UNLOCKING MOLECULAR INSIGHTS: A MULTIMODAL DEEP LEARNING APPROACH FOR ENHANCED MOLECULAR ANALYSIS

Hatice K. Erdogan

Department of Mathematics and Statistics, American University, Washington, DC, 20016

ABSTRACT

In the realm of energetics and molecular research, the integration of diverse data modalities is increasingly crucial for comprehensive analysis and effective modeling. This project proposes a novel approach leveraging three distinct modalities: textual data comprising molecular descriptions, image data representing molecular shapes, and molecular information data characterized by SMILES representations. The objective is to establish a shared latent space while maximizing variance across all modalities, facilitating the creation of a query system capable of identifying the most similar molecules based on specific characteristics outlined in a query sentence. Through a multimodal deep learning framework employing Autoencoders, this study aims to explore synergies among these modalities to significantly enhance the capabilities of molecular query systems. As an application project with an unsupervised learning approach, this endeavor promises to advance the field by enabling more efficient and comprehensive molecular analysis.

***Keywords* —** Query Systems, Autoencoders, Data fusion, Multimodal representation

1. INTRODUCTION

Understanding the intricate properties and behaviors of molecules is paramount in various scientific domains, including drug discovery, material science, and environmental research. Traditional approaches have often focused on individual data modalities, such as textual descriptions or structural representations. However, recent advancements in machine learning and data integration techniques have opened avenues for exploiting the synergies among diverse modalities, thereby enabling a more comprehensive understanding of molecular characteristics.

2. PRIOR RESEARCH

In recent years, researchers have recognized that multimodal data often contain both complementary and common information across different modalities. Multimodal machine learning involves algorithms that learn from diverse datasets to enhance performance. Multimodal Deep Learning, a subfield of machine learning, focuses on training AI models to understand and relate different types of data (modalities) like images, video, audio, and text. By combining these modalities, models can better understand their environment. [1] discusses emotion recognition, as facial expressions (visual modality) may not fully convey emotions but considering tone and pitch in a person's voice (audio modality) can provide valuable insights. [1] emphasizes the importance of multimodal deep learning in capturing comprehensive insights from different data sources.

This concept has been further explored in subsequent studies. [2] introduced the use of autoencoders (AE) in this context, proposing models that extract lower-dimensional hidden features from individual modalities. Their first model, ConcatAE, trains AEs separately on each modality before concatenating the learned features. In contrast, their second model, CrossAE, utilizes cross-modality AEs trained to recover both modalities from each modality, capturing common information inherent across modalities.

[3] discusses the various fusion strategies, including early, joint, and late fusion. Early fusion involves merging data modalities into a common feature vector at the input layer, with preprocessing techniques like dimensionality reduction aiding in aligning input data dimensions. In contrast, joint fusion enables merging modalities at different depths of the model, allowing for the learning of latent feature representations before fusion. This approach, often referred to as end-to-end learning, integrates feature extraction from raw data within the model architecture. Finally, late fusion, akin to decision-level fusion, combines predictions from models trained on each data modality separately. Each model contributes independently to the final prediction, with aggregation methods such as averaging or majority voting employed.

[4] proposes three different architectures for fusing data from multiple modalities aim to exploit both intra- and inter-modality semantic correlations. The Shared Single-Layered Autoencoder architecture (S-SLAE) involves utilizing a single autoencoder shared among modalities, where input data from textual and visual modalities are concatenated. This approach expects the hidden layer of the autoencoder to capture inter-modality semantic relations. In contrast, the

Independent Single-Layered Autoencoder architecture (I-SLAE) employs separate autoencoders for each modality to focus on learning intra-modality semantic relations. The representations learned from these autoencoders are then concatenated to form the new item representation. Combining aspects of both S-SLAE and I-SLAE, the Two-Layered Autoencoder architecture (TLAE) incorporates distinct autoencoders for intra-modality relations in the bottom layer, while concatenating the learned representations for inter-modality relations in the top layer. These architectures provide comprehensive approaches to fusion, catering to the diverse semantic relationships within and across modalities.

