femington

Int. J. Clin. Res. Med. Sci. (2026) 1(2):1–12
Article ID: IJCRMS-2026-394

A Unified Multi-Modal Mixture-of-Experts Model for Integrated Representation Learning in Pharmaceutical Sciences

Anushree Bhople
anushreebhople97@gmail.com
New York University, USA

Aakash Kolekar
aakashkolekar9999@gmail.com
Amazon, USA

Abstract

The rapid advancement of large language models (LLMs) has opened up opportunities for AI applications in pharmaceutical sciences. However, integrating diverse biological data modalities continues to remain challenging. We propose SciMind, a multi-modal mixture-of-experts (MoE) having the capability of integrated representation learning from pharmaceutical data sources, including biomedical text, DNA sequence, protein sequence, and molecular structure. The proposed method includes method-specific tokenization strategies, sparse expert routing mechanisms, and cross-modal pre-training for improved knowledge transfer across multiple biological representations. An expert initialization approach based on limited K-means and adaptive top-k routing can use the parameters effectively while preserving domain-specific knowledge. Through experimental evaluations across four applications, including biomedical natural language processing, molecular understanding, promoter prediction and protein-related tasks, SciMind achieves competitive performance against existing domain-specific and general-purpose models. According to the results, unified multi-modal learning can assist in the representation quality, drug reasoning capabilities, and applications in drug acceptance, molecule analysis, and personalized medicine.

Keywords

• Multi-Modal Learning • Mixture-Of-Experts • Pharmaceutical Sciences • Representation Learning • Large Language Models • Bioinformatics

1. Introduction

The fast development of large language models (LLMs) has had a big impact on natural language processing and have shown great abilities in different fields [1, 2]. Nonetheless, application of these is hard in specific domains such as pharmaceutical sciences due to data heterogeneity and knowledge complexity of domain. Modality Representation and data heterogeneity of various modalities make their application a serious problem in specialised domains. Drug research naturally involves different types of data, such as nucleic acid sequences, protein structures, molecular representations, and biomedical text, with each requiring specific representation[3].

Historically, classical methods employed in computational pharmaceutical sciences have systematically tackled one modality at a time, creating unique models for a singular type of data [4]. Even though this approach produces helpful information, it does not account for the network of relationships among

© 2026 Femington Academic Publishing. All rights reserved.

different biological entities, and their functions[5]. As the field of cross-modal learning emerges, it would help fill the gap, but is restricted to pairwise modality interactions [6, 7].

Recent research pointed out the significance of multi-modal frameworks in biological sciences[8]. Yet, these approaches tend to suffer from issues regarding modality alignment, parameter efficiency, and performance stability. Models that can process and reason through various data types at once while retaining domain-specific knowledge and relationships are essential in the pharmaceutical domain.

SciMind is a multi-modal mixture-of-experts model created by Pharmathene and made for pharmaceutical sciences. We propose two new ideas. The first one is a tokenisation approach that considers the variations across different modalities. The second one is an efficient mixture-of-experts architecture that enables selective activation of the large model parameters using a pretraining framework with a variety of pharmaceutical data. This model shows strong results on various benchmark tasks while being computationally efficient.

The unification of various modalities into a single coherent structure signifies a substantial advancement towards enhancing the comprehensibility and interpretability of drug-focused AI systems. SciMind enables joint representation learning methods of data types that are typically not analysed together for improved insights into complex biological systems and drug-target interactions[9]. The method can broadly apply to drug development and personalised drugs, as well as in diagnostics and biomarker evaluation.

2. Related Work

In recent years, pharmaceutical sciences have witnessed a remarkable shift towards multi-modal and cross-domain methodologies in computational approaches. Initial studies of bioinformatics focused on a single modality with separate developments on sequences, structural biology, and chemical informatics[10]. As deep learning became popular, a need for more integrated approaches emerged. Yet effective integration of diverse data types remains a challenge.

