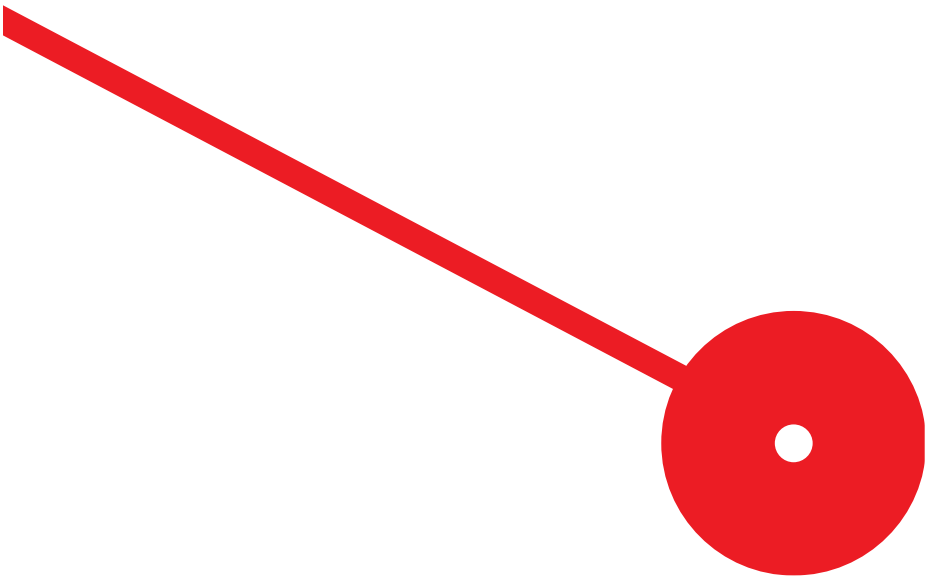
INSTITUTO
SUPERIOR
DE CONTABILIDADE
E ADMINISTRAÇÃO
DO PORTO
POLITÉCNICO
DO PORTO

M MESTRADO
BUSINESS INTELLIGENCE AND ANALYTICS

Machine Learning for Single-Modality and Multi-Modality Data Integration in the Materials Industry

Vítor José Figueiredo Costa

10/2024



Vítor José Figueiredo Costa. Machine Learning for Single-Modality and Multi-Modality Data Integration in the Materials Industry 10/2024

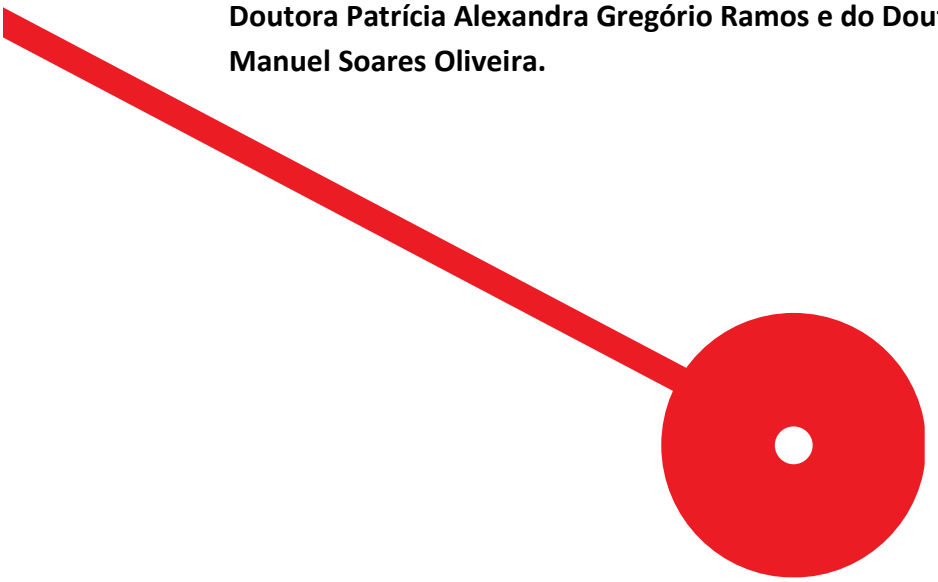
INSTITUTO
SUPERIOR
DE CONTABILIDADE
E ADMINISTRAÇÃO
DO PORTO
POLITÉCNICO
DO PORTO

M MESTRADO
BUSINESS INTELLIGENCE AND ANALYTICS

Machine Learning for Single-Modality and Multi-Modality Data Integration in the Materials Industry

Vítor José Figueiredo Costa

Dissertação de Mestrado apresentada ao Instituto Superior de Contabilidade e Administração do Porto para a obtenção do grau de Mestre em Business Intelligence and Analytics, sob orientação da Doutora Patrícia Alexandra Gregório Ramos e do Doutor José Manuel Soares Oliveira.



Vítor José Figueiredo Costa. Machine Learning for Single-Modality and Multi-Modality Data Integration in the Materials Industry 10/2024

Resumo:

Esta tese explora a aplicação de técnicas de aprendizagem automática para prever as propriedades dos materiais utilizando a integração de dados multimodais. O aparecimento de técnicas computacionais avançadas e a disponibilidade de grandes conjuntos de dados abriram novos caminhos para acelerar a descoberta de materiais utilizando modelos de previsão. No entanto, a previsão exata das propriedades dos materiais continua a ser um desafio complexo devido à natureza intrincada dos dados dos materiais. Os modelos de aprendizagem automática de modalidade única, embora eficazes para determinadas propriedades, não conseguem frequentemente captar toda a complexidade das caraterísticas dos materiais.

Esta tese aborda esta limitação investigando o impacto da integração de dados multimodais, centrando-se especificamente na forma como as combinações de texto, imagem e dados tabulares melhoram a precisão da previsão das propriedades dos materiais. O estudo utiliza o conjunto de dados Alexandria, um recurso abrangente que oferece dados pormenorizados sobre as composições químicas e as propriedades de milhões de materiais. Um subconjunto de 1000 materiais deste conjunto de dados foi utilizado para construir um conjunto de dados multimodal que incorpora: composição química representada como uma sequência de elementos e as respetivas contagens de átomos (modalidade de texto); visualizações 2D da estrutura cristalina 3D de cada material, geradas com o Crystal Toolkit e captadas através de uma aplicação Web personalizada (modalidade de imagem); e *embeddings* estruturais de tamanho fixo gerados com a arquitetura PotNet, um modelo de rede neural gráfica concebido para captar interações atómicas complexas (modalidade tabular).

O estudo utilizou o AutoGluon-Multimodal (AutoMM), uma estrutura de aprendizagem automática de máquinas, para treinar e avaliar modelos utilizando várias combinações de modalidades. O Erro Absoluto Médio (MAE) e o Erro Escalado Absoluto Médio (MASE) foram utilizados como métricas de avaliação. Os resultados demonstram que as abordagens multimodais, especialmente a combinação de dados de texto e imagem, superam consistentemente os modelos de modalidade única. Este facto realça a importância da integração de diversos tipos de dados para captar uma compreensão mais abrangente das propriedades dos materiais. Nomeadamente, a combinação de Texto e Imagem revelou-se particularmente eficaz na previsão de caraterísticas complexas como o intervalo de banda (Gap), que requer informações estruturais e de composição

complexas. Por outro lado, os modelos de modalidade única, particularmente os que se baseiam apenas em dados tabulares, apresentaram a menor precisão na maioria das caraterísticas. Esta investigação fornece provas convincentes dos benefícios da integração de dados multimodais na previsão das propriedades dos materiais. Estabelece uma base para trabalhos futuros que explorem a incorporação de tipos de dados adicionais, o desenvolvimento de modelos mais avançados e a expansão de conjuntos de dados para melhorar ainda mais a precisão da previsão e acelerar a descoberta de novos materiais com as propriedades desejadas.

Palavras-chave: *Machine Learning*, Multimodalidades, Modelos Multimodais, Ciência de Materiais

Abstract:

This dissertation explores the application of machine learning techniques for predicting material properties using multimodal data integration. The emergence of advanced computational techniques and the availability of large datasets have opened new avenues for accelerating material discovery using predictive models. However, accurately predicting material properties remains a complex challenge due to the intricate nature of material data. Single-modality machine learning models, while effective for certain properties, often fail to capture the full complexity of material characteristics.

This dissertation addresses this limitation by investigating the impact of multimodal data integration, specifically focusing on how combinations of Text, Image, and Tabular data enhance material property prediction accuracy. The study utilizes the Alexandria dataset, a comprehensive resource offering detailed data on the chemical compositions and properties of millions of materials. A subset of 1,000 materials from this dataset was used to construct a multimodal dataset incorporating: chemical composition represented as a sequence of elements and their corresponding atom counts (Text modality); 2D visualizations of each material's 3D crystal structure generated using Crystal Toolkit and captured via a custom-built web application (Image modality); and fixed-size structural embeddings generated using the PotNet architecture, a graph neural network model designed for capturing complex atomic interactions (Tabular modality).

The study employed AutoGluon-Multimodal (AutoMM), an automated machine learning framework, to train and evaluate models using various modality combinations. Mean Absolute Error (MAE) and Mean Absolute Scaled Error (MASE) were used as evaluation metrics. Results demonstrate that multimodal approaches, especially the combination of text and image data, consistently outperform single-modality models. This highlights the importance of integrating diverse data types to capture a more comprehensive understanding of material properties. Notably, the Text and Image combination proved particularly effective for predicting complex features like band gap (Gap), which requires intricate compositional and structural information. Conversely, single-modality models, particularly those relying solely on Tabular data, exhibited the lowest accuracy across most features. This research provides compelling evidence for the benefits of multimodal data integration in material property prediction. It lays a foundation for future work exploring the incorporation of additional data types, the development of more advanced

models, and the expansion of datasets to further enhance predictive accuracy and accelerate the discovery of novel materials with desired properties.

Keywords: Machine Learning, Multimodalities, Multimodal Models, Materials Science

Index

Chapter – Introduction 1

1 Introduction 2

Chapter II – Modalities 6

2 Modalities 7

2.1 Data Modalities 9

2.1.1 Tabular Data 9

2.1.2 Image Data 10

2.1.3 Text Data 11

2.2 Single Modality Machine Learning Models 12

2.2.1 Composition Based Models (Text) 13

2.2.2 Graph Neural Networks (GNNs) 16

2.2.3 Image Based Models 20

Chapter III – Multimodal Learning Models and Frameworks 23

3 Multimodal Learning Models and Frameworks 24

3.1 Fusion Techniques 24

3.1.1 Early Fusion (Feature-Level Fusion) 25

3.1.2 Late Fusion (Decision-Level Fusion) 25

3.1.3 Hybrid Fusion 26

3.2 CLIP 27

3.3 MultiMat 28

3.4 AutoGluon-Multimodal (AutoMM) 30

3.5 Challenges in Multimodal Learning 32

3.5.1 Computational Complexity 32

3.5.2 Data Integration 33

3.5.3 Interpretability 33

Chapter IV – Empirical Study 35

4 Empirical Study 36
4.1 Data Understanding 36
4.2 Dataset Construction and Multimodal Generation 39
4.2.1 Image Generation 40
4.2.2 Tabular Generation 41
4.2.3 Text Generation 45
4.2.4 Features 47
4.3 Dataset Alignment 48
4.4 Model Building 48
4.5 Evaluation and Analysis 49
Chapter V – Conclusion 57
5 Conclusion 58
References 60

List of Figures

Figure 1 - Ac_2IrCu Material from the Materials Project Database (Jain et al., 2013). ... 10
Figure 2 - Overall pre-training and fine-tuning procedures for BERT (Devlin et al., 2019). 14
Figure 3 - Overall CrabNet architecture and prediction of material property and uncertainty (A. Y.-T. Wang et al., 2021). 15
Figure 4 - Some applications where the information is represented by graphs: (a) a chemical compound (adrenaline), (b) an image (Scarselli et al., 2009). 16
Figure 5 - Illustration of the crystal graph convolutional neural networks (Xie & Grossman, 2018a). 17
Figure 6 - Schematic illustrations of how complete interatomic interactions are captured in PotNet (Lin et al., 2023). 18
Figure 7 - The developed network architecture for PotNet (Lin et al., 2023). 19
Figure 8 - MultiMat Framework Diagram. 29
Figure 9 - AutoMM architecture. 31
Figure 10 - Materials Project Ac2AgIr 3D rendered object. 40
Figure 11 - 100 of the materials 2D representations of the 3D renderization generated. 41
Figure 12 - Graphic representation of MAE for each target. 55
Figure 13 - Heatmap of MASE Values Across Modality Combinations and Features. . 56

List of Tables

Table 1 - MAE results for modalities combination. 50
Table 2 - MASE results for modalities combination. 51

List of Code Snippets

Code Snippet 1 - JSON object for the material $Ba(SrPd)_2$. 37
Code Snippet 2 - CIF generation function. 43
Code Snippet 3 - Node Attributes and Edge Indices calculations. 44
Code Snippet 4 - Setup of Convultional and Transformer layer. 45
Code Snippet 5 - Autogluon's MultiModal predictor configuration. 49

List of Abbreviations

CGCNN - Crystal Graph Convolutional Neural Network

CIF - Crystallographic Information File

CLIP - Contrastive Language-Image Pre-Training

CNN - Convolutional Neural Network

DL - Deep Learning

GNN - Graph Neural Network

LLM - Large Language Model

LMM - Large Multimodal Model

MAE - Mean Absolute Error

MASE - Mean Absolute Scaled Error

ML - Machine Learning

MLLM - Multimodal Large Language Model

NLP - Natural Language Processing

PNG - Portable Network Graphics

VAE - Variational Auto Encoder

1 Introduction

Materials Science is a multidisciplinary field that explores the properties, performance, and synthesis of materials, aiming to innovate and optimize materials for various applications. The advent of advanced computational techniques and the explosion of available data have opened new frontiers in this field, enabling the development of predictive models and simulations that can accelerate the discovery and design of novel materials. Traditional methods of materials discovery are often time-consuming and resource-intensive, necessitating a shift towards more efficient, data-driven approaches.

The integration of machine learning into materials research offers an opportunity to overcome these limitations, providing predictive models that can streamline the materials discovery pipeline.

However, accurately predicting material properties remains a complex challenge due to the diverse and intricate nature of material data. While many machine learning models have achieved success with single-modality data (e.g., tabular data, text, or images), recent studies suggest that multimodal data integration—combining various types of data—may enhance predictive accuracy by capturing complementary insights. Multimodal approaches allow for a richer representation of material properties, potentially capturing compositional, structural, and spatial information that single-modality models may overlook.

This dissertation explores the impact of multimodal data on predictive accuracy, examining how combinations of text, images, and tabular data can improve material property predictions.

The significance of this study lies in its potential to transform the field of Materials Science through the use of advanced machine learning and data integration techniques. Despite the promise of multimodal machine learning in materials science, there is limited understanding of how different modality combinations influence predictive performance across a range of material properties.

Single-modality models, while effective in certain contexts, often fail to capture the full complexity of material characteristics, leading to suboptimal predictions for features with intricate dependencies.

This dissertation addresses this gap by systematically evaluating different modality combinations, investigating their effect on predictive accuracy for key material properties.

By integrating diverse data types and leveraging advanced machine learning techniques, this research aims to significantly enhance the efficiency and effectiveness of materials discovery and innovation.