3. METHODOLOGY

The methodology of this research is grounded in a multimodal deep learning framework designed to harness the complimentary of different data modalities. The primary objective is to develop a query system capable of finding the most similar molecules based on their textual descriptions, image representations, and molecular information given a query sentence. This methodology unfolds through the following steps:

3.1. Feature Extraction

Feature extraction is a crucial step in processing diverse modalities, transforming raw data into meaningful representations for further analysis.

For textual data, which encompasses molecular descriptions, GPT word embeddings are employed to capture intricate semantic nuances. Textual data comprises several key aspects related to each molecule's chemical structure, energetic properties (such as high heat of formation, detonation velocity, and pressure) explosive potential, applications, safety considerations (highlighting sensitivity to shock, friction, and heat), and regulatory aspects. This textual data is structured into 462 rows, each corresponding to a unique molecule, and 1536 features vectorized, providing a rich representation of the textual information associated with each molecule.

The query data, which is used to identify the most similar molecules based on a given a query sentence, is also represented as GPT embeddings to ensure consistency. It is structured into 10 rows, each corresponding to a unique query sentence, and 1536 features vectorized. A representative example of a query sentence is: "Give me a set of molecules that have similar detonation velocity and detonation pressure with DNTF".

In the case of image data, representing molecular structures, pre-trained ImageNET model is used to get the feature representations. The image data is structured into 462 rows, each corresponding to a unique molecule, and 1024 features.

Molecular information data, characterized by SMILES representations, transformed using mol2vec embeddings. This dataset is structured into 462 rows, each corresponding to a unique molecule, and 300 features. These embeddings encode detailed structural and chemical information, furnishing rich contextual features essential for modeling.

Ensuring numerical representations of each modality is a pivotal step in crafting effective neural network architectures. Following representation, all features are normalized to fall within the range [0, 1], ensuring uniformity and facilitating optimal model performance.

3.2 Autoencoder Integration

Autoencoders are neural network architectures commonly utilized for unsupervised learning tasks, where they aim to reconstruct input data. In this research, autoencoders were employed to capture the intrinsic structure and features of each modality. This study focuses on investigating the encoded space, also known as latent representations, of autoencoders (AEs).

In this research, I adopt the approach proposed by [4] for integrating autoencoders (AE) into the query system pipeline, Specifically, I'll consider three Autoencoder models: I-SLAE, S-SLAE, TLAE.

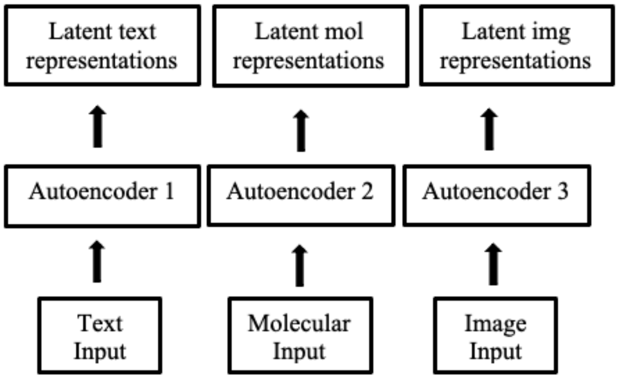
Latent text representations
Latent mol representations
Latent img representations
Autoencoder 1
Autoencoder 2
Autoencoder 3
Text Input
Molecular Input
Image Input

Fig. 1. I-SLAE Model Architecture

In the Independent Single-Layered Autoencoder architecture (I-SLAE), individual autoencoders are trained independently for each modality, including text, image, and molecular descriptions, as well as for query data comprising query sentences. Each AE aims to learn to encode the input data into lower-dimensional representations (300 dimensions) specific to its modality. Subsequently, the latent representations obtained from each distinct autoencoder are utilized within the query system.

As [3] emphasizes on how each modality contributes independently to the final prediction by introducing aggregation methods. I introduce the usage of different weights applied to the cosine similarity of each encoded modality and the encoded query sentences depending on the number of modalities used.