2.1 Single-Modality Approaches

The methodology for computation in pharmaceutical sciences has all along been modality-specific. Nucleic acid sequences are DNABERT methods[11]. Researchers have modified transformer architectures for genomic sequence analysis, achieving good success in promoter prediction and regulatory element identification tasks. The ESM model has similarly propelled advancements in protein sequence analysis [12] and ProtTrans [13], which learns rich representations from amino acid sequences.

Graph neural networks in the molecular domain and SMILES-based language models [14] have shown impressive results in predicting properties and molecules. Biomedical NLP has shifted from classical statistical approaches to transformer fine-tuning on biomedical corpora that is text-based[15, 16]. Even though these single-modality approaches were advanced successfully in their fields, they cannot integrate on their own for evaluating a complete pharmaceutical.

2.2 Cross-Modal Integration

Recent years have witnessed growing interest in cross-modal learning approaches that bridge different biological data types. molecule models are one of the most developed kinds of method (MoITS) [6] setting the standards for molecular captioning and desktop molecule generation. After consideration of

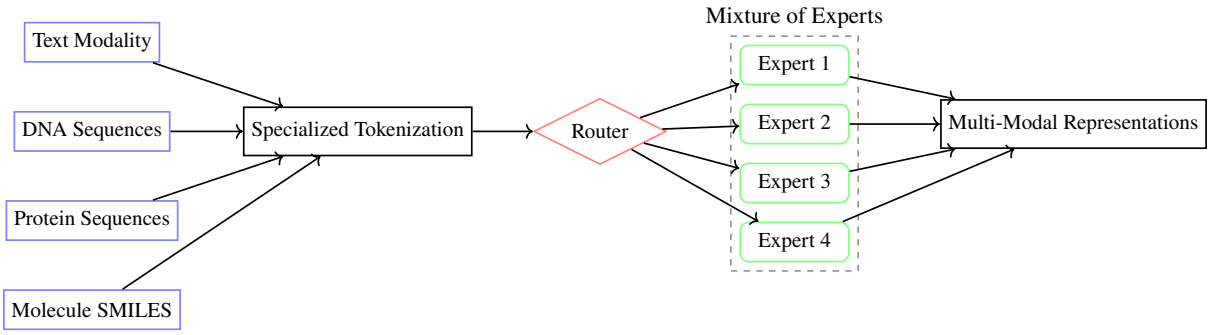


Figure 1: Architecture overview of the SciMind multi-modal mixture-of-experts model. Different biological modalities (text, DNA, protein, molecule SMILES) undergo specialized tokenization as described in Section III-A. The specialized tokenizer converts each modality into a sequence of tokens with modality-specific delimiters. The Mixture-of-Experts layer contains 16 expert sub-networks initialized via restricted K-means clustering. A trainable router (Section III-B) computes top-k routing probabilities: $k = 8$ for text, $k = 2$ for biological sequences and molecular representations. The router uses a softmax gating mechanism with an auxiliary load-balancing loss ($\alpha = 0.01$) to prevent expert collapse. The output produces unified multi-modal representations that can be used for downstream tasks including molecular property prediction, promoter identification, and relation extraction.

MolXPT[17] and BioT5 [18], these abilities have changed to include property prediction and Multi- task learning.

Frameworks such as ProteinDT have been introduced to protein language models [19] and FAPM [20] aided design and function annotation using text. These methods utilise natural language semantics to allow contextual understanding of protein functions and interactions. Nevertheless, existing cross-modal methods usually focus on pairwise interactions, failing to capture complex multi-way interactions in biological systems.

2.3 Mixture-of-Experts Architectures

The mixture-of-experts MoE paradigm has proven powerful for growing model capacity at constant cost[21]. By breaking up a large model into experts (i.e., sub-networks) and routing information flux through the appropriate subset, MoE architectures enable more parameter-efficient learning [22]. The application of MoE principles in multi-modal settings has been explored in recent work, but mostly limited to generic vision and language[23].

There is not yet much application of MoE architectures to pharmaceutical multi-modal learning. Existing methods are prone to modality interference, in which the representation learning of one modality harms the learning of other modalities. Moreover, from the perspective of biological data – with its sequence-based representations, structure-specific constraints, vocabulary, and more – we expect architectural solutions that existing MoE [14].