In traditional machine learning workflows, methodologies such as CRISP-DM (Chapman et al., 1999) guide the work through a sequence of structured steps, from data understanding to deployment. However, this work deviated from these traditional methodologies to adopt a workflow specifically designed to address the unique requirements of multimodal data integration and the automatic optimization capabilities offered by AutoGluon.

We used a custom methodology developed for this work, which focused on multimodal data creation, streamlined data preparation, and automated fine-tuning and hyperparameter selection. By leveraging a structured multimodal dataset (Alexandria) and automated model tuning, this approach allowed us to prioritize building the multimodal dataset and applying targeted fusion techniques.

Our approach can be summarized in the following phases:

1. Data Understanding;
2. Dataset Construction and Multimodal Generation;
3. Data Alignment;
4. Model Building;
5. Evaluation and Analysis.

These phases align with the work's emphasis on dataset creation, minimal preprocessing, and efficient model tuning rather than the more iterative, exploratory steps typically required in traditional machine learning pipelines.

Dissertation Outline

This dissertation outlines a comprehensive approach to leveraging machine learning for improved material property prediction through multimodal data integration, addressing

the limitations of traditional discovery methods and single-modality models in capturing the complexity of material characteristics.

Chapter 1 introduces the study's foundation in materials science, underscoring the time-intensive nature of traditional discovery methods and the promise of machine learning to streamline material discovery. While single-modality machine learning has shown limited success, multimodal data integration is identified as a more comprehensive approach. The dissertation employs a custom methodology to address challenges specific to multimodal data, incorporating data understanding, dataset construction, model building, and evaluation using AutoGluon.

Chapter 2 outlines the core data modalities: tabular data for numerical properties, image data for visual atomic structures, and text data for chemical composition. It describes single-modality machine learning models, such as BERT for text, graph neural networks like CGCNN and PotNet for structural data, and CNNs and vision transformers for image analysis.

Chapter 3 focuses into multimodal learning models and frameworks, explaining the benefits and challenges of fusion techniques—early, late, and hybrid fusion—and introduces models such as CLIP, MultiMat, and AutoGluon-Multimodal (AutoMM), highlighting their applicability to materials science. Challenges discussed include computational complexity, data integration, and interpretability issues.

Chapter 4 describes the construction of the multimodal dataset using the Alexandria dataset, detailing the steps for creating text, image, and tabular representations, feature selection, and data alignment. PotNet embeddings, a web application for image generation, and methods for formatting text data as chemical compositions were employed. The model-building phase used AutoGluon’s MultiModalPredictor to automate the workflow, optimize models, and evaluate performance.

Chapter 5 concludes by affirming that multimodal approaches, especially integrating text and image data, consistently enhance predictive accuracy compared to single-modality models. However, single-modality models relying solely on tabular data were less accurate. Notably, certain features, such as band gap, remain challenging to predict. Future research directions include expanding datasets, incorporating additional data types, and exploring specialized models for materials science. The study ultimately

reinforces multimodal integration's promise in advancing material property predictions and accelerating novel materials discovery.

In the evolving landscape of representation learning and artificial intelligence, the concept of multimodality has a pivotal role, aiming to mimic the human cognitive ability to process and integrate information from various sources.

2 Modalities

The world surrounding us involves multiple modalities — we see objects, hear sounds, feel texture, smell odors, and so on (Baltrušaitis et al., 2017). As mentioned by Summaira (2021) our world is, indeed, inherently multimodal.

In general terms, a modality refers to the way in which something happens or is experienced (Baltrušaitis et al., 2017). To convey the comprehensive information about objects in the world, various cognitive signals describing different aspects of the same object are recorded in different kinds of media such as text, image, video, sound, and graph (Guo et al., 2019). Having this in mind, multimodality becomes one of the main research directions to be pursued.

Unimodal or Single Modality Machine Learning

Unimodal or Single Modality machine learning is a learning paradigm that focuses on representations of the same type. It is an approach that has been widely used in machine learning and AI research in recent years. Examples of unimodal machine learning include systems that can learn from text, images, or graphs (Škrlj, 2024) .

Researchers are now better able to comprehend the nuances of individual modalities thanks to unimodal machine learning. This method has made it possible to create algorithms that can learn from large amounts of data and use that knowledge to generate predictions. In this way, unimodal machine learning offers a strong basis for creating increasingly intricate and advanced machine learning algorithms, paving the way for the integration of two or more types of modalities into a single model.

Multimodal Machine Learning

Unlike unimodal machine learning, multimodal machine learning incorporates two or more types of data (modalities) into a single model. The use of multiple modalities in machine learning has become increasingly important since many real-world applications require the integration of information from multiple sources. For example, in autonomous driving, it is necessary to process information from cameras, lidar, and radar sensors to

make decisions in real time. In speech recognition, audio and text modalities are often combined to improve accuracy (Škrlj, 2024).

A key challenge in multimodal machine learning lies in effectively integrating information from various modalities. Solutions to this challenge include fusion-based, graph-based, and attention-based approaches, each offering unique strategies for combining multimodal data.

By effectively combining information from different modalities, researchers can develop powerful machine learning algorithms that can handle complex real-world scenarios.

A recent overview on multimodal learning identifies several core challenges central to this domain:

1. Multimodal Alignment: The goal is to define universal principles governing interactions across various modalities, with cross-modal interaction identification emerging as a major focus (Liang et al., 2023).
2. Compositionality: This principle, foundational to neural networks, supports hierarchical learning. Neural networks transform raw data to higher-level representations, enabling reasoning and extending to out-of-distribution scenarios (Škrlj, 2024).
3. Representation Generation: Research here focuses on efficient modality-specific representation creation, often through encoder-decoder architectures (Y. Chen et al., 2023).
4. Transferability of Representations: Modern multimodal frameworks prioritize leveraging pre-trained models to jump-start new tasks with minimal data, promoting transferability within or across modalities (Y. Chen et al., 2023).
5. Representation Fusion: One of the most complex tasks, representation fusion, seeks to combine diverse semantic levels of data (e.g., images, text, or knowledge graphs) seamlessly. This requires normalization and embedding processes before integration, crucial for scalable multimodal learning applications (Škrlj, 2024).

These open challenges underscore the potential of multimodal learning to expand machine learning applications by integrating varied data forms and improving model adaptability and efficiency.

2.1 Data Modalities

Understanding and predicting material properties requires data that can capture the complex interactions and characteristics inherent to different materials. In materials science, data comes in various **modalities**—distinct types of information that each provide unique insights into the properties, structure, and behaviour of materials. Machine learning models can leverage these modalities to make more accurate predictions and enable the discovery of novel materials with desired properties.

We examine some of the primary data modalities used in materials science, highlighting how each modality contributes distinct information. The main modalities discussed include **tabular data**, **images**, and **textual composition data**. Each of these data types holds valuable information that, when used in isolation, offers a limited view of a material but becomes powerful when integrated within multimodal learning frameworks.

2.1.1 Tabular Data

Tabular data is one of the foundational modalities in materials science, encapsulating structured, numerical information about a material's atomic and physical properties. This type of data is typically organized in a table format, where each row represents a unique material, and each column corresponds to specific properties or features of that material, such as atomic coordinates, bond lengths, elemental compositions, or calculated thermodynamic properties. Due to its structured nature, tabular data is widely used in machine learning as it enables precise, quantitative analysis and can be easily fed into models that are well-suited for numerical data processing, like linear regression, support vector machines, and neural networks.

In materials science, tabular data often originates from computational chemistry calculations or experimental measurements. For example, large datasets such as the Materials Project database contain tabular representations of millions of materials, including properties derived from Density Functional Theory (DFT) calculations (Jain et al., 2013). These datasets capture vital information about each material, such as:

1. Atomic Structure: Including details like atomic coordinates, bond types, and lattice parameters that define the spatial arrangement of atoms within a crystal.
2. Electronic Properties: Such as band gaps, electronic densities, and magnetic properties.

3. Thermodynamic and Mechanical Properties: Like enthalpy of formation, elasticity, and thermal conductivity, which are crucial for predicting material stability and suitability for various applications.

Models can utilize tabular data to predict properties directly or to create embeddings-representations that capture essential patterns within the data. For instance, models such as PotNet (Lin et al., 2023) use interatomic potentials to create tabular embeddings of crystal structures.

2.1.2 Image Data

Image data in materials science captures the spatial and visual representation of materials, particularly their atomic and crystal structures.

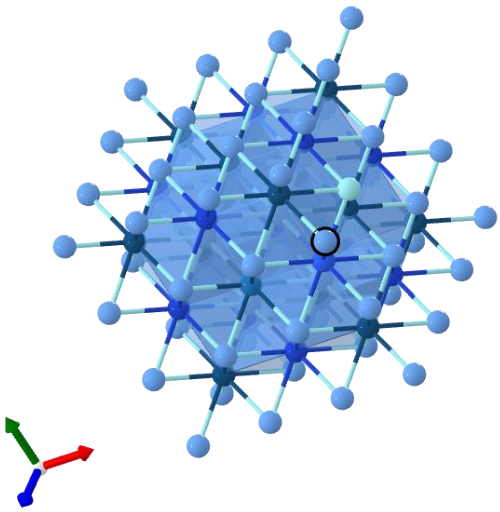


Figure 1 - Ac_2IrCu Material from the Materials Project Database (Jain et al., 2013).

Images such as the one in Figure 1 provide a way to visually observe the arrangement and geometry of atoms within a material, revealing crucial insights into its properties and behaviors.

Unlike tabular or textual data, which represent materials in structured formats or as compositional descriptions, image data enables machine learning models to directly analyze and interpret the spatial context and structural patterns inherent in materials.

Image data often comes from simulations or visualization tools, such as Crystal Toolkit, that render 3D structures into 2D image formats, such as PNG or JPEG (Horton et al., 2023). These images can represent materials at different scales, from the microscopic level (showing atomic bonds and lattice configurations) to the macroscopic level (illustrating grain boundaries or phase structures in metals and alloys). Through these visualizations, researchers gain insight into lattice symmetry, bond lengths, angles, and defects—features that can be indicative of a material's mechanical, thermal, and electrical properties.

2.1.3 Text Data

Text data in materials science typically represents the chemical composition and elemental makeup of a material. This modality conveys critical information about which elements are present in a material, their quantities, and sometimes even the structural details in the form of descriptive text. Unlike image data, which provides a visual representation, or tabular data, which organizes quantitative properties, text data serves as a straightforward means of conveying compositional information, often in the form of formulas, symbols, or chemical notations.

In ML applications, text data can be used to predict material properties based solely on the elemental composition. Models process this data as sequences of symbols or tokens (for example, "Fe2O3" for iron oxide) to learn relationships between different compositions and their corresponding material properties. Text data is crucial in materials science because it provides a high-level view of the material's composition that can help in identifying basic properties, such as whether a material is metallic, non-metallic, or likely to exhibit magnetic behaviour.

The Alexandria Dataset is a large-scale, open-source resource offering detailed data on the chemical compositions and properties of millions of materials. It provides, on the same footing, calculations for one-dimensional (1D), two-dimensional (2D), and three-dimensional (3D) materials, primarily derived from density functional theory (DFT) calculations (Schmidt et al., 2024).

As one of the most extensive datasets available in materials science, Alexandria includes composition-based data, structural properties, and calculated material properties, making it a valuable resource for both single-modality and multimodal machine learning applications.

This dataset offers a comprehensive source of text data, with each entry's chemical formula represented as text, allowing for text-based machine learning models to leverage millions of examples. Its extensive compositional coverage helps models generalize across a wide range of materials, enabling predictions even for new or less common compounds. Additionally, the availability of calculated properties in Alexandria allows researchers to validate models and improve accuracy by training on high-quality, computationally derived data.

Alexandria's combination of text (composition) and tabular (structural properties) data makes it ideal for multimodal models, where text data on composition is combined with structural embeddings or images to achieve a more holistic understanding of materials.

The vast size of Alexandria makes it suitable for data augmentation and transfer learning. Machine learning models can pre-train on Alexandria's data and transfer knowledge to smaller or more specialized datasets, expanding the applicability and accuracy of predictions across a wide range of material types.

2.2 Single Modality Machine Learning Models

Single modality machine learning models in materials science rely on a single type of data to make predictions about material properties. For example, **composition-based models** focus exclusively on **chemical composition data** (text modality) to infer properties, such as stability, hardness, conductivity, or elasticity, based solely on the elements and their ratios within a material. This approach allows for simpler models that are computationally efficient and effective, especially in situations where structural or visual data may not be readily available (Škrlj, 2024).

Composition-based models use text representations of materials, such as elemental formulas, to identify patterns between the types of elements in a material and its expected properties. Despite their limitations—mainly the lack of structural information—these models have shown remarkable predictive power by using large datasets and advanced neural network architectures like **transformers** (P. Xu et al., 2023). Two leading models in this category are **CrabNet** which leverages compositional data to make accurate predictions and **BERT** a foundation model in natural language processing.

2.2.1 Composition Based Models (Text)

2.2.1.1 BERT

BERT, introduced by Devlin (2019), stands for "Bidirectional Encoder Representations from Transformers" and marks a departure from traditional, unidirectional language models (e.g., OpenAI's GPT (Radford et al., 2018)), which predict tokens based on either left-to-right or right-to-left context (Vaswani et al., 2023). BERT's architecture, in contrast, uses a bidirectional transformer encoder that leverages context from both directions simultaneously. This capability allows BERT to achieve a deeper understanding of language context, enhancing performance across a range of NLP tasks like question answering, language inference, and named entity recognition.

BERT uses two primary pre-training tasks to capture linguistic patterns and relationships:

1. **Masked Language Modeling (MLM):** This objective, inspired by the Cloze task (Taylor, 1953), randomly masks 15% of tokens in each sequence. The model then attempts to predict these masked tokens, leveraging context from both directions. MLM thus forces BERT to develop a nuanced, bidirectional understanding of language, which distinguishes it from traditional, unidirectional language models that process context in one direction only.
2. **Next Sentence Prediction (NSP):** This task is designed to capture relationships between sentence pairs. BERT is trained on paired sentences, where 50% of pairs are actual consecutive sentences, and 50% are randomly paired sentences from the dataset. The model learns to predict whether the second sentence follows the first, which is particularly useful for tasks involving sentence pair relationships, like question answering or natural language inference.

Once pre-trained, BERT can be fine-tuned on specific NLP tasks by adding minimal task-specific layers, making it highly adaptable. During fine-tuning, BERT learns task-specific patterns with the pre-trained bidirectional embeddings, which significantly enhances performance without extensive task-specific architecture engineering. BERT achieves state-of-the-art results across diverse benchmarks, including GLUE (A. Wang et al., 2019), SQuAD (Rajpurkar et al., 2016), and SWAG (Zellers et al., 2018).

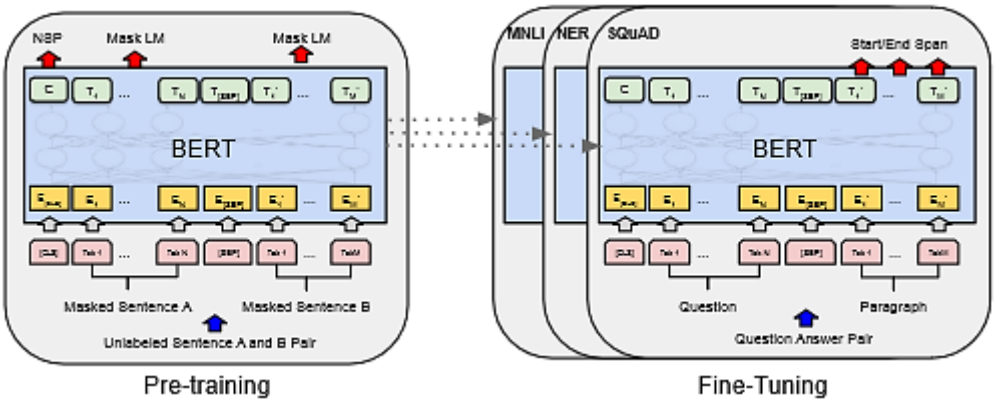


Figure 2 - Overall pre-training and fine-tuning procedures for BERT (Devlin et al., 2019).