In the query process, when employing single encoders, cosine similarity scores are computed between each modality and the query sentence encodings. For queries involving two modalities, the cosine similarity scores between each modality and the query sentences are calculated, with a weight of 0.50 applied to each modality (text and molecules, molecules and image, text and image) to obtain a combined similarity score. Lastly, for the query involving all three modalities, cosine similarity scores for each modality with the query sentences are calculated, and each modality's score is multiplied by 0.33 for aggregation. This approach enables the comprehensive integration of multiple modalities in the query system, leveraging the strengths of individual encoders while ensuring effective fusion of information for enhanced query processing.

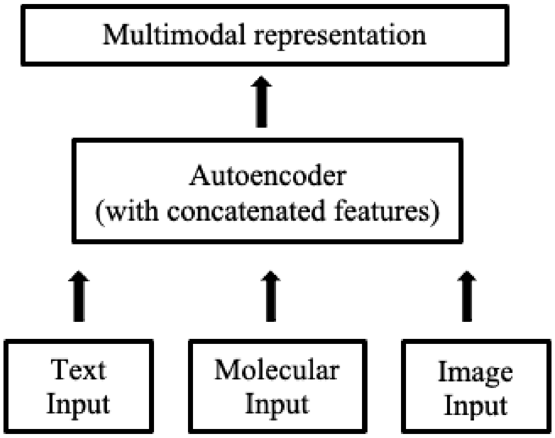

Fig. 5. S-SLAE Model Architecture

In the Shared Single Layered Autoencoder architecture (S-SLAE), the input data consists of the concatenated form of text, image, and molecular descriptions. This concatenated input is fed into a single autoencoder, which encodes the data into a latent representation. This latent representation (multimodal representation) is then used to compute the cosine similarity between the concatenated encodings and the encoded query sentences.

To explore the impact of different dimensionalities in the latent representation, multiple models were developed within the S-SLAE architecture. The simplest model contains 300 neurons in the encoder layer, resulting in a 300-dimensional latent representation. This model aims to capture essential features while maintaining simplicity and efficiency.

Additionally, deeper models were constructed with 512 neurons in the encoder layer, yielding a higher-dimensional latent representation of 512 dimensions. To prevent overfitting and improve generalization, a dropout rate of 0.50 was applied for regularization in these deeper models. By varying the dimensionality of the latent representation and incorporating dropout regularization, the study aims to assess the performance and robustness of the S-SLAE architecture in capturing and representing information from multiple modalities.

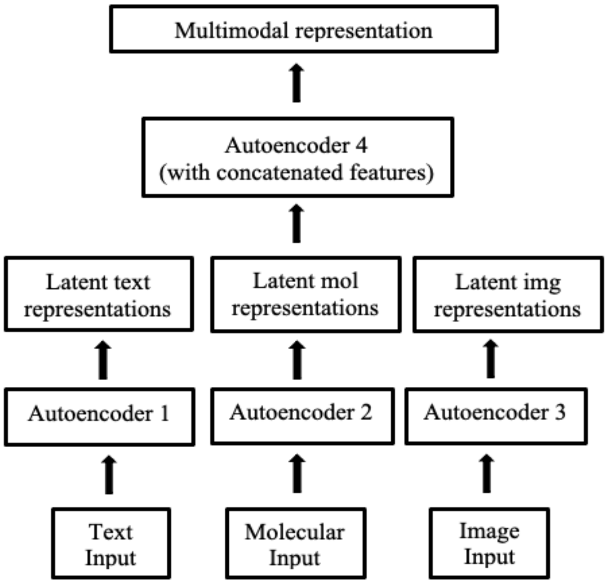

Fig. 6. TLAE Model Architecture

In the Two-Layered Autoencoder (TLAE) architecture, distinct autoencoders are employed in the bottom layer to individually capture intra-modality relations for each modality, including text, images, and molecular descriptions. Subsequently, in the top layer, the latent representations acquired from each modality's autoencoder are concatenated into a unified feature vector. This concatenated representation is then utilized by another autoencoder to collectively capture inter-modality semantic relations. The resulting encoded representations, obtained from the top layer's autoencoder, are used for calculating cosine similarity scores with the query sentences to retrieve the top five most similar molecules corresponding to each query sentence.

