarial generative embedding) model that learns a joint embedding space across the three aforementioned modalities: sequence, interaction, and localization (Box 1). Importantly, our model does not require full information, allowing us to represent proteins for which information on one or two modalities is missing. Our approach draws inspiration from CycleGAN,[7] adapting its concept of bidirectional translation to the domain of biological data modalities. In our model, different modalities are encoded into a shared latent space, from which we can generate other modalities. This creates a cycle of translations: modality A can be used to generate modality B, and the generated B can be used to reconstruct A, ensuring consistency and information preservation across modalities. This methodology enables the translation and generation of one modality from another, offering a solution to the pervasive issue of data scarcity and incompleteness in biological research.

We demonstrate the effectiveness of our multi-modal generative model by integrating protein sequence data, PPI information, and subcellular localization images to construct a hierarchical map of subcellular organization. Our results show that our approach successfully learns a joint embedding space that

Please cite this article in press as: Nasser et al., An adversarial scheme for integrating multi-modal data on protein function, Cell Systems (2025), https://doi.org/10.1016/j.cels.2025.101444

Box 1. Progress and potential progress

Proteins are complex biological entities that can be understood through multiple lenses: their amino acid sequence, their interactions with other proteins, and their location within cells. While each of these "modalities" provides valuable insights, the true understanding of protein function emerges from integrating all three perspectives. However, this integration has been challenging because comprehensive datasets containing all three modalities for the same proteins are extremely rare—most proteins have information for only one or two modalities. MIRAGE (multi-modal integrative representation using adversarial generative embedding) solves this problem by translating between modalities rather than requiring all modalities to be present for every protein. Drawing inspiration from CycleGAN, MIRAGE creates a continuous cycle of translations: modality A can generate modality B, which can then regenerate A, ensuring consistency across modalities. This approach allows MIRAGE to learn from unaligned data—proteins with information in only one or two modalities—vastly expanding the available training data. The key innovation of MIRAGE lies in its ability to encode different protein modalities into a shared embedding space without requiring complete information across all modalities. This enables representation of proteins even when information on one or two modalities is missing. This further allows the generation of missing modality data. Unlike previous approaches that require aligned data that drastically reduce sample size, MIRAGE integrates all three modalities simultaneously while leveraging much larger datasets. Potential: MIRAGE represents a significant advancement in our ability to understand protein function through integrated multi-modal analysis. Beyond outperforming existing methods in protein module detection and function prediction, MIRAGE offers several promising directions for future research and applications. First, MIRAGE's ability to generate missing modality data could help overcome the persistent challenge of incomplete biological information. For example, when protein localization images are unavailable, MIRAGE can generate high-quality image representations from sequence or interaction data, expanding the scope of proteins that can be analyzed. Second, the hierarchical cell map created using MIRAGE embeddings provides richer insights into cellular organization across multiple scales, from large compartments like the nucleus down to specific protein complexes. This multi-scale perspective could enhance our understanding of how cellular components work together in health and disease. Finally, MIRAGE's framework is inherently scalable to additional modalities beyond the three demonstrated here. As new types of protein data become available (such as 3D structures or post-translational modifications), they can be incorporated into the MIRAGE framework, potentially leading to even more comprehensive protein representations. This scalability makes MIRAGE a valuable platform for integrating the increasingly diverse landscape of biological data, ultimately advancing our understanding of cellular systems and protein function.

captures the complex relationships between these diverse modalities. We evaluate our model's performance against existing methods for protein representation learning, including those that focus on a single modality or on pairs of modalities. Our framework demonstrates superior performance in several key tasks, including protein function prediction, module detection, and data generation for missing modalities.

RESULTS

We developed MIRAGE, a multi-modal generative model for integrating protein sequence, PPI, and protein localization data to construct a hierarchical map of the cell (Figure 1). The model receives as input protein data from multiple modalities. MIRAGE adversarially learns to embed these modalities into a shared latent space that captures the complex relationships between these diverse modalities and allows the generation of missing modalities Figure 2). Model training relies on three objective or loss components: (1) adversarial loss, which ensures that the distribution of the generated modality matches the data distribution in the target domain; (2) reconstruction loss, which enforces consistency between the latent space representation and the original domain; and (3) latent-cycle-consistency loss, ensuring alignment in the latent space for the same protein encoded from different modalities.

One of the key strengths of MIRAGE is its ability to generate representations for missing modalities. During training, each input modality is mapped to a shared latent space using an encoder, and from this shared representation, the model learns to generate the other modalities using dedicated generators. This cross-modal translation is enabled by adversarial training and cycle-consistency constraints, allowing MIRAGE to infer missing modality embeddings from available ones at inference time. The MIRAGE framework is described in detail in the STAR Methods.

We applied MIRAGE to integrate sequence, interaction, and localization information in HEK293T cells to construct a hierarchical map of protein subcellular organization. As a sanity check, we compared the MIRAGE joint embedding to those obtained from each modality separately. As illustrated in Figure 3, our method consistently outperformed single-modality approaches across all benchmark datasets. For clarity, the single-modality baselines evaluated here correspond to pretrained embeddings obtained from widely used models: ESM-2 for sequence, node2-vec for PPI, and DenseNet for image data, as detailed in STAR Methods. The joint embeddings produced by MIRAGE better align with known biological structures compared with those derived from any individual modality. This superior performance underscores the effectiveness of our approach in capturing complementary information from diverse data types, resulting in more biologically relevant and informative representations.

While MIRAGE seamlessly integrates all three input modalities, methods such as MUSE[3] and DICE[2] are constrained to integrating only two modalities at a time. To compare to those methods, we performed three distinct experiments, each combining a pair out of the three modalities. For each modality pair, we applied the same evaluation protocol described earlier. The results of these pairwise integration experiments are presented in Figure 4. Remarkably, MIRAGE consistently outperformed the benchmark

Please cite this article in press as: Nasser et al., An adversarial scheme for integrating multi-modal data on protein function, Cell Systems (2025), https://doi.org/10.1016/j.cels.2025.101444