The procedures for pre-training and fine-tuning, illustrated in Figure 2, show that apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. [CLS] is a special symbol added in front of every input example, and [SEP] is a special separator token (e.g. separating questions/answers).

BERT's architecture and training setup set a new standard for pre-trained language models, achieving state-of-the-art results across multiple NLP benchmarks and inspiring a wave of similar transformer-based models like MatBERT (MatBERT, 2021) for the materials science field. Its adaptability across tasks and reduced need for extensive task-specific engineering have cemented BERT as a foundational model in NLP, fostering research in transfer learning and model interpretability.

2.2.1.2 CrabNet

CrabNet (Compositionally Restricted Attention-Based Network) is a state-of-the-art model for predicting material properties based solely on chemical composition data.

This approach introduces the self-attention mechanism to the task of materials property predictions, and dynamically learns and updates individual element representations based on their chemical environment, enabled by featurization scheme that represents and preserves individual element identities while sharing information between elements. (A. Y.-T. Wang et al., 2021).

This network is unique in that it employs the transformer architecture, which was originally developed for natural language processing (NLP) tasks. Transformers use an attention mechanism that allows the model to "focus" on specific parts of the input sequence, in this case, elements within the chemical formula, to capture relevant relationships among them.

Transformers are especially effective for sequence-based data, and in CrabNet, this architecture interprets the elements and their proportions within a chemical formula as a sequential arrangement.

By learning to focus on different "tokens" (i.e., elements) in the formula, CrabNet can capture complex interactions, like how a language model interprets sentences.

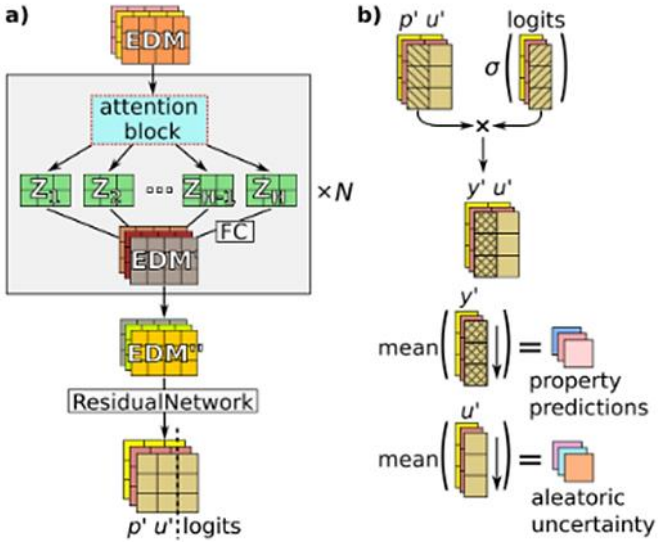


Figure 3 - Overall CrabNet architecture and prediction of material property and uncertainty (A. Y.-T. Wang et al., 2021).

As shown in Figure 3, **a** represents the Schematic of the CrabNet architecture including the input EDM, the self-attention layers (repeated N times), the updated and final element representations (EDM0 and EDM″), the residual network, and the final model output and in **b** the calculation steps for element contributions and prediction of the targets and uncertainties (A. Y.-T. Wang et al., 2021).

By analyzing only the elemental composition, CrabNet has demonstrated high accuracy in predicting a range of material properties, such as band gaps, formation energies, and mechanical strength.

This model is especially useful when structural information is unavailable, as it can still provide insights based solely on compositional data, however, when this type of data is available, GNNs offer a better, more well-rounded method for dealing with structures.

2.2.2 Graph Neural Networks (GNNs)

Graph Neural Networks (GNNs) have emerged as a powerful framework for handling data structured as graphs, where nodes represent entities and edges denote the relationships between these entities (Scarselli et al., 2009). This structure makes GNNs particularly effective for applications that require modelling complex interactions and dependencies, which are commonplace in multimodal data (Wu et al., 2021).

Graph neural networks capture dependencies in graphs through message passing between nodes. In recent years, variants such as Graph Convolutional Networks (GCNs), Graph Attention Networks (GATs), and Graph Recurrent Networks (GRNs) have demonstrated groundbreaking performances on many deep learning tasks (Kipf & Welling, 2017; Veličković et al., 2018).

Unlike composition-based models that rely solely on elemental formulas, GNNs are uniquely suited to handle **structured data** that captures the spatial arrangement of atoms, treating the material as a **graph of nodes and edges**.

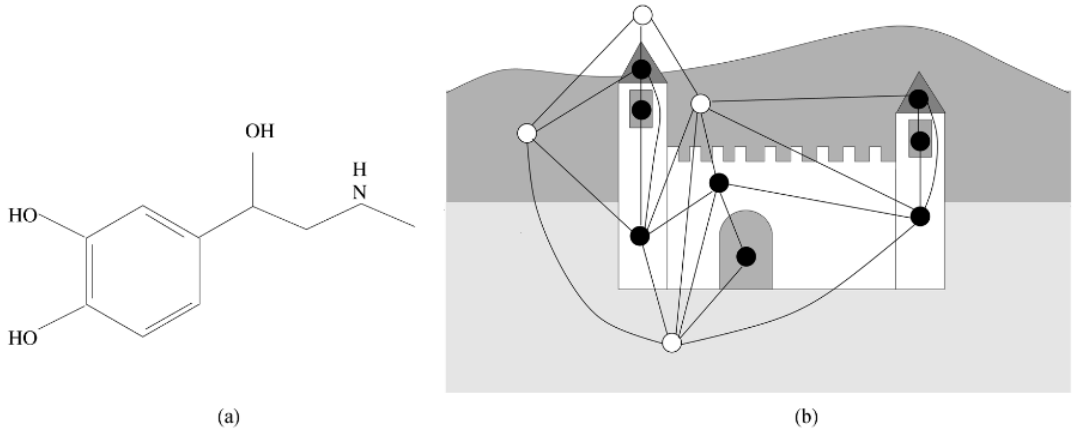


Figure 4 - Some applications where the information is represented by graphs: (a) a chemical compound (adrenaline), (b) an image (Scarselli et al., 2009).

For example, a chemical compound can be modelled by a graph, the nodes of which stand for atoms (or chemical groups) and the edges of which represent chemical bonds as shown on Figure 4 linking together some of the atoms.

Two prominent GNN-based models in materials science are the **Crystal Graph Convolutional Neural Network (CGCNN)** and **PotNet**.

2.2.2.1 CGCNN

CGCNN (Crystal Graph Convolutional Neural Network) is one of the first and most widely used GNN architectures developed specifically for materials science (Xie & Grossman, 2018a). It represents each material as a graph based on its **crystal structure**, where atoms serve as nodes and bonds serve as edges. CGCNN applies convolutional layers over this graph structure, allowing the model to learn from spatial and relational information among atoms in a crystal lattice.

Each node i is represented by a feature vector vi, encoding the property of the atom corresponding to node i. Similarly, each edge (i,j)k is represented by a feature vector u(i,j)k corresponding to the kth bond connecting atom i and atom j.

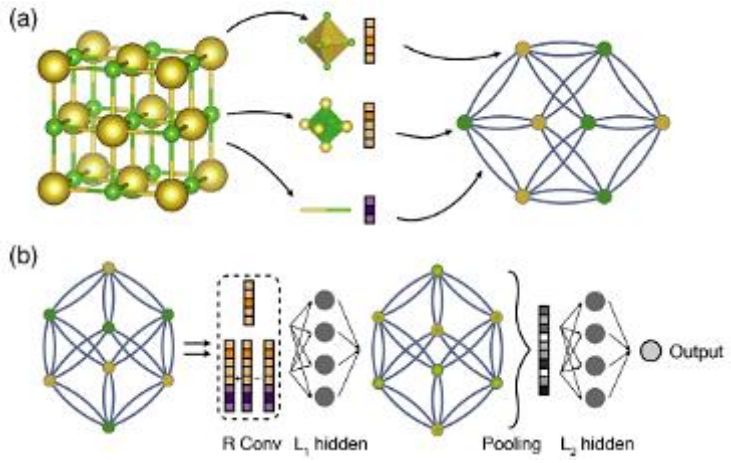


Figure 5 - Illustration of the crystal graph convolutional neural networks (Xie & Grossman, 2018a).

Figure 5 illustrates the crystal graph convolutional neural networks. In (**a)** the construction of the crystal graph where crystals are converted to graphs with nodes representing atoms in the unit cell and edges representing atom connections. Nodes and edges are characterized by vectors corresponding to the atoms and bonds in the crystal, respectively. In (**b)** the structure of the convolutional neural network on top of the crystal graph. R convolutional layers and L1 hidden layers are built on top of each node, resulting in a new graph with each node representing the local environment of each atom. After pooling, a vector representing the entire crystal is connected to L2 hidden layers, followed by the output layer to provide the prediction (Xie & Grossman, 2018a).

By treating crystal structures as graphs, CGCNN can directly process the spatial and connectivity patterns of atoms, capturing the geometry of the lattice and the relationships between atoms. This graph structure allows the model to inherently capture symmetry and periodicity, both of which are critical in crystalline materials.

2.2.2.2 PotNet

PotNet (Potential Network) is a GNN-based model that builds upon the idea of modeling atomic structures as graphs but introduces a unique approach by focusing on interatomic potentials rather than atomic bonds alone (Lin et al., 2023). Based on the physical modeling of crystal energy, PotNet explicitly uses interatomic potentials and complete interatomic potentials as input features. The complete interatomic potentials are incorporated into the message passing mechanism of graph neural networks and efficiently approximated by an efficient algorithm. This approach provides a more physics-driven representation of atomic interactions, making PotNet particularly valuable for capturing complex behaviors in materials.

Unlike typical GNNs that consider only direct, nearest-neighbor interactions, PotNet uses a complete set of interatomic potentials to represent atomic interactions, allowing it to capture long-range forces that affect material properties.

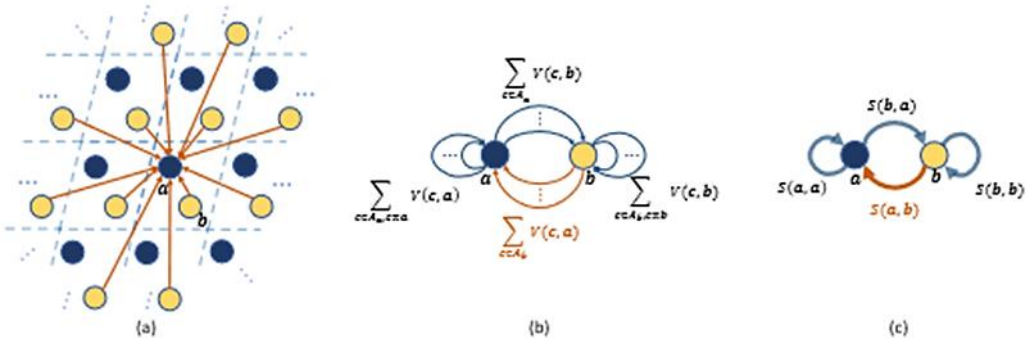


Figure 6 - Schematic illustrations of how complete interatomic interactions are captured in PotNet (Lin et al., 2023).

As shown in Figure 6, in **a** is represented an example crystal in which each unit cell contains two atoms a and b. In PotNet, the potentials between all pairs of atoms are captured , in (b) the complete set of potentials in (a) can be grouped into four categories, including a → b, b → a, a → a, and b → b and in (c) the proposal to compute an approximate summation for each category of potentials (Lin et al., 2023).

In PotNet, each node (atom) is connected to every other atom in the structure through potential-based edges, creating a fully connected graph. These connections are weighted according to calculated potentials, giving a more nuanced representation of atomic interactions. PotNet employs message passing algorithms where each atom's representation is updated based on both short- and long-range interactions, leading to embeddings that capture subtle variations in potential energy landscapes. This process provides a detailed spatial understanding of atomic forces and their influence on material properties.

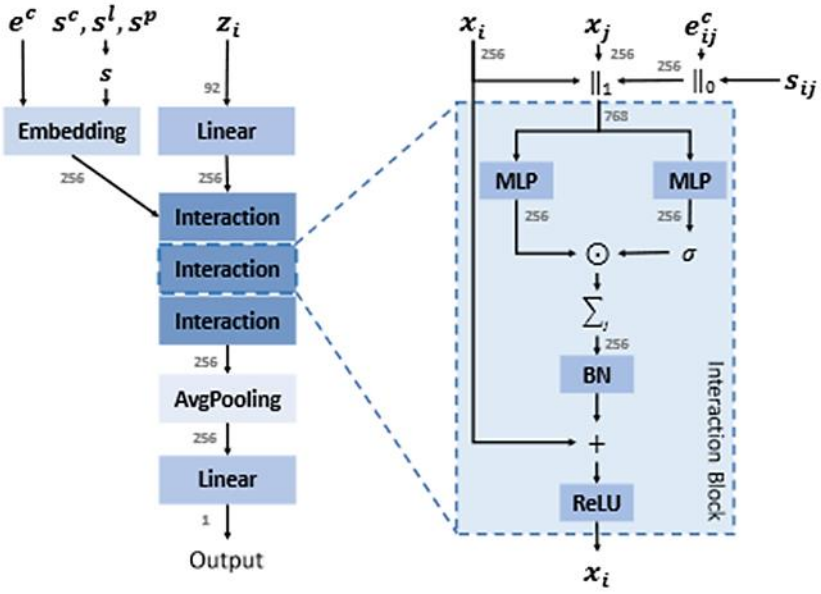


Figure 7 - The developed network architecture for PotNet (Lin et al., 2023).

The employed network architecture in PotNet is shown in Fig. 7 and it is designed following the commonly used settings, sharing a similar architecture to existing methods for 3D graphs (Gasteiger et al., 2021, 2022; Schütt et al., 2017; Xie & Grossman, 2018b).

This architecture contains an input block, an interaction block, and an output block. The inputs contain atomic features and potentials where zi is the 92-dimensional atomic feature for any atom i following CGCNN (Xie & Grossman, 2018a).

- **Input Block** includes a Linear layer and an Embedding layer. Each node's input features are generated as 256-dimensional vectors using the Linear layer. For edges, an Embedding layer maps Coulomb potentials and infinite potential summations to 256-dimensional embeddings.

- **Interaction Block** comprises several interaction layers that update each node's feature vector by incorporating features from neighbouring nodes and edge embeddings. For each neighbouring node, embeddings are concatenated along the edge dimension, then with node features along the feature dimension like in the structure in CGCNN (Xie & Grossman, 2018b).
- **Readout Block** includes an AvgPooling layer followed by a Linear layer. AvgPooling aggregates features from all nodes, and the Linear layer maps the 256-dimensional hidden features to a final scalar output.

