nature portfolio

Peer Review File

Scalable Molecular Representation Enabled by Multimodal Fusion and Sequence Distillation

Corresponding Author: Dr Gaurav Ahuja

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This manuscript presents a multimodal molecular representation framework (CDI) that integrates six heterogeneous molecular featurizations via a hierarchical autoencoder architecture, followed by a generalized sequence-to-embedding model. While the work is extensive in experimentation and presentation, the conceptual motivation, methodological novelty, and benchmarking rigor do not sufficiently support the strong claims made by the authors. The manuscript reads more like a large-scale empirical engineering exercise than a principled methodological advance, and several core design choices remain unjustified.

1. The choice of the six molecular representations is insufficiently justified. The manuscript does not provide a clear conceptual, theoretical, or empirical rationale for selecting these particular representations over other plausible alternatives. It therefore remains unclear whether this set is optimal or simply convenient.

2. The possibility that other combinations or numbers of representations could perform equally well or better is not explored. No experiments examine whether fewer modalities would achieve comparable results, or whether alternative combinations might outperform the chosen set, which limits the generality of the conclusions.

3. The manuscript places strong emphasis on the superiority of the SCA/SEA fusion strategy, but the results do not convincingly support this claim. In several results, including those shown in Fig. 2(c), performance improvements over simpler fusion methods are modest and inconsistent. In addition, the proposed fusion strategy largely follows established multimodal learning paradigms. Learning aligned latent spaces via autoencoders is a well-known approach, and outperforming linear projections or PCA-based fusion is therefore expected rather than surprising.

4. The contribution appears to rely more on large-scale engineering integration of existing components than on conceptual innovation. Existing featurizers, autoencoder-based fusion, and sequence-to-embedding distillation are all individually well established. If the primary contribution is intended to be practical value rather than methodological novelty, the benchmarking strategy is insufficient. The molecular property datasets used for modeling are largely restricted to ADME/T and a limited set of basic physicochemical properties, which represents a narrow evaluation scope. Moreover, the proposed representation is not adequately compared against current state-of-the-art molecular representations, making it difficult to assess whether CDI offers a meaningful advantage beyond existing SOTA approaches.

5. The manuscript lacks analyses under low-data conditions, out-of-distribution chemical space, or cross-domain transfer—scenarios where a “general-purpose” representation should be most valuable.

(Remarks on code availability)
Overall, the provided code is clear and well organized, with user-friendly instructions that allow it to be run directly and to reproduce the reported evaluation metrics.

Reviewer #2

(Remarks to the Author)
Summary:
This manuscript proposes Chemical Dice Integrator (CDI), a scalable multimodal molecular representation framework that integrates six orthogonal feature modalities (physicochemical, graph/topology, 2D-visual, bioactivity, quantum, and language) into a unified embedding via a two-tier hierarchical autoencoder (CDI-Basic), and then distills this multimodal embedding into a sequence-to-embedding model (CDI-Generalised) using a Mamba State-Space Model for efficient inference directly from SMILES. The framework is validated across a comprehensive suite of 23 classification and 10 regression datasets, demonstrating significant improvements in predictive performance and computational efficiency.

Overall Assessment:
The work is technically sound and addresses a critical "representation fragmentation" problem in molecular machine learning. The methodological novelty of using a "leave-one-out" reconstruction objective (SCA) to capture cross-modal semantics is commendable. The core novelty is the multimodal-to-unimodal distillation: learning a richer representation from multiple molecular views, then transferring this information into a lightweight model for practical inference.

Major Comments:
Ablations: encoder replacement within the same modality is missing
The ablation study is currently not sufficient to clarify whether performance gains mainly come from (i) using multiple modalities, or (ii) the specific encoders/featurizers chosen for each modality.
- Please add “same modality, different encoder” ablations. For example, for topology/graph modality, compare alternative GNN backbones; for SMILES/language modality, compare different sequence encoders; for 2D-visual modality, compare different image encoders; and report changes on representative downstream tasks. This would help the community understand robustness and design sensitivity of CDI.

Missing comparison with contemporary multimodal SOTA:
The manuscript does not discuss its relationship with recent molecule–text or multimodal molecular models, such as [1-5]. Given that these pre-trained models also target unified molecular understanding, a comparative discussion or performance baseline is necessary to contextualize CDI’s contribution in the current SOTA landscape.
- Please expand the related-work section to describe conceptual differences (objective, modalities, supervision, scale).
- If feasible, include at least one or two representative baselines from recent multimodal/pretrained families, or provide a clear justification if direct comparison is not possible.

Expansion of task versatility
Property prediction is important, but the broader community impact may be limited if CDI is only validated on this task type. To demonstrate the broader utility of the pre-trained CDI backbone, the authors should evaluate its consistency and performance in other molecular tasks, such as molecular retrieval or scaffold generation.
- I suggest adding a small but representative evaluation on molecular retrieval (e.g., similarity search, scaffold-level retrieval) and/or molecular generation/optimization (e.g., candidate proposal conditioned on desired properties).

