Bi-semantic Chemical Embedder for Joint Representation Learning of SMILES and Natural Language

Anonymous Authors[1]

Abstract

Transformer models have revolutionized natural language processing (NLP), and text-based molecular representations like SMILES have successfully extended these architectures to chemistry. However, domain-adaptive pre-training often causes models to overfit to chemical syntax, catastrophically forgetting their foundational semantic capabilities. To address this challenge, we introduce CheMatE, a chemistry-oriented embedding model that jointly captures molecular structure and domain-specific natural language within the same representation space. Initialized from a ModernBERT backbone, CheMatE learns bi-semantic representations through a two-stage training procedure: continued masked language modeling (MLM) followed by a Matryoshka contrastive learning via Multiple Negative Ranking Loss (MNRL). First, we train the model using MLM on a novel, large-scale corpus of SMILES-annotated, long-context scientific documents that were constructed and curated from FineWeb and ChemPile (comprising 10.4B and 11.5B tokens, respectively). Subsequently, the model undergoes contrastive learning using a synthetic dataset of SMILES-text pairs algorithmically derived from our original training corpus. This design exposes the model to SMILES-enriched scientific literature, enabling bi-semantic understanding. We evaluate CheMatE across a range of downstream tasks covering molecular property prediction and scientific language understanding. Our results demonstrate that coupling our custom-curated datasets with this sequential training strategy yields robust, highly transferable representations. By effectively unifying structural and contextual signals within a single text-based framework, CheMatE achieves competitive performance against both specialized chemistry models and general-purpose language model baselines.

1. Introduction

The rise of machine learning (ML) techniques has led to widespread applications in chemistry problems, supporting chemists in reaction planning, property prediction, optimization, and more generally, in similarity search applications (Coley et al., 2018; Bian & Xie, 2021; Jablonka et al., 2024; Ranković et al., 2025). Groundbreaking contributions in the field range from the earliest neural networks for supervised molecular property prediction using graph representations to transformer-based models with molecular text representations for diverse applications. Central to the success of these computational approaches is the choice of underlying data representation. Molecules can be represented through various modalities, from 2D graphs (Montanari et al., 2020), 3D conformers (Xu et al., 2021), structural fingerprints (Rogers & Hahn, 2010), specific natural language descriptions (Edwards et al., 2021), or even through Simplified Molecular Input Line Entry System (SMILES) (Weininger, 1988). The linearization of molecular structures into SMILES strings unlocked a new spectrum of possibilities and led to the birth of powerful sequence-based modeling techniques (Liu et al., 2017; Gómez-Bombarelli et al., 2018; Schwaller et al., 2018; 2019), leveraging the inherent capabilities of transformer architectures with textual data (Vaswani et al., 2017; Schwaller et al., 2018; Devlin et al., 2019). Additionally, self-supervised transformers have already established a robust paradigm for learning SMILES representations (Wang et al., 2019; Fabian et al., 2020; Irwin et al., 2022). Notably, the RXNFP model successfully maps reaction SMILES into continuous embeddings to capture complex chemical transformations (Schwaller et al., 2020). Similarly, the ChemBERTa family of models has set foundational baselines by pre-training on extensive corpora of molecular strings (Chithrananda et al., 2020; Ahmad et al., 2022).

However, while these models serve as strong baselines for SMILES-centered representation tasks, they suffer from

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Submitted to the AI for Science workshop (ICML 2026). Do not distribute.

two critical limitations. First, models that undergo intensive domain-specific pre-training on these syntactic molecular representations often suffer from domain over-specialization. While they excel at specific featurisation tasks such as property prediction, they struggle to generalize across diverse semantic applications (Chithrananda et al., 2020; Gururangan et al., 2020; Li et al., 2024; Mukhoti et al., 2024). Models pre-trained from scratch on SMILES inherently lack the capability to represent general scientific natural language, while the ones adapted from general language models see this capacity degrade through domain-specific continuous pre-training (Christofidellis et al., 2023; Tan et al., 2025). Secondly, because molecular SMILES strings are inherently concise representations of isolated entities, embedding models dedicated to molecular entities are historically optimized for short inputs (Honda et al., 2019; Wang et al., 2019; Chithrananda et al., 2020; Fabian et al., 2020; Irwin et al., 2022). Consequently, they are architecturally constrained by a shorter token context limit.

We address these limitations through a dual strategy: combining bi-semantic text learning with the integration of long-context documents. By simultaneously mapping SMILES and textual scientific descriptions into a shared semantic space, we avoid the model becoming confined to domain-specific syntax. At the same time, by providing an enlarged context window that handles entire scientific documents rather than isolated strings, the model learns to ground structural chemical tokens within their broader, long-form scientific context. While recent bi-semantic efforts have explored this space for generative applications (Christofidellis et al., 2023; Tan et al., 2025), the development of encoder-centric models dedicated to bi-semantic representation learning remains largely unexplored.

We introduce CheMatE (Chemical Embedder with Matryoshka Embedding), a long-context bi-semantic chemistry encoder built upon a modern transformer architecture (Warner et al., 2024). Our model supports sequences up to 8,192 tokens and was extensively trained on an in-house annotated scientific database. Our training strategy relies on a novel, hybrid dataset specifically designed to intertwine molecular structures with rich textual context. We curated more than **14M long-context documents** (spanning up to 8,192 tokens) primarily consisting of scientific articles and educational materials from FineWeb and ChemPile (Penedo et al., 2024; Mirza et al., 2025) (See Figure 2). These texts were then annotated via an automated SMILES-injection pipeline (See section 2.1), capable of detecting chemical entities present within the text and inserting a chemical SMILES translation at the next position (Lowe et al., 2011; Mavračić et al., 2021; Kim et al., 2025; Landrum et al., 2025). By tightly weaving structural notation into scientific text, we yield a highly contextualized training corpus directly optimized for bi-semantic representation learning.

Ultimately, CheMatE is designed to serve as a versatile foundation for both molecule and language embedding applications. Across our evaluation pipeline for predictive tasks, CheMatE is the only model that ranks among the best-performing group across both modalities simultaneously, consistently generating high-quality embeddings for both SMILES-based molecular tasks and scientific NLP benchmarks, whereas baseline models typically excel in only one domain. We summarize our core methodological and architectural contributions as follows (Figure 1):

(i) Large-scale SMILES injection pipeline. We describe a multi-stage automated pipeline that combines chemical named-entity recognition (NER), structure resolution from public databases, and canonicalization using RDKit (Landrum et al., 2025), applied across ∼14.4 million deduplicated documents drawn from various complementary scientific sources (see Section 2.1). Our pipeline is applied at a large scale, producing a pre-training corpus of 21.9 billion tokens.

(ii) Cost-budget distributed batch sampler. We develop a new batch sampler (*BalancedTokenBatchSampler*) to address the severe length variability across our heterogeneous dataset collection. Rather than enforcing a fixed number of samples per batch, which is not efficient in our case at large-scale with a multi-GPU setup, or ensuring batch homogeneity by sorting the texts by token counts, which would bias the training and reduce batch variety, our sampler enforces three simultaneous constraints for the batch construction: a raw-token budget, a padded-token memory cap, and a quadratic cost ceiling. These constraints are combined with a DDP-aware group-balancing policy that guarantees that all GPU ranks process comparable workloads per synchronization step (see Algorithm 1). This reduces mean padding overhead by up to 50% relative to naive batching and eliminates out-of-memory errors on extremely uneven sequences within a single batch, enabling efficient training at an 8,192-token context length with up to a 30% increase in GPU Utilization across multi-node GH200 hardware.

(iii) Bi-semantic sequential training. We introduce a *bi-semantic* training pipeline in which SMILES are not a separate semantic modality but are woven into natural-language documents at the positions of chemical entity mentions. The standard MLM objective then operates on these mixed token sequences, forcing the model to predict masked SMILES tokens from the textual context and vice versa. We subsequently use a contrastive learning objective to refine the embeddings learned during the first training phase by creating artificial SMILES-annotated text pairs through a similarity-based approach on chemical content. This training pipeline enables a single model to produce consistent representations across heterogeneous chemical and scientific domains.

(iv) Multi-purpose benchmarking. To assess the flexibil-

ity of our model across the two semantics, we construct a unified benchmarking pipeline that covers 48 diverse evaluation datasets, comprising 26 classification and 22 regression tasks. These datasets span a broad range of chemistry-relevant problem settings, including molecular property prediction, materials science, and chemistry-related scientific language understanding drawn from established community benchmarks validated across literature (see Section A.1).

2. Methods

In this section, we introduce the two-stage training pipeline that sequentially integrates masked language modeling (MLM) with a contrastive learning objective. To fuel this bi-semantic training, we processed our entire collection of 14.4 million source documents through our automated SMILES-injection pipeline. This massive source corpus comprises 6.6 million scientific texts (5.9 million abstracts and 0.9 million full-text articles) from ChemPile, as well as 7.8 million high-quality educational documents from FineWeb-EDU.

2.1. Data Filtering & SMILES Annotation Pipeline

2.1.1. Data Filtering

We collected datasets from the ChemPile corpus; some chemistry papers from ChemPile-Papers (11.45B tokens), a curated, open-access collection of chemistry-related scientific articles, as well as some educational texts from ChemPile-Education (75M tokens) (Mirza et al., 2025). Additionally, our main source of educational texts is FineWeb-Edu, a 1.3T tokens subset of the FineWeb web crawl filtered using an educational content classifier trained on Llama3-70B-Instruct annotations (Penedo et al., 2024). Since we needed only chemistry-related content, we filtered the texts from FineWeb-Edu using a custom scoring pipeline: each text is first standardized (lowercase, removal of special characters) and then split into a list of words, and secondly, each text's relevance to chemistry is scored using a word frequency-based approach, using a simple word classifier algorithm. This classification allows us to compute a text chemistry score (TCS) for each text, a metric that measures the text's chemical relevance (Bran et al., 2026). The TCS formulas can be found in Equation (1). The classification is based on the word frequencies found in two manually labeled text corpora: chemistry texts (positive) and non-chemistry texts (negative). The word k frequencies in chemistry texts and non-chemistry texts are written f_k^c and f_k^n respectively:

$$TCS(\text{text}) := \frac{1}{N_{\text{words}}} \sum_{\substack{k=\text{word} \\ \text{in text}}} w_k$$
$$\text{with} \quad w_k = \begin{cases} f_k^c / f_k^n, & \text{if } f_k^c / f_k^n > 1 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

We filtered FineWeb-Edu by retaining texts with a TCS above 1, resulting in a dataset of 10.41B tokens after deduplication.