PotNet is particularly effective for properties that depend on long-range atomic interactions, such as thermal conductivity, electrical behavior, and mechanical resilience.

2.2.3 Image Based Models

Image-based models in materials science leverage visual data to analyze and predict material properties based on the spatial and structural characteristics captured in images. Unlike text-based or tabular data, which represent materials through compositional or numerical formats, image data provides a visual representation of the material's atomic or microscopic structure. These images often show the geometric arrangement of atoms, grains, or crystal lattices, capturing essential information about spatial orientation, symmetry, molecular patterns and defects that can significantly impact material properties.

In materials science, image data is typically generated through electron microscopy, X-ray diffraction patterns, or visualizations of crystal structures created with software like Crystal Toolkit. This data modality is highly useful for capturing characteristics that are challenging to represent in other formats, such as grain boundaries, surface morphologies, and structural defects, which are often vital for understanding material strength, conductivity, and stability.

Regarding Image Based Models, some of the primary architectures for image analysis are Convolutional Neural Networks (CNNs) and Transformer-based models.

1. **CNNs**: Designed for local feature extraction, use convolutional layers to capture hierarchical patterns, moving from simple textures in early layers to complex structures in deeper layers (Lecun et al., 1998). This makes CNNs ideal for detecting fine-grained details, such as atomic bonds and lattice arrangements,

which are crucial for analyzing material properties at a microscopic level. ResNet (He et al., 2015) and EfficientNet (Tan & Le, 2020) are popular CNN architectures in materials science for capturing such structural nuances.

2. **Transformers**: Initially developed for natural language processing, transformers have been adapted for vision tasks and use self-attention mechanisms to capture global dependencies across the entire image. By processing images as sequences of patches, transformers excel at understanding broad contextual relationships and long-range patterns that CNNs may miss (Vaswani et al., 2023). Vision Transformer (ViT) (Dosovitskiy et al., 2021) and Swin Transformer (Liu et al., 2021) are commonly used to capture both local and global patterns, which is valuable for comprehensive material analysis.

2.2.3.1 TIMM

TIMM (Torchvision Image Models) is a comprehensive library for PyTorch that provides access to a wide range of pre-trained image models and utilities for training and fine-tuning models on custom image data. TIMM is particularly useful in the context of materials science, as it supports diverse model architectures and pre-trained weights, making it highly adaptable for applications involving high-resolution images of material structures.

This library includes over 600 pre-trained models, such as ResNet (He et al., 2015), EfficientNet (Tan & Le, 2020), ViT (Vision Transformer) (Dosovitskiy et al., 2021), and Swin Transformer (Liu et al., 2021). These models are fine-tuned on large image datasets like ImageNet (Deng et al., 2009), making them effective for transfer learning in specialized fields.

This framework allows easy customization, including architecture changes, layer freezing, and model fine-tuning, which are valuable for tailoring the models to materials science images. Researchers can leverage the transfer learning capabilities of TIMM to adapt these models to high-resolution 3D images representing crystal structures.

In materials science, pre-trained models from TIMM can be fine-tuned on domain-specific images. For example, ResNet or EfficientNet architectures pre-trained on ImageNet can be adapted to classify and analyze crystal structure images, capturing spatial relationships crucial for property prediction.

Features extracted from these models can be integrated with other modalities (e.g., composition and structure) in a multimodal framework, enabling more comprehensive property predictions. For instance, a Vision Transformer (ViT) from TIMM could be used to generate embeddings that reflect spatial arrangements in crystal images, complementing compositional and structural embeddings.

CHAPTER III – MULTIMODAL LEARNING MODELS AND FRAMEWORKS

3 Multimodal Learning Models and Frameworks

Multimodal machine learning aims to build models that can process and relate information from multiple modalities (Baltrušaitis et al., 2017). Given the heterogeneity of the data, the research field of Multimodal Machine Learning brings some unique challenges for computational researchers.

In materials science, multimodal learning can be transformative, as it allows models to combine data from **composition (text)**, **structure (tabular)**, and **visual representations (images)** to capture a more comprehensive understanding of materials. By leveraging the strengths of each modality, multimodal models can provide deeper insights into the relationships between material composition, atomic structure, and physical properties.

The integration of these diverse data types, however, requires specialized techniques known as **fusion techniques**. Fusion techniques define how and at what stage the data from different modalities is combined within the model, ensuring that the information from each modality contributes meaningfully to the final prediction. Effective fusion strategies can greatly enhance the model's ability to generalize, handle missing data, and provide accurate predictions, even for complex materials with intricate property relationships.

3.1 Fusion Techniques

Multimodal fusion is one of the original topics in multimodal machine learning, with previous surveys emphasizing early, late and hybrid fusion approaches (D'mello & Kory, 2015; Zhihong Zeng et al., 2009)

In technical terms, multimodal fusion is the concept of integrating information from multiple modalities with the goal of predicting an outcome measure: a class (e.g., happy vs. sad) through classification, or a continuous value (e.g., positivity of sentiment) through regression.

Such approaches can be split into early (i.e., feature-based), late (i.e., decision-based) and hybrid fusion (Atrey et al., 2010). Early fusion integrates features immediately after they are extracted (often by simply concatenating their representations). Late fusion on the other hand performs integration after each of the modalities has decided (e.g., classification or regression). Finally, hybrid fusion combines outputs from early fusion and individual unimodal predictors.

3.1.1 Early Fusion (Feature-Level Fusion)

Early fusion, also known as feature-level fusion, could be seen as an initial attempt by multimodal researchers to perform multimodal representation learning — as it can learn to exploit the correlation and interactions between low level features of each modality. Furthermore, it only requires the training of a single model, making the training pipeline easier compared to late and hybrid fusion.

This method combines different modalities at the initial stage of the model. In this approach, data from each modality is transformed into a common representation, often through embeddings or feature vectors, and then concatenated or integrated as a single input into the model. This fused representation is then processed through the network as one cohesive input, allowing the model to learn interactions between modalities from the start.

In materials science, early fusion might involve transforming composition data (text), structural embeddings (tabular), and image data into feature vectors and concatenating them into a unified input. By training the model on this combined input, it learns to recognize relationships across modalities, such as how compositional elements interact with structural features or how atomic arrangements influence visual patterns.

Early fusion is particularly effective for models that need to capture correlations between modalities right from the start, making it ideal for tasks where interactions between composition, structure, and images contribute to the material's properties.

Early fusion allows the model to capture cross-modal interactions early in the learning process, leading to potentially richer feature representations. This approach also allows each modality to influence the other during training, which can improve prediction accuracy for complex tasks.

3.1.2 Late Fusion (Decision-Level Fusion)

In contrast, late fusion , or decision-level fusion, uses unimodal decision values and fuses them using a fusion mechanism such as averaging, weighted summation, or using a meta-classifier to merge the outputs into a final prediction or a learned model (Baltrušaitis et al., 2017). It allows for the use of different models for each modality as different predictors can model each individual modality better, allowing for more flexibility. Moreover, it makes it easier to make predictions when one or more of the modalities are

missing and even allows for training when no parallel data is available. However, late fusion ignores the low-level interaction between the modalities.

In materials science, late fusion could involve training one model on compositional data, another on structural data, and a third on image data. Each model produces an output or prediction independently, and these outputs are then combined to form a final prediction. This approach is beneficial when the relationships between modalities are independent or weakly correlated, allowing each modality to contribute separately to the final decision.

Late fusion allows each modality to be fully processed before combining, making it more computationally efficient for high-dimensional inputs. It also reduces the risk of overfitting, as each model specializes in one modality, and the final fusion leverages each modality's strengths. However, since interactions between modalities are not learned until the final layer, late fusion may miss complex interdependencies across modalities. This approach might be less effective for tasks where deep interactions between data types are essential for accurate predictions.

3.1.3 Hybrid Fusion

Hybrid fusion attempts to exploit the advantages of both of the above-described methods in a common framework. It has been used successfully for multimodal speaker identification (Wu et al., 2005) and multimedia event detection (MED) (Lan et al., 2012).

This method combines elements of both early and late fusion by integrating modalities at multiple points within the model architecture. In this approach, data from each modality is first processed individually, and intermediate representations are fused at various stages of the model to create shared representations. Hybrid fusion allows for both independent feature extraction and joint learning across modalities, making it a flexible option that captures interactions at different levels of the model.

In materials science, hybrid fusion can be particularly advantageous for multimodal models that require both modality-specific processing and cross-modal interactions. For example, composition and structural data might be processed separately in the initial layers, and their intermediate representations are fused with image data at deeper layers to learn joint representations.

3.2 CLIP

One of the most used techniques is CLIP (Contrastive Language-Image Pre-Training), a groundbreaking model developed by OpenAI that leverages contrastive learning to align textual and visual representations. By training on a large dataset of images and their corresponding text descriptions, CLIP learns a shared embedding space where images and their associated textual descriptions are close together. This enables powerful capabilities in cross-modal retrieval, zero-shot learning, and various other applications (Radford et al., 2021).

1. **Contrastive Learning Objective**

CLIP uses a contrastive learning objective to align image and text pairs in a shared latent space. The model is trained to maximize the similarity between the embeddings of paired images and texts (positive pairs) while minimizing the similarity between embeddings of mismatched pairs (negative pairs). This objective is typically implemented using a contrastive loss function, such as the InfoNCE loss, which encourages the model to distinguish between correct and incorrect pairs (T. Chen et al., 2020).

2. **Dual-Encoder Architecture**

CLIP employs a dual-encoder architecture, consisting of:

- **Image Encoder**: A convolutional neural network (CNN) or a vision transformer (ViT) that processes images and maps them into the shared latent space.
- **Text Encoder**: A transformer-based model that processes text descriptions and maps them into the same latent space as the images.

The dual-encoder setup allows CLIP to independently process images and texts while ensuring that their embeddings are comparable in the shared space.

3. **Large-Scale Training Data**

CLIP is trained on a large-scale dataset of image-text pairs collected from the internet. The diversity and scale of this dataset are crucial for the model's ability to generalize across a wide range of concepts and perform well in zero-shot learning scenarios. This extensive dataset enables CLIP to understand a vast array of visual and textual contexts, making it highly versatile (Radford et al., 2021).

This method offers several applications in the following areas:

1. **Zero-Shot Learning:** CLIP excels in zero-shot learning by using its shared embedding space to connect unseen classes to seen ones through textual descriptions. This allows it to retrieve relevant images for new class descriptions without explicit training on those classes.
2. **Cross-Modal Retrieval:** CLIP performs well in cross-modal retrieval, retrieving images based on text queries and finding text descriptions based on images. This is useful for search engines, content recommendation systems, and other applications needing effective cross-modal interaction (Zheng et al., 2020).
3. **Content Moderation and Filtering:** CLIP aids in content moderation by matching inappropriate text with corresponding images, helping platforms manage large volumes of user-generated content and ensure compliance with guidelines.
4. **Enhanced Image Captioning:** By aligning images and text in a shared space, CLIP enhances image captioning quality, producing more accurate and contextually relevant captions that closely match the visual content (K. Xu et al., 2015).

CLIP represents a significant advancement in multimodal representation learning, offering a powerful framework for aligning textual and visual data in a shared embedding space. Ongoing research continues to explore ways to improve CLIP's robustness, efficiency, and interpretability, ensuring its applicability across a wide range of real-world scenarios.

Contrastive representation learning, as represented in Figure 3, is a foundational technique in multimodal embedding, enabling the integration and alignment of diverse data sources in a shared latent space. By leveraging mechanisms such as SimCLR (T. Chen et al., 2020), CMC (Tian et al., 2020), and CLIP (Li et al., 2024), models can effectively learn representations that enhance performance across various multimodal tasks.

3.3 MultiMat

The MultiMat framework, as detailed in Viggo Moro's paper "Multimodal Learning for Materials," (Moro et al., 2024) is a comprehensive system designed to integrate and process multimodal data specifically for materials science applications.

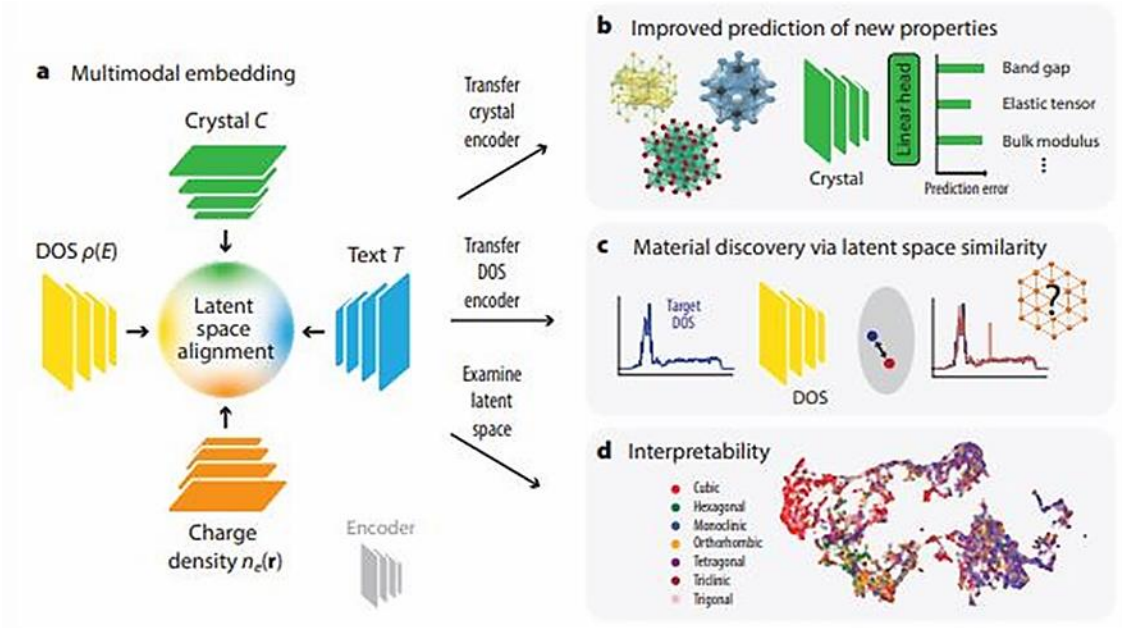


Figure 8 - MultiMat Framework Diagram.

MultiMat, as represented in Figure 8, is a framework for training a foundation model for crystalline materials that allows for the incorporation of several modalities. The basis for MultiMat is a multimodal pre-training method that connects high-dimensional material properties (i.e., modalities) in a shared latent space to produce highly effective material representations that can then be transferred to various downstream tasks.

The MultiMat framework trains a foundation model for materials by aligning the latent spaces of encoders of different information-rich modalities, such as the crystal structure, DOS, charge density, and textual description, as shown in Figure 8a.

This alignment process produces shared latent spaces and effective material representations which can then be leveraged for a series of downstream tasks (Figure 8b–d). For instance, the crystal encoder can be transferred and fine-tuned for material property prediction, enabling improved predictive performance compared to traditional training techniques. Since MultiMat aligns the latent spaces of different modalities, it can also be used in a novel material discovery strategy by screening large crystal-structure databases with comparisons between target properties and candidate crystals based on the latent-space similarity. Finally, we demonstrate the interpretability enabled by the MultiMat approach, by exploring the latent space from MultiMat using a dimensionality reduction approach.