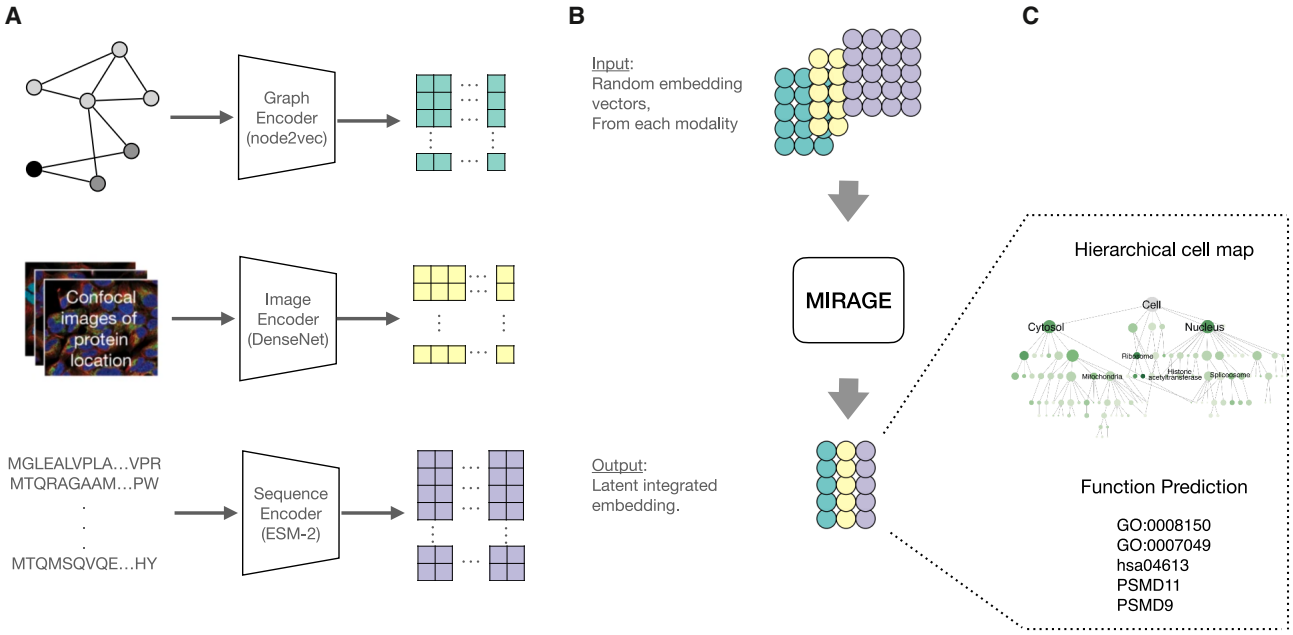


Figure 1. MIRAGE workflow for multi-modal protein data integration

(A) Data preprocessing: MIRAGE inputs three types of protein data: amino acid sequences (processed through ESM-2 to generate embeddings), PPI networks (processed through node2vec), and subcellular localization images (processed through DenseNet). Each modality is encoded into vector representations before entering the MIRAGE pipeline. Importantly, proteins with incomplete data across modalities can still be included.

(B) MIRAGE adversarial model. The model learns a latent embedding for each modality.

(C) Application to downstream tasks: a hierarchical map of cellular structures and protein function prediction.

methods across almost all modality combinations and evaluation benchmarks. All pairwise evaluations in Figure 4 were performed only on proteins for which both corresponding modalities were available to ensure fairness and comparability.

We observed that MIRAGE outperforms baseline methods (MUSE and DICE) across most modality pairs, particularly when the modalities carry complementary information. For example, combining sequence and image data leverages distinct sources of information—sequence features encode structural and functional motifs, while image embeddings capture subcellular localization. By contrast, for tasks like GO cellular component (CC) prediction, image embeddings alone already carry strong spatial information, leaving limited room for improvement through integration. Additionally, we note that MIRAGE was not fine-tuned for specific tasks or label sets, and further performance gains may be achievable through task-specific optimization.

To comprehensively evaluate MIRAGE in the context of integrating three or more modalities, we conducted a comparative analysis against alternative approaches capable of handling multi-modal data. Due to the scarcity of well-established models designed to operate on three or more modalities simultaneously, we selected two baseline methods: simple feature concatenation (CONCAT) and a multi-modal autoencoder (MMAE), which is considered a strong baseline for multi-modal integration.[8] The CONCAT method serves as a naive baseline, directly combining features from different modalities without learning inter-modal relationships. By contrast, the MMAE represents a more sophisticated approach, combining multi-modal data through a bottleneck layer that can reconstruct original features, a technique that has shown promise in various multi-modal learning tasks.[9] The results of this comparison, constrained to proteins with data in all modalities, are presented in Figure 5. A further evaluation with respect to GO molecular function categories appears in Figure S5, showing a significant advantage to MIRAGE compared with the other methods. Overall, MIRAGE demonstrated superior performance compared with both CONCAT and MMAE in the vast majority of benchmarks and evaluation metrics (with a single exception where

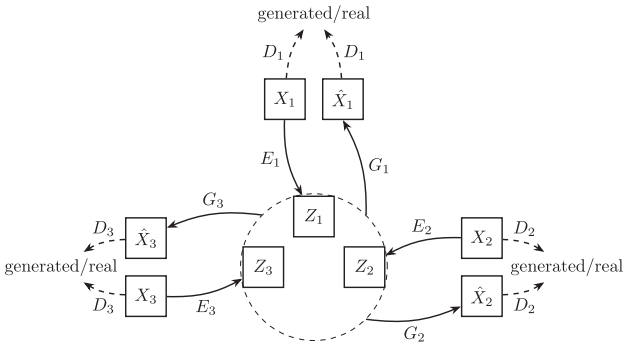
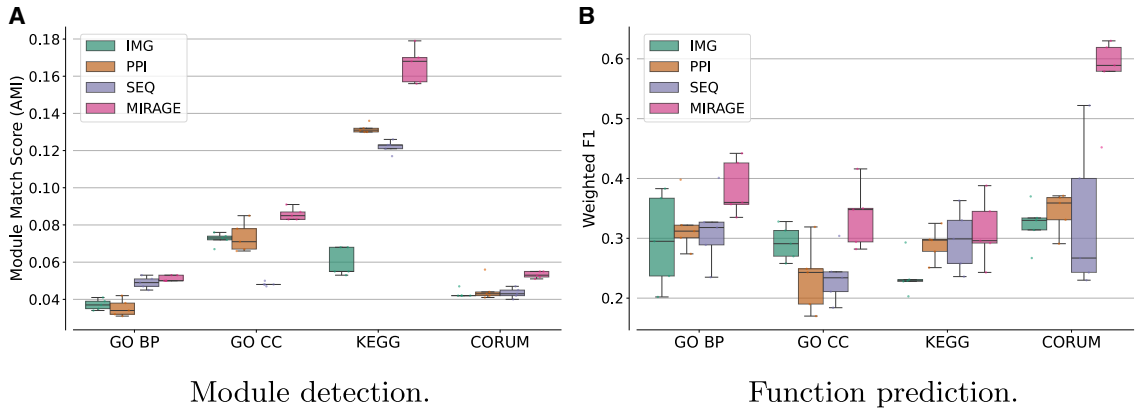


Figure 2. The MIRAGE scheme

For each modality, it learns a mapping $E : X \rightarrow Z$ to latent space (dashed circle) and a generator from latent space to modality $G : Z \rightarrow X$. Real and generated samples are classified using a parameterized discriminator $D : X, \hat{X} \rightarrow 0, 1$.