2.1.2. SMILES Annotation

Each text is then processed through a four-stage SMILES injection pipeline to annotate chemical entities with their canonical SMILES representations. First, Chemical Data Extractor 2 (CDE2) (Swain & Cole, 2016; Mavračić et al., 2021) is used to perform chemistry-aware named entity recognition, detecting all chemical entities in the texts. Second, identified chemical spans are passed into OPSIN (Lowe et al., 2011) for deterministic IUPAC-to-SMILES conversion or standard name conversions. Third, each flagged chemical span that OPSIN was unable to resolve is passed to a PubChem API call via PubChemPy as a fallback lookup to obtain the associated SMILES representations (Kim et al., 2025). Then, all recovered SMILES are canonicalized and validated through RDKit (Landrum et al., 2025); structures that fail valence checking are discarded. Finally, the valid SMILES representations are inserted immediately after the original chemical entity in the source text, yielding our final annotated dataset.

2.2. Bi-semantic training strategies

2.2.1. MLM Strategy

Each text sample consists of natural language containing SMILES strings injected using our annotation pipeline. The number of SMILES per sample is proportional to sequence length, as shown in Figure 2: the longest samples contain more than 200 SMILES within the full 8,192-token context window, providing rich co-occurrence signals among molecular species that appear in similar chemical contexts.

To infuse bi-semantic knowledge, we started by conducting a continuous pre-training of ModernBERT base (Warner et al., 2024) under a standard Masked Language Modeling (MLM) (see equation 2) objective with a 15% masking rate, using Flash Attention 2 (Dao, 2023) and BF16 precision across 16 GH200 GPUs (4 nodes $\times$ 4 GPUs).

$$\mathcal{L}_{\text{MLM}}(\theta) \;=\; -\frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \log P_\theta\big(x_i \;\big|\; \widetilde{X}\big) \tag{2}$$

where $\mathcal{L}_{\text{MLM}}(\theta)$ is the masked language modeling loss for model weights θ, $\mathcal{M}$ denotes the set of masked token positions, and $P_\theta(x_i \mid \widetilde{X})$ is the predicted probability of predicting the original token x_i given the corrupted bidirectional context $\widetilde{X}$ (Devlin et al., 2019).

A key challenge arises from the highly variable sequence lengths in our corpus under a multi-GPU training setup; the computational load is not evenly distributed across GPUs. Standard transformer training typically pads sequences

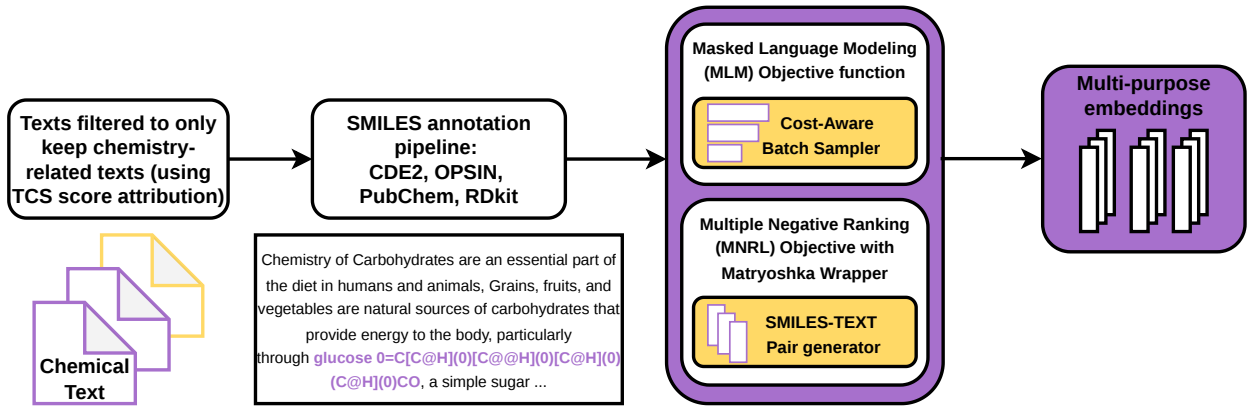

Figure 1. Overview of the multi-task training pipeline for learning bi-semantic chemistry embeddings. Chemical texts are first filtered for domain relevance using TCS score attribution, then processed through a SMILES annotation pipeline (CDE2, OPSIN, PubChem, RDKit) to produce SMILES-injected text corpora. The model is trained sequentially using two objective functions: a Masked Language Modeling (MLM) objective with a Cost-Aware Batch Sampler to improve training efficiency, and a Multiple Negative Ranking (MNRL) objective with a Matryoshka Wrapper with SMILES-TEXT pairs generated from our annotated text. This combined training strategy yields general-purpose embeddings suitable for a range of downstream chemistry tasks. Full training hyperparameters and architectural details are detailed in Section 2.2.1 and 2.2.2.

within each batch, which implicitly homogenizes per-device compute costs. In the case of ModernBERT training, the sequences are unpadded and concatenated to optimize the training efficiency with jagged attention masks (Warner et al., 2024). While effective in settings with homogeneous sequence lengths or limited device parallelism (e.g., single-GPU training), this approach can lead to substantial load imbalance in large-scale multi-GPU setups when sequence lengths vary significantly. In these cases, overall training efficiency decreases because some GPUs remain idle while waiting for the device with the most computationally expensive batch to complete. To address this, we created a custom batch sampler: `BalancedTokenBatchSampler` (Algorithm 1). The objective of this sampler is to construct batches with balanced per-device computational costs, enabling efficient workload distribution across GPUs. The core principles behind the batch construction strategy is that, in ModernBERT training, the computational cost of a batch is approximately proportional to $\sum_i L_i^2$ due to the use of unpadding and jagged attention masks, whereas standard transformer training with padding scales approximately as $\sum_i L_{\max}^2$, where $L_{\max}$ denotes the length of the longest sequence in the batch. Consequently, our sampler aims to construct per-device batches such that the values of $\sum_i L_i^2$ remain balanced across GPUs at each training step.

The details of our custom batch sampler (`BalancedTokenBatchSampler`) are provided in Algorithm 1. Samples are first sorted by token length, split into B bins, shuffled within each bin, and then concatenated back together. Batch stochasticity is preserved as long as $N\_samples/B$ remains sufficiently large. If B becomes too large, the bin assignment becomes increasingly deterministic; however, this is not problematic in our setting, given the scale of the dataset (millions of samples) and the relatively small value of B. After this sample shuffling, per-device batches are formed greedily under three simultaneous constraints: a raw-token budget $T_{\max}$, a padded-token memory cap $T_{\max}^{\text{pad}}$, and a quadratic cost ceiling $C_{\max}$. As a result, the sampler naturally produces variable-length per-device batches containing different numbers of sequences depending on their token lengths and estimated computational cost. The resulting per-device batches are sorted by quadratic cost, grouped into blocks of size W (corresponding to the number of GPUs), and the blocks are shuffled to mitigate ordering bias. This results in a batch construction strategy that assigns per-device batches with comparable computational loads to each training step, ensuring balanced workload distribution across GPUs.

2.2.2. Contrastive Training Strategy

To improve the bi-semantic understanding in our model and further refine the quality of the embeddings, we algorithmically constructed an artificial contrastive dataset directly from our annotated corpora by leveraging the chem-

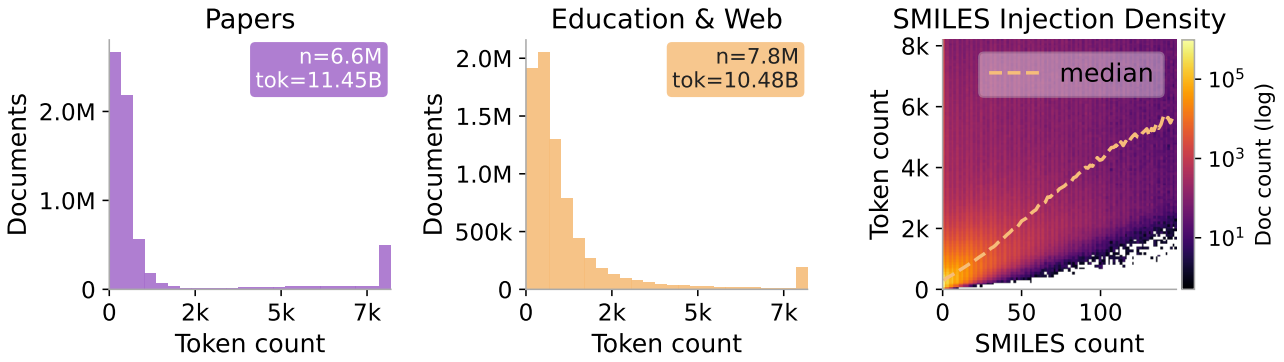

Figure 2. Distribution of sample lengths and SMILES content in the CheMatE MLM training corpus. The left and middle panels show token-count histograms for the Papers subset (6.6M documents, 11.45B tokens) and the Education/Web subset (7.8M documents, 10.48B tokens), respectively, after truncating the texts at 8,192 tokens. The right panel shows a log-scaled 2D histogram of SMILES count versus token count across the combined corpus. The dashed line represents the median token count within each SMILES-count bin.