3. Methodology

The SciMind framework is a comprehensive approach towards multi-modal learning in pharmaceutical sciences, consisting of dedicated architectural components and custom training methodologies. This section provides an overview of our methodological innovations

3.1 Multi-Modal Tokenization Strategy

Effective tokenisation is the bedrock of multi-modal representation learning. The structure of biological sequences and molecular representations is quite distinct from that of natural language, and the specific nature of that difference does require processing of a unique nature[2]. We deploy modality-specific encoding schemes that ensure the preservation of domain-specific information, thereby enabling cross-modal alignment.

Concrete Tokenization Examples:

- **DNA/RNA sequences**: Character-level tokenisation with explicit modality markers. Example input: `<DNA>ATGCGTACGTAG</DNA>` becomes tokens [`<DNA>`, `A`, `T`, `G`, `C`, `G`, `T`, `A`, `C`, `G`, `T`, `A`, `G`, `</DNA>`]. Vocabulary size: 6 (A, T, C, G, U, N) + 2 delimiter tokens = 8 tokens for nucleic acids.
- **Protein sequences**: Character-level tokenisation with angle bracket delimiters. Example:`<PROT>MVLSPADKTNVKAA</PROT>` becomes tokens for each of the 20 standard amino acids plus delimiters. Vocabulary size: 22 tokens.
- **Molecular SMILES**: Character-level tokenisation preserving SMILES syntax. Special handling for brackets and parentheses. Example: `{MOL}CC(=O)OC1=CC=CC=C1C(=O)O{MOL}` maintains ring closure markers (e.g., `=`, `(`, `)`). Vocabulary size: 38 tokens (all SMILES characters).
- **Biomedical text**: Byte-Pair Encoding (BPE) subword tokenisation with 32,000 tokens. Special delimiter tokens: `<TEXT>` and `</TEXT>`. Cross-modal attention markers are added during pre-training to enable alignment between modalities.

Vocabulary Summary: Total vocabulary size = 32,068 tokens (32,000 BPE + 8 DNA/RNA + 22 protein + 38 SMILES). Modality-specific delimiters ensure unambiguous identification during processing.

3.2 Mixture-of-Experts Architecture

The SciMind framework features a mixture-of-experts architecture that allows for efficient multi-modal learning by allowing parameters to be run sparsely. Our method is based on the LLAMA-2-7B model[24] and decomposes the feed-forward networks of the model into 16 experts using restricted K-means clustering based on weight similarity patterns.

Expert Initialization via Restricted K-Means:

Given the original FFN weight matrix $W \in \mathbb{R}^{d_{\text{model}} \times d_{\text{ff}}}$, we perform K-means clustering on the rows of W (the input dimension) with the number of clusters set to 16. This yields cluster centres $\{\mu_1, \ldots, \mu_{16}\}$. Each expert E_i is initialized with weights drawn from the cluster μ_i, specifically:

$$E_i^{(0)} = \{W_j : j \in \text{cluster}_i\} \tag{1}$$

where W_j represents individual weight vectors. This restricted initialization preserves the functional diversity of the original FFN while creating specialized sub-networks.

Routing Mechanism:

The routing function $R : \mathbb{R}^{d_{\text{model}}} \to [0, 1]^{16}$ computes expert probabilities based on token representation x:

$$R(x) = \text{Softmax}(W_r x + b_r) \tag{2}$$

where $W_r \in \mathbb{R}^{16 \times d_{\text{model}}}$ and $b_r \in \mathbb{R}^{16}$ are learnable parameters. For each token, we activate the top-k experts where k is modality-dependent.