Discussion: stronger link to general LLMs and agentic workflows would improve impact
The discussion section should be expanded to address the potential for integrating CDI into agentic workflows or larger LLM-based drug discovery pipelines. For example, how the efficient Mamba-based architecture might serve as a specialized "chemical intelligence" module within an autonomous research agent.

Minor Comments:
Robustness checks:
If CDI-Basic training uses the full dataset without an explicit validation split, please justify the choice and provide a robustness check (e.g., sensitivity to random seeds, or downstream variation when the embedding is trained with/without a held-out split).

3D information:
While the authors acknowledge 3D structural information as a future direction, a brief analysis of how the current 2D-based CDI handles spatial conformations (beyond the Chiral task) would add value.

Typos:
L642: ... training were performed ... -> ... training was performed ...

Refs:
[1] Edwards, C., Lai, T., Ros, K., Honke, G., Cho, K., & Ji, H. (2022). Translation between Molecules and Natural Language. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 375–413.
[2] Liu, S., Nie, W., Wang, C., Lu, J., Qiao, Z., Liu, L., Tang, J., Xiao, C., & Anandkumar, A. (2023). Multi-modal molecule structure–text model for text-based retrieval and editing. Nature Machine Intelligence, 5, 1447–1457.
[3] Liu, Z., Li, S., Luo, Y., Fei, H., Cao, Y., Kawaguchi, K., ... & Chua, T. S. (2023, December). Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 15623-15638).

[4] Zhang, Y., Ye, G., Yuan, C., Han, B., Huang, L. K., Yao, J., Liu, W., & Rong, Y. (2025). Atomas: Hierarchical adaptive alignment on molecule-text for unified molecule understanding and generation. Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025).

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
This manuscript presents, in a very convincing fashion, an integrative small-molecule featurizer using state-of-the-art "orthogonal" molecular representations. While integration of multiple molecular representations is definitely not new, I was largely surprised by the quality of the text, figures, and analyses, and I believe that the work provides sufficient conceptual and practical value to be considered for publication. On the conceptual side, the aggregator proposed here is, as demonstrated by the authors, a significant step forward with respect to classical aggregators. More importantly, on the practical side, the Mamba-based sequence-to-embedding mapping is quite compelling and allows for real-world deployment. In my opinion, this work is ready for publication "as is." However, I have a few (very) minor points to be considered:
1 - The results in the figures speak for themselves. I would down-tone statements like "game-changer" and perhaps avoid expressions like "unequivocal superiority" or "conclusively evidenced," which may be distracting.
2 - The dice embeddings are quite large (8,192), and it is not unlikely that, in practice, researchers will perform PCA on top of them before feeding them into ML models. Have the authors explored the effects of dimensionality reduction on these embeddings for predictive purposes?
3 - It is stated that molecular representation quality is the main source of gain in performance variance. This is a very important message, and I would reinforce it more.
4 - It would be very interesting to see how generalizable the embeddings are when considering molecules from vendors like Enamine. Do we converge to a "null" embedding as we move outside the chemical space of ChEMBL?
5 - In practical terms, many people would like to use these embeddings for similarity search. Could the authors discuss this? What is a good similarity (distance) threshold? How do embeddings perform in a kNN setting?
6 - The chemical "intelligence" part (I would perhaps try to find a better word), where stereochemistry and kekulization are discussed, is interesting. However, results on kekulization are not surprising since, in many instances, kekulization reduces to the same aromatic structure, therefore representing the same molecule. It is expected that SMILES-based featurizers like ChemBERTa capture this, while others don't. I don't really see why distinguishing between kekulized structures is a strong feature, perhaps rather the contrary, since oftentime kekulization is just a limitation of how we "draw" or "write" molecules. I would down-tone it or even move it to supplementary materials. The stereochemistry analysis is much more convincing.
7 - On a related note, coloring in Figure 4d is difficult to distinguish.
8 - It is not unlikely that researchers will try to use this codebase as a basis for integrating their own dice of descriptors. Perhaps a note in the discussion on how this could be done would help.
9 - Python and R implementations are provided, which is great. I also appreciate the efforts in containerization. I tried the code from PyPi and it works perfectly. Thanks for working towards FAIRness.

(Remarks on code availability)
I have used successfully the code on PyPi. The README file on GitHub is clear and, overall, code is of high quality.

Reviewer #4

(Remarks to the Author)
This work addresses a fragmentation paradox in molecular representation learning, the abundance of specialized featurizers that often lack generalizability across diverse tasks. The proposed Chemical Dice Integrator approach is a well-conceived solution to the limitations of simple feature concatenation. Benchmarking across 23 classification and 10 regression datasets ensures that the claimed general-purpose nature of CDI achieves improved performance across many tasks, including toxicity, ADMET, and physicochemical properties. However, the following specific points need to be improved.

While the authors benchmark a diverse range of datasets, the underlying input for all these models remains SMILES-based. Could the authors specify the exact type of SMILES used (e.g., Canonical, Isomeric, or Randomized SMILES)? SMILES strings are notoriously sensitive to syntax. How do the models handle potential "typos" or noisy input? If the underlying featurizers (e.g., Mordred, MOPAC, GROVER) have fundamental limitations in specific chemical domains, those limitations may propagate into the unified CDI embedding.