1. **Graph-Based Integration:**
    - MultiMat uses graph-based structures to represent multimodal data, where nodes correspond to material properties and edges represent the relationships between these properties. This structure allows for the natural integration of diverse data types, capturing complex dependencies in materials science (Moro et al., 2024).
2. **Node and Edge Features:**
    - Each node and edge in the MultiMat framework is associated with specific features that encapsulate the properties and relationships of materials. These features are crucial for LMMs to learn effective representations that encode multimodal information (Moro et al., 2024).
3. **Advanced GNN Architectures:**
    - MultiMat employs advanced GNN architectures, such as Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs), to process and learn from the graph-structured data. These architectures enable the model to propagate information through the graph, capturing intricate dependencies and interactions between different material properties (Moro et al., 2024).

3.4 AutoGluon-Multimodal (AutoMM)

AutoGluon Multimodal or AutoMM, is a Python based open-source AutoML framework tailored for multimodal learning with foundation models developed by Amazon Web Services (AWS) that enables efficient multimodal learning by integrating diverse data types—such as text, tabular, image, and even time series—into a single predictive model (Tang et al., 2024). Designed to streamline the model-building process, AutoGluon automatically handles data preprocessing, model selection, hyperparameter tuning, and multimodal fusion, making it accessible for both experts and non-experts to develop powerful machine learning models.

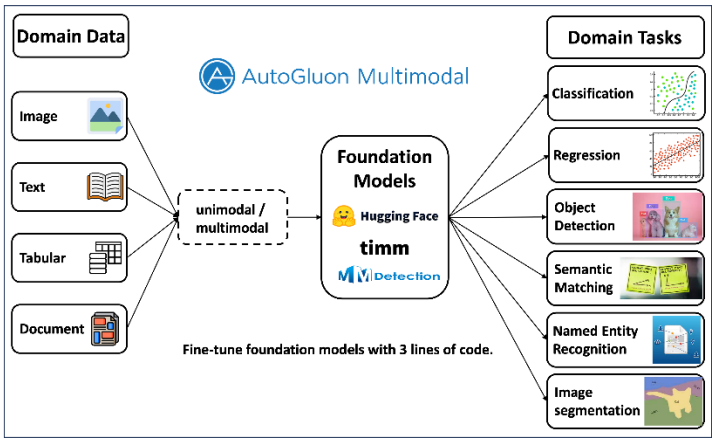


Figure 9 - AutoMM architecture.

As shown in Figure 9, supporting both unimodal and multimodal data (left), AutoMM enables seamless fine-tuning of foundation models (middle) for basic classification/regression as well as advanced tasks (right).

In the context of materials science, AutoGluon's multimodal capabilities are especially valuable, as researchers can integrate various forms of data—such as composition (text), structural properties (tabular data), and crystal structure images—without needing to manually engineer complex fusion mechanisms. This enables researchers to capture a more holistic view of materials, ultimately improving predictive accuracy for complex properties like thermal stability, electrical conductivity, and mechanical strength.

AutoGluon automates the process of model selection and hyperparameter tuning by training multiple models in parallel and selecting the best-performing model through an ensemble approach. This process eliminates the need for manual tuning, allowing researchers to focus on input data quality and interpretation.

The automated tuning feature is particularly useful in materials science, where selecting and optimizing multimodal architectures can be time-consuming. By testing a range of model architectures and parameters, AutoGluon finds the optimal configuration for a given multimodal dataset, whether it involves compositional, structural, or visual data.

AutoGluon's MultimodalPredictor is specifically designed to handle and fuse data from multiple modalities. Each type of data is encoded through modality-specific encoders (such as CNNs for images, transformers for text, and dense layers for tabular data) before being combined into a unified representation.

This framework is designed to handle large datasets, making it suitable for materials science applications that involve extensive databases like the Materials Project or Alexandria. Its ability to parallelize training across multiple data types enables efficient processing and model optimization, even when handling large multimodal datasets with high-dimensional features.

3.5 Challenges in Multimodal Learning

While multimodal learning brings significant benefits in combining diverse data types for improved predictive power, it also presents unique challenges that can limit its application, particularly in complex fields like materials science. These challenges include high computational complexity, difficulties in data integration across modalities, and issues with interpretability. Addressing these challenges is crucial for developing robust multimodal models that are scalable, accurate, and understandable.

3.5.1 Computational Complexity

Multimodal models are inherently complex due to the need to process and fuse multiple types of data, such as text, images, and structured data. Each modality often requires a specialized encoder (e.g., CNNs for images, transformers for text, dense layers for tabular data), which increases the model's size and computational requirements. This complexity can lead to high memory usage and longer training times, especially when dealing with large datasets.

Training multimodal models on large datasets is challenging because of the increased volume of data and the diverse preprocessing requirements for each modality. Scaling up these models requires distributed computing resources and advanced hardware (e.g., GPUs or TPUs), which may be inaccessible to many researchers, especially in resource-constrained environments.

Optimizing multimodal models to balance accuracy with computational efficiency is challenging. These models require tuning for each modality's encoder, the fusion layers, and the overall architecture, which increases the complexity of hyperparameter optimization. Techniques such as knowledge distillation, where a smaller model is trained to mimic a larger one, and model pruning can help reduce model size and improve efficiency, but these approaches add extra steps to the training process.

3.5.2 Data Integration

One of the primary challenges in multimodal learning is aligning diverse data types that have inherently different structures and scales. For example, text data (composition) is sequential, tabular data (structure) is numerical, and image data (crystal structures) is spatial. Developing an effective fusion strategy that combines these varied data types into a coherent representation without losing important information is difficult..

In real-world datasets, it is common to encounter missing data in one or more modalities. For instance, some materials may lack high-quality images or detailed structural data. Multimodal models need to either handle these missing modalities gracefully or use data imputation methods, but such solutions add to the model's complexity and may lead to biased predictions if not handled properly.

Solutions like zero-imputation (filling in missing data with zeros) or cross-modal inference (using other modalities to predict missing data) can help but require careful implementation to avoid negatively impacting model performance.

Ensuring consistency across modalities is essential, especially when one data type is more informative than others. In cases where certain modalities are more relevant for predicting a property, the model should be able to weigh the contribution of each modality dynamically. This dependency management requires adaptive fusion techniques, such as attention mechanisms, to prioritize important modalities without ignoring the potential contributions of other data types.

3.5.3 Interpretability

Multimodal models are often "black boxes" due to their complexity, making it challenging to interpret how different modalities contribute to the final prediction. For example, a model predicting material stability may integrate information from composition, structure, and images, but understanding the exact role each modality plays in the prediction is difficult.

Interpretability is crucial in materials science, where researchers need to understand which factors drive material properties to inform material design. However, with multimodal models, tracing the decision-making process across modalities is challenging, especially when using deep learning techniques like neural networks, which inherently lack transparency.

In multimodal models, it's essential to understand not only the influence of individual modalities but also the interactions between them. Cross-modal interpretability is complex, as the model may rely on subtle relationships between modalities that are difficult to disentangle.

Techniques like attention maps and SHAP values (SHapley Additive exPlanations) are used to interpret the contributions of individual modalities, but adapting these for multimodal settings is not straightforward and often requires further research to make these explanations reliable and accessible.

In materials science, domain-specific explanations are essential for validating model predictions. For instance, if a multimodal model predicts high thermal conductivity, researchers need to understand whether this prediction is driven by certain atomic arrangements (structure), specific compositional patterns, or visual features in crystal images. Developing explanations that align with domain knowledge helps build trust in the model but is challenging in multimodal frameworks.

For example, in materials discovery, automated interpretability could reveal which features (such as specific atomic arrangements or structural characteristics) contribute to a material's predicted stability or conductivity, providing scientists with actionable insights.

4 Empirical Study

At the outset of this work, several key data sources were considered to build a dataset suitable for multimodal machine learning in materials science. Options included established repositories such as Materials Project (Jain et al., 2013) and a next-gen dataset Alexandria (Schmidt et al., 2024). Each of these databases provides comprehensive data on materials, including chemical composition, structural properties, and various material characteristics. However, a common limitation was that all these sources stored data primarily in JSON format, resulting in tabular representations that were rich in numerical, structural and categorical information but lacked diversity in data modalities.

The challenge was that these JSON-based formats inherently limited the data to tabular modalities excluding other modalities, such as visual representations. For multimodal learning, where it was intended to integrate diverse data types (e.g., text, images, and structured numerical data) it was essential to extend beyond purely tabular data. Therefore, a critical focus of the work became transforming the JSON-based tabular data from the Alexandria dataset into a fully multimodal dataset that incorporated text-based, tabular, and visual modalities.

Considering the scale of the Alexandria dataset, which contains millions of materials, we chose a random sample of 1,000 materials from the complete 3D JSON database due to limitations in computational power and processing capacity required to handle the full dataset.

4.1 Data Understanding

After the dataset selection and sampling, the next step in the development process was to understand and interpret the structure and content of the Alexandria dataset. This dataset was provided in JSON format, with each entry corresponding to a specific material and containing various fields that describe the material’s composition, structural attributes, electronic properties, and thermodynamic stability.

```
{
  "@module": "pymatgen.entries.computed_entries",
  "@class": "ComputedStructureEntry",
  "energy": -17.83768098,
  "composition": {"Ba": 1.0, "Sr": 2.0, "Pd": 2.0},
  "entry_id": null,
  "correction": 9.8e-07,
  "parameters": {},
  "data": {
    "mat_id": "agm003399846",
    "prototype_id": "AB2C2_11_spg12",
    "formula": "Ba(SrPd)2",
    "elements": ["Ba", "Sr", "Pd"],
    "spg": 12,
    "nsites": 5,
    "stress": [[1.4, 0.0, 0.0], [0.0, 1.39, -0.05], [0.0, -0.05, 1.43]],
    "energy_total": -17.83768,
    "total_mag": 1.78e-05,
    "band_gap_ind": 0.0,
    "band_gap_dir": 0.0676,
    "dos_ef": 2.6769836,
    "energy_corrected": -17.83768,
    "e_above_hull": 0.088,
    "e_form": -0.433,
    "decomposition": "Ba2SrPd SrPd"
  },
  "structure": {
    "@module": "pymatgen.core.structure",
    "@class": "Structure",
    "lattice": {
      "matrix": [[4.59, 0.0, 0.0], [0.0, 6.24, 0.06], [-2.29, -1.73, 6.06]],
      "a": 4.59,
      "b": 6.24,
      "c": 6.71,
      "volume": 174.39
    },
    "sites": [
      {"species": [{"element": "Ba", "occu": 1}], "abc": [0.0, 0.0, 0.0]},
      {"species": [{"element": "Sr", "occu": 1}], "abc": [0.34, 0.38, 0.68]},
      {"species": [{"element": "Pd", "occu": 1}], "abc": [0.71, 0.12, 0.42]}
    ]
  }
}
```


Code Snippet 1 - JSON object for the material $Ba(SrPd)_2$.

In the Code Snippet 1, is presented the structure of the JSON object for the material $Ba(SrPd)_2$, and all the information given in this structure.

Focusing on this JSON object, we denote that the entries begin with metadata fields, such as *@module* and *@class*, which indicate that the data structure originates from Python's **pymatgen** package, a widely used tool in materials science for processing and analyzing crystal structures. While these fields won't be directly used in the modeling, they provide context for the data's format and origins, ensuring compatibility with pymatgen-based workflows if needed.

Among the core properties of each material, the **energy** field is especially important. Representing the computed total energy (in electron volts, ***eV***), this field gives insight into the material's stability, as lower energy values typically correspond to more stable structures. Another key attribute is the **composition**, which lists the elements present in the material and their respective quantities, such as {"Ba": 1.0, "Sr": 2.0, "Pd": 2.0} for the example in Code Snippet 1.

Then we have the data structure, a nested object that holds additional physical and electronic properties specific to each material such as:

- **mat_id**: Unique identifier for the material, used for indexing and referencing entries.
- **prototype_id**: Describes the prototype structure of the material, indicating its crystal type and symmetry.
- **formula**: A text representation of the chemical formula, e.g., $Ba(SrPd)_2$, which is used in the textual modality.
- **elements**: A list of chemical elements in the material.
- **spg**: The space group number, which provides information about the symmetry of the material's crystal structure.
- **nsites**: The number of atomic sites in the material's unit cell, representing its complexity.
- **stress**: A 3x3 matrix representing the stress tensor (in GPa), providing details on internal forces within the material's structure.
- **band_gap_ind** and **band_gap_dir**: The indirect and direct band gaps of the material, respectively, which are essential properties for understanding its electronic behavior.
- **total_mag**: The total magnetic moment, representing the magnetic properties of the material.
- **dos_ef**: Density of states at the Fermi level, offering insights into the material's conductive properties.
- **e_above_hull**: Energy above the convex hull, indicating stability relative to potential phase separations.
- **decomposition**: Possible phase-separated components of the material, providing information on stability under different conditions.

This object is followed by the Structure object, another nested JSON but with a special format, in line with the pymatgen's class: Structure describes the atomic structure of the material in detail. This object is composed of:

- **lattice**: A dictionary specifying the lattice parameters and volume of the unit cell. It contains:
    - **matrix**: A 3x3 matrix that represents the lattice vectors.
    - **a, b, c**: Lengths of the unit cell edges.
    - **volume**: Total volume of the unit cell.
- **sites**: A list detailing each atomic site in the material, where each entry represents an atom. For each atom:
    - **species**: The atomic species (element) and occupancy.
    - **abc**: Fractional coordinates of the atom within the unit cell.

In summary, these JSON objects are a comprehensive digital representation of materials, detailing their chemical composition, physical properties, and atomic structure.

By thoroughly understanding and parsing these features, we can extract relevant information to build a rich, multimodal dataset that reflects both the chemical composition and spatial arrangement of each material.

4.2 Dataset Construction and Multimodal Generation

Considering the information provided by the dataset, we decided to generate three modalities: a text modality from the composition data and both image and tabular modalities from the structure object.

To begin, we created a dataset containing a sample of 1,000 randomly selected materials from the Alexandria 3D dataset. Using a Python script, we loaded a JSON file with 100,000 materials from the Alexandria database. We then set a random seed to ensure reproducibility and sampled 1,000 random materials from this file. After sampling, we saved the selected subset as a pickle file—a binary format that allows us to efficiently store and later load the data, ensuring consistency for the training and testing phases. Pickling (serialization) and unpickling (deserialization) is often more efficient for both storage and loading speeds compared to re-processing raw files or loading from CSV or JSON formats. This approach provided a stable foundation for generating and aligning other data modalities.

4.2.1 Image Generation

To create the image modality for this work, we generated 2D visualizations of each material's 3D crystal structure. These images provide a spatial representation of atomic arrangements within the materials, capturing essential structural details that can complement the text and tabular data modalities.

To generate the images, we used the same framework used to render the 3D objects in the Materials Projects Web Interface, as shown in Figure 10.