Please cite this article in press as: Nasser et al., An adversarial scheme for integrating multi-modal data on protein function, Cell Systems (2025), https://doi.org/10.1016/j.cels.2025.101444



Module detection. Function prediction.

Figure 3. Performance evaluation of multi-modal protein embeddings in HEK293T cells using the BIONIC benchmark

(A) Module detection performance measured by adjusted mutual information (AMI) score comparing image-based (IMG), protein-protein interaction (PPI), sequence-based (SEQ), and integrated MIRAGE embeddings across GO biological process (GO BP), GO cellular component (GO CC), KEGG pathways, and CORUM protein complexes.

(B) Supervised function prediction performance measured by weighted F1 score across the same functional annotation databases. Single-modality baselines use pre-aligned embeddings: ESM-2 for sequence (n = 20,218 proteins from UniProt), node2vec for PPI (n = 14,032 proteins from the BioPlex network), and DenseNet for images (n = 1,125 proteins from the Human Protein Atlas), as described in STAR Methods. Box plots show median, quartiles, and individual data points from n = 5 cross-validation folds.

MMAE slightly outperformed MIRAGE on GO BP function prediction).

To assess the robustness of multi-modal models under stricter generalization criteria, we also performed a homology-aware 5-fold cross-validation. Protein sequences were clustered using CD-HIT[10] at 40% sequence identity, and entire clusters were assigned to separate folds to prevent homologous proteins from appearing in both training and test sets. This evaluation was applied to MIRAGE, CONCAT, and MMAE. The results of this analysis are shown in Figure S6, which reflects model performance under these homology-aware splits.

Hierarchical cell map

After establishing the utility of MIRAGE, we applied it to construct a hierarchical map of cell structure. To this end, we clustered the proteins according to the similarities between their integrative embeddings at multiple resolutions.[11] For comparison purposes, we focused on the 907 proteins present in all three modalities. Different protein modalities capture information at different scales. For example, the imaging data reveal information on a protein's localization in the cell, while the AP-MS data (BioPlex network) reveal the protein's specific interaction partners and complexes. Aligning these two modalities using MIRAGE into a unified embedding enables capturing information from all modalities, and clustering at multiple resolutions enables resolving protein assemblies across scales (Figure 6).

The resulting hierarchy contains 111 clusters, including 62 clusters that overlap significantly with a component in the Gene Ontology, CORUM, or HPA (hypergeometric test, false discovery rate [FDR] < 10%, and Jaccard index > 10%). We recovered assemblies across scales, including large compartments (e.g., nucleus and cytosol) and small compartments (e.g., histone acetyltransferase complex and ribosomal complex). In comparison, a previous integration of interaction and image information using DICE[2] identified only 46 clusters that overlapped with known components at the same thresholds. This comparison suggests that MIRAGE embeddings better capture known biological systems.

Embedding alignment and robustness to missing information

Despite MIRAGE's foundation in generative adversarial networks (GANs), we observed a notable alignment property in MIRAGE's joint embeddings. This phenomenon is reminiscent of the characteristics typically associated with contrastive learning approaches, where matching pairs are drawn closer together in the embedding space, while non-matching pairs are more uniformly distributed.[12] To investigate the alignment properties of our joint embedding, we used the UMAP dimensionality reduction algorithm[13] for visualization. The results, presented in Figure 7, reveal a striking alignment among the three modalities for each protein. By contrast, the three modalities are completely separated in the raw embedding. This observed alignment is particularly notable given that MIRAGE employs L_{cyc} loss, which ensures that different modalities for the same protein are mapped nearby in latent space. The emergence of such alignment suggests that our model has successfully captured the intrinsic relationships between different modalities of the same protein, effectively learning a shared representation space.

To test the robustness of the generated embeddings, we focused on proteins with at least two modalities, and for each modality of a given protein, we measured the distance between its real embedding and its generated embedding based on the other modalities of that protein. The results are shown in Figure S7, demonstrating the potential of our model to fill in information gaps. Notably, when comparing generation quality between proteins with complete versus partial modality coverage, we observed no significant difference (Figure S8).

Please cite this article in press as: Nasser et al., An adversarial scheme for integrating multi-modal data on protein function, Cell Systems (2025), https://doi.org/10.1016/j.cels.2025.101444

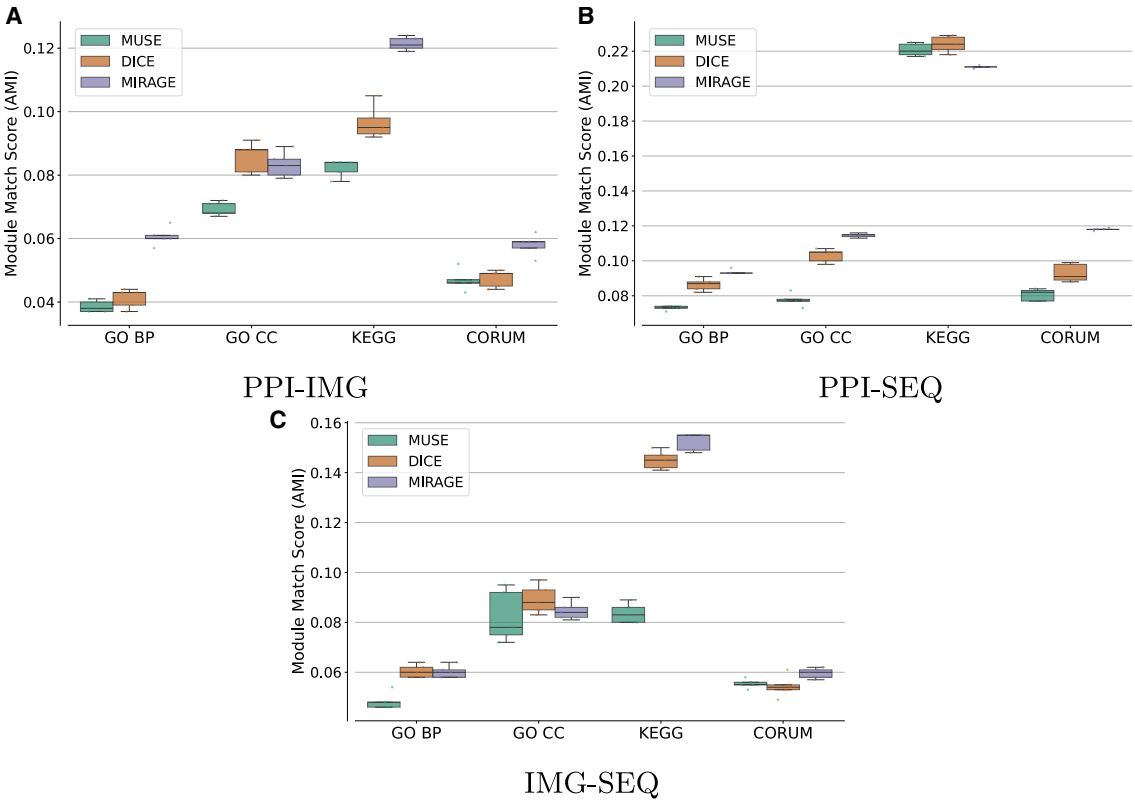


Figure 4. Comparison of multi-modal integration methods for pairwise modality combinations in HEK293T protein data

(A) PPI and imaging data (PPI-IMG, n = 921 proteins with both modalities).