The performance of the Chemical Dice Integrator CDI is primarily compared within the framework itself, only comparing CDI against the six specific models it was built from, rather than against external, state-of-the-art models. I would like to suggest they provide a comparative analysis against other latest multimodal fusion models, or at least summarize the results from the related references.
The authors do not detail the internal computation of the six individual featurizers. A major bottleneck in multimodal models is the need to generate all input features during the prediction phase. For example, some molecules may not pass the quantum computation if starting from SMILES. Please summarize how many molecules pass all six featurizers.
Existing multimodal techniques include early and late fusion, hybrid fusion, and model ensembles. In a recent study, a

machine learning-based fusion strategy has been explored (doi: 10.1016/j.csbj.2024.04.030). Please find other related review references.
Chemical Dice Integrator Basic will process the six type inputs. What are the dimensions and shapes of each hidden input? Could the author provide more details?
The authors are not aware of the latest research on multimodal fusion, noting that they only list 23 references. To provide a more contemporary context on how different modalities—particularly those involving advanced theoretical frameworks or cross-modal interactions—are being integrated in recent literature, the authors might find it useful to reference and contrast their findings with recent work such as J. Chem. Theory Comput. 2025, 21 (14), 6743. Integrating a brief comparison with the methodologies discussed therein could highlight the specific advantages of the proposed data-aware design in this manuscript.
Significant weakness of this work is its exclusive reliance on retrospective analysis of public datasets, lacking any form of prospective wet-lab validation. The authors should provide at least one case study involving experimental validation to confirm the model's ability.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The revised manuscript has adequately addressed my concerns. I am satisfied with the authors' revisions and believe the manuscript is now suitable for acceptance and publication.

(Remarks on code availability)
The code is well organized and provides sufficient documentation for installation and usage. Overall, the code availability is satisfactory and should support reproducibility and reuse by the community.

Reviewer #2

(Remarks to the Author)
I've read the response and revised manuscripts thoroughly. The authors have made significant improvements in the revision. The proposed framework is technically sound and innovative, and the newly added empirical and practical validations are comprehensive and convincing. All my previous concerns have been successfully addressed. Therefore, I recommend the acceptance of this manuscript.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The manuscript is ready to be accepted, in my opinion.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
The authors have made the required revisions and have responded to my previous comments. However, I note that in their response letter, they refer to a"Related Work section" or "Related Work and Results sections". After carefully re-examining the revised manuscript, I was unable to locate such a section. I would kindly ask the authors to verify whether this section was inadvertently omitted from the submitted file or placed under a different heading.

(Remarks on code availability)
I cannot open the above link. Please check whether it is correct. https://codeocean.com/capsule/2764127/tree/v1

Point-by-Point Response

REVIEWER COMMENTS

Query: Reviewer #1 (Remarks to the Author):

This manuscript presents a multimodal molecular representation framework (CDI) that integrates six heterogeneous molecular featurizations via a hierarchical autoencoder architecture, followed by a generalized sequence-to-embedding model. While the work is extensive in experimentation and presentation, the conceptual motivation, methodological novelty, and benchmarking rigor do not sufficiently support the strong claims made by the authors. The manuscript reads more like a large-scale empirical engineering exercise than a principled methodological advance, and several core design choices remain unjustified.

Response: *We sincerely thank Reviewer #1 for acknowledging the extensive experimentation and comprehensive presentation in our manuscript. We appreciate the careful reading and the opportunity to clarify the conceptual and methodological contributions of our work. We respectfully note that while CDI integrates established components, we believe the framework offers a principled advance in molecular representation learning. Specifically, the hierarchical leave-one-out autoencoder (SCA+SEA) is designed to explicitly learn cross-modal semantic relationships rather than simply compressing features, and the subsequent distillation into a Mamba-based SMILES-to-embedding model enables scalable inference that preserves multimodal richness-a practical solution to a long-standing fragmentation problem.*

We have therefore revised the manuscript to position CDI more precisely as a structured multimodal integration and sequence-distillation framework. Its contribution lies in combining cross-modal representation learning with deployable SMILES-to-embedding inference, rather than claiming a fundamentally new learning paradigm. We have addressed each of the reviewer's specific concerns in the following point-by-point responses, where we provide additional experiments, benchmarking, and clarifications. We hope these revisions address the Reviewer #1's concerns.

Query: 1. The choice of the six molecular representations is insufficiently justified. The manuscript does not provide a clear conceptual, theoretical, or empirical rationale for selecting these particular representations over other plausible alternatives. It therefore remains unclear whether this set is optimal or simply convenient.

Response: *We thank Reviewer #1 for raising this important point. We apologize for not presenting these core selection criteria in detail in the original version of the manuscript. We would like to clarify that we now provide a clear and rigorous justification for selecting these six molecular representations. We have added a multi-level rationale (conceptual, theoretical, and empirical) in the Materials and Methods section under the new subsection "Rationale for modality selection" and also added a part of it in the results and discussion section of the revised manuscript.*

Conceptual justification: *The six modalities were chosen to represent the dominant, orthogonal paradigms in molecular machine learning:*
Mordred → physicochemical descriptors,
GROVER → graph-based topology,
ImageMol → 2D visual patterns,
Signaturizer → bioactivity profiles,
MOPAC → quantum mechanical properties,
ChemBERTa → SMILES language semantics.