Algorithm 1 Smart Batching BalancedTokenBatchSampler

Input: token counts $\{l_i\}_{i=1}^{N}$; B bins; $T_{\max}$, $T_{\max}^{\text{pad}}$, $C_{\max}$; world size W, rank r
Output: ordered batch list for rank r
Sort indices by l_i; split into B equal bins; shuffle within each bin; concatenate to obtain $\mathcal{I}$
$\mathcal{B} \leftarrow [\,]$, $b \leftarrow [\,]$
for each index $i \in \mathcal{I}$ **do**
  if $\sum_{j\in b} l_j + l_i > T_{\max}$ **or** $\max(\max_{j\in b} l_j, l_i) \cdot (|b|+1) > T_{\max}^{\text{pad}}$ **or** $\sum_{j\in b} l_j^2 + l_i^2 > C_{\max}$ **then**
    $\mathcal{B} \leftarrow \mathcal{B} \,\|\, [b]$; $b \leftarrow [\,]$
  end if
  $b \leftarrow b \,\|\, [i]$
end for
Sort $\mathcal{B}$ by cost; group into $\lfloor |\mathcal{B}|/W \rfloor$ groups of size W
Shuffle groups; assign group$[\cdot][r]$ to rank r
return rank r's batch list

ical content naturally embedded in each document. The core intuition is that two text passages discussing chemically similar molecules should be close in embedding space, while passages whose molecular content is chemically unrelated should be pushed apart. Rather than relying on document-level labels, we compute this signal at the SMILES level. At each iteration, we draw K documents from our collection of SMILES annotated texts $\mathcal{D}_s$ and chunk them into sentence-level segments to form a candidate pool $\mathcal{C}$. We proceed by sampling a single canonical SMILES s^* present in a random text from $\mathcal{C}$ and use it as the anchor. For every segment $c \in \mathcal{C}$, each inline SMILES it contains is individually compared to s^* via Tanimoto similarity over Morgan fingerprints (radius 2, 2048 bits), and the scores are aggregated (e.g. max) into a single segment-level score $c.score$ (Rogers & Hahn, 2010; Landrum et al., 2025). The top-P segments whose total score exceeds τ^+ become the positive set $\mathcal{C}^+$, and the bottom-Q segments whose score are below τ^- become the negative set $\mathcal{C}^-$. Together with the anchor they form a triple $(s^*, \mathcal{C}^+, \mathcal{C}^-)$, and the procedure is repeated until the contrastive dataset $\mathcal{P}$ contains N such triples. The model is then fine-tuned on these triples or pairs using the Multiple Negative Ranking Loss (see Equation 3) contrastive objective with Matryoshka loss (Kusupati et al., 2024), simultaneously optimizing embedding alignment across sub-dimensions (768, 512, 256, 128, 64) to produce representations that remain informative even when truncated. In practice for CheMatE, we train for a single epoch on a 20k subset of anchor-positive pairs, which we found sufficient to achieve meaningful embedding refinement without overfitting.

The Multiple Negatives Ranking Loss (MNRL) at dimension d_k is a cross-entropy objective, where corresponding texts are pulled together, and non-matching ones are pushed apart (Henderson et al., 2017):

$$\mathcal{L}_{\text{MNRL}}^{(k)} = -\frac{1}{N}\sum_{i=1}^{N}\left[S_{x_i,y_i}^{(k)} - \log\sum_{j=1}^{N} e^{S_{x_i,y_j}^{(k)}}\right] \tag{3}$$

where $S_{x_i,y_j}^{(k)}$ is the similarity score (cosine similarity) computed between the text embeddings x_i and y_i at dimension d_k (truncated embeddings for the Matryoshka loss). This objective encourages each anchor embedding x_i to be more similar to its corresponding positive y_i than to negatives y_j (when $i \neq j$). The final Matryoshka loss is the weighted sum across all dimensions (Kusupati et al., 2024):

$$\mathcal{L}_{\text{Matryoshka}} = \sum_{k=1}^{K} w_k \, \mathcal{L}_{\text{MNRL}}^{(k)} \tag{4}$$

2.2.3. CHEMATE TRAINING DETAILS

Our model is built on the ModernBERT-base architecture (Warner et al., 2024), a long-context encoder-only model pre-trained with a masked language modeling objective. We perform continued MLM pre-training using `answerdotai/ModernBERT-base` native tokenizer, on our chemically-annotated corpus for 3 epochs at LR=5×10^{-4} with a linear scheduler and 1,000 warmup steps, in `bf16` mixed precision. We use AdamW with β_1=0.9, β_2=0.999, ϵ=10^{-8}, weight decay 0.01, and gradient clipping at norm 1.0. The MLM training is run on 16 GPUs with DDP using dynamic batch sizes. We use the final epoch-3 checkpoint as the initialization for contrastive fine-tuning. The resulting MLM checkpoint is then fine-tuned through Matryoshka contrastive learning with a Multiple Negatives Ranking Loss (MNRL) objective on pairs generated by our in-house data generation pipeline, using in-batch negatives and mean token pooling to obtain sequence embeddings during training. The synthetic pair dataset generated was filtered to keep only anchor SMILES of length ≥ 64 characters (19,638 pairs out of an initial 100,000), and the model was trained for 3 epochs at LR=2×10^{-5} with 500 linear warmup steps followed by cosine decay. The contrastive training is run on 4 GPUs with DDP using a per-rank batch size of 16 and gradient accumulation of 2, yielding an effective batch size of 128 (around 15 in-batch negatives per anchor per rank). The released CheMatE checkpoint corresponds to the end of the first contrastive epoch. For evaluation, we employ mean-token pooling to match our training methodology.

3. Results

We evaluate CheMatE against a total of 11 baselines on 48 benchmark datasets (26 classification, 22 regression) drawn from a diverse collection spanning molecular property prediction (see Section A.1), materials science, and chemistry-related NLP tasks (Gurulingappa et al., 2012; Ramakrishnan et al., 2014; Bravo et al., 2015; Baker et al., 2016; Wu et al., 2018; Cohan et al., 2019; He et al., 2019; Jin et al., 2019; Kotonya & Toni, 2020; Huang et al., 2021; Singh et al., 2022; Wognum et al., 2024; Herck et al., 2025; Kasmaee et al., 2025b). For comparison purposes, we select baselines containing both general-purpose language models (ModernBERT (Warner et al., 2024), SciBERT (Beltagy et al., 2019), Nomic-v1.5 (Nussbaum et al., 2025), GTE-base-v1.5 (Zhang et al., 2024), Nomic-MLM, and GTE-MLM (Zhang et al., 2024)) and chemistry-specialized encoder-based models (ChemBERTa-77M-MLM (Chithrananda et al., 2020), MoLFormer-c3-1.1B (Singh et al., 2026), BERT-SMILES (Jouary et al., 2025), and ChEmbed-full (Kasmaee et al., 2025a). Additionally, we include a standard Morgan fingerprint (radius 2, 2048 bits) baseline (Rogers & Hahn, 2010) for SMILES-only tasks to compare our results against established cheminformatics representations (Rogers & Hahn, 2010). To assess the embedding quality of the text provided, we adopt a frozen-embedding evaluation protocol. A linear probe is initialized on top of the fixed embedding, ensuring that the transformer backbones act purely as feature extractors without any task-specific fine-tuning. For classification tasks, we employ a multinomial logistic regression classifier with balanced class weights and report balanced accuracy as our main evaluation metric. For regression tasks, we train a Ridge regressor ($\alpha = 1.0$) and report R^2 as our primary metric.

Figure 3 displays aggregated Critical Difference (CD) diagrams for both SMILES and scientific natural language modalities, pooling classification and regression tasks to maximize statistical power per panel. We use the Friedman omnibus test to reject the null hypothesis of equal model performance, and the Nemenyi post-hoc test (α=0.05) to identify pairwise equivalences (Friedman, 1937; Demšar, 2006). Models with no significant difference are connected by vertical brackets. Across both semantic modalities, CheMatE achieves the lowest mean rank (1.9 on SMILES, 3.2 on NLP) and is the only chemistry-specialized model in the top Nemenyi equivalence group on both panels.

This result demonstrates the strong bi-semantic capabilities of our model without the modality trade-off present in SMILES-only encoders such as MoLFormer and BERT-SMILES. In addition, our mid-stage checkpoint (CheMatE-MLM) ranks significantly lower than our final contrastively tuned model on the SMILES panel (rank 6.6 vs 1.9), indicating that the second sequential training step crucially refines embedding quality. Furthermore, CheMatE outranks Morgan fingerprints as well as chemistry-specialized encoders such as MoLFormer-c3 and ChemBERTa-MLM. On the NLP panel (n=17), CheMatE also achieves the lowest mean rank (3.2) and is not significantly worse than any dedicated scientific language model, supporting that chemical specialization does not come at the cost of natural scientific language representational quality.

However, the true advantage of CheMatE is revealed by the cross-modal asymmetry between the two panels. While some text-trained encoders manage to retain competitive performances on SMILES tasks (serving as CheMatE's closest competitors), every chemistry-specialized SMILES baseline, including MoLFormer-c3-1.1B (11.4), ChemBERTa-MLM (9.8), and BERT-SMILES (9.5), is located at the bottom of the NLP leaderboard. CheMatE is the only model that avoids this semantic trade-off, confirming that our joint contrastive training objective on our in-house annotated data effectively produces a single encoder capable of robust performance across both representations.

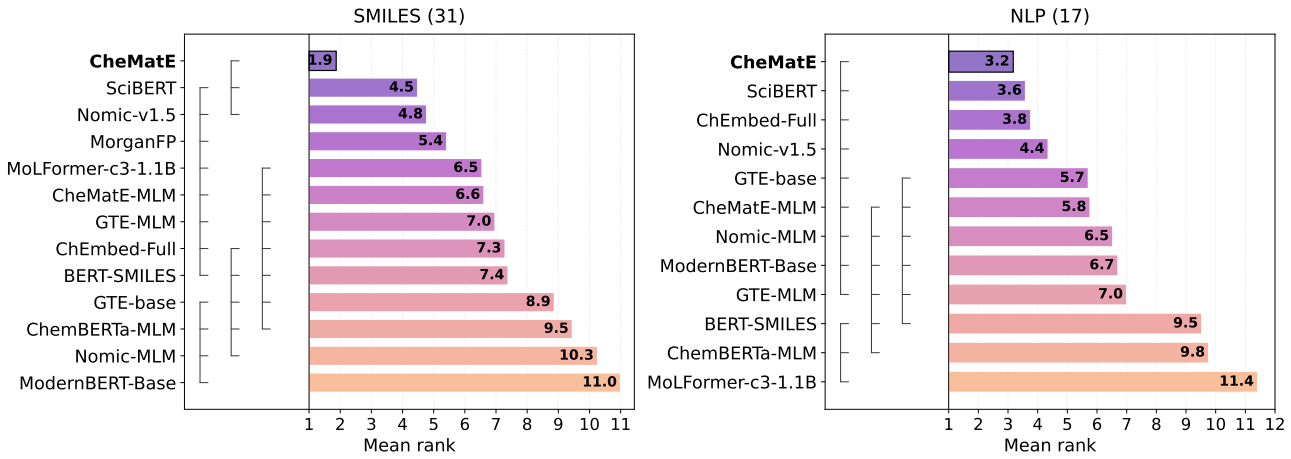

Figure 3. Per-modality Demšar critical-difference (CD) diagrams over 20 cross-validation folds (Friedman omnibus + post-hoc Nemenyi, α=0.05) (Friedman, 1937; Demšar, 2006). Each bar shows a model's mean rank on fold-averaged scores across the given datasets (lower is better, where 1 is the best). Vertical connector bars join models whose pairwise rank differences are not significant under the Nemenyi test. The metrics used for classification and regression are, respectively, balanced accuracy and R^2. Both plots show aggregated results per representation with 13 regression and 18 classification SMILES-based tasks (left panel), and 4 regression with 13 classification datasets for scientific natural language (right panel).