(B) PPI and sequence data (PPI-SEQ, n = 13,674 proteins with both modalities).

(C) Imaging and sequence data (IMG-SEQ, n = 1,070 proteins with both modalities). Module detection performance (AMI score) comparing MIRAGE against established integration methods MUSE and DICE across GO biological process (GO BP), GO cellular component (GO CC), KEGG pathways, and CORUM protein complexes using Louvain clustering. Box plots show median, quartiles, and individual data points from n = 5 cross-validation folds.

As image information is the scarcest, we particularly evaluated our model's ability to generate it using data from the other two modalities. Since MIRAGE generates image embeddings in the latent space of a pretrained image encoder (DenseNet), predicting the original (raw) image content from the embedding would require a separate decoder trained to invert the embedding, which is beyond the scope of this work. Instead, to assess the quality of these generated image embeddings, we employed the Fréchet inception distance (FID) metric[16] (see STAR Methods). We compared the FID scores of our generated image embeddings against those produced by a K-nearest neighbors (KNNs) approach, using the true image embeddings as a reference. Our model consistently achieved lower FID scores compared with KNN, indicating superior performance (see Table S3). This shows the effectiveness of our method in generating high-quality protein image representations from PPI, sequence, or a combination of both.

Last, we evaluated the downstream utility of MIRAGE-generated image representations by testing their performance in a protein function prediction task. As shown in Figure S9, classifiers trained on generated image embeddings achieve predictive accuracy comparable to those using real image embeddings.

DISCUSSION

MIRAGE is a multi-modal generative model for integrating protein sequence, PPI, and protein localization data for constructing a hierarchical map of subcellular organization. Our approach successfully learns a joint embedding space that captures the complex relationships between these diverse modalities. By enabling the use of unaligned data, our model can exploit a broader range of information, potentially leading to more robust and generalizable representations. This unaligned training paradigm offers several advantages. First, it substantially increases the quantity of data that can be utilized for training, as it removes the constraint of finding perfectly matched samples across modalities. Second, it enhances the model's ability to learn more flexible and diverse mappings between modalities, potentially capturing a wider range of inter-modal relationships. In particular, MIRAGE allows representation learning even when some modalities are missing, leveraging adversarial and cycle-consistent objectives to translate between available data types. This is especially valuable in biological datasets with sparse modality coverage. Importantly, we demonstrate that the generated embeddings retain sufficient biological signal to support downstream tasks, such as protein function prediction.

Please cite this article in press as: Nasser et al., An adversarial scheme for integrating multi-modal data on protein function, Cell Systems (2025), https://doi.org/10.1016/j.cels.2025.101444

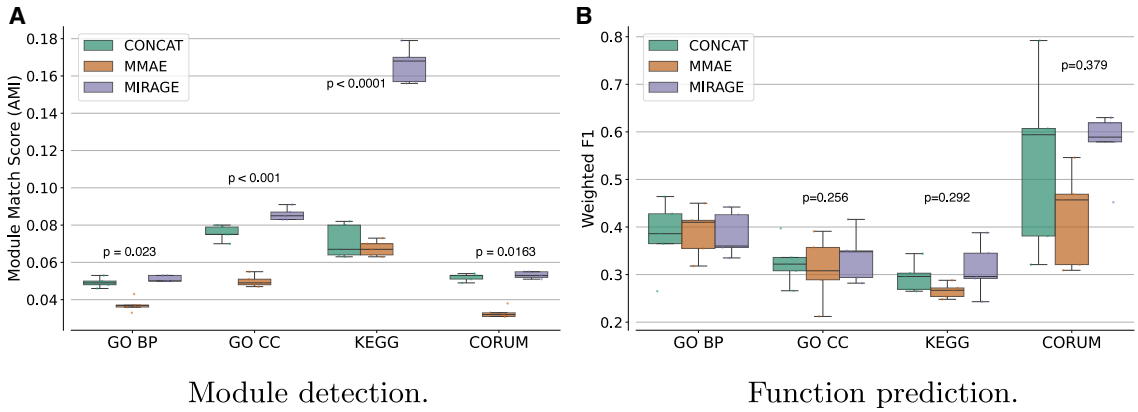


Figure 5. Performance comparison of multi-modal integration methods in HEK293T protein data

(A) Module detection performance measured by adjusted mutual information (AMI) score comparing concatenation baseline (CONCAT), multi-modal autoencoder (MMAE), and MIRAGE across GO biological process (GO BP), GO cellular component (GO CC), KEGG pathways, and CORUM protein complexes. (B) Supervised function prediction performance measured by weighted F1 score across the same functional annotation databases. Analysis performed on proteins with all three modalities available (n = 907 proteins with PPI, imaging, and sequence data). Box plots show median, quartiles, and individual data points from n = 5 cross-validation folds. Statistical significance was determined by a paired t test comparing MIRAGE to the second-best-performing method. For GO BP prediction, where MIRAGE is slightly outperformed by MMAE, no p value is shown.

While MIRAGE achieves state-of-the-art results in multi-modal protein representation learning, several important limitations must be acknowledged. One limitation is that MIRAGE generates embeddings in the latent space rather than reconstructing the original modalities themselves. For instance, when generating missing image information, MIRAGE produces embeddings in DenseNet's feature space rather than actual immunofluorescence images. This implies that while these generated embeddings retain biological signal sufficient for downstream tasks like function prediction, they cannot be directly interpreted or visualized as cellular images. Additionally, MIRAGE's performance is inherently bounded by the quality of the pretrained encoders it employs. Errors or biases in ESM-2, DenseNet, or node2vec will propagate through the integration process, potentially limiting the biological fidelity of the joint representations. Finally, our evaluation focuses primarily on HEK293T cells, and the generalization to other cell types or species remains to be established. Future work should address these constraints through end-to-end architectures that can generate interpretable outputs across diverse biological systems.