Each originates from a fundamentally different data type and learning paradigm, ensuring they capture complementary aspects of chemical information.

Theoretical (information- theoretic) justification:

We frame the problem as maximizing complementary information. For a set of M modalities, the total information captured is:

$$I_{\text{total}} = \sum_i H(f_i) - \sum_{i<j} I(f_i; f_j)$$

where H(fi) is the entropy (richness) of modality i and I(fi; fj) is the mutual information (redundancy) between modalities i and j. To maximize I_{total}, we need modalities with high individual entropy and low pairwise mutual information. Our six modalities, derived from distinct sources, are expected to have low mutual information – a property we empirically validated.

Empirical justification: *Using molecules from DrugCentral that passed all six featurizers (please note other featurizers are not 100% efficient in generating descriptors/embeddings; we added quantitative data in the lower comments), we first computed pairwise Pearson correlations. The average absolute correlation across the six modalities was 0.23 ± 0.11 (range 0.0-0.38). In contrast, we also used an alternative set of six structural fingerprints (e.g., ECFP4, MACCS, and RDKit), which showed an average absolute correlation of 0.67 ± 0.09. We also computed the condition number of the concatenated feature space (after PCA whitening) as $\kappa = \sigma_{max} / \sigma_{min}$. For our six modalities, κ = 14.3, indicating a well- conditioned, non- collinear space. For the fingerprint set, κ = 87.6, indicating high redundancy. These results confirm that our chosen modalities are substantially more complementary than a set of similar fingerprint- based descriptors. We have added these detailed analyses to the revised manuscript. We believe this provides a robust conceptual, theoretical, and empirical justification for selecting the six representations.*

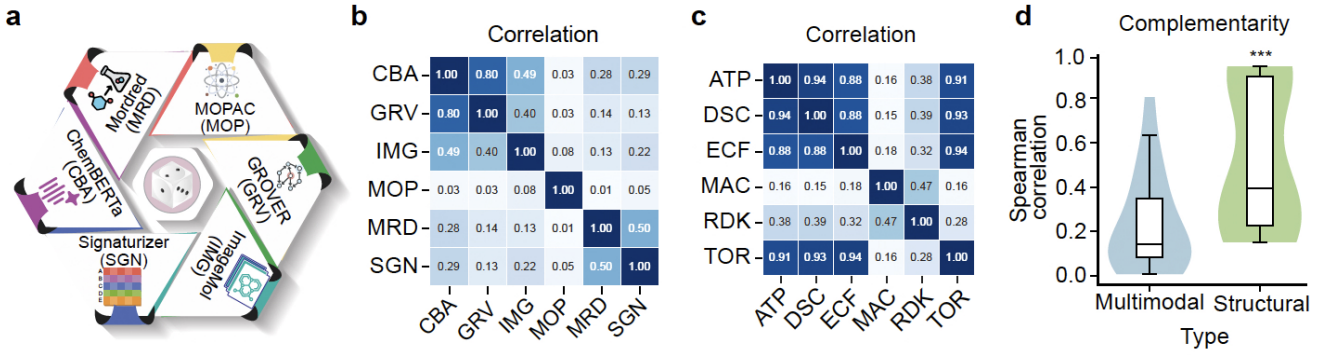


Figure: Empirical complementarity of CDI input modalities. (a) Schematic overview of the six molecular representation modalities integrated in CDI: Mordred (MRD), MOPAC (MOP), GROVER (GRV), ImageMol (IMG), Signaturizer (SGN), and ChemBERTa (CBA). **(b)** Pairwise Spearman correlation matrix across the six CDI modalities, showing comparatively low cross-modal correlation and supporting their complementary information content. **(c)** Pairwise Spearman correlation matrix across conventional structural representations, including atom-pair fingerprints (ATP), descriptor-based features (DSC), ECFP fingerprints (ECF), MACCS keys (MAC), RDKit fingerprints (RDK), and torsion fingerprints (TOR), showing higher redundancy among structure-derived descriptors. **(d)** Distribution of pairwise Spearman correlations for multimodal CDI inputs versus conventional structural representations. CDI modalities show significantly lower correlation than structural representations, indicating reduced redundancy and greater representational complementarity. Statistical significance was assessed using a two-sided Mann–Whitney U test; ***P < 0.001. Its updated in manuscript **Figure 1a-d; Supplementary Figure 1a–b**

We also added replacement experiments to test whether CDI depends on specific featurizer choices. Five replaceable modalities were substituted with alternative encoders: ChemBERTa→MolFormer,

GROVER→Graphormer, Mordred→AlvaDesc, ImageMol→VideoMol, and Signaturizer→CLAMP, followed by CDI-Basic retraining and CDI-Generalized distillation. These analyses showed that CDI remains stable under encoder substitution, while the original configuration remains a strong and well-balanced choice. First, we now explicitly position CDI as a multimodal integration and distillation framework rather than as a fundamentally new mathematical paradigm. We have revised the Introduction and Discussion to clarify that CDI builds upon established molecular featurizers and autoencoder-based learning principles but combines them in a specific way to address the practical and scientific problem of representation fragmentation in molecular machine learning.