Three different architectures were developed with TLAE. The first model consists of autoencoders with 100 neurons in each encoder layer, resulting in a 300-unit representation focusing on capturing a more concise representation. The second model is constructed with dropout of 0.25 with 300 dimensions. In contrast, the third model employs autoencoders with 300 neurons in each encoder layer, yielding a 900-unit representation aiming to capture a richer and more detailed representation of the input data with dropout of 0.50 to prevent overfitting.

3.3 Fusion Approaches

The methodology will involve integrating information from different modalities using two different fusion approaches: early and late fusion.

3.3.1. Early Fusion Approach

In the context of early fusion, where data from different modalities are integrated into a unified feature representation at the beginning of the model. This latent representation serves as a unified feature representation that combines information from all modalities. Therefore, models developed within the S-SLAE architecture, such as those with 300 and 512 neurons in the encoder layer, fall under the early fusion approach.

3.3.2. Late Fusion Approach

In contrast to early fusion, the late fusion approach involves training separate models for each modality independently. Each modality-specific model is optimized individually to capture the unique characteristics and patterns inherent in its respective data modality. Once trained, predictions from these individual models are aggregated using ensemble learning techniques such as averaging.

In the I-SLAE architecture, each modality is processed independently using distinct autoencoders. This enables each autoencoder to capture the unique characteristics and patterns inherent in its respective data modality. Subsequently, the latent representations generated by each modality-specific autoencoder are calculated in the query system with respect to the number of modalities used. Therefore, the I-SLAE architecture falls under the late fusion approach as it involves training separate models for each modality and then aggregating their representations.

Similarly, the TLAE architecture combines both intra-modality and inter-modality semantic relations. In the bottom layer, distinct autoencoders are employed to learn intra-modality relations independently for each modality. Then, in the top layer, the representations learned from each modality are concatenated and used by a different autoencoder to capture inter-modality semantic relations. Therefore, both the I-SLAE and TLAE architectures are examples of late fusion approaches as they involve training separate models for each modality and then combining their representations.

The evaluation of these fusion approaches will be conducted using Mean Squared Error (MSE) metric and cosine similarity scores achieved from the top five molecules given a query sentence. This comparative analysis informs our understanding of how early and late fusion techniques, demonstrated by the I-SLAE, S-SLAE, and TLAE architectures, perform in capturing and aggregating modality-specific information to predict similar molecules based on query sentences. By comparing the results obtained from different approaches, I aim to determine the most effective strategy for predicting the most similar molecules given a query sentence.

4. RESULTS AND DISCUSSION

The results are compared using the Query Sentence 8 from the set of queries used to evaluate the model performance.

Query 8: Give me a set of molecules that have similar formulation process with CL-16 and have low Gurney energy.

Fusion Approach	**Model**	**Similarity Score**
Late Fusion	I-SLAE (text)	0.213
	I-SLAE (image)	0.452
	I-SLAE (molecules)	0.408
	I-SLAE (text & mol)	0.307
	I-SLAE (mol & img)	0.423
	I-SLAE (text & img)	0.367
	I-SLAE (text, img, mol)	0.335
	TLAE (0.25 Dropout, 900D)	0.346
	TLAE (0.25 Dropout, 300D)	0.348
	TLAE (No Dropout, 300D)	0.381
Early Fusion	S-SLAE Simple Model (300D)	0.373
	S-SLAE Deeper Model (512D)	0.290
	S-SLAE Further Deepened Model (0.5 Dropout, 512D)	0.157

Table 1 cosine similarity score of the most similar molecule with each model

The model comparison table (Table 1) illustrates the cosine similarity scores obtained by various fusion approaches and architectures in response to Query Sentence 8. In the late fusion approach, the I-SLAE architecture demonstrates varying degrees of effectiveness across different modalities. Notably, the I-SLAE models trained on individual modalities, namely text, image, and molecular descriptions, exhibit diverse performance, with the image-based I-SLAE model achieving the highest similarity score of 0.452. However, combining modalities in the I-SLAE architecture yields mixed results, with the combination of text and image achieving a moderate similarity score of 0.367 and with using all the modalities 0.335. Similarly, the TLAE architecture, with different dropout rates and dimensionalities, provides comparable results, with the model employing no dropout and 300 dimensions achieving the highest score of 0.381.