To complement this view, we employ a per-dataset frequency-of-best group analysis following the principles of Ash et al. (2025). For each dataset in our collection, we conduct a one-way ANOVA across all models over the 20 CV folds, and apply a Tukey's honestly significant difference (HSD) post-hoc test on those datasets where the omnibus test rejects equal-means at α=0.05 (Tukey, 1949; Ash et al., 2025). When the test fails to reject the null hypothesis, we conservatively treat all models as tied, counting each as belonging to the best-performing group for that dataset.

Table 1 reports the proportion of datasets where each model ranks in the top tier, which is defined as being statistically indistinguishable from the best-performing model. While the CD diagram illustrates average ordering across datasets, this analysis offers a view of the frequency at which each model is among the dataset's top performers. We design a new evaluation metric, named *Bi-semantic Score %*, to account for dataset modality imbalance, by weighting the two modalities equally.

Across all baselines, CheMatE achieves a bi-semantic score of 86.7% and belongs to the best-performing group on 87.8% of datasets (i.e., the best group in 43/48 datasets), with consistent top-tier performance across modalities: 30/31 on SMILES and 13/17 on NLP. This produces a single encoder that is competitive on chemistry and natural scientific language with little observable specialization trade-off. Comparing CheMatE to its MLM-only counterpart enables isolating the contribution of the contrastive training stage: removing the contrastive supervision reduces the bi-semantic score to 55.5% (27/48 datasets), a 31.2%

Table 1. Frequency of models in the best-performing statistical group across all evaluated datasets. The best-performing group is defined as the top-scoring model and any models statistically indistinguishable from it under a Tukey's HSD post-hoc test (α = 0.05, 20 CV folds). If the initial ANOVA omnibus test lacked significance, all models were considered tied. Results are displayed as raw counts (in-best group/$N_{benchmarks}$) per modality, along with a bi-semantic score (%) built by weighting the two modalities equally. We denote our models with an asterisk (*).

	In-best group		
Model	SMILES	NLP	Bi-sem. %
CheMatE*	30/31	13/17	**86.7**
SciBERT	23/31	13/17	75.4
Nomic-v1.5	23/31	10/17	66.5
ChEmbed-Full	17/31	10/17	56.8
CheMatE-MLM*	18/31	9/17	55.5
MoLFormer-c3-1.1B	24/31	4/17	50.5
GTE-MLM	20/31	6/17	49.9
GTE-base-v1.5	14/31	8/17	46.2
BERT-SMILES	18/31	4/17	40.8
Nomic-BERT-MLM	10/31	8/17	39.7
ModernBERT-Base	11/31	7/17	38.4
ChemBERTa-MLM	10/31	5/17	30.9
Morgan FP (r=2, 2048)	21/31	—	—

performance difference. This ablation indicates that contrastive training over scientific corpora is also a key ingredient driving CheMatE's cross-modality performance.

4. Discussion

We presented CheMatE, a bi-semantic chemistry encoder that jointly represents molecular SMILES and scientific natural language within a unified embedding space. Our model mitigates the trade-off between chemical and natural language specialization that is observed in current encoder models by combining a large-scale SMILES injection pipeline to generate high-quality annotated texts and a two-stage sequential training strategy. CheMatE obtains a bi-semantic score of 86.7% and has the lowest mean rank across both SMILES and natural language modalities when evaluated on 48 benchmark datasets covering molecular property and scientific NLP predictive tasks. Our results demonstrate that strong SMILES and scientific language understanding are not mutually exclusive in encoder-only models. Furthermore, our ablation study shows that the contrastive stage is critical as the MLM-only checkpoint (CheMatE-MLM) ranks significantly lower despite sharing the same backbone.

In addition to the model itself, this work provides an automatic SMILES annotation pipeline for scientific texts and introduces a collection of 14.4 million annotated documents spanning from chemistry papers to educational web content. To train efficiently on this heterogeneous corpus with a multi-GPU setup, we developed the `BalancedTokenBatchSampler`, a cost-aware dynamic batch sampler that enforces simultaneous constraints on token budget, memory, and quadratic computational cost across distributed GPU ranks, reducing padding overhead by up to 50% and enabling stable MLM training at 8,192-token context lengths. We further introduce an automated method for constructing contrastive training pairs directly from annotated corpora by scoring the chemical content of annotated text chunks, thereby reducing the reliance on expensive labeling methodologies. Despite these contributions, several limitations remain and motivate future work. First, our SMILES annotation pipeline depends on a cascade of NER (CDE2) (Mavračić et al., 2021), rule-based parsing (OPSIN) (Lowe et al., 2011), and database lookup (PubChem)(Kim et al., 2025), which can fail on novel compounds, ambiguous names, unusual text formatting, or entities described only by partial structural fragments. Although RDKit canonicalization (Landrum et al., 2025) filters invalid structures, silent annotation errors (e.g., the wrong tautomer or stereoisomer) can propagate into the training corpus and are challenging to quantify at scale. Second, our contrastive stage uses a relatively modest 20k filtered anchor-positive pairs. The trade-off between dataset size, similarity threshold for building positive/negative pairs, and downstream performance has not been systematically explored. Third, while CheMatE supports an 8,192-token context window, most downstream benchmarks considered in this study consist primarily of short SMILES strings or short scientific passages, meaning that the model's long-context capabilities are not directly stressed in our evaluation.

Nevertheless, we believe CheMatE supports the idea that merging molecular notation directly into scientific text is a fruitful route toward general-purpose chemistry representation learning for encoders and that structural and contextual chemical understanding are not mutually exclusive goals. Beyond feature extraction, the contrastive training stage further positions CheMatE as a promising backbone for retrieval applications in mixed-semantic environments. We also hypothesize that bi-semantic encoder representations of this kind could serve as a foundation for future generative chemistry models, where grounding molecular structures in rich scientific contexts through Retrieval-Augmented Generation (RAG) pipelines may prove particularly valuable.

Software and Data

All data and custom code developed for this work are implemented in Python and will be made publicly available in an open-source GitHub repository under the MIT license upon acceptance of this manuscript.

Impact Statement

This work advances representation learning at the interface of chemistry and natural language, with potential applications in drug discovery, materials science, and scientific literature. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ahmad, W., Simon, E., Chithrananda, S., Grand, G., and Ramsundar, B. ChemBERTa-2: Towards Chemical Foundation Models, September 2022. URL http://arxiv.org/abs/2209.01712. arXiv:2209.01712 [cs].

Ash, J. R., Wognum, C., Rodríguez-Pérez, R., Aldeghi, M., Cheng, A. C., Clevert, D.-A., Engkvist, O., Fang, C., Price, D. J., Hughes-Oliver, J. M., and Walters, W. P. Practically Significant Method Comparison Protocols for Machine Learning in Small Molecule Drug Discovery. *Journal of Chemical Information and Modeling*, 65(18): 9398–9411, September 2025. ISSN 1549-9596. doi: 10.1021/acs.jcim.5c01609. URL https://doi.org/10.1021/acs.jcim.5c01609. Publisher: American Chemical Society.

Baker, S., Silins, I., Guo, Y., Ali, I., Högberg, J., Stenius, U., and Korhonen, A. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics*, 32(3):432–440, February 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv585. URL https://doi.org/10.1093/bioinformatics/btv585.

Beltagy, I., Lo, K., and Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text, September 2019. URL http://arxiv.org/abs/1903.10676. arXiv:1903.10676 [cs].

Bian, Y. and Xie, X.-Q. Generative chemistry: drug discovery with deep learning generative models. *Journal of Molecular Modeling*, 27(3):71, March 2021. ISSN 1610-2940, 0948-5023. doi: 10.1007/s00894-021-04674-8. URL http://arxiv.org/abs/2008.09000. arXiv:2008.09000 [q-bio].

Bran, A. M., Xie, T., Pranesh, S., Meng, J., Nguyen, X. V., Goumaz, J., Segura, D. M., Xu, R., Zhou, D., Zhang, W., and Schwaller, P. MiST: Understanding the Role of Mid-Stage Scientific Training in Developing Chemical Reasoning Models. March 2026. URL https://openreview.net/forum?id=363e8WyvLm.

Bravo, , Piñero, J., Queralt-Rosinach, N., Rautschka, M., and Furlong, L. I. Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC Bioinformatics*, 16(1):55, February 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0472-9. URL https://doi.org/10.1186/s12859-015-0472-9.

Chithrananda, S., Grand, G., and Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction, October 2020. URL http://arxiv.org/abs/2010.09885. arXiv:2010.09885 [cs].

Christofidellis, D., Giannone, G., Born, J., Winther, O., Laino, T., and Manica, M. Unifying Molecular and Textual Representations via Multi-task Language Modelling, May 2023. URL http://arxiv.org/abs/2301.12586. arXiv:2301.12586 [cs:LG, cs:cs:CL].

Cohan, A., Ammar, W., van Zuylen, M., and Cady, F. Structural Scaffolds for Citation Intent Classification in Scientific Publications, September 2019. URL http://arxiv.org/abs/1904.01608. arXiv:1904.01608 [cs:CL].