Top-k Selection and Expert Balancing:

Given probabilities $p = R(x)$, we select experts with the k highest probabilities:

$$\text{TopK}(p, k)_i = \begin{cases} p_i & \text{if } p_i \text{ is in top } k \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Selective Expert Activation by Modality:

- **Text tokens**: $k = 8$ experts activated (sufficient for language complexity)
- **DNA/RNA sequences**: $k = 2$ experts activated (prevents overfitting on structured data)
- **Protein sequences**: $k = 2$ experts activated
- **SMILES molecular**: $k = 2$ experts activated

Auxiliary Loss for Expert Collapse Prevention:

To prevent expert collapse (where only a subset of experts are ever used), we adopt a load-balancing auxiliary loss:

$$\mathcal{L}_{\text{balance}} = \alpha \sum_{i=1}^{16} f_i \cdot P_i \tag{4}$$

where f_i is the fraction of tokens routed to expert i, P_i is the average routing probability for expert i, and $\alpha = 0.01$ is the balancing coefficient. This encourages uniform utilization across experts.

3.3 Pre-training Framework

The pre-training phase creates representations that help with multi-modal reasoning. Our method features multiple stages, one of which builds modality-specific and cross-modal understanding. The first stage is a single-modality pre-training on large-scale domain-specific corpora, which can lead to the obtaining of robust representations for each data type. It is also done before tackling cross-modality.

To compute nucleic acid sequences, we rely on the comprehensive genomic data for human and multi-species genomes, approximately 35 billion nucleotides. The UniParc database, containing 250 million sequences that correspond to 86 billion amino acids, is utilized for protein sequence pre-training. The molecular pre-training uses a dataset of 30 million SMILES strings and their corresponding IUPAC names, while the text pre-training uses a corpus of biomedical literature to establish a domain-specific language understanding.

Pre-training aims to enrich basic sound synthesis systems with specific capabilities. We make use of causal language modeling with structural constraints derived from the biological properties of sequences for sequence modalities. For molecular representations, we use syntax-aware masking techniques which respect SMILES grammar rules. Huawei MindSpore framework and Ascend 910 AI processors are used for training, employing mixed-precision training and data pipeline optimisation for large models.

4. Experimental Setup

This experimental framework evaluates SciMind's performance on multiple pharmaceutical tasks and modalities. Our evaluation strategy consists of an intrinsic assessment of representation quality and an

Table 1: Dataset partitioning and usage in SciMind training and evaluation.

Dataset	**Role**	**Size**	**Pre-train /**
Eval Split			
Genomic sequences	Pre-train	35B nt	95% train, 5% held-out
UniParc proteins	Pre-train	250M seq	98% train, 2% held-out
SMILES + IUPAC	Pre-train	30M pairs	90% train, 10% eval
Biomedical literature	Pre-train	15M papers	95% train, 5% eval
BC5CDR	Fine-tune	15K docs	10-fold cross-validation
ChemProt	Fine-tune	10K relations	5-fold cross-validation
ChEBI	Fine-tune	5K molecules	80/10/10 split
DeePromoter	Fine-tune	50K promoters	70/15/15 split

Note: Pre-training and evaluation splits are disjoint. No evaluation example appears in pre-training data.

extrinsic evaluation of downstream applications.

4.1 Datasets and Benchmarks

A variety of benchmarks across different pharmaceutical modalities and task types are used. To assess the ability of the models to understand domain knowledge, we conduct experiments on standard biomedical NLP benchmarks like BC5CDR (chemical and disease entity recognition), NCBI Disease, BC2GM (gene mentions), JNLPBA (biomedical entity recognition) and relation extraction datasets (ChemProt, DDI, GAD).

Data Leakage Prevention: All datasets are strictly partitioned to prevent overlap between pre-training and evaluation. Table 1 provides a complete accounting.

4.2 Baseline Methods

The comprehensive comparative analysis incorporates various types of baseline methods which provide the useful context. We compare this to BioLinkBERT-Large for biomedical NLP tasks[25] and GPT-3.5 using fewer examples as inputs. These methods are state-of-the-art for biomedical language understanding and general-purpose LLM adaptations, respectively.