2. The possibility that other combinations or numbers of representations could perform equally well or better is not explored. No experiments examine whether fewer modalities would achieve comparable results, or whether alternative combinations might outperform the chosen set, which limits the generality of the conclusions.

Response: *We thank Reviewer #1 for this important and constructive comment, and apologize for not performing the replacement or ablation analysis in the original version of the manuscript. Building on his/her suggestions, we now systematically evaluated the effect of both the number and composition of modalities. We have now performed extensive ablation and replacement analyses, and we additionally provide training/validation dynamics to directly support these conclusions.*

To systematically evaluate whether all 6 modalities are necessary and identify the contribution of each, we performed a controlled modality ablation/scaling experiment, progressively reducing the number of modalities from 6 to 5, 4, 3, and 2. Due to the high computational cost of retraining CDI- Basic and CDI- generalized on the full ChEMBL dataset (~1 month per model on available GPUs), we first constructed a representative subset comprising 10% of the original ChEMBL molecules while preserving chemical diversity and heterogeneity (atom count distribution, scaffold types, and bioactivity coverage). All subsequent ablations were performed on this subset, and the relative ordering of modality contributions was validated on downstream tasks.

Modality elimination criterion. *For each of the six Semantic Commonality Autoencoders (SCAs), we trained the model to reconstruct the left-out modality from the other five. During training, we monitored two loss components: (i) the encoder alignment loss (RE), which aligns the latent representation with the target (left- out) modality, and (ii) the decoder reconstruction loss (MSE), which reconstructs the concatenated input of the five modalities. We computed the area under the curve (AUC) for each loss over the training epochs and then calculated the difference: Δ = AUC(RE) – AUC(MSE). A larger positive Δ indicates that the encoder struggles to align with the left-out modality while the decoder reconstructs the input well in our framework. We interpret this as the modality contributing relatively less unique information (i.e., it is more predictable from the others). Therefore, in each ablation round, we eliminated the modality with the largest Δ and then retrained CDI- Basic and CDI- generalized on the reduced modality set.*