Coley, C. W., Green, W. H., and Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Accounts of Chemical Research*, 51(5):1281–1289, May 2018. ISSN 0001-4842. doi: 10.1021/acs.accounts.8b00087. URL https://doi.org/10.1021/acs.accounts.8b00087. Publisher: American Chemical Society.

Dao, T. FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning, July 2023. URL http://arxiv.org/abs/2307.08691. arXiv:2307.08691 [cs].

Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7(1):1–30, 2006. ISSN 1533-7928. URL http://jmlr.org/papers/v7/demsar06a.html.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs].

Edwards, C., Zhai, C., and Ji, H. Text2Mol: Cross-Modal Molecule Retrieval with Natural Language Queries. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 595–607, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.47. URL https://aclanthology.org/2021.emnlp-main.47/.

Fabian, B., Edlich, T., Gaspar, H., Segler, M., Meyers, J., Fiscato, M., and Ahmed, M. Molecular representation learning with language models and domain-relevant auxiliary tasks, November 2020. URL http://arxiv.org/abs/2011.13230. arXiv:2011.13230 [cs:LG, cs:cs:AI].

Friedman, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *Journal of the American Statistical Association*, 32(200):675–701, December 1937. ISSN 0162-1459. doi: 10.1080/01621459.1937.10503522. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1937.10503522. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1937.10503522.

Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M., and Toldo, L. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of Biomedical Informatics*, 45(5):885–892, October 2012. ISSN 1532-0464. doi: 10.1016/j.jbi.2012.04.008. URL https://www.sciencedirect.com/science/article/pii/S1532046412000615.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks, May 2020. URL http://arxiv.org/abs/2004.10964. arXiv:2004.10964 [cs].

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, 4(2):268–276, February 2018. ISSN 2374-7943. doi: 10.1021/acscentsci.7b00572. URL https://doi.org/10.1021/acscentsci.7b00572. Publisher: American Chemical Society.

He, J., Wang, L., Liu, L., Feng, J., and Wu, H. Long document classification from local word glimpses via recurrent attention learning, 2019.

Henderson, M., Al-Rfou, R., Strope, B., Sung, Y.-h., Lukacs, L., Guo, R., Kumar, S., Miklos, B., and Kurzweil, R. Efficient Natural Language Response Suggestion for Smart Reply, May 2017. URL http://arxiv.org/abs/1705.00652. arXiv:1705.00652 [cs].

Herck, J. V., Victoria Gil, M., Maik Jablonka, K., Abrudan, A., S. Anker, A., Asgari, M., Blaiszik, B., Buffo, A., Choudhury, L., Corminboeuf, C., Daglar, H., Mohammad Elahi, A., T. Foster, I., Garcia, S., Garvin, M., Godin, G., L. Good, L., Gu, J., Hu, N. X., Jin, X., Junkers, T., Keskin, S., J. Knowles, T. P., Laplaza, R., Lessona, M., Majumdar, S., Mashhadimoslem, H., D. McIntosh, R., Mohamad Moosavi, S., Mouriño, B., Nerli, F., Pevida, C., Poudineh, N., Rajabi-Kochi, M., L. Saar, K., Saboor, F. H., Sagharichiha, M., J. Schmidt, K., Shi, J., Simone, E., Svatunek, D., Taddei, M., Tetko, I., Tolnai, D., Vahdatifar, S., Whitmer, J., Florian Wieland, D. C., Willumeit-Römer, R., Züttel, A., and Smit, B. Assessment of fine-tuned large language models for real-world chemistry and material science applications. *Chemical Science*, 16(2):670–684, 2025. doi: 10.1039/D4SC04401K. URL https://pubs.rsc.org/en/content/articlelanding/2025/sc/d4sc04401k. Publisher: Royal Society of Chemistry.

Honda, S., Shi, S., and Ueda, H. R. SMILES Transformer: Pre-trained Molecular Fingerprint for Low Data Drug Discovery, November 2019. URL http://arxiv.org/abs/1911.04738. arXiv:1911.04738 [cs:LG, stat:stat:ML].

Huang, K., Fu, T., Gao, W., Zhao, Y., Roohani, Y. H., Leskovec, J., Coley, C. W., Xiao, C., Sun, J., and Zitnik, M. Therapeutics Data Commons: Machine Learning Datasets and Tasks for Drug Discovery and Development. June 2021. URL https://openreview.net/forum?id=8nvgnORnoWr.

Irwin, R., Dimitriadis, S., He, J., and Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, January 2022. ISSN 2632-2153. doi: 10.1088/2632-2153/ac3ffb. URL https://doi.org/10.1088/2632-2153/ac3ffb. Publisher: IOP Publishing.

Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A., and Smit, B. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169, February 2024. ISSN 2522-5839. doi: 10.1038/s42256-023-00788-1. URL https://www.nature.com/articles/s42256-023-00788-1. Publisher: Nature Publishing Group.

Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., and Lu, X. PubMedQA: A Dataset for Biomedical Research Question Answering, September 2019. URL http://arxiv.org/abs/1909.06146. arXiv:1909.06146 [cs].

Jouary, A., Mata, J. M., Rance, D., Polavieja, G. G. d., Machens, C. K., and Orger, M. Bridging scales between chemical space and behavioral phenotype. July 2025. URL https://openreview.net/forum?id=DJI46O06tF.

Kasmaee, A. S., Khodadad, M., Astaraki, M., Saloot, M. A., Sherck, N., Mahyar, H., and Samiee, S. ChEmbed: Enhancing Chemical Literature Search Through Domain-Specific Text Embeddings, August 2025a. URL http://arxiv.org/abs/2508.01643. arXiv:2508.01643 [cs].

Kasmaee, A. S., Khodadad, M., Saloot, M. A., Sherck, N., Dokas, S., Mahyar, H., and Samiee, S. ChemTEB: Chemical Text Embedding Benchmark, an Overview of Embedding Models Performance & Efficiency on a Specific Domain, January 2025b. URL http://arxiv.org/abs/2412.00532. arXiv:2412.00532 [cs].

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B., Thiessen, P., Yu, B., Zaslavsky, L., Zhang, J., and Bolton, E. PubChem 2025 update. *Nucleic Acids Research*, 53(D1):D1516–D1525, January 2025. ISSN 1362-4962. doi: 10.1093/nar/gkae1059. URL https://doi.org/10.1093/nar/gkae1059.

Kotonya, N. and Toni, F. Explainable Automated Fact-Checking for Public Health Claims, October 2020. URL http://arxiv.org/abs/2010.09926. arXiv:2010.09926 [cs:CL, cs:cs:AI].

Kusupati, A., Bhatt, G., Rege, A., Wallingford, M., Sinha, A., Ramanujan, V., Howard-Snyder, W., Chen, K., Kakade, S., Jain, P., and Farhadi, A. Matryoshka Representation Learning, February 2024. URL http://arxiv.org/abs/2205.13147. arXiv:2205.13147 [cs].

Landrum, G., Tosco, P., Kelley, B., Rodriguez, R., Cosgrove, D., Vianello, R., Gedeck, P., Jones, G., Kawashima, E., Nealschneider, D., et al. rdkit/rdkit: 2025_03_1 (q1 2025) release. *Zenodo*, 2025.

Li, H., Ding, L., Fang, M., and Tao, D. Revisiting Catastrophic Forgetting in Large Language Model Tuning. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4297–4308, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.249. URL https://aclanthology.org/2024.findings-emnlp.249/.

Liu, B., Ramsundar, B., Kawthekar, P., Shi, J., Gomes, J., Nguyen, Q. L., Ho, S., Sloane, J., Wender, P., and Pande, V. Retrosynthetic reaction prediction using neural sequence-to-sequence models, June 2017. URL https://arxiv.org/abs/1706.01643v1.

Lowe, D. M., Corbett, P. T., Murray-Rust, P., and Glen, R. C. Chemical Name to Structure: OPSIN, an Open Source Solution. *Journal of Chemical Information and Modeling*, 51(3):739–753, March 2011. ISSN 1549-9596. doi: 10.1021/ci100384d. URL https://doi.org/10.1021/ci100384d.

Mavračić, J., Court, C. J., Isazawa, T., Elliott, S. R., and Cole, J. M. ChemDataExtractor 2.0: Autopopulated Ontologies for Materials Science. *Journal of Chemical Information and Modeling*, 61(9):4280–4289, September 2021. ISSN 1549-9596. doi: 10.1021/acs.jcim.1c00446. URL https://doi.org/10.1021/acs.jcim.1c00446.

Mirza, A., Alampara, N., Ríos-García, M., Abdelalim, M., Butler, J., Connolly, B., Dogan, T., Nezhurina, M., Şen, B., Tirunagari, S., Worrall, M., Young, A., Schwaller, P., Pieler, M., and Jablonka, K. M. ChemPile: A 250GB Diverse and Curated Dataset for Chemical Foundation Models, May 2025. URL http://arxiv.org/abs/2505.12534. arXiv:2505.12534 [cs].

Montanari, F., Kuhnke, L., Ter Laak, A., and Clevert, D.-A. Modeling Physico-Chemical ADMET Endpoints with Multitask Graph Convolutional Networks. *Molecules*, 25(1):44, January 2020. ISSN 1420-3049. doi: 10.3390/molecules25010044. URL https://www.mdpi.com/1420-3049/25/1/44. Publisher: Multidisciplinary Digital Publishing Institute.

Mukhoti, J., Gal, Y., Torr, P. H. S., and Dokania, P. K. Fine-tuning can cripple your foundation model; preserving features may be the solution, July 2024. URL http://arxiv.org/abs/2308.13320. arXiv:2308.13320 [cs].

Nussbaum, Z., Morris, J. X., Duderstadt, B., and Mulyar, A. Nomic Embed: Training a Reproducible Long Context Text Embedder, February 2025. URL http://arxiv.org/abs/2402.01613. arXiv:2402.01613 [cs].

Penedo, G., Kydlíček, H., allal, L. B., Lozhkov, A., Mitchell, M., Raffel, C., Werra, L. V., and Wolf, T. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale, October 2024. URL http://arxiv.org/abs/2406.17557. arXiv:2406.17557.