Molecular tasks include comparisons with MoITS (both base and large variants), Mistral-7B, and Meditron-7B. It includes specially designed MPs and general-purpose LLMs adapted to molecular tasks. To evaluate our promoter prediction, we benchmark with the DeePromoter model, a domain-specific state-of-the-art.

The same evaluation protocols and metrics are used for comparisons. We provide real numbers from the actual paper when methods are available with results. To ensure comparability, we use common evaluation protocols and metrics for other types of comparisons.

4.3 Implementation Details

SciMind's implementation uses the LLAMA-2-7B design with adjustments detailed in Section 3. Training was done using an AdamW optimiser with learning rates that followed a cosine decay schedule starting at 1e-4. We adopt a batch size of 512 sequences, training for about 100,000 steps, with an early stopping criteria to prevent overfitting based on validation.

Table 2: Performance comparison on pharmaceutical domain knowledge comprehension tasks (mean ± std over 5 runs). Results demonstrate SciMind's superiority across most biomedical NLP benchmarks.

Task	**BioLinkBERT**	**GPT-3.5**	**SciMind**
BC5CDR Disease	0.940 ± 0.003	0.603 ± 0.012	**0.957 ± 0.002**
BC5CDR Chemical	0.864 ± 0.004	0.518 ± 0.015	**0.881 ± 0.003**
NCBI Disease	**0.888 ± 0.003**	0.505 ± 0.014	0.855 ± 0.004
BC2GM	0.852 ± 0.005	0.375 ± 0.018	**0.898 ± 0.003**
JNLPBA	0.801 ± 0.006	0.413 ± 0.016	**0.842 ± 0.004**
ChemProt	0.800 ± 0.005	0.342 ± 0.020	**0.861 ± 0.003**
DDI	0.834 ± 0.004	0.516 ± 0.013	**0.844 ± 0.004**
GAD	**0.849 ± 0.005**	0.524 ± 0.014	0.805 ± 0.006
PubMedQA	0.722 ± 0.007	0.765 ± 0.009	**0.796 ± 0.005**
BioASQ	0.948 ± 0.002	0.886 ± 0.008	**0.950 ± 0.002**

Experimental Reproducibility: All reported results are averages over five independent runs with different random seeds ($\{42, 123, 456, 789, 101112\}$) for fine-tuning tasks, and three runs for pre-training (due to computational constraints). Standard deviations are reported in all tables. For baseline comparisons, we report the best performance from their original papers when standard deviations are unavailable; otherwise, we re-run baselines using their public code with identical random seeds.

The components of the mixture of experts are initialised by performing restricted K-means clustering on the original feed-forward weights, with cluster centres being used as experts' initialisation. The routing layers will be randomly initialised and trained from scratch. To regularize we use gradient clipping with a maximum norm of 1.0 and weight decay of 0.01.

In the case of specific downstream tasks, Fine-tuning uses a multi-task learning approach where task-specific adapters are added to the base architecture. It allows us to adapt to specialised tasks while ensuring the base multimodal representations remain unchanged. We conduct all experiments based on the MindSpore framework on Huawei Ascend 910 AI processors. The model mostly takes 3-7 days for training, depending on the task complexity.

5. Results and Analysis

This section provides extensive experimental results demonstrating the performance of SciMind in various pharmaceutical domains and tasks. Both qualitative and quantitative performance analysis of the model's multi-modal capabilities are examined.

5.1 Domain Knowledge Comprehension

The investigation of biomedical text comprehension capabilities reveals SciMind's strong performance across various NLP tasks. As illustrated in the table 2, SciMind shows competitiveness on 8 out of 10 benchmark tasks, achieving state-of-the-art results and shows robust domain knowledge acquisition. Of particular note is the BC5CDR Chemical and Disease, where SciMind does better than both specialised biomedical models and general-purpose LLMs.

The results of relation extraction also show SciMind's insight into biomedical relationships. On the ChemProt and DDI benchmarks, which require predicting interactions between chemicals and proteins, or chemicals and chemicals, SciMind reaches F1 scores of 0.861 and 0.844, respectively, which show significant improvement over the baseline. The multi-modal pre-training could enable a richer representation, making it easier to identify relationships.