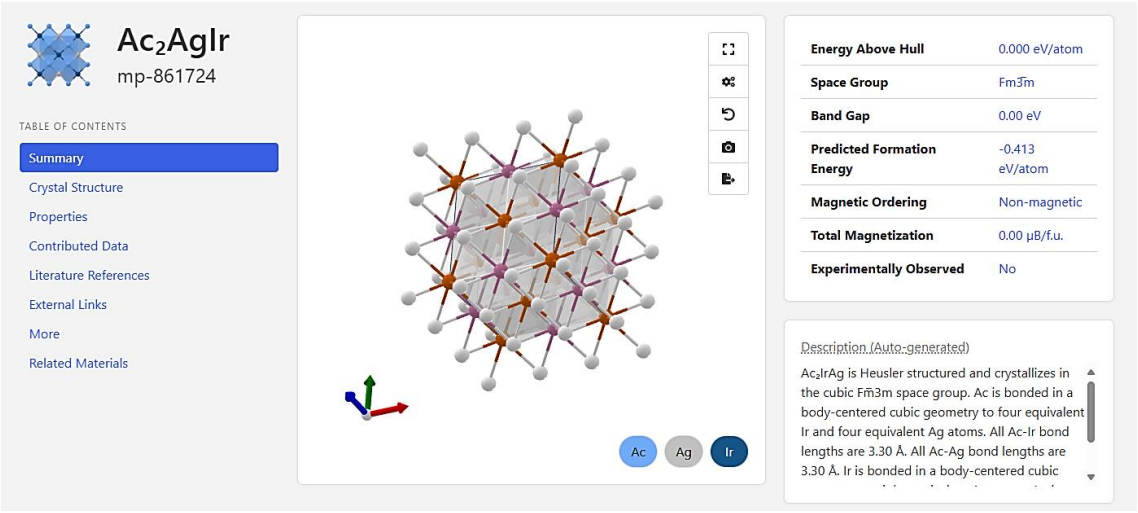


Figure 10 - Materials Project Ac2AgIr 3D rendered object.

The process of generating the images was indeed intricate, so we needed to be creative and come up with a solution. The solution required the development of a web application to render and capture the 3D visualizations of each material. This approach involved using **Dash** for the web interface and **Crystal Toolkit** integrated with **pymatgen** to display the 3D structures interactively. Each material was iteratively rendered as a 3D model on the web interface, which was then automatically screenshotted and saved as a 2D PNG file in a specified directory.

In more detail, we started by loading the pickle file, with the dataset, and the initializing the Dash App. Then we created a Crystal Toolkit Component that renders 3D structures in the web app, and sets the settings which include disabling a compass, showing bonds outside the unit cell, and ensuring the structure fits within the 2D view. After that we defined the layout for the structure and set intervals of 5 seconds for automatic updates. After 5 seconds, the app will update the displayed structure to the next one in the dataset.

Then, for each interval, a callback function triggers when data for the structure changes and requests a PNG image from the rendered scene after a 1-second delay, setting up the structure for saving. A screenshot of the rendered object is then saved to a specified directory, with the mat_id, to allow the alignment with the dataset after. The image generation process, took approximately 70 minutes to produce 1,000 images, resulting in some of the materials shown in Figure 11.

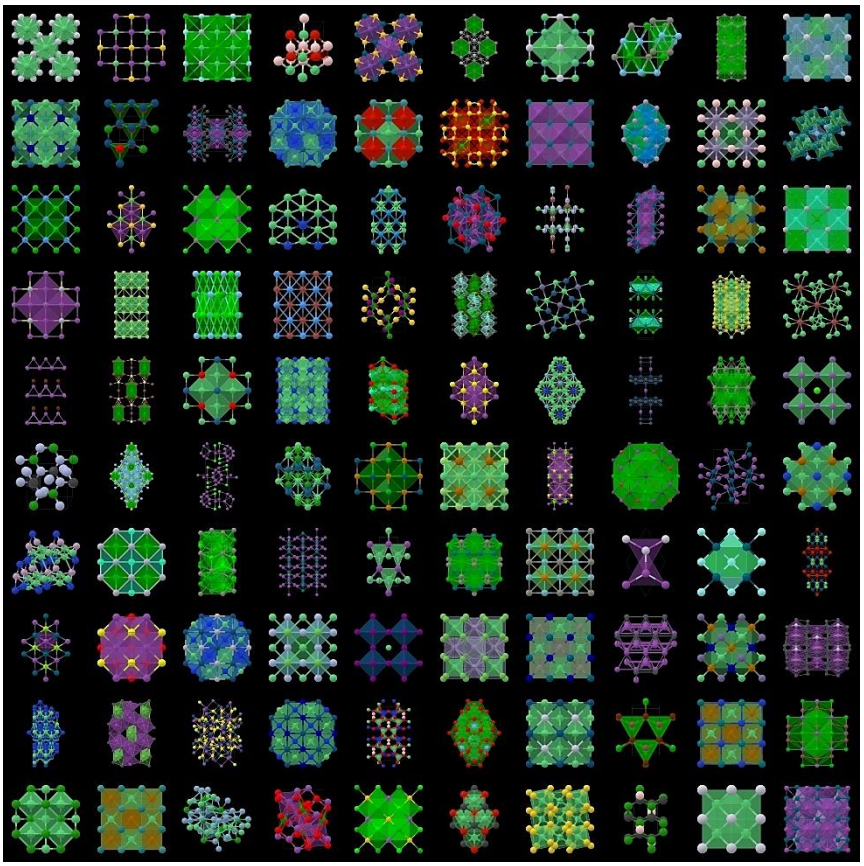


Figure 11 - 100 of the materials 2D representations of the 3D renderization generated.

Using this approach enabled a high degree of automation, which was critical for processing the dataset efficiently and ensuring uniformity across all generated images. The final output was a set of 2D images, which referenced the mat_id in the name, each corresponding to a material's 3D atomic structure.

4.2.2 Tabular Generation

For the tabular modality, we initially had a structured format that could be relatively easily converted into tabular data. However, when incorporating the structure object—one of the most critical sources of information—we encountered a significant challenge. This

structure object varied in terms of the number of columns and entries for each material, as different materials had different numbers of atomic sites and coordination environments.

This variability made it impractical, and effectively impossible, to generate a consistent tabular format, as the number of columns would fluctuate widely from one material to another. Without a fixed column structure, standard tabular models would struggle to process the data effectively, leading us to explore alternative methods for representing this structural information in a more consistent way across all materials.

The PotNet architecture offered a promising solution to address the variability issue within the structure object. PotNet is designed to generate a fixed-size embedding of a material's structure, capturing essential information about atomic positions and interatomic relationships. While PotNet typically uses these embeddings directly for property predictions, we adapted this approach by stopping the process before the prediction phase. This allowed us to extract a fixed-length vector representing the structure.

This fixed-length vector could then be seamlessly integrated into our tabular dataset, providing a consistent and comprehensive representation of the structure object. By using these structural embeddings as a standardized feature, we were able to incorporate the structure information into the tabular modality effectively, overcoming the limitations posed by the varying numbers of atomic sites and coordination environments in different materials.

To generate the embeddings using PotNet, additional steps were required. PotNet relies on interatomic potentials as an input feature to capture the spatial and bonding characteristics of a material. These potentials describe the interactions between atoms based on their relative positions, which provide essential insights into the structure's physical and chemical properties. To obtain these interatomic potentials, we leveraged the atomic structure data within each material.

To run PotNet and the associated libraries, we began by configuring our environment. This included installing necessary libraries for materials science and graph processing in PyTorch, such as pymatgen, jarvis-tools, torch-geometric, and dgl. These packages enabled us to work with crystal structures, build graph representations of atomic structures, and use advanced graph neural network functionalities.

Each of these libraries serves a specific purpose:

- pymatgen: Used for handling crystal structures and converting structural data.
- jarvis-tools: Provides additional tools for working with CIF files and atomic data.
- torch-geometric, dgl, torch-sparse, torch-scatter: Essential for creating and manipulating graph data structures, which PotNet relies on for learning structural representations.

We needed a CIF (Crystallographic Information File) for each material. CIF files are a standard format for storing crystal structure data, capturing lattice parameters, atomic coordinates, and symmetry information in a compact format. Using the Structure object in pymatgen, we converted each material's structure dictionary into a CIF file, as shown in Code Snippet 2.

```
1. from pymatgen.core.structure import Structure
2. from pymatgen.io.cif import CifWriter
3.
4. # Convert each structure into a CIF file
5. for mat in data_1000:
6.     structure = Structure.from_dict(mat['structure'])
7.     w = CifWriter(structure)
8.     w.write_file(CIFs_dir + mat['data']['mat_id'] + '.cif')
```

Code Snippet 2 - CIF generation function.

These CIF files were saved in a specified directory, with filenames corresponding to each material's unique ID, making them easy to reference and retrieve.

To work with the structural data, we used jarvis-tools to convert each CIF file into the JARVIS Atoms format. This format is compatible with various calculations and allows easy conversion of atomic information into a dictionary. We loaded each CIF file, converted it to the Atoms format, and stored it in a Pandas DataFrame for further processing. By storing the atomic data in a DataFrame, we ensured that each entry could be easily accessed for subsequent calculations, and this format allowed efficient data manipulation.

We used atomic information from each material to calculate node attributes (representing each atom's properties) and edge indices (representing bonds or potential interactions between atoms). This step was critical for transforming each material's structure into a graph format compatible with PotNet.

```
from jarvis.core.specie import chem_data, get_node_attributes
import torch

# Calculate node features for each atom
sps_features = []
for ii, s in enumerate(structure.elements):
        feat = list(get_node_attributes(s, atom_features="atomic_number"))
        sps_features.append(feat)

sps_features = np.array(sps_features)
node_features = torch.tensor(sps_features).type(
        torch.get_default_dtype()
)

u = torch.arange(0, node_features.size(0), 1).unsqueeze(1).repeat((1,
node_features.size(0))).flatten().long()
v = torch.arange(0, node_features.size(0), 1).unsqueeze(0).repeat((node_features.size(0),
1)).flatten().long()

edge_index = torch.stack([u, v])

lattice_mat = structure.lattice_mat.astype(dtype=np.double)

vecs = structure.cart_coords[u.flatten().numpy().astype(int)] -
structure.cart_coords[v.flatten().numpy().astype(int)]
```

Code Snippet 3 - Node Attributes and Edge Indices calculations.

In Code Snippet 3 we have:

- **Node Features:** Each atom is represented by a feature vector, typically based on its atomic number or other relevant properties. These features serve as input data for PotNet's graph-based layers.
- **Edge Indices**: We defined connections between all pairs of atoms, creating an edge index tensor, which PotNet uses to understand relationships between atoms.

PotNet requires certain dependencies, including the GNU Scientific Library (GSL), for advanced numerical computations. We downloaded, configured, and installed GSL, ensuring the environment was ready for PotNet operations. This involved setting up the LD_LIBRARY_PATH and sourcing a .bashrc file to configure system paths. Then we ran these CIF files through PotNet and generated the potentials.

After this step, we needed to generate the PotNet embeddings for each material. We began by loading the dataset and associated CIF files, followed by initializing and configuring

PotNet. We set up PotNet with three convolution layers, defined the dimensional sizes for both atom and edge features, and specified an output vector size. For our work, we determined that an output size of 128 was sufficient to capture the structural complexity of each material. Additionally, we constructed an attribute tensor for each atom within the structure, which provided essential input features for PotNet's embedding process.

Then we had to construct an attribute tensor for each atom in the structure. This tensor uses atomic numbers or other properties to represent each atom as a vector of features, and using PotNetConv and TransformerConv, we create the convolutional and transformer layers to generate embeddings from node and edge features, as shown in Code Snippet 4.

```
edge_embedding = nn.Sequential(
    RBFExpansion(vmin=rbf_min, vmax=rbf_max, bins=fc_features),
    nn.Linear(fc_features, fc_features),
    nn.SiLU(),
)
edge_features = edge_embedding(-0.75 / data.edge_attr)

conv_layers_nn = nn.ModuleList([PotNetConv(fc_features) for _ in range(conv_layers)])
transformer_conv_layers_nn = nn.ModuleList([TransformerConv(fc_features, fc_features) for _ in range(conv_layers)])

for i in range(conv_layers):
    local_node_features = conv_layers_nn[i](node_features, edge_index, edge_features)
    inf_node_features = transformer_conv_layers_nn[i](node_features, inf_edge_index, inf_edge_feat)
    node_features = local_node_features + inf_node_features

```

Code Snippet 4 - Setup of Convultional and Transformer layer.

After the convolutional and transformer layers, global mean pooling is applied to the node features, creating a single vector representation for the entire structure. Following pooling, the features are passed through a fully connected layer, further refining the structure’s representation.

The embeddings generated for each material are saved as rows in a tensor called final_tensor, which is then converted into a DataFrame. This DataFrame is stored in a pickle file (df_features_1000.pkl), creating a tabular representation of structural features with 128 positions.

4.2.3 Text Generation

We began by creating the text modality using RoboCrystallographer to generate detailed descriptions of each material's crystal structure and formula. RoboCrystallographer is a

tool designed to provide textual descriptions of crystal structures, much like the analysis a real-life crystallographer would conduct. It integrates seamlessly with the Materials Project and is compatible with pymatgen formats, producing descriptions like those shown in the bottom right of Figure 10.

The descriptions generated by RoboCrystallographer included advanced structural details that were too intricate for the model to leverage effectively. Attributes such as specific space groups, coordination polyhedra, and complex bonding motifs may have introduced noise rather than useful patterns, as the model struggled to interpret these specialized terms and complexity. The model likely found it challenging to generalize across materials with such specific descriptors. The high degree of specificity meant that each material's description was unique, providing little common ground for the model to learn patterns that could generalize to new data points.

While this level of detail is useful in contexts where human interpretation is needed, we found that these overly detailed descriptions did not contribute effectively to the model's training.

Following our initial experiments with the text modality, we pivoted to a more straightforward approach for text representation by using the composition as a generic text description of each material. We explored CrabNet's architecture as inspiration, focusing on how it represented compositional data. Given that we were working with an NLP transformer architecture, we decided to structure the composition text by separating each element along with its atom count in the material, with elements and counts separated by spaces, so we created a function to dynamically make this conversion.

This approach aimed to simplify tokenization for the generic BERT model used within AutoGluon, ensuring that the model could effectively interpret each component of the composition. Additionally, this approach aimed to increase generalization and similarity across composition texts, promoting the model's ability to identify useful patterns. By standardizing the format, with elements followed by their counts and separated by spaces, we intended to enhance the model's ability to recognize common elemental compositions across materials.

For example, for a material such as Ac_2HgCd, composed of 5 atoms with the composition {"Ac": 1.0, "Hg": 2.0, "Cd": 2.0}, we converted this to the text representation "Ac Hg 2 Cd 2".

4.2.4 Features

From the initial dataset, it was essential to select specific target features for prediction. These features not only served as benchmarks for our models but also allowed for consistent comparison across different modalities and models. By focusing on a core set of properties, we ensured that each model—whether unimodal or multimodal—had a clear, standardized objective, facilitating performance assessment and enabling robust evaluation across all predicted features.

To create our target features, we used specific columns from the initial dataset and applied transformations to make the data more standardized and meaningful for our models.

- Gap (eV): Represents the band gap in electron volts (eV). This feature is typically related to a material's electronic behavior (whether it's a conductor, insulator, or semiconductor), directly retrieved from band_gap_ind;
- Eform (eV/atom): The formation energy per atom in eV, indicating the thermodynamic stability of the material, directly sourced from e_form;
- Ehull (eV/atom): Energy above the convex hull per atom, in eV. This value indicates stability relative to possible phase separation, retrieved from e_above_hull;
- Etot/atom (eV/atom): Total energy per atom in eV, representing the cumulative stability and binding energy of the atomic configuration, calculated by dividing energy_total by nsites(atoms);
- Mag/vol (μB/Å³): Magnetic moment per unit volume in micro-Bohr magnetons per cubic angstrom (μB/Å³). This provides a normalized measure of the material's magnetism, calculated by dividing total_mag by volume.
- Vol/atom (Å³/atom): Atomic volume per atom in cubic angstroms (Å³/atom), which gives insights into atomic packing density, computed by dividing the volume by nsites;
- DOS (eV): Density of states at the Fermi level, in eV, giving insight into the material's electronic and conductive properties, computed by dividing the dos_ef by nsites;

These features provide a comprehensive view of the material properties, each normalized or structured in a way that allows for robust comparisons and effective model training across the selected modalities.