Ramakrishnan, R., Dral, P. O., Rupp, M., and von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, August 2014. ISSN 2052-4463. doi: 10.1038/sdata.2014.22. URL https://www.nature.com/articles/sdata201422. Number: 1 Publisher: Nature Publishing Group.

Ranković, B., Griffiths, R.-R., and Schwaller, P. Large language models as uncertainty-calibrated optimizers for experimental discovery, November 2025. URL http://arxiv.org/abs/2504.06265. arXiv:2504.06265 [cs].

Rogers, D. and Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, May 2010. ISSN 1549-9596. doi: 10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t.

Schwaller, P., Gaudin, T., Lányi, D., Bekas, C., and Laino, T. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical Science*, 9(28):6091–6098, July 2018. ISSN 2041-6539. doi: 10.1039/C8SC02339E. URL https://pubs.rsc.org/en/content/articlelanding/2018/sc/c8sc02339e. Publisher: The Royal Society of Chemistry.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science*, 5(9):1572–1583, September 2019. ISSN 2374-7943. doi: 10.1021/acscentsci.9b00576. URL https://doi.org/10.1021/acscentsci.9b00576. Publisher: American Chemical Society.

Schwaller, P., Probst, D., Vaucher, A. C., Nair, V. H., Kreutter, D., Laino, T., and Reymond, J.-L. Mapping the Space of Chemical Reactions using Attention-Based Neural Networks, December 2020. URL https://chemrxiv.org/engage/chemrxiv/article-details/60c753a0bdbb89acf8a3a4b5.

Singh, A., D'Arcy, M., Cohan, A., Downey, D., and Feldman, S. SciRepEval: A Multi-Format Benchmark for Scientific Document Representations, November 2022. URL https://arxiv.org/abs/2211.13308v4.

Singh, R., Barsainyan, A. A., Irfan, R., Amorin, C. J., He, S., Davis, T., Thiagarajan, A., Sankaran, S., Chithrananda, S., Ahmad, W., Jones, D., McLoughlin, K., Kim, H., Bhutani, A., Sathyanarayana, S. V., Viswanathan, V., Allen, J. E., and Ramsundar, B. ChemBERTa-3: an open source training framework for chemical foundation models. *Digital Discovery*, 5(2):662–685, February 2026. ISSN 2635-098X. doi: 10.1039/D5DD00348B. URL https://pubs.rsc.org/en/content/articlelanding/2026/dd/d5dd00348b. Publisher: RSC.

Swain, M. C. and Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *Journal of Chemical Information and Modeling*, 56(10):1894–1904, October 2016. ISSN 1549-9596. doi: 10.1021/acs.jcim.6b00207. URL https://doi.org/10.1021/acs.jcim.6b00207.

Tan, Q., Zhou, D., Xia, P., Liu, W., Ouyang, W., Bai, L., Li, Y., and Fu, T. ChemMLLM: Chemical Multimodal Large Language Model, August 2025. URL http://arxiv.org/abs/2505.16326. arXiv:2505.16326 [cs:LG].

Tukey, J. W. Comparing Individual Means in the Analysis of Variance. *Biometrics*, 5(2):99–114, 1949. ISSN 0006-341X. doi: 10.2307/3001913. URL https://www.jstor.org/stable/3001913.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention Is All You Need, August 2017. URL http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs].

Wang, S., Guo, Y., Wang, Y., Sun, H., and Huang, J. SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '19, pp. 429–436, New York, NY, USA, September 2019. Association for Computing Machinery. ISBN 978-1-4503-6666-3. doi: 10.1145/3307339.3342186. URL https://doi.org/10.1145/3307339.3342186.

Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Cooper, N., Adams, G., Howard, J., and Poli, I. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference, December 2024. URL http://arxiv.org/abs/2412.13663. arXiv:2412.13663 [cs].

Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, February 1988. ISSN 0095-2338. doi: 10.1021/ci00057a005. URL https://doi.org/10.1021/ci00057a005. Publisher: American Chemical Society.

Wognum, C., Ash, J. R., Aldeghi, M., Rodríguez-Pérez, R., Fang, C., Cheng, A. C., Price, D. J., Clevert, D.-A., Engkvist, O., and Walters, W. P. A call for an industry-led initiative to critically assess machine learning for real-world drug discovery. *Nature Machine Intelligence*, 6(10):1120–1121, October 2024. ISSN 2522-5839. doi: 10.1038/s42256-024-00911-w. URL https://www.nature.com/articles/s42256-024-00911-w. Publisher: Nature Publishing Group.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning, October 2018. URL http://arxiv.org/abs/1703.00564. arXiv:1703.00564 [cs].

Xu, M., Luo, S., Bengio, Y., Peng, J., and Tang, J. Learning Neural Generative Dynamics for Molecular Conformation Generation, February 2021. URL https://arxiv.org/abs/2102.10240v3.

Zhang, X., Zhang, Y., Long, D., Xie, W., Dai, Z., Tang, J., Lin, H., Yang, B., Xie, P., Huang, F., Zhang, M., Li, W., and Zhang, M. mGTE: Generalized Long-Context Text Representation and Reranking Models for Multilingual Text Retrieval, October 2024. URL http://arxiv.org/abs/2407.19669. arXiv:2407.19669 [cs].

A. Appendix

A.1. Details on benchmarks collected

Table 2. Full list of the 48 datasets in our evaluation suite, organized by source collection. Each dataset is evaluated under 20-fold cross-validation (StratifiedKFold for classification, KFold for regression).

Collection	Task	Datasets
MoleculeNet (SMILES)	Cls.	BACE, HIV
MoleculeNet (SMILES)	Reg.	ESOL, FreeSolv, Lipophilicity, QM9 (gap), QM9 (zpve)
Polaris TDC (SMILES)	Cls.	Ames Mutagenicity, BBB Penetration, CYP2C9 Inhibition, CYP2D6 Inhibition, CYP3A4 Inhibition, CYP3A4 Substrate, DILI, hERG Cardiotoxicity
Polaris TDC (SMILES)	Reg.	Caco2 Permeability, LD50 (Zhu), Lipophilicity (AZ)
Polaris Certified (SMILES)	Reg.	ASAP ADMET - LogD, ASAP ADMET - MDR1-MDCKII, ASAP ADMET - MLM, Biogen ADME - HLM CLint, Biogen ADME - MDR1-MDCK ER
GPT Challenge (SMILES)	Cls.	Cycloadd. Energy, Densities (Monomer), Synthesability
GPT Challenge (SMILES)	Reg.	Free E. cyclo, MP tryg., Monomer Densities, Monomer Ecoh, Monomer Tg
GPT Challenge (NLP)	Cls.	CO_2 ads. of Bio-Adsorbents, Carbondioxide Adsorption, Gasification Biomass, Thermal Desalination 2
GPT Challenge (NLP)	Reg.	Carbondioxide Adsorption
NLP Scientific	Cls.	ADE Corpus, ArXiv Categories, GAD Gene-Disease, Hallmarks of Cancer, Health Fact, SciCite Citation Intent, SciRepEval Field of Study
NLP Biomedical	Cls.	PubMedQA
NLP Materials	Reg.	AFLOW Band Gap (eV), TextEdge Band Gap (eV), TextEdge Volume (log A^3)
ChemTEB	Cls.	WikipediaChemFields

A.2. Tukey's HSD Test Individual Dataset Performances

The supplementary Tukey's HSD test plots can be found in Figures 4, 5, 6, and 7.

A.3. Ranking results for each semantic modalities and each objective task

The detailed ranking results can be found in Figure 8.

A.4. Baseline model loading

We compare CheMatE against 11 baselines from BERT-style chemistry encoders, BERT-style general-purpose encoders, modern open sentence-embedding models, and a non-transformer molecular featuriser. All transformer baselines are loaded through HuggingFace's `AutoModel` functionality, retain their released weights, and have *no* additional layers added at evaluation time. For each baseline, we apply a single pooling operation to the final hidden states, followed by L2 normalization. Pooling is selected by a per-model rule. When the released model card or repository explicitly mentions a pooling strategy (e.g., contrastive-trained sentence encoders), we use the documented choice. Otherwise, we follow the convention for the model's architecture, applying CLS-token pooling to masked-language-model checkpoints and mean-token pooling to sentence-embedding models. The models used as baselines for MLM are the following: **ModernBERT-base** (Warner et al., 2024) (`answerdotai/ModernBERT-base`) is the long-context encoder we build upon. **ChemBERTa-MLM** (`DeepChem/ChemBERTa-77M-MLM`) is 77M-parameter RoBERTa-style chemistry encoders pre-trained respectively on SMILES MLM (Chithrananda et al., 2020). **BERT-SMILES** (`unikei/bert-base-smiles`) is a BERT-base model pre-trained from scratch on SMILES strings (Jouary et al., 2025). **SciBERT** (`allenai/scibert_scivocab_uncased`) is BERT-base pre-trained on scientific text with a domain-specific vocabulary (Beltagy et al., 2019). **MoLFormer-c3-1.1B** (`DeepChem/MoLFormer-c3-1.1B`) is a 1.1B SMILES encoder with linear attention (Singh et al., 2026). **Nomic-BERT-MLM** (`nomic-ai/nomic-bert-2048`) is the 2048-token Nomic-BERT MLM checkpoint (Nussbaum et al., 2025). **GTE-MLM** (`Alibaba-NLP/gte-en-mlm-base`) is the base-MLM checkpoint released with the GTE family (Zhang et al., 2024). For contrastive models, **Nomic-v1.5**