A key application of MIRAGE is the derivation of a hierarchical map of subcellular organization based on the learned integrated protein embeddings. Importantly, MIRAGE managed to recover known protein assemblies across scales, including large compartments (e.g., nucleus and cytosol) and small compartments (e.g., histone acetyltransferase complex and ribosomal complex). In addition, it suggests dozens of novel assemblies whose validation and interpretation require further research. As more data become available, MIRAGE could be applied to study how protein function and organization evolve across different conditions and species, providing insights into the fundamental principles governing cellular architecture.

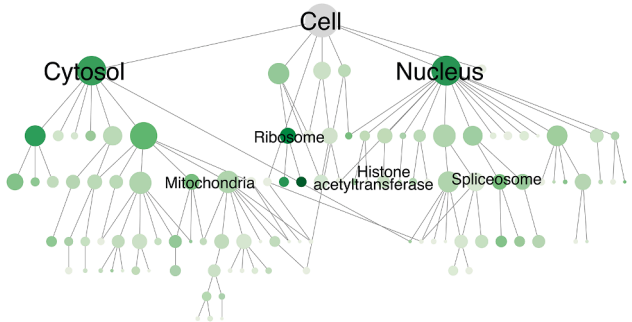


Figure 6. Hierarchy of protein assemblies constructed using the MIRAGE embeddings

Nodes represent protein assemblies, and edges represent hierarchical containment. Node size is proportional to the number of proteins. Nodes are shaded based on overlap with known cellular components.

RESOURCE AVAILABILITY

Lead contact

Requests for further information should be directed to and will be fulfilled by the lead contact, Roded Sharan (roded@tauex.tau.ac.il).

Materials availability

This study did not generate new materials.

Data and code availability

This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table. All original code has been deposited at https://github.com/raminass/MIRAGE and is publicly available as of the date of publication. The repository is archived on Zenodo (https://doi.org/10.5281/zenodo.15569121). Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

R.N. was supported by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. R.S. was supported by a research grant from the Israel Science Foundation (grant no. 1692/24).

Please cite this article in press as: Nasser et al., An adversarial scheme for integrating multi-modal data on protein function, Cell Systems (2025), https://doi.org/10.1016/j.cels.2025.101444

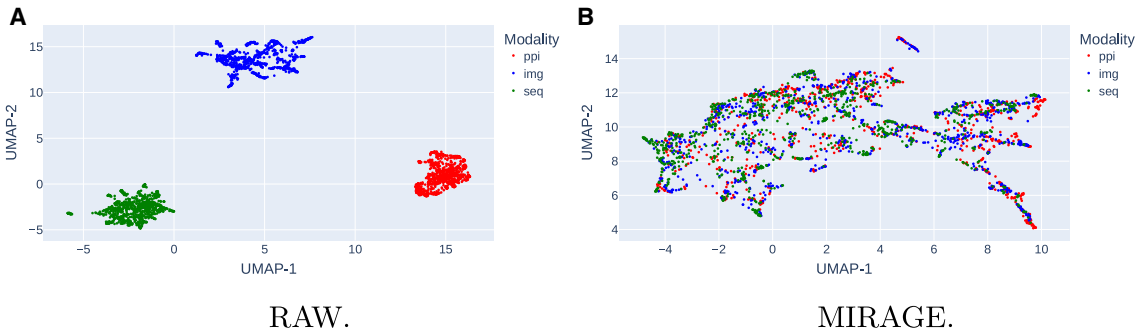


Figure 7. UMAP projections of protein embeddings from three data modalities in HEK293T cells

(A) RAW embeddings represent input to the model before integration.

(B) MIRAGE embeddings represent output after integration. Network-based embeddings (red, n = 14,032 proteins from the BioPlex protein-protein interaction network[14]), image-based embeddings (blue, n = 1,125 proteins from the Human Protein Atlas immunofluorescence images[15]), and sequence-based embeddings (green, n = 20,218 proteins from UniProt) are shown before (RAW) (A) and after (MIRAGE) (B) integration. UniProt sequences are publicly available.

AUTHOR CONTRIBUTIONS

Conceptualization, R.N. and R.S.; methodology, R.N., L.V.S., T.I., and R.S.; software, R.N.; writing – original draft, R.N. and R.S.; writing – review & editing, R.N., L.V.S., T.I., and R.S.; funding, T.I. and R.S.; supervision, T.I. and R.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - Application to HEK293T cells
  - Performance evaluation
  - Adversarial loss
  - Cycle Consistency Loss
  - Reconstruction Loss
  - Fréchet Inception Distance (FID)
  - Adjusted Mutual Information (AMI)
  - Implementation details

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2025.101444.

Received: March 28, 2025
Revised: July 28, 2025
Accepted: October 14, 2025

REFERENCES

1. Kondratyeva, L., Alekseenko, I., Chernov, I., and Sverdlov, E. (2022). Data incompleteness may form a hard-to-overcome barrier to decoding life's mechanism. Biology *11*, 1208. https://doi.org/10.3390/biology11081208.
2. Nasser, R., Schaffer, L.V., Ideker, T., and Sharan, R. (2024). Multi-modal contrastive learning of subcellular organization using DICE. Bioinformatics *40*, ii105–ii110. https://doi.org/10.1093/bioinformatics/btae387.
3. Bao, F., Deng, Y., Wan, S., Shen, S.Q., Wang, B., Dai, Q., Altschuler, S.J., and Wu, L.F. (2022). Integrative spatial analysis of cell morphologies and transcriptional states with MUSE. Nat. Biotechnol. *40*, 1200–1209. https://doi.org/10.1038/s41587-022-01251-z.
4. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022). Flamingo: a visual language model for few-shot learning. Adv. Neural Inf. Process. Syst. *35*, 23716–23736. https://doi.org/10.48550/arXiv.2204.14198.
5. Arandjelovic, R., and Zisserman, A. (2017). Look, listen and learn. In Proceedings of the IEEE International Conference on Computer Vision (IEEE), pp. 609–617. https://doi.org/10.1109/ICCV.2017.73.
6. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., and Misra, I. (2023). Imagebind: One embedding space to bind them all. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (IEEE), pp. 15180–15190. https://doi.org/10.1109/CVPR52729.2023.01457.
7. Zhu, J.-Y., Park, T., Isola, P., and Efros, A.A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In 2017 IEEE International Conference on Computer Vision (ICCV) (IEEE), pp. 2242–2251. https://doi.org/10.1109/ICCV.2017.244.
8. Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. *41*, 423–443. https://doi.org/10.1109/TPAMI.2018.2798607.
9. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A.Y. (2011). Multimodal deep learning. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 689–696. https://doi.org/10.48550/arXiv.2301.04856.
10. Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659. https://doi.org/10.1093/bioinformatics/btl158.
11. Zheng, F., Zhang, S., Churas, C., Pratt, D., Bahar, I., and Ideker, T. (2021). HiDeF: identifying persistent structures in multiscale 'omics data. Genome Biol. *22*, 21. https://doi.org/10.1186/s13059-020-02228-4.
12. Wang, T., and Isola, P. (2020). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In International Conference on Machine Learning. PMLR, pp. 9929–9939. https://doi.org/10.48550/arXiv.2005.10242.
13. McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. Preprint at arXiv. https://doi.org/10.48550/arXiv.1802.03426.
14. Huttlin, E.L., Bruckner, R.J., Navarrete-Perea, J., Cannon, J.R., Baltier, K., Gebreab, F., Gygi, M.P., Thornock, A., Zarraga, G., Tam, S., et al. (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Cell *184*, 3022–3040.e28. https://doi.org/10.1016/j.cell.2021.04.011.