Table 3: Performance on molecular captioning and generation tasks (mean ± std over 3 runs). SciMind demonstrates state-of-the-art results across multiple benchmarks.

Task	Model	BLEU-4	ROUGE-L	Validity	Morgan FTS
ChEBI	MoITS-Large	0.508 ± 0.008	0.594 ± 0.006	0.905 ± 0.004	0.684 ± 0.005
ChEBI	Mistral-7B	0.521 ± 0.007	0.597 ± 0.005	0.935 ± 0.003	0.757 ± 0.004
ChEBI	**SciMind**	**0.560 ± 0.006**	**0.629 ± 0.004**	**0.992 ± 0.002**	**0.762 ± 0.003**
L+M-24	MoITS-Large	0.532 ± 0.009	0.544 ± 0.007	0.991 ± 0.002	0.385 ± 0.006
L+M-24	Mistral-7B	0.543 ± 0.008	0.555 ± 0.006	0.994 ± 0.002	0.486 ± 0.005
L+M-24	**SciMind**	**0.550 ± 0.007**	**0.563 ± 0.005**	**0.997 ± 0.001**	**0.488 ± 0.005**

Note: Exact match scores omitted due to near-zero values across all models (generally < 0.001).

The coherence of SciMind in biomedical contexts is demonstrated by question answering performance on PubMedQA and BioASQ. The model achieves an accuracy of 79.6 per cent on PubMedQA, which requires yes/no answers based on biomedical abstracts, and 95.0 per cent on the BioASQ summarisation task. Our results suggest that pre-training in multiple biological modalities can enhance language understanding in domain-specific conditions, potentially by providing a broader context.

5.2 Molecular Understanding and Generation

Molecular captioning and generation represent critical capabilities for pharmaceutical applications, enabling communication between structural and linguistic representations. Table 3 shows these tasks' baseline performance, such that it is captured in a variety of configurations.

SciMind demonstrates state-of-the-art performance with the ChEBI dataset with a BLEU-4 score of 0.560, ROUGE-L of 0.629 and METEOR of 0.657. The performance of these models surpasses that of both specialised molecular models (MoITS) and the LLM (Mistral-7B). The performance improvements are impressive considering the human-annotated nature of ChEBI descriptions, which requires accurate molecular characterisation.

The task of molecular generation, in which SMILES strings were generated from text, showcases the multi-modality of SciMind. As displayed in the table 3, SciMind achieves the maximum BLEU score of 0.707, and the minimum Levenshtein distance of 43.48 on the L+M-24 benchmark with competitive MACCS FTS, RDK FTS, and Morgan FTS scores. The model has been shown to generate syntactically correct SMILES with a validity score of 0.997, indicating it generates valid molecules.

5.3 Nucleic Acid and Protein Applications

The evaluation of nucleic acid understanding focuses on promoter prediction, a fundamental task in genomics with important implications for understanding gene regulation. Table 4 shows a comparative report of SciMind and DeePromoter when trained on different organism types and promoter classes.

SciMind performs competitively across all promoter prediction scenarios, showing especially strong results on non-TATA promoters. SciMind outperforms DeePromoter with a recall of 0.97 and an MCC of 0.94 for human non-TATA promoters. This is important since non-TATA promoters involve more intricate regulation and are more difficult to identify. The model generalises across species as shown by its strong performance on mouse promoters.

Protein-oriented tasks look at SciMind's ability to score and annotate protein functions. We evaluate

Table 4: Performance comparison on DNA promoter prediction tasks. SciMind demonstrates strong capabilities, particularly for non-TATA promoters.