In contrast, the early fusion approach, represented by various S-SLAE models with different dimensionalities, shows a less consistent performance. While the simple model with 300 dimensions achieves a relatively high similarity score of 0.373, the more complex models exhibit lower scores, indicating a potential trade-off between model complexity and performance.

Fusion Approaches	**Model**	**Training Loss**	**Validation Loss**
Late Fusion	I-SLAE (text)	0.0203	0.0263
	I-SLAE (image)	0.0641	0.0853
	I-SLAE (molecules)	0.0046	0.0064
	I-SLAE (queries)	0.2161	0.2441
	TLAE (0.25 Dropout, 900D)	0.0369	0.0480
	TLAE (0.25 Dropout, 300D)	0.0226	0.0391
	TLAE (No Dropout, 300D)	0.0390	0.0509
Early Fusion	S-SLAE Simple Model (300D)	0.0402	0.0521
	S-SLAE Deeper Model (512D)	0.0384	0.0515
	S-SLAE Further Deepened Model (0.5 Dropout, 512D)	0.0530	0.0615

Table 2 Mean Squared Error (training and validation) loss of each model

The Mean Squared Error (MSE) loss values in Table 2 offer crucial insights into the training and validation performance of each model. Lower MSE values indicate better alignment between predicted and actual similarity scores, reflecting the models' efficacy in capturing semantic similarities within the multimodal data. In the late fusion approach, the I-SLAE models demonstrate relatively low MSE values across all modalities, indicating effective learning of semantic relations. Specifically, the I-SLAE model trained on molecular descriptions achieves the lowest MSE for both training and validation, suggesting robust learning of molecular features. However, the text-based I-SLAE model exhibits slightly higher MSE values, indicating some challenges in capturing semantic relations from textual data.

The TLAE architecture also exhibits competitive performance, suggesting successful fusion of information from different modalities while minimizing prediction error. Conversely, the early fusion approach, particularly the deeper and more complex S-SLAE models, shows higher MSE values, possibly indicating overfitting or difficulty in generalizing to unseen data. These results highlight the trade-offs between model complexity, generalization ability, and predictive performance.

Despite the overall good performance, it's essential to note that the cosine similarity scores are relatively low compared to the models' learning abilities. This suggests a limitation of cosine similarity as a sole evaluation metric. Therefore, it's crucial to develop supplementary assessment criteria to ensure a thorough evaluation of model performance. Other metrics to evaluate the performance of the query system should be explored to provide a more comprehensive understanding of its capabilities and limitations.

5. REFERENCES

[1] Poulinakis, K.: Multimodal Deep Learning: A Complete Guide. https://www.v7labs.com/blog/multimodal-deep-learning-guide Accessed Access Date

[2] Tong, C., Li, J., Lang, C., Kong, F., Niu, J., Rodrigues, J.J.: An efficient deep model for day-ahead electricity load forecasting with stacked denoising auto-encoders. Journal of parallel and distributed computing 117, 267–273 (2018)

[3] Akkus, C., Chu, L., Djakovic, V., Jauch-Walser, S., Koch, P., Loss, G., Marquardt, C., Moldovan, M., Sauter, N., Schneider, M., et al.: Multimodal deep learning. arXiv preprint arXiv:2301.04856 (2023)

[4] Conceic ao, F.L., Padua, F.L., Lacerda, A., Machado, A.C., Dalip, D.H.: Multimodal data fusion framework based on autoencoders for top-n recommender systems. Applied Intelligence 49, 3267–3282 (2019)