15. Thul, P.J., Åkesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., Alm, T., Asplund, A., Björk, L., Breckels, L.M., et al. (2017). A subcellular map of the human proteome. Science *356*, eaal3321. https://doi.org/10.1126/science.aal3321.

16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. Preprint at arXiv. Adv. Neural Inf. Process. Syst. *30*. https://doi.org/10.48550/arXiv.1706.08500.

17. Grover, A., and Leskovec, J. (2016). node2vec: Scalable Feature Learning for Networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. (Association for Computing Machinery)), pp. 855–864. https://doi.org/10.1145/2939672.2939754.

18. Ouyang, W., Winsnes, C.F., Hjelmare, M., Cesnik, A.J., Åkesson, L., Xu, H., Sullivan, D.P., Dai, S., Lan, J., Jinmo, P., et al. (2019). Analysis of the human protein atlas image classification competition. Nat. Methods *16*, 1254–1261. https://doi.org/10.1038/s41592-019-0658-6.

19. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science *379*, 1123–1130. https://doi.org/10.1126/science.ade2574.

20. Mao, X., Li, Q., Xie, H., Lau, R.Y.K., Wang, Z., and Smolley, S.P. (2017). Least Squares Generative Adversarial Networks. In 2017 IEEE International Conference on Computer Vision (ICCV) (IEEE), pp. 2813–2821. https://doi.org/10.1109/ICCV.2017.304.

21. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A.C. (2017). Improved training of Wasserstein GANs. Adv. Neural Inf. Process. Syst. *30*. https://doi.org/10.48550/arXiv.1704.00028.

22. Lahat, D., Adali, T., and Jutten, C. (2015). Multimodal data fusion: an overview of methods, challenges, and prospects. Proc. IEEE *103*, 1449–1477. https://doi.org/10.1109/JPROC.2015.2460697.

23. Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. *35*, 1798–1828. https://doi.org/10.1109/TPAMI.2013.50.

24. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

25. Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In International Conference on Machine Learning. PMLR, pp. 1597–1607. https://doi.org/10.48550/arXiv.2002.05709.

26. Radford, A. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1511.06434.

27. Forster, D.T., Li, S.C., Yashiroda, Y., Yoshimura, M., Li, Z., Isuhuaylas, L.A.V., Itto-Nakama, K., Yamanaka, D., Ohya, Y., Osada, H., et al. (2022). BIONIC: biological network integration using convolutions. Nat. Methods *19*, 1250–1261. https://doi.org/10.1038/s41592-022-01616-x.

28. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J Stat Mech.: Theory Exp. *2008*, P10008. https://doi.org/10.1088/1742-5468/2008/10/P10008.

29. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.

30. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29. https://doi.org/10.1038/75556.

31. Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Ruepp, A. (2019). CORUM: the comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res. *47*, D559–D563. https://doi.org/10.1093/nar/gky973.

32. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. Preprint at arXiv. Adv. Neural Inf. Process. Syst. *27*. https://doi.org/10.48550/arXiv.1406.2661.

33. Almahairi, A., Rajeshwar, S., Sordoni, A., Bachman, P., and Courville, A. (2018). Augmented cyclegan: Learning many-to-many mappings from unpaired data. In International Conference on Machine Learning. PMLR, pp. 195–204. https://doi.org/10.48550/arXiv.1802.10151.

34. Vinh, N.X., Epps, J., and Bailey, J. (2009). Information theoretic measures for clusterings comparison: is a correction for chance necessary? In Proceedings of the 26th Annual International Conference on Machine Learning (ACM), pp. 1073–1080. https://doi.org/10.1145/1553374.1553511.

35. Glorot, X., and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, pp. 249–256.

36. Kingma, D.P., and Ba, J. (2014). Adam: A method for stochastic optimization. Preprint at arXiv. https://doi.org/10.48550/arXiv.1412.6980.

Please cite this article in press as: Nasser et al., An adversarial scheme for integrating multi-modal data on protein function, Cell Systems (2025), https://doi.org/10.1016/j.cels.2025.101444

STAR★METHODS

KEY RESOURCES TABLE

REAGENTS or RESOURCES	SOURCE	IDENTIFIER
Deposited data		
Cell images	Thul et al.[15]	https://www.proteinatlas.org
Protein Sequence	UniProt	https://www.uniprot.org
Bioplex Network	Huttlin et al.[14]	https://www.ndexbio.org
Protein Labels	GO, KEGG, CORUM	https://doi.org/10.5281/zenodo.15553132
Software and algorithms		
MIRAGE	This Paper	https://doi.org/10.5281/zenodo.15569121
node2vec	Grover et al.[17]	https://snap.stanford.edu/node2vec
DenseNet	Ouyang et al.[18]	https://github.com/CellProfiling/densenet
ESM-2	Lin et al.[19]	https://huggingface.co/facebook/esm2/_t33/_650M/_UR50D

METHOD DETAILS

Let $X = \{X_1, X_2, \ldots, X_M\}$ represent the set of M different modalities for protein representation (e.g., sequence (SEQ), interaction (PPI) and localization (IMG)), where $X_i \in \mathbb{R}^{d_i}$. The goal is to embed $X_i \underset{E_i}{\rightarrow} Z$ into a shared latent space $Z \in \mathbb{R}^l$, and to generate $Z \underset{G_i}{\rightarrow} X_i$ modality vector from any latent modality j, where E_i and G_i are parameterized mapping functions.