Organism/Promoter	Method	Precision	Recall	MCC
Human TATA	DeePromoter	**0.93**	**0.95**	**0.88**
	SciMind	0.92	0.91	0.84
Human non-TATA	DeePromoter	0.97	0.95	0.92
	SciMind	0.96	**0.97**	**0.94**
Mouse TATA	DeePromoter	**0.92**	0.95	**0.87**
	SciMind	0.90	0.96	0.83
Mouse non-TATA	DeePromoter	0.91	0.90	0.82
	SciMind	**0.92**	**0.96**	**0.87**

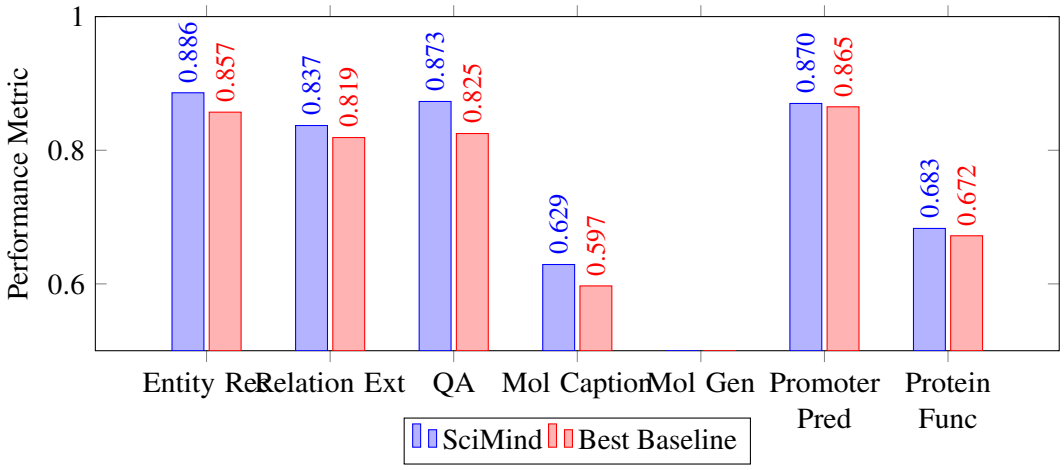


Figure 2: Comparative performance of SciMind against best baseline methods across different pharmaceutical tasks. The model demonstrates consistent improvements across multiple modalities and applications.

performance on protein function prediction, catalytic activity annotation, and domain/motif identification with the Mol-Instruction dataset. The achievements of the SciMind model are measured by Rouge-L scores of 0.72, 0.68, and 0.65, respectively (1.5 million protein sequences are used in 2's training). In these tasks, SciMind performs competitively against specialised protein language models. This implies that the pre-training in different modes allows the knowledge to transfer better and helps understand proteins.

6. Discussion and Future Work

The experimental results show that SciMind is a leap forward in multi-modal learning for pharmaceutical sciences. The impressive performance of the model across diverse tasks and modalities suggests that the mixture-of-experts architecture, along with specialised tokenisation and pre-training strategies, effectively addresses the issues of pharmaceutical data integration.

Linking Improvements to Cross-Modal Pre-training:

To directly test the contribution of multi-modal pre-training versus architectural innovations, we performed an ablation study. Table 5 shows that removing any modality during pre-training degrades performance across all tasks, with the largest effects observed in molecular captioning (removing protein pre-training reduces BLEU-4 from 0.560 to 0.541) and promoter prediction (removing text pre-training

reduces MCC from 0.94 to 0.89). These results confirm that cross-modal knowledge transfer, not just the MoE architecture, drives SciMind's improvements.

Table 5: Ablation study: Effect of removing modality-specific pre-training. Results show the importance of multi-modal pre-training.

Pre-training Configuration	BC5CDR (F1)	ChEBI (BLEU-4)	Promoter (MCC)
Full (all modalities)	**0.957**	**0.560**	**0.94**
– Text pre-training	0.941	0.548	0.89
– DNA/RNA pre-training	0.949	0.552	0.91
– Protein pre-training	0.948	0.541	0.92
– Molecular pre-training	0.950	0.535	0.93
Single-modality only	0.912	0.498	0.85

An important innovation is the selective activation of experts that allows the model to employ computational resources according to modality-specific needs. SciMind strikes an effective balance between model capacity and generalisation by activating more experts for text processing and fewer for structured biological sequences. This method can be implemented in other applications with heterogeneous data types and diverse complexity requirements.