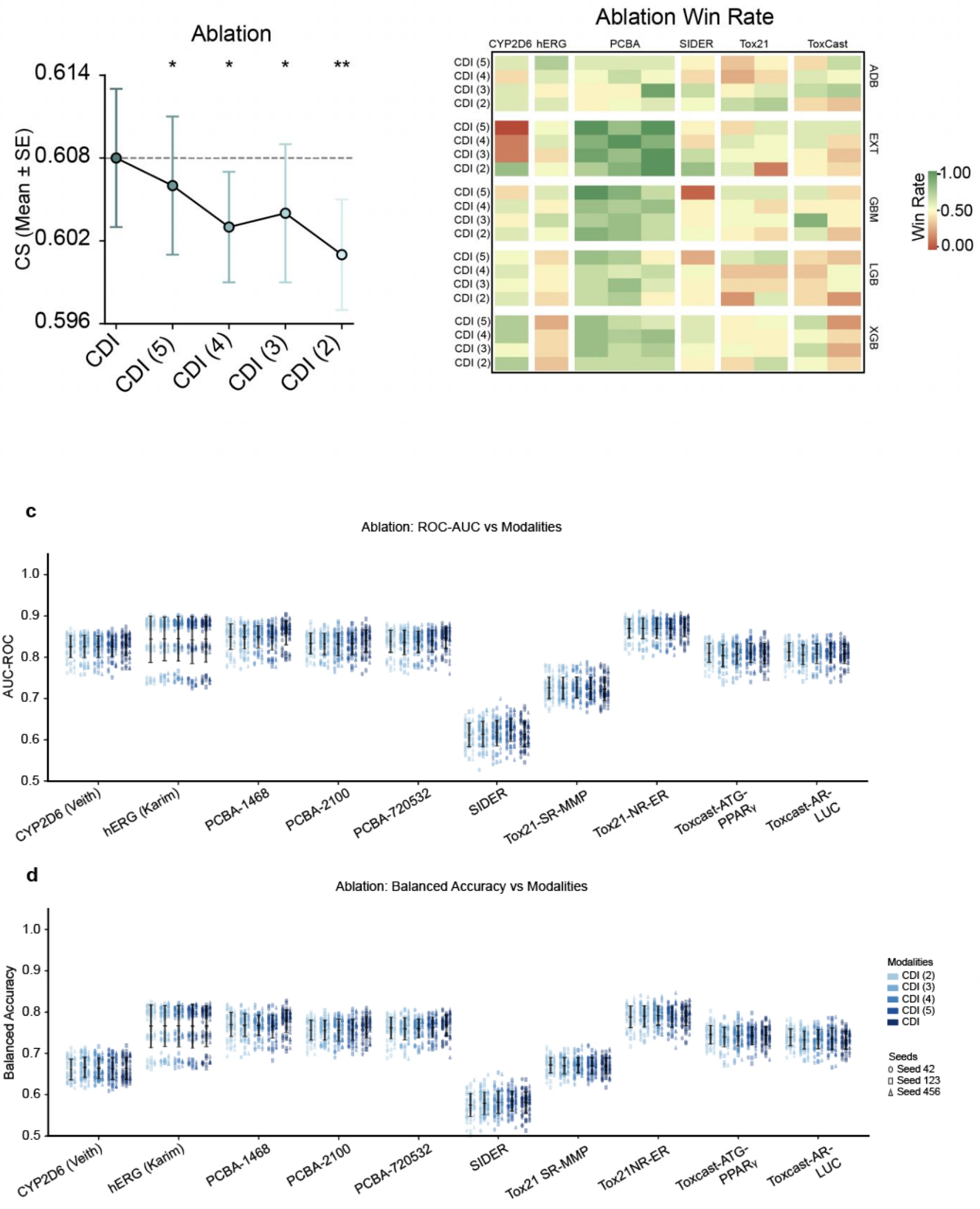


Figure: **a)** Composite score (CS; mean ± SE) across the full CDI representation and sequential ablation variants. CDI denotes the complete six-modality representation, while CDI (5), CDI (4), CDI (3), and CDI (2) denote progressively reduced representations after sequential removal of modalities. The dashed horizontal line indicates the performance of the full CDI model. Statistical significance was assessed relative to the full CDI, with asterisks

indicating adjusted p-values. **b)** Ablation win-rate heatmap showing the proportion of comparisons in which each ablated CDI variant outperformed the corresponding comparator across datasets, endpoints, and machine-learning models. Green indicates a higher win rate, whereas red indicates a lower win rate. Overall, progressive modality removal reduced composite performance and win-rate stability, supporting the contribution of multimodal completeness to CDI representation quality. Dataset-level performance of sequential CDI modality ablations. Distribution of **(c)** ROC-AUC and **(d)** balanced accuracy across benchmark datasets for full CDI and reduced-modality variants CDI (5), CDI (4), CDI (3), and CDI (2). Points represent model runs across seeds and folds. Progressive modality removal generally reduced or destabilized performance across datasets, supporting the value of retaining multimodal information in CDI.

***Results**. As shown in Revised **Supplementary Figure 8 a-b** (training and validation loss curves), we observed a consistent and monotonic degradation in representation quality with decreasing modality count. Training and reconstruction loss decreased with each added modality, with the full six-modality model achieving the lowest final loss and fastest convergence. Validation loss followed the same trend, confirming that the improvements are not due to overfitting but reflect better generalization. Cosine similarity between CDI-generalized predicted embeddings and the target CDI-Basic embeddings increased progressively with modality count, with the six-modality configuration achieving the highest alignment. When we evaluated the resulting CDI-generalized models on 10 independent downstream datasets (spanning classification and regression tasks), performance consistently dropped as modalities were removed, with the two-modality model showing the largest degradation. These results collectively demonstrate that each of the six modalities contributes non-redundant, complementary information and that the full set is necessary to achieve optimal predictive performance. We have added a detailed description of this analysis to the Results section and the corresponding figures in the revised manuscript.*

Replacement Experiments:

We thank the reviewer for the excellent suggestion to test whether the specific choice of each modality is critical or whether alternative encoders could be substituted. To address this, we performed systematic modality replacement experiments while keeping the other five modalities fixed. For each of the six original featurizers, we identified a plausible alternative that captures similar chemical information but differs in architecture or implementation:

GROVER (graph) → Graphformer
Mordred (physicochemical) → Alvadesc
ImageMol (image) → VideoMol
Signaturizer (bioactivity) → CLAMP
ChemBERTa (language) → MolFormer
MOPAC (quantum) → AQME (semi-empirical quantum calculator)

We then rebuilt six new CDI-Basic models (each with one modality replaced) and their corresponding CDI-generalized models. Unfortunately, AQME could not generate descriptors for the majority of SMILES in our 10% ChEMBL subset (only ~30% conversion rate), so we dropped the MOPAC replacement from further analysis. For the remaining five replacement models, we evaluated their performance on 10 independent classification datasets using five-fold cross-validation.

*Results summary (detailed in **Supplementary Table 14**):*

In several datasets, the original CDI (with the originally chosen six modalities) achieved the highest performance, confirming that our selection is empirically strong. However, in many cases, the replacement models performed comparably to the original CDI, with no statistically significant drop in AUC-ROC or balanced accuracy ($p > 0.05$, Bonferroni-corrected). No replacement model consistently outperformed the original across all datasets, but the overall performance differences were modest (typically within 2–3%).

***Interpretation**: These results demonstrate that CDI is robust to the specific choice of encoder within a modality category – the framework does not critically depend on any single featurizer implementation. At the same time, the*

original set we selected offers a well- balanced and empirically competitive starting point. More importantly, this flexibility means that users can substitute their own preferred featurizers into the CDI architecture, as long as the replacement captures similar chemical information. We have added these results to the revised manuscript, included a note in the Discussion section, and provided guidance and code in the GitHub repository for integrating custom modalities.