We employ adversarial learning, where discriminators D_i are trained to distinguish between real and generated samples of each modality. The key innovation lies in feeding unaligned modality embeddings to the framework, allowing the model to learn cross-modal relationships without requiring full M-tuple data from the same protein. Formally, for any pair of modalities (i,j), we can perform translations $X_i \rightarrow E_i Z \rightarrow G_j X_j'$, where X_j' is the generated version of modality j from modality i. Our approach, MIRAGE, is illustrated in Figure 2. Importantly, MIRAGE's translation process is independent of the specific proteins used in the input modality, enabling flexible cross-modal generation even when data for all modalities is not available for every protein. This overcomes the requirement for aligned data that often drastically reduces the volume of available training samples as demonstrated in Figure S1.

Our objective function consists of three main components, which are described in detail below: (i) adversarial loss $\mathcal{L}_{gan}$, which ensures that the distribution of the generated modality matches the data distribution in the target domain; (ii) reconstruction loss $\mathcal{L}_{rec}$, which enforces consistency between the latent space representation and the original domain; and (iii) latent-cycle-consistency loss $\mathcal{L}_{cyc}$, introduced to prevent contradictions between the learned mappings E and G, ensuring alignment in the latent space for the same protein encoded from different modalities. The full loss function is a weighted sum of these loss terms:

$$\mathcal{L}_{total} = \lambda_{gan}\mathcal{L}_{gan} + \lambda_{cyc}\mathcal{L}_{\text{cyc}} + \lambda_{rec}\mathcal{L}_{\text{rec}}, \quad \text{(Equation 1)}$$

where $\lambda_{gan}, \lambda_{cyc}, \lambda_{rec}$ control the relative importance of each loss term. Computationally, this setup requires learning $2\times M$ parameters, making the complexity linear with respect to the number of modalities, in contrast to a direct mapping between all modalities, which would lead to quadratic complexity. We evaluate key training dynamics in Figures S2 and S3. The former illustrates that discriminator buffering helps prevent overfitting, while the latter shows that smaller batch sizes destabilize adversarial training—a critical issue for relatively small biological datasets. Our design follows established practices in GAN training, including the use of least-squares GAN loss[20] and gradient penalty.[21]

To assess the impact of the latent cycle-consistency loss, we conducted an ablation experiment varying its weight $(\lambda_{\text{cyc}} = 0,5,10)$. As shown in Figure S4, higher values of λ_{cyc} result in greater alignment among the three modality representations in the shared latent space, as visualized via UMAP. This supports the role of the cycle-consistency loss in encouraging cross-modal coherence.

Application to HEK293T cells

We applied MIRAGE to data coming from three different modalities measured in HEK293T cells. Specifically, we integrated protein sequence data from UniProt for 20,218 human proteins, alongside immunofluorescence images from the Human Protein Atlas[15] and protein-protein interaction data from the Bioplex network.[14] The protein interaction network contains 14,032 proteins and 127,732 interactions, while the image dataset includes 1,125 immunofluorescence images. Dataset sizes and overlaps are depicted in Figure S1.

After learning, in order to construct a representative embedding for each protein that accounts for its image-based, network-based and sequence-based embeddings, we fuse those embeddings into a joint representation. While there are many potential fusion functions,[22] we simply use concatenation: $Z = [Z_1|Z_2|Z_3]$.

Please cite this article in press as: Nasser et al., An adversarial scheme for integrating multi-modal data on protein function, Cell Systems (2025), https://doi.org/10.1016/j.cels.2025.101444

In recent works, there has been a growing shift toward using embedded or encoded features instead of raw data, enabling models to extract rich, abstract representations that capture essential aspects of the data.[23] By breaking large tasks into smaller, manageable ones, intermediate representations can be learned and then leveraged for the main objective. For example, in ESM-2,[19] masked language modeling (MLM) was first applied to learn sequence embeddings, which were subsequently used to train a 3D structure prediction model, in contrast to end-to-end approaches like AlphaFold.[24] Another example is in computer vision, where image representation is learned for downstream tasks such as: classification, segmentation.[25,26] Following this practice, MIRAGE also utilizes embeddings as input, incorporating multiple modalities in a joint embedding space rather than working directly with raw data. This approach enhances flexibility and improves performance in complex tasks.

For sequence encoding, we utilized the ESM-2[19] model to obtain sequence embeddings from the raw sequence data. We excluded 460 proteins with sequences longer than 2000 amino acids due to input length limitations of the ESM-2 model. For image encoding, we use DenseNet.[18] For network encoding, we employed node2vec.[17] These encodings allow for direct comparison with previous works that have used similar methods.[3]

Performance evaluation

To assess the performance of our method, we utilized the recently developed BIONIC benchmark[27] with the following tasks: (i) protein module detection and (ii) supervised protein function prediction. For module detection, BIONIC uses the Adjusted Mutual Information (AMI) score (see Adjusted Mutual Information (AMI)) to compare clustering outputs against known protein modules. AMI measures the agreement between two clusterings while correcting for chance, making it robust for evaluating unsupervised methods. This metric was also used in the BIONIC benchmark,[27] ensuring consistency with prior evaluations. While BIONIC originally employed hierarchical clustering for module evaluation, we opted for the popular Louvain clustering algorithm[28] due to its computational efficiency and effectiveness in identifying community structures in large networks. For supervised function prediction we followed BIONIC and used a One-vs-Rest scheme with a linear support vector classifier.

Our evaluation was conducted using human protein module benchmarks derived from multiple well-established biological databases. These included KEGG pathways,[29] excluding metabolic pathways to focus on signaling and regulatory modules; Gene Ontology (GO) annotations,[30] specifically targeting Cellular Components (CC) and Biological Processes (BP); and CORUM complexes,[31] which provide a curated set of mammalian protein complexes. Following BIONIC, we used 5-fold cross-validation for all evaluations. For module detection, we consider labels that occur in at least 5 proteins, and for supervised function prediction, we restrict to frequent labels that appear in at least 20 proteins, consistent with BIONIC. This diverse set of benchmarks allowed us to evaluate our method's performance across various biological contexts and scales, ranging from specific protein complexes to broader functional modules and pathways.

A breakdown of protein coverage across different functional annotations and modality combinations is provided in Tables S1 and S2 reports the number of unique GO, KEGG and CORUM terms available for each modality or modality combination. Together, these tables summarize the dataset composition and term diversity used throughout our experiments. All datasets and annotation standards used in our experiments are publicly available (key resources table) to support reproducibility.

Adversarial loss

Generative Adversarial Networks (GANs)[32] are a class of deep learning models consisting of two neural networks, a generator and a discriminator, trained simultaneously through adversarial learning. In traditional GANs, the discriminator is typically trained using binary cross-entropy loss to distinguish between real and generated samples, while the generator aims to produce samples that can fool the discriminator. However, this approach can lead to training instability and vanishing gradients. To address these limitations, we adopted the Least Squares Generative Adversarial Network (LSGAN)[20] approach in our work. LSGAN replace the cross-entropy loss function in the discriminator with a least squares loss. This modification provides several advantages: (i) it helps mitigate the vanishing gradients problem often encountered in GAN training; (ii) and it addresses the issue of the discriminator learning too quickly and overpowering the generator, which can lead to training instability. This results in improved stability during the training process compared to regular GANs.