4.3 Dataset Alignment

To create the final dataset with aligned modalities, we followed these steps:

- Starting with the Initial Dataset: We loaded the initial sample dataset, converted it into a Pandas DataFrame, and calculated the predefined feature columns essential for model predictions (e.g., Gap, Eform, Ehull, etc.). All unnecessary columns were removed to focus solely on the target features and identifiers required for merging.
- Adding Image Data: We loaded and unzipped a .zip file containing images representing the 3D structures of each material. Using each material's mat_id as a unique identifier, we merged the image data with the feature dataset, creating a new column that references the image file paths. This column served as the image modality for each material entry.
- Integrating Text Data: The dataset containing text representations of compositions was then joined with the main DataFrame. This text-based composition data provides a descriptive modality for each material's chemical makeup, allowing models to leverage textual insights along with numerical features.
- Including PotNet Embeddings: Finally, we added the 128-dimensional embeddings generated by PotNet, which represent the structural information in a fixed-size vector. These embeddings were incorporated as part of the tabular modality, capturing atomic relationships and structural patterns.

With all modalities—tabular, image, text, and structural embeddings—aligned within a single DataFrame, we created a unified dataset that fully represents each material in multiple dimensions. This final Dataframe was exported to a pickle containing the complete information about the three modalities.

4.4 Model Building

For the Model Building phase, we developed an automated pipeline that streamlined the processes of training, predicting, and evaluating our model. This pipeline allowed for flexibility, adapting to different combinations of selected modalities and target features.

We split the dataset into 85% training and 15% test subsets. This split ensured a robust sample for training while preserving a separate test set for unbiased evaluation.

Leveraging AutoGluon's MultiModalPredictor, the pipeline required minimal configuration for training. By specifying the target feature to predict and selecting the modalities (e.g., tabular, image, text, or combinations thereof), the pipeline could dynamically adapt to various input types. The high_quality preset was chosen to optimize model performance by employing advanced training techniques and hyperparameter tuning, as shown in Code Snippet 5.

```
1.  from autogluon.multimodal import MultiModalPredictor
2. predictor = MultiModalPredictor(label=label_col)
3. predictor.fit(
4.     train_data=train_df[colnames_modalities],
5.     presets='high_quality', # 'best_quality', 'high_quality'
6.     # time_limit=120, # 120 seconds
7. )
```

Code Snippet 5 - Autogluon's MultiModal predictor configuration.

With AutoGluon handling much of the underlying complexity, the pipeline streamlined the training process. AutoGluon's framework automatically optimized architectures, handling data processing and model selection within the specified modality and preset configurations and also saves the best model.

The infrastructure used for model building and training was Google Colab, supported by an NVIDIA T4 GPU, enabling training times ranging from a few minutes to approximately 2 hours, depending on the number of modalities used.

This approach provided an efficient, adaptable way to build and evaluate models across multiple combinations of input data, enabling comprehensive multimodal learning without extensive manual setup.

4.5 Evaluation and Analysis

Following model training, we used the predictor to generate predictions on the test dataset for each combination of modalities and target features. The predictions were saved for further analysis and evaluation.

Using the trained AutoGluon predictor, predictions were made on the test set for each modality combination. These predictions were then saved, allowing us to easily reference and compare results across different models and configurations.

We used evaluation metrics like Mean Absolute Error (MAE) and Mean Absolute Scaled Error (MASE) to quantify the model's performance and to achieve a standardized comparison across different configurations.

- **Mean Absolute Error (MAE)**: We calculated the MAE to assess the average magnitude of prediction errors, providing a straightforward measure of prediction accuracy.
- **Mean Absolute Scaled Error (MASE):** To understand the predictive performance we calculated MASE by scaling the MAE using the Mean Absolute Deviation (MAD) of the target feature.

Following our analysis of model performance using Mean Absolute Error (MAE) and Mean Absolute Scaled Error (MASE), we observed distinct variations in prediction accuracy based on the combination of modalities used.

Table 1 and Table 2 provide a detailed summary of the MAE and MASE values across different modalities and target features, where the lower values indicate higher prediction accuracy, enabling a direct comparison of the effectiveness of each configuration.

Modalities Combination	**Gap (eV)**	**Eform (eVatom-1)**	**Ehull (eVatom-1)**	**Etot/atom (eVatom-1)**	**Mag/vol (ubA-3)**	**Vol/atom (A3atom-1)**	**DOS (eV)**
Tabular	0.1472	0.4941	0.4606	10.3838	0.0085	6.5876	0.2985
Images	0.1093	0.3773	0.3307	13.7607	0.0077	2.6787	0.2396
Text	0.1301	0.3764	0.3557	**5.4159**	0.0059	2.7284	0.2389
Tabular + Images	0.1202	0.3827	0.3236	11.5278	0.0064	2.4272	0.2416
Tabular + Text	0.1112	0.4109	0.3102	5.4299	0.0057	2.7542	0.2245
Images + Text	**0.0933**	0.3358	**0.2833**	9.0845	**0.0054**	**2.2857**	**0.2085**
Tabular + Images + Text	0.1686	**0.3344**	0.2925	5.6597	0.0055	2.2877	0.2171

Table 1 - MAE results for modalities combination.

Modalities Combination	Gap (eV)	Eform (eVatom-1)	Ehull (eVatom-1)	Etot/atom (eVatom-1)	Mag/vol (ubA-3)	Vol/atom (A3atom-1)	DOS (eV)
Tabular	2.2512	0.9530	0.8608	0.5859	1.0767	1.0448	0.9832
Images	1.6718	0.7278	0.6181	0.7764	0.9733	0.4248	0.7892
Text	1.9884	0.7260	0.6648	**0.3056**	0.7442	0.4327	0.7867
Tabular + Images	1.8379	0.7381	0.6048	0.6504	0.8082	0.3850	0.7959
Tabular + Text	1.6998	0.7925	0.5797	0.3064	0.7195	0.4368	0.7395
Images + Text	**1.4268**	0.6476	**0.5294**	0.5126	**0.6811**	**0.3625**	**0.6866**
Tabular + Images + Text	2.5779	**0.6450**	0.5466	0.3193	0.6869	0.3628	0.7151

Table 2 - MASE results for modalities combination.

Based on the expected alignment between MAE and MASE values, we can derive the following facts from the results:

1. Gap (eV)

For the Gap (eV) feature, the results reveal that the Images + Text modality combination achieves the lowest Mean Absolute Scaled Error (MASE) at 1.4268. The presence of visual and textual data appears to add complementary information, enhancing the model's ability to predict Gap accurately.

Conversely, the Tabular + Images + Text combination results in the highest MASE at 2.5779, indicating that incorporating all three modalities does not yield improved accuracy. This elevated error might be attributed to the possibility of overfitting or redundant information, where the additional data fails to contribute meaningful new insights for this specific feature and instead introduces complexity that the model cannot fully leverage.

The deviation between the best and worst MASE values for Gap is 1.1511, a substantial difference that highlights the feature's high sensitivity to the modality combination. This spread underscores the difficulty in predicting Gap accurately, as even the best-performing combination still has a relatively high MASE compared to other features, pointing to the unique challenges associated with modelling this property effectively.

2. Eform (eV/atom)

In this feature, the Images + Text combination yields the best predictive performance with a MASE of 0.5294. This low MASE value suggests that the integration of both image and text data effectively captures the structural or compositional details essential for accurate predictions. The dual-modality approach likely enhances the model's understanding of Eform by providing complementary insights that neither modality alone could fully capture.

In contrast, using Tabular data alone results in the highest MASE at 0.9530, indicating that this single modality may lack the necessary depth or context to predict Eform with high accuracy.

The difference between the best and worst MASE values is 0.4236, a relatively narrow spread compared to other features like Gap. This smaller deviation suggests that Eform predictions are less sensitive to the choice of modality. Most modality combinations provide a moderate level of accuracy, indicating a certain robustness in predicting Eform across different data types.

3. Ehull (eV/atom)

In this target, the Images + Text modality combination once again achieves the lowest MASE, at 0.5294. This result reinforces the strength of combining both image and text data for this feature, as it appears that the multimodal approach successfully captures essential information needed for accurate predictions. The complementary insights provided by visual and textual data seem particularly effective for modelling Ehull.

On the other hand, Tabular data alone yields the highest MASE for Ehull at 0.8608, similar to Eform.

The deviation between the best and worst MASE values is 0.3314, a relatively small range compared to other features. This suggests that predictive performance remains fairly consistent across different modality combinations for Ehull.

4. Etot/atom (eV/atom)

Regarding the Etot/atom (eV/atom) feature, the Text alone modality combination achieves the lowest MASE at 0.3056. This suggests that textual information by

itself effectively captures the essential trends or properties needed for accurate predictions of Etot.

In contrast, the Images alone modality yields a higher MASE of 0.7764, indicating that visual information, when used without supplementary text or tabular data, may not provide enough detail for accurate predictions.

The spread between the best and worst MASE values is 0.4708, suggesting a moderate level of variability across modality combinations. This range underscores the particular benefit of using textual data for Etot predictions, as text provides a more comprehensive basis for accurate predictions compared to other single-modality options.

5. Mag/vol (μB/Å³)

As for the Mag/vol (μB/Å³) feature, the Images + Text modality combination achieves the lowest MASE at 0.6811.

In contrast, using Tabular data alone yields the highest MASE at 1.0767, highlighting its limitations in predicting Mag/vol accurately.

The spread between the best and worst MASE values is 0.3956, representing a modest level of variation across modality combinations. This indicates that while multimodal combinations, particularly Images + Text, improve predictive accuracy, Mag/vol is not highly sensitive to modality changes. However, the multimodal approach still provides a noticeable accuracy boost, emphasizing its value for this feature.

6. Vol/atom (Å³/atom)

In the case of the Vol/atom (Å³/atom) feature, using Images alone yields the lowest MASE at 0.4248. This suggests that volume-related properties may be more effectively captured visually, with image data providing spatial or compositional details that are critical for accurate predictions. The model appears to benefit from the direct visual representation of volume, allowing it to discern patterns that might be harder to capture in other data formats.

On the other hand, Tabular data alone, like in other cases has the highest MASE at 1.0448, indicating that without visual context, tabular data falls short in accurately predicting this feature.

The spread between the best and worst MASE values is 0.6200, which is relatively substantial. This suggests that certain data types, especially images, have a strong impact on prediction accuracy for Vol/atom.

7. DOS (eV)

As for the DOS (eV) feature, the Images + Text modality combination achieves the lowest MASE at 0.6866.

In contrast, Tabular data alone results in the highest MASE at 0.9832, which suggests that relying solely on tabular information is insufficient for precise DOS predictions.

The deviation between the best and worst MASE values is 0.2966, which is relatively small. This limited spread suggests moderate consistency in prediction accuracy across different modality combinations for DOS, although multimodal approaches, especially Images + Text, clearly offer an advantage.

Overall, as illustrated in the bar charts in Figure 12, it is evident that the Text and Image combination consistently produces the most accurate results across various features.

Moreover, the analysis confirms that the use of multiple modalities generally leads to improved accuracy over single-modality approaches.

We also observe, as illustrated in Figure 13, that the Gap feature posed the greatest challenge for accurate prediction, likely due to its complex dependencies, while Vol/atom emerged as the easiest to predict, showing more consistent accuracy across models. This finding highlights the inherent variability in prediction difficulty across features, with some properties requiring richer data inputs for accurate modelling.

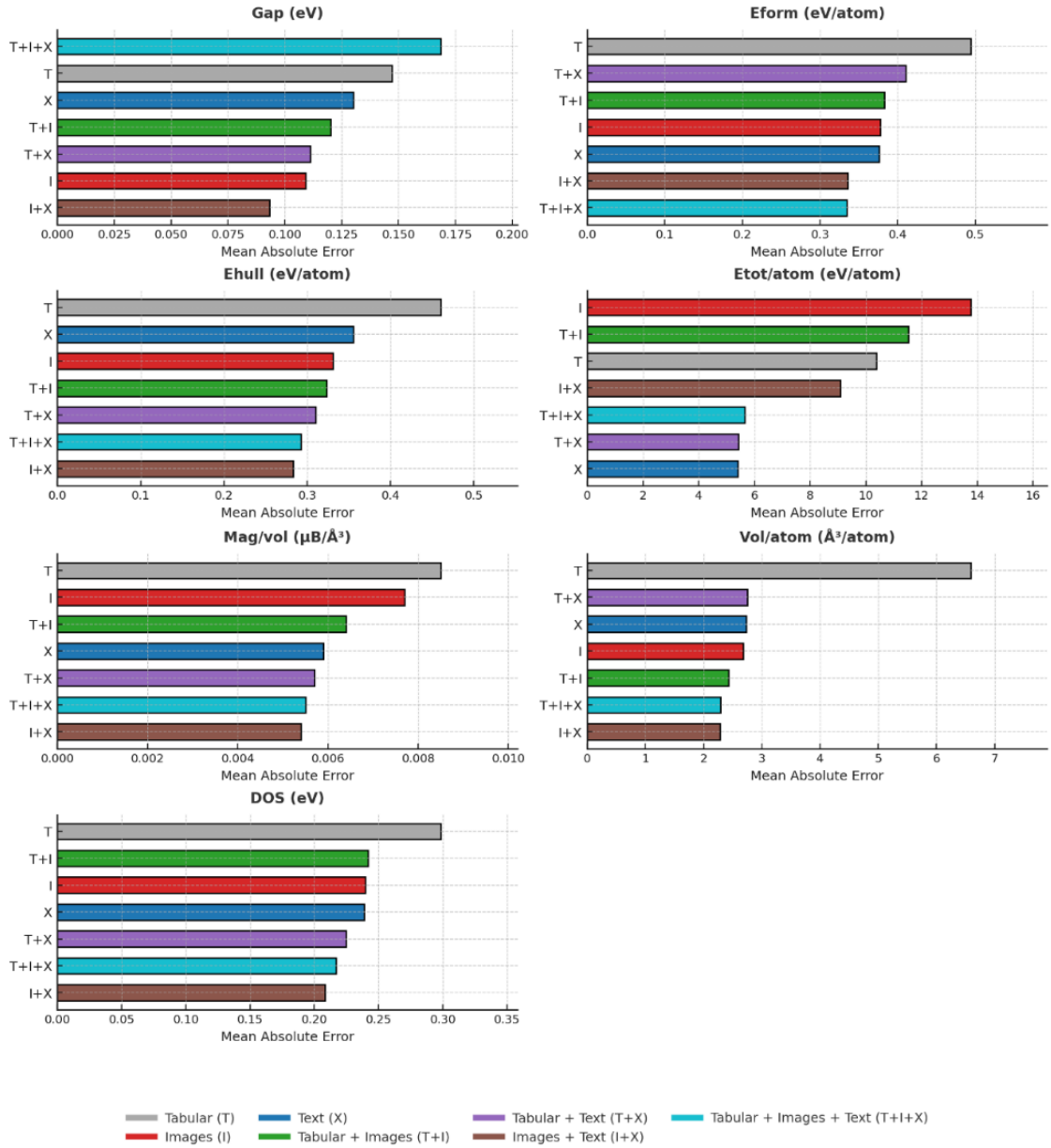


Figure 12 - Graphic representation of MAE for each target.