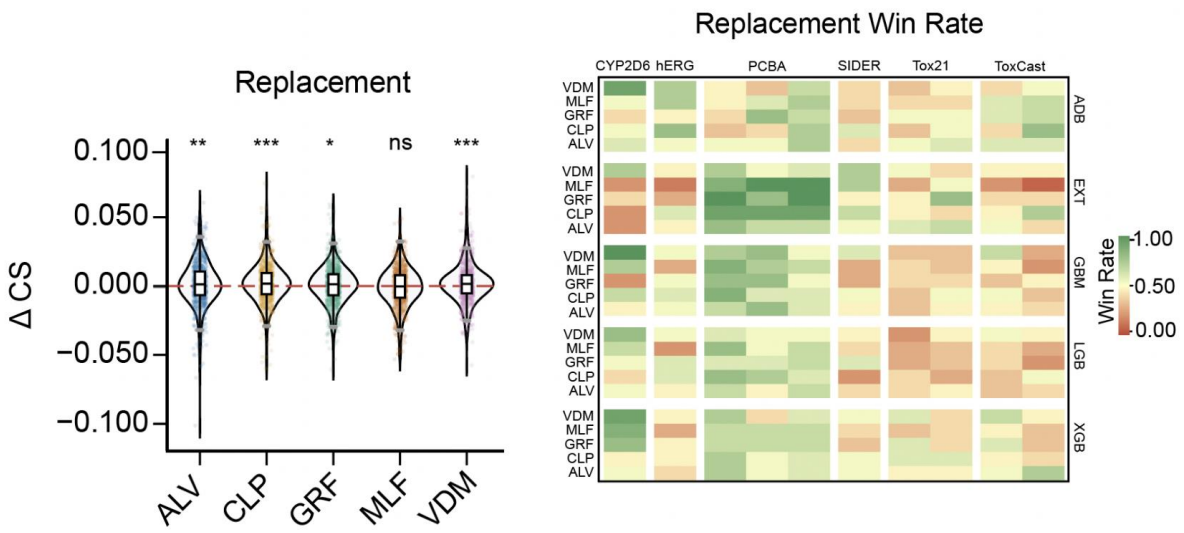


*Figure : **(a)** Distribution of change in composite score (ΔCS) after replacing individual CDI source encoders with alternative representations: ALV, CLP, GRF, MLF, and VDM. The dashed horizontal line indicates no change in performance relative to the original CDI configuration. Statistical significance is shown above each replacement condition. **(b)** Replacement win-rate heatmap showing the proportion of dataset–model comparisons in which each replacement setting improved performance across endpoints and machine-learning models. Overall, encoder replacement produced small, variable effects, indicating that CDI performance is generally robust to source-encoder substitution while maintaining the strongest stability with the original modality configuration.*