For the discriminator the loss is:

$$\mathcal{L}_D = \left[(D(x) - 1)^2\right] + \left[D(G(z))^2\right], \quad \text{(Equation 2)}$$

where 1 represents the target label for real samples, and only the discriminator D is updated during this step.

For the generator the loss is:

$$\mathcal{L}_G = \left[(D(G(z)) - 1)^2\right], \quad \text{(Equation 3)}$$

where 1 represents the target label for generated samples to fool the discriminator, and only the generator G is updated during this step.

We further incorporate the gradient penalty[21] approach, which enhances the stability of GAN training by regulating how quickly the discriminator's output can change with respect to its input. Specifically, the gradient penalty adds a term to the loss function that discourages the discriminator from making overly confident predictions based on small changes in input. This helps to prevent

Please cite this article in press as: Nasser et al., An adversarial scheme for integrating multi-modal data on protein function, Cell Systems (2025), https://doi.org/10.1016/j.cels.2025.101444

the discriminator from becoming too powerful too quickly, which can lead to training instability. The gradient penalty is added to the original $\mathcal{L}_D$ loss, resulting in the following adversarial objective:

$$\mathcal{L}_{gan} = \mathcal{L}_D + \mathcal{L}_G + \lambda_{gp}\mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}\left[\left(\|\nabla_{\hat{x}}D(\hat{x})\|_2 - 1\right)^2\right], \quad \text{(Equation 4)}$$

where λ_{gp} is a constant that controls the strength of the gradient penalty,[8] $\hat{x}$ is the input sampled along the line between real and generated samples, sampled only from real, $\nabla_{\hat{x}}D(\hat{x})$ is the gradient of the discriminator with respect to $\hat{x}$, $\|\cdot\|_2$ is the L2 norm, and the term $(\|\nabla_{\hat{x}}D(\hat{x})\|_2 - 1)^2$ encourages the gradient norm to be close to 1.

Cycle Consistency Loss

The cycle consistency loss, originally introduced for CycleGAN,[7] helps ensure that the translation process is bijective and preserves important features of the input. By extending this concept to the latent space, we aim to enforce consistency in the learned representations (Figure S10), which can lead to more stable and meaningful translations. This latent cycle consistency loss, allows many to many mapping without fixing pairs to perform a full cycle as in Zhu et al.[7] also it contribute to better alignment of translations in the latent space. Formally, the loss is:

$$\mathcal{L}_{\text{cyc}} = \left\|Z_i - E_j(G_j(Z_i))\right\|_2^2, \quad \text{(Equation 5)}$$

where $Z_i = E_i(X_i)$ is the latent representation of input modality i. We randomly select a target modality j, generate that modality and then encode back to latent space (Figure S10).

Reconstruction Loss

In addition to the cycle consistency loss, we utilize a reconstruction loss from the latent space back to the original space. This additional constraint helps to preserve important features of the input during the translation process and encourages the model to learn a meaningful and diverse mapping between domains. By enforcing this reconstruction, we aim to avoid the problem of mode collapse, by generator produces a limited variety of outputs regardless of the input. This approach is inspired by the original CycleGAN paper,[7] and further explored in the augmented CycleGAN model.[33] We use the following reconstruction loss:

$$\mathcal{L}_{\text{rec}} = \|X_i - G_i(Z_i)\|_2^2, \quad \text{(Equation 6)}$$

where $Z_i = E_i(X_i)$ is the latent representation of modality i (Figure S11).

Fréchet Inception Distance (FID)

The Fréchet Inception Distance (FID) score between two distributions P_r (real data) and P_g (generated data) is given by:

$$\text{FID}(P_r, P_g) = \left\|\mu_r - \mu_g\right\|_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{\frac{1}{2}}\right), \quad \text{(Equation 7)}$$

where μ_r and Σ_r are the mean and covariance of the real data features, and μ_g and Σ_g are the mean and covariance of the generated data features.

FID is particularly suitable for this task as it captures both the quality and diversity of generated images, providing a comprehensive measure of how well the generated distributions match the real data distribution.[16]

Adjusted Mutual Information (AMI)

Adjusted Mutual Information (AMI) measures how similar two sets of cluster assignments are, while correcting for similarities that could occur just by chance.[34] It is particularly helpful when comparing predicted clusters—such as protein modules—to known biological groupings that may differ in size or number.

The core idea is to compute how much the predicted and true clusterings overlap (called Mutual Information), and then adjust that score to account for random chance. If two clusterings agree perfectly, AMI will be 1. If they are no better than random, AMI will be close to 0.

Implementation details

We implemented MIRAGE using PyTorch and conducted experiments on a Linux machine equipped with an NVIDIA TITAN Xp GPU. The source code is available at https://github.com/raminass/MIRAGE. Our network architecture utilizes a latent dimension of 128 and a hidden dimension of 512. We initialized the model weights using Xavier initialization.[35] We employed a batch size of 32 for all experiments, as it provided optimal stability for adversarial training. Smaller batch sizes such as 8 were explored but resulted in unstable training dynamics, as illustrated in Figure S3. For optimization, We employed the Adam optimizer[36] with a learning rate of 0.0002 and $\beta_1 = 0.5$, following established practices in GAN training.[26] Our training schedule maintained a constant learning rate for the first 100 epochs, followed by a linear decay to zero over the subsequent 100 epochs, a strategy that has been shown to enhance stability and convergence in GAN training.[7] Notably, we update the discriminator using only the samples produced by the most recent generator iteration. This design choice was made after observing that including older or buffered samples—i.e., previously generated outputs stored in a sample history buffer—led to an overpowered discriminator that quickly overfits to stale generator outputs, destabilizing

Please cite this article in press as: Nasser et al., An adversarial scheme for integrating multi-modal data on protein function, Cell Systems (2025), https://doi.org/10.1016/j.cels.2025.101444

the adversarial training. While buffering is sometimes used to regularize discriminator updates,[7] in our setting with limited data, it had the opposite effect. This observation is illustrated in Figure S2.

In our loss function, we employed multiple components weighted by specific hyperparameters to balance their contributions. The total loss is computed as a weighted sum of the GAN loss, cycle consistency loss, reconstruction loss, and gradient penalty, specifically we use the following weighting parameters same as in Zhu et al.[7]: $\lambda_{gan} = 1, \lambda_{cyc} = 10, \lambda_{rec} = 10, \lambda_{gp} = 1$.