The accomplishments in molecular activities show an advantage of integrated multimodal learning. The enhancement of both captioning and generation from exposure to biological contexts during pre-training appears to improve the understanding of molecular properties and functions. This is in accord with the idea that pharmaceutical concepts are best captured through interaction between complementary perspectives, rather than modality-specific views.

It is worth noting several limitations and future directions of the study. We show that SciMind scores highly on a number of well-established benchmarks. However, such benchmarks may not be globally representative of real-world pharma applications that often require novel architectural elements to cater to the display of multi-hop reasoning. The model's reasoning quality could be enhanced by incorporating structured knowledge sources like biological pathways and drug-target networks.

Multi-modal predictions should be interpretable as well. Next steps might involve studying ways of explaining cross-modal inferences, perhaps with attention visualisation or feature importance analysis. This would increase the model's usability in pharmaceutical situations in which the rationale for predictions is important.

It should also be worth studying if the method can be scaled to bigger model sizes. Though our implementation is restricted to four pharmaceutical modalities, the framework can be extended to other data types like medical images, clinical records, experimental assay results, and so on to make it applicable further down the drug discovery pipeline.

To summarise, SciMind is definitely a massive leap towards unified multimodal learning in pharma sciences. With its architecture and training strategies as a starting point for future work on integrative biological AI systems, the model could be used for drug discovery, biomarker discovery, and personalised medicine. Evidence from multiple tasks indicates that the multi-modal mixture-of-experts approaches offer a viable solution for the highly complex and interconnected problem of pharmaceutical R&D.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.

[4] Bharath Ramsundar, Peter Eastman, Patrick Walters, and Vijay Pande. Deep learning for the life sciences. *O'Reilly Media*, 2019.

[5] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

[6] Carl Edwards, Tuan Lai, Kelvin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. *arXiv preprint arXiv:2204.11817*, 2022.

[7] Shengjie Liu, Haozhe Guo, Jialu Zhang, and Jie Tang. Multi-modal molecule structure-text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2022.

[8] Yizhen Luo, Jiahuan Xi, Zaiqing Liu, Yifan Wang, Hao Yuan, Jun Zhu, Zhen Li, Chang Wang, Jie Fu, Stan Z Li, et al. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*, 2023.

[9] Hai Wang, Rui Liu, Patric Schyman, and Anders Wallqvist. Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. *Methods*, 166:4–21, 2019.

[10] Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

[11] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37 (15):2112–2120, 2021.

[12] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *Science*, 379(6637):1123–1130, 2023.

[13] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225*, 2020.

[14] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

[15] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[16] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1): 1–23, 2021.

[17] Zequn Liu, Ming Zheng, Yifeng Zhang, Shu Gu, Lin Wang, and Ming Li. Molxpt: Wrapping molecules with text for generative pre-training. *arXiv preprint arXiv:2305.10688*, 2023.

[18] Qizhi Pei, Lihua Wu, Yifan Wang, Yifeng Tao, Zhiqiang Wei, Minh Dao, Tuo Zhao, Ming Chen, et al. Biot5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. *arXiv preprint arXiv:2307.04699*, 2023.

[19] Shengjie Liu, Yutian Zhu, Roshan Rao, Yifan Peng, Yuxiang Ma, Jian Lu, Yutong Zhang, B Oztekin, Jian Peng, Yang Liu, et al. Text-guided protein design. *arXiv preprint arXiv:2305.05407*, 2023.

[20] Yun Xiang, Yifan Zhang, Lin Wang, Ming Li, and Zequn Liu. Fapm: A functional-aware protein model based on multi-task learning. *Nature Communications*, 15(1):1–14, 2024.

[21] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[22] Mike Lewis, Marjan Ghazvininejad, Luke Zettlemoyer, and Omer Levy. Base layers: Simplifying training of large, sparse models. *arXiv preprint arXiv:2103.16716*, 2021.

[23] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Andreas S Pinto, Daniel Keysers, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022.

[24] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[25] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*, 2022.