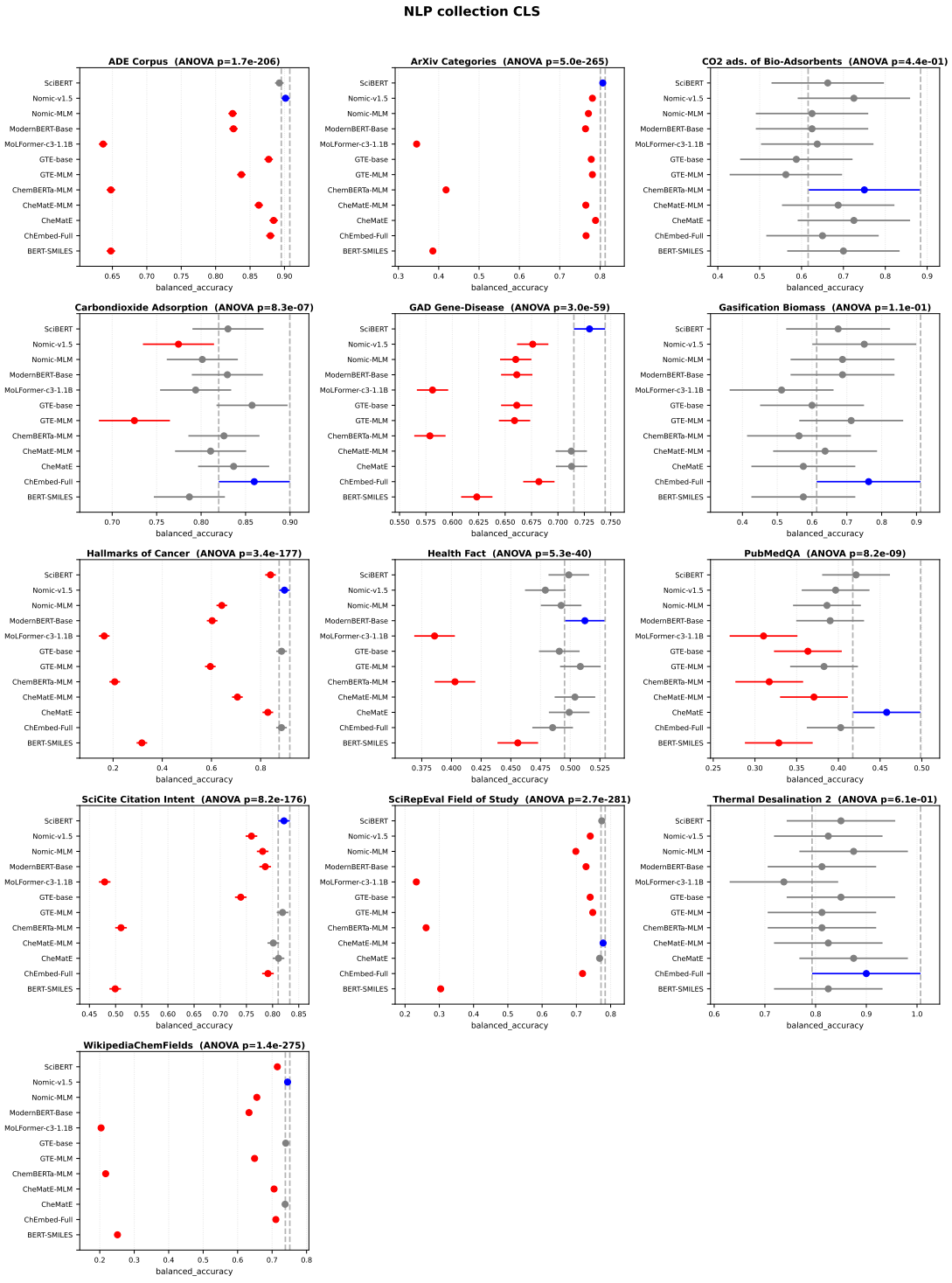

Figure 4. Tukey HSD simultaneous confidence-interval plots for the 13 NLP classification datasets, using balanced accuracy over 20 CV folds. Each panel shows the average performance for each model and its Tukey HSD interval relative to the best mean model for that dataset (Ash et al., 2025). Blue marks the best mean, gray indicates models not statistically significantly different from the best, and red indicates models significantly worse than the best. The sub-plot titles report the one-way ANOVA p-value used to justify the post-hoc comparison.

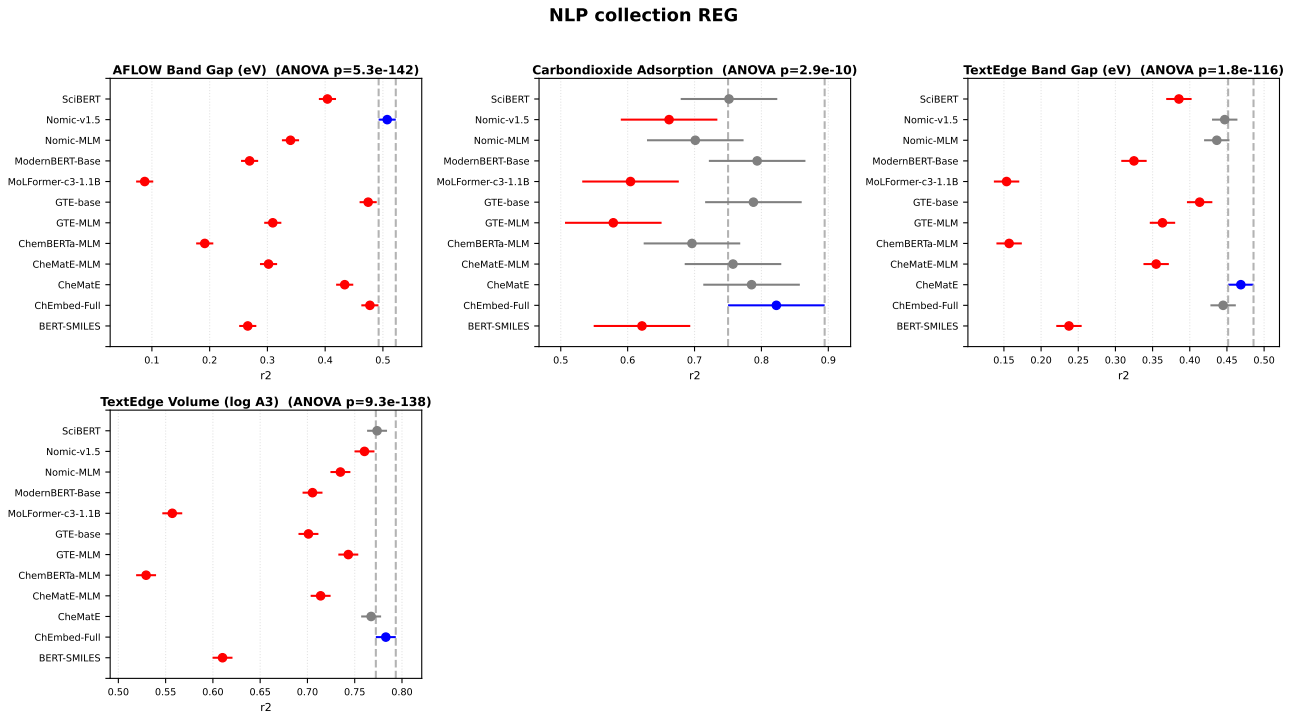

Figure 5. Tukey HSD simultaneous confidence-interval plots for the 4 NLP regression datasets, using balanced accuracy over 20 CV folds. Each panel shows the average performance for each model and its Tukey HSD interval relative to the best mean model for that dataset (Ash et al., 2025). Blue marks the best mean, gray indicates models not statistically significantly different from the best, and red indicates models significantly worse than the best. The sub-plot titles report the one-way ANOVA p-value used to justify the post-hoc comparison.

(`nomic-ai/nomic-embed-text-v1.5`) is a trained through contrastive learning 137M-parameter sentence encoder (Nussbaum et al., 2025). **GTE-base-v1.5** (`Alibaba-NLP/gte-base-en-v1.5`) is the base sentence encoder of the GTE family (Zhang et al., 2024). **ChEmbed-full** (`BASF-AI/ChEmbed-full`) is a chemistry sentence-embedding model (Kasmaee et al., 2025a). **CheMatE-MLM** is our own MLM-only checkpoint (the contrastive ablation, with the same architecture as CheMatE but without contrastive fine-tuning.