Additionally, the results confirm that using multiple modalities yields better predictive accuracy overall compared to single-modality approaches. Among single modalities, Tabular data consistently produced the least accurate results, underscoring its limitations in capturing the depth of information needed for complex features. These insights reinforce the advantage of multimodal approaches, where combining diverse data types provides a more comprehensive understanding and leads to improved performance across features.

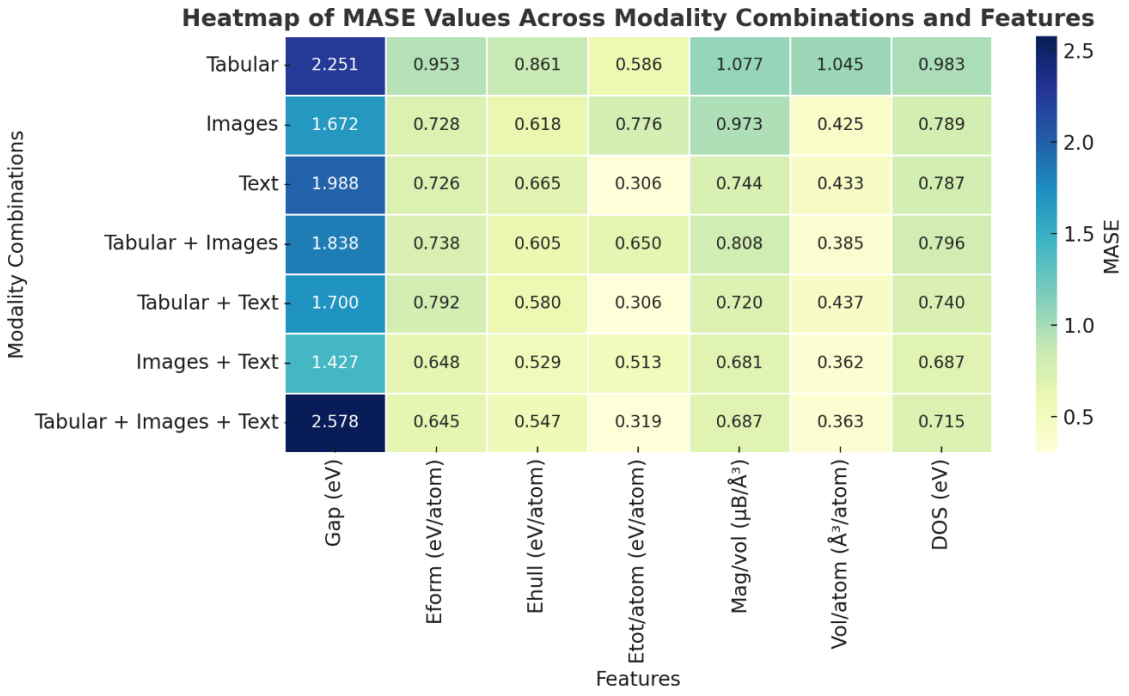


Figure 13 - Heatmap of MASE Values Across Modality Combinations and Features.

5 Conclusion

The primary objective of this dissertation was to investigate the impact of various data modalities versus single modalities on the predictive accuracy of material property models. By exploring combinations of Text, Image, and Tabular data, this study aimed to determine the modality configurations that yield the highest accuracy for predicting key material properties. Using advanced machine learning algorithms, the predictive models were trained on a multimodal dataset encompassing a range of material features, with Mean Absolute Error (MAE) and Mean Absolute Scaled Error (MASE) serving as the primary evaluation metrics.

The findings demonstrate that multimodal approaches significantly enhance predictive accuracy compared to single-modality models. Notably, the Text and Image combination consistently produced the lowest MASE and MAE values, indicating its effectiveness across a variety of features. This combination was particularly beneficial for complex features such as Gap (eV), which require intricate compositional and structural information that a single modality could not fully capture.

On the other hand, single-modality models, particularly Tabular data alone, exhibited the lowest accuracy across most features. Tabular data, while useful in capturing numeric trends, lacked the compositional and structural depth needed for complex property predictions. Additionally, certain features, such as Vol/atom (Å^3/atom), emerged as easier to predict across models, while Gap (eV) remained challenging due to its inherent complexity.

The results of this study underscore the critical role that multimodal data integration plays in predictive modeling for materials science. By combining diverse data sources, models can capture a broader spectrum of information, leading to more accurate and generalized predictions. These findings align with the growing trend toward multimodal machine learning, which leverages various data types to improve model robustness and predictive power.

While the results are promising, this work encountered several limitations that warrant consideration. First, the dataset used, though multimodal, may have lacked sufficient diversity or quantity in certain modalities, potentially impacting model performance. The use of the Alexandria full database would be very beneficial. Additionally, the machine learning models applied were limited to current architectures, which may not fully capture

the complexity of material properties. Computational constraints also limited the exploration of more extensive hyperparameter tuning and the testing of advanced models, which could further optimize prediction accuracy. Also, the foundational models used by AutoGluon are pre-trained with generic data, making them not the best suited for Materials Science.

Building on the insights gained from this study, future research could focus on incorporating additional data types, such as molecular simulations or spectroscopic data, which could provide even deeper insights and enhance predictive accuracy, particularly for complex features like Gap.

Exploring more advanced or specialized architectures, such as Transformer-based models for Text and Image data, may better capture the multimodal interactions needed for accurate material property predictions.

Developing tailored models for particularly challenging features, like Gap, could improve prediction accuracy. These models could incorporate domain-specific information to address the unique complexities of certain properties. For instance, the use of domain-specific models like MatBERT, which is tailored to material science language, may outperform general-purpose models like BERT, used in AutoGluon by capturing field-specific terminology and nuances more effectively.

Expanding the dataset with diverse material samples across a broader range of property values would improve model robustness and generalization.

By leveraging the complementary strengths of Text, Image, and Tabular data, this work underscores the potential of machine learning models to achieve higher accuracy and generalization. As materials science increasingly adopts data-driven approaches, the integration of multimodal data will likely play a pivotal role in advancing predictive accuracy and accelerating the discovery of novel materials.

In conclusion, this dissertation provides a foundation for future research in multimodal materials science, offering insights and recommendations that will help shape more accurate, versatile, and comprehensive predictive models for complex material properties.

Atrey, P. K., Hossain, M. A., El Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, *16*(6), 345–379. https://doi.org/10.1007/s00530-010-0182-0

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2017). *Multimodal Machine Learning: A Survey and Taxonomy* (arXiv:1705.09406). arXiv. http://arxiv.org/abs/1705.09406

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). CRISP-DM 1.0 step-by-step data mining guide. Em *Springer*.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). *A Simple Framework for Contrastive Learning of Visual Representations* (arXiv:2002.05709). arXiv. https://doi.org/10.48550/arXiv.2002.05709

Chen, Y., Wei, F., Sun, X., Wu, Z., & Lin, S. (2023). *A Simple Multi-Modality Transfer Learning Baseline for Sign Language Translation* (arXiv:2203.04287). arXiv. https://doi.org/10.48550/arXiv.2203.04287

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

D'mello, S. K., & Kory, J. (2015). A Review and Meta-Analysis of Multimodal Affect Detection Systems. *ACM Computing Surveys*, *47*(3), 1–36. https://doi.org/10.1145/2682899

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N.

(2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale* (arXiv:2010.11929). arXiv. https://doi.org/10.48550/arXiv.2010.11929

Gasteiger, J., Becker, F., & Günnemann, S. (2021). GemNet: Universal Directional Graph Neural Networks for Molecules. *Advances in Neural Information Processing Systems*, *34*, 6790–6802. https://proceedings.neurips.cc/paper/2021/hash/35cf8659cfcb13224cbd47863a34fc58-Abstract.html

Gasteiger, J., Groß, J., & Günnemann, S. (2022). *Directional Message Passing for Molecular Graphs* (arXiv:2003.03123). arXiv. https://doi.org/10.48550/arXiv.2003.03123

Guo, W., Wang, J., & Wang, S. (2019). *Deep Multimodal Representation Learning: A Survey*. *7*.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition* (arXiv:1512.03385). arXiv. https://doi.org/10.48550/arXiv.1512.03385

Horton, M., Shen, J.-X., Burns, J., Cohen, O., Chabbey, F., Ganose, A. M., Guha, R., Huck, P., Li, H. H., McDermott, M., Montoya, J., Moore, G., Munro, J., O'Donnell, C., Ophus, C., Petretto, G., Riebesell, J., Wetizner, S., Wander, B., … Persson, K. A. (2023). *Crystal Toolkit: A Web App Framework to Improve Usability and Accessibility of Materials Science Research Algorithms* (arXiv:2302.06147). arXiv. https://doi.org/10.48550/arXiv.2302.06147

Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., & Persson, K. A. (2013). Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, *1*(1), 011002. https://doi.org/10.1063/1.4812323

Kipf, T. N., & Welling, M. (2017). *Semi-Supervised Classification with Graph Convolutional Networks* (arXiv:1609.02907). arXiv. https://doi.org/10.48550/arXiv.1609.02907

Lan, Z., Bao, L., Yu, S.-I., Liu, W., & Hauptmann, A. G. (2012). Double Fusion for Multimedia Event Detection. Em K. Schoeffmann, B. Merialdo, A. G. Hauptmann, C.-W. Ngo, Y. Andreopoulos, & C. Breiteneder (Eds.), *Advances in Multimedia Modeling* (Vol. 7131, pp. 173–185). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-27355-1_18

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324. Proceedings of the IEEE. https://doi.org/10.1109/5.726791

Li, Z., Xie, C., & Cubuk, E. D. (2024). *Scaling (Down) CLIP: A Comprehensive Analysis of Data, Architecture, and Training Strategies* (arXiv:2404.08197). arXiv. http://arxiv.org/abs/2404.08197

Liang, P. P., Zadeh, A., & Morency, L.-P. (2023). *Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions* (arXiv:2209.03430). arXiv. https://doi.org/10.48550/arXiv.2209.03430

Lin, Y., Yan, K., Luo, Y., Liu, Y., Qian, X., & Ji, S. (2023). *Efficient Approximations of Complete Interatomic Potentials for Crystal Property Prediction* (arXiv:2306.10045). arXiv. http://arxiv.org/abs/2306.10045

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows* (arXiv:2103.14030). arXiv. https://doi.org/10.48550/arXiv.2103.14030

MatBERT. (2021). *MatBERT: A pretrained BERT model on materials science literature*. https://github.com/lbnlp/MatBERT

Moro, V., Loh, C., Dangovski, R., Ghorashi, A., Ma, A., Chen, Z., Kim, S., Lu, P. Y., Christensen, T., & Soljačić, M. (2024). *Multimodal Learning for Materials* (arXiv:2312.00111; Versão 3). arXiv. https://doi.org/10.48550/arXiv.2312.00111

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). *Learning Transferable Visual Models From Natural Language Supervision* (arXiv:2103.00020). arXiv. https://doi.org/10.48550/arXiv.2103.00020

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). *SQuAD: 100,000+ Questions for Machine Comprehension of Text* (arXiv:1606.05250). arXiv. https://doi.org/10.48550/arXiv.1606.05250

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2009). The Graph Neural Network Model. *IEEE Transactions on Neural Networks, 20*(1), 61–80. IEEE Transactions on Neural Networks. https://doi.org/10.1109/TNN.2008.2005605

Schmidt, J., Cerqueira, T. F. T., Romero, A. H., Loew, A., Jäger, F., Wang, H.-C., Botti, S., & Marques, M. A. L. (2024). Improving machine-learning models in materials science through large datasets. *Materials Today Physics, 48*, 101560. https://doi.org/10.1016/j.mtphys.2024.101560

Schütt, K. T., Kindermans, P.-J., Sauceda, H. E., Chmiela, S., Tkatchenko, A., & Müller, K.-R. (2017). *SchNet: A continuous-filter convolutional neural network for modeling quantum interactions* (arXiv:1706.08566). arXiv. https://doi.org/10.48550/arXiv.1706.08566

Škrlj, B. (2024). *From Unimodal to Multimodal Machine Learning: An Overview*. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-57016-2

Summaira, J., Li, X., Shoib, A. M., Li, S., & Abdul, J. (2021). *Recent Advances and Trends in Multimodal Deep Learning: A Review* (arXiv:2105.11087). arXiv. http://arxiv.org/abs/2105.11087

Tan, M., & Le, Q. V. (2020). *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks* (arXiv:1905.11946). arXiv. https://doi.org/10.48550/arXiv.1905.11946

Tang, Z., Fang, H., Zhou, S., Yang, T., Zhong, Z., Hu, T., Kirchhoff, K., & Karypis, G. (2024). *AutoGluon-Multimodal (AutoMM): Supercharging Multimodal AutoML with Foundation Models* (arXiv:2404.16233). arXiv. http://arxiv.org/abs/2404.16233

Taylor, W. L. (1953). "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly*, *30*(4), 415–433. https://doi.org/10.1177/107769905303000401

Tian, Y., Krishnan, D., & Isola, P. (2020). *Contrastive Multiview Coding* (arXiv:1906.05849). arXiv. https://doi.org/10.48550/arXiv.1906.05849

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). *Graph Attention Networks* (arXiv:1710.10903). arXiv. https://doi.org/10.48550/arXiv.1710.10903

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language*

Understanding (arXiv:1804.07461). arXiv. https://doi.org/10.48550/arXiv.1804.07461

Wang, A. Y.-T., Kauwe, S. K., Murdock, R. J., & Sparks, T. D. (2021). Compositionally restricted attention-based network for materials property predictions. *Npj Computational Materials*, *7*(1), 77. https://doi.org/10.1038/s41524-021-00545-1

Wu, Z., Cai, L., & Meng, H. (2005). Multi-level Fusion of Audio and Visual Features for Speaker Identification. Em D. Zhang & A. K. Jain (Eds.), *Advances in Biometrics* (Vol. 3832, pp. 493–499). Springer Berlin Heidelberg. https://doi.org/10.1007/11608288_66

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A Comprehensive Survey on Graph Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(1), 4–24. IEEE Transactions on Neural Networks and Learning Systems. https://doi.org/10.1109/TNNLS.2020.2978386

Xie, T., & Grossman, J. C. (2018a). Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters*, *120*(14), 145301. https://doi.org/10.1103/PhysRevLett.120.145301

Xie, T., & Grossman, J. C. (2018b). Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters*, *120*(14), 145301. https://doi.org/10.1103/PhysRevLett.120.145301

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *Proceedings of the 32nd International Conference on Machine Learning*, 2048–2057. https://proceedings.mlr.press/v37/xuc15.html

Xu, P., Zhu, X., & Clifton, D. A. (2023). *Multimodal Learning with Transformers: A Survey* (arXiv:2206.06488). arXiv. https://doi.org/10.48550/arXiv.2206.06488

Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y. (2018). *SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference* (arXiv:1808.05326). arXiv. https://doi.org/10.48550/arXiv.1808.05326

Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M., & Shen, Y.-D. (2020). Dual-Path Convolutional Image-Text Embeddings with Instance Loss. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *16*(2), 1–23. https://doi.org/10.1145/3383184

Zhihong Zeng, Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(1), 39–58. https://doi.org/10.1109/TPAMI.2008.52