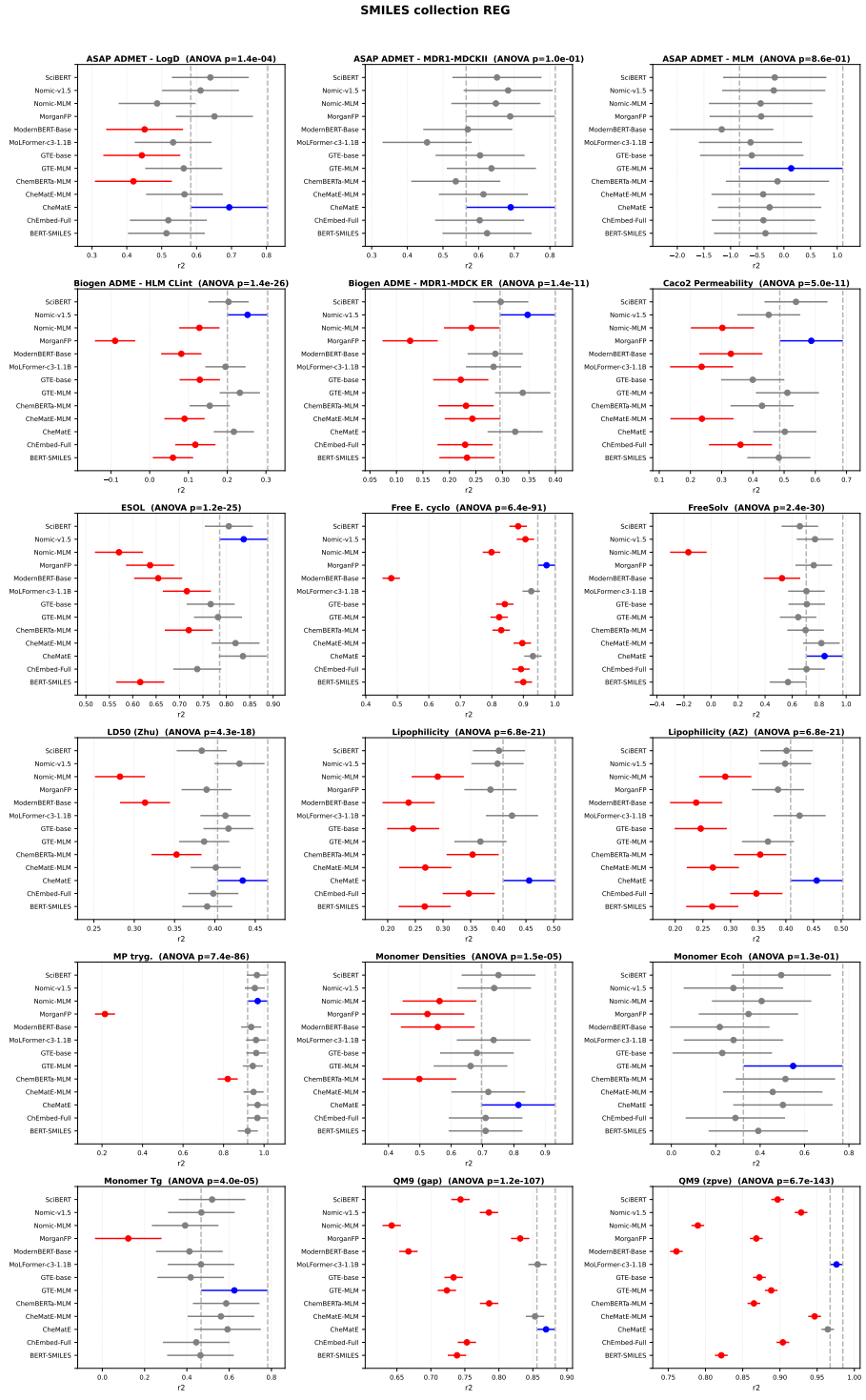

Figure 6. Tukey HSD simultaneous confidence-interval plots for the 18 SMILES regression datasets, using balanced accuracy over 20 CV folds. Each panel shows the average performance for each model and its Tukey HSD interval relative to the best mean model for that dataset (Ash et al., 2025). Blue marks the best mean, gray indicates models not statistically significantly different from the best, and red indicates models significantly worse than the best. The sub-plot titles report the one-way ANOVA p-value used to justify the post-hoc comparison.

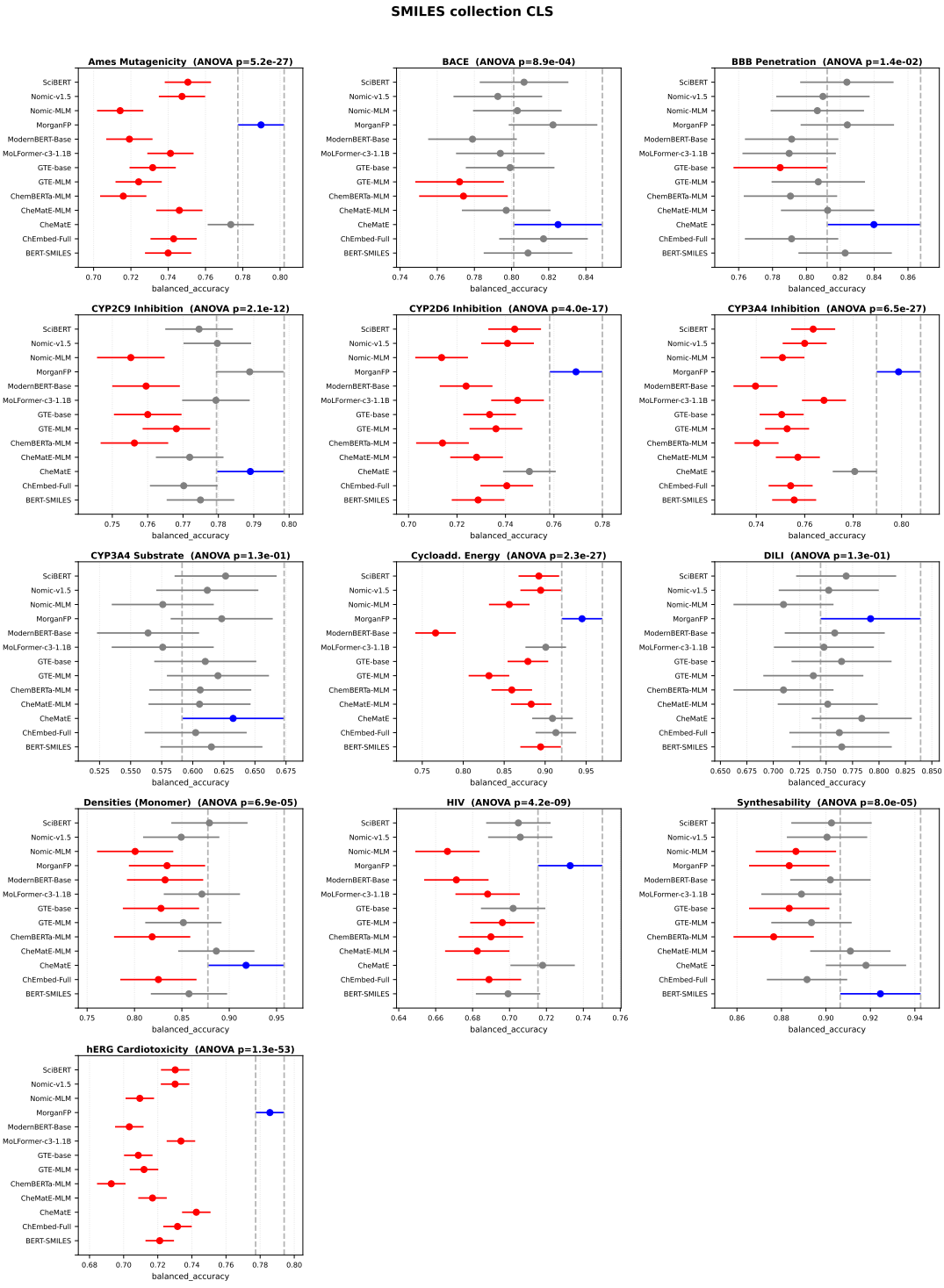

Figure 7. Tukey HSD simultaneous confidence-interval plots for the 13 SMILES classification datasets, using balanced accuracy over 20 CV folds. Each panel shows the average performance for each model and its Tukey HSD interval relative to the best mean model for that dataset (Ash et al., 2025). Blue marks the best mean, gray indicates models not statistically significantly different from the best, and red indicates models significantly worse than the best. The sub-plot titles report the one-way ANOVA p-value used to justify the post-hoc comparison.

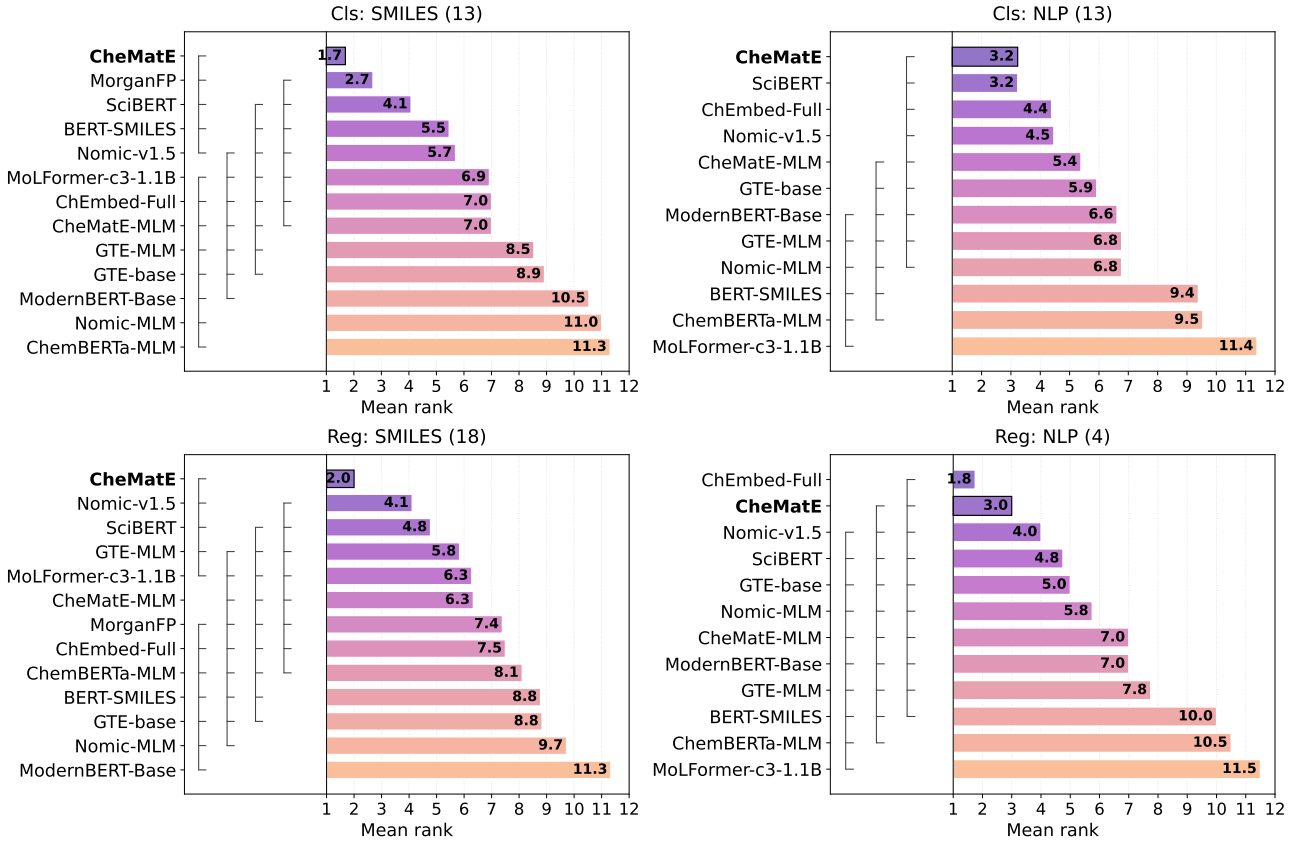

Figure 8. Per-modality Demšar critical-difference (CD) diagrams over 20 Cross-Validation folds (Friedman omnibus + post-hoc Nemenyi, α=0.05). Each bar shows a model's mean rank on fold-averaged scores across the given datasets (lower is better, where 1 is the best). Vertical connector bars join models whose pairwise rank differences are not significant under the Nemenyi test. The metrics used for classification and regression are, respectively, balanced accuracy and R^2. Panel sizes; *Cls: SMILES* 13 datasets, *Cls: NLP* 13 datasets, *Reg: SMILES* 18 datasets, *Reg: NLP* 4: datasets (N=48 in total).