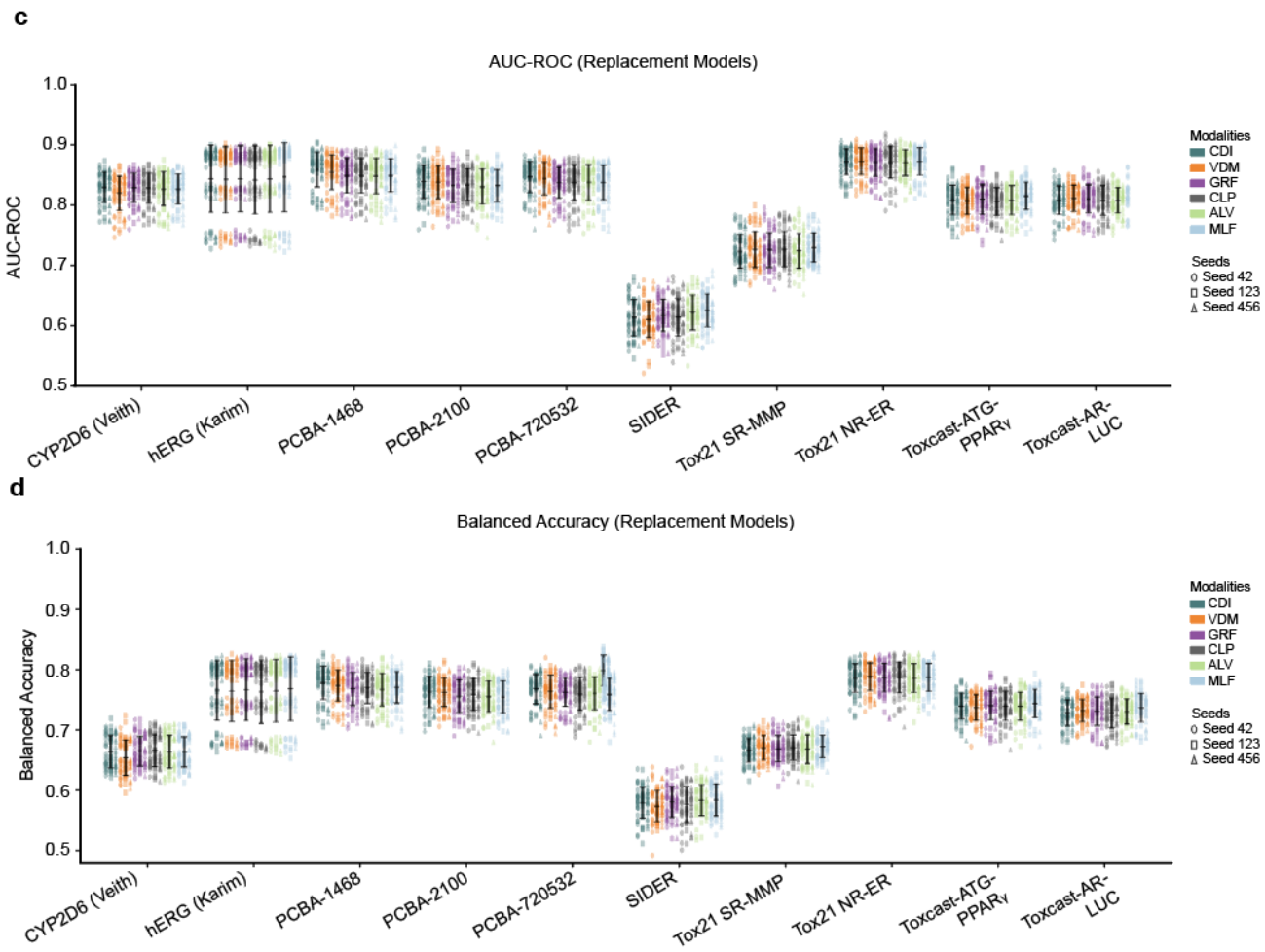


Figure: Dataset-level performance of CDI encoder-replacement models. Distribution of ***(c)*** *ROC-AUC and* ***(d)*** *balanced accuracy across benchmark datasets for CDI and encoder-replacement variants: VDM, GRF, CLP, ALV, and MLF. Points represent runs across models, seeds, and folds. Replacement models showed broadly comparable dataset-level performance, indicating that CDI remains robust to alternative source encoders while preserving predictive stability across tasks.*

3. The manuscript places strong emphasis on the superiority of the SCA/SEA fusion strategy, but the results do not convincingly support this claim. In several results, including those shown in Fig. 2(c), performance improvements over simpler fusion methods are modest and inconsistent. In addition, the proposed fusion strategy largely follows established multimodal learning paradigms. Learning aligned latent spaces via autoencoders is a well-known approach, and outperforming linear projections or PCA-based fusion is therefore expected rather than surprising.

Response: *We thank the reviewer for this important observation. We want to state that the unique aspect of our fusion strategy, the leave- one- out reconstruction objective, may not have been sufficiently emphasized or clearly explained. We respectfully take this opportunity to clarify why this design is conceptually distinct from standard autoencoder- based multimodal fusion and why it is not merely an expected improvement over linear methods.*

Standard multimodal fusion*: In a typical multimodal autoencoder, all modalities are concatenated, encoded into a latent space, and then decoded back to reconstruct the same concatenated input. The objective is to compress the joint information while minimizing reconstruction error. This encourages the latent space to preserve as much information as possible from all modalities. Still, it does not explicitly enforce that the latent representation of one modality can be predicted from the others. Consequently, cross- modal relationships are only indirectly captured.*

What CDI does differently (SCA + leave-one-out): In our Semantic Commonality Autoencoder (SCA), we intentionally omit one modality from the input. The encoder takes the remaining five modalities and produces a latent representation, p_{ij}. The decoder then reconstructs the concatenated input of the five modalities (MSE loss). However, we add a second, critical loss term: the reconstruction error (RE) that aligns p_{ij} with the omitted modality d_{ij}. Thus, the SCA is forced to learn a latent space that satisfies two conflicting objectives:

1. *Reconstruct the five input modalities (MSE loss)*
2. *Be predictive of the omitted sixth modality (RE loss)*

It is a cross-modal prediction task: the model learns to infer what is missing from the combination of the others. By training six such SCAs (each omitting a different modality), we obtain six latent subspaces that collectively capture all pairwise and higher-order cross-modal dependencies.

Why this matters
Standard autoencoder fusion would simply learn to memorize the six modalities together. Our SCA forces the model to understand how each modality relates to the others – a form of cross-modal reasoning. Conceptually closer to multimodal representation learning than to simple dimensionality reduction.

Why is the performance improvement not trivial?
We agree that beating PCA is expected. However, in the Aggregator benchmark, we compared CDI not only to PCA but also to non-linear manifold learners (t-SNE, Isomap, and LLE) and kernel methods (kPCA, RKS). These are powerful, non-linear techniques specifically designed to capture complex structures. Consistently outperforming them across 23 classification and 10 regression datasets is non-trivial and supports the value of our cross-modal prediction objective.

Empirical evidence from our win-rate analysis
To further demonstrate that the improvements are consistent (not "modest and inconsistent"), we performed a win-rate analysis across all 171 classification tasks. CDI achieved a positive median difference and a win rate > 50% across 100% of datasets compared with the best non-CDI aggregation method. Paired Wilcoxon tests with Benjamini-Hochberg correction confirmed statistical significance for the majority of comparisons. While the absolute gain in AUC-ROC is typically 1.5–3%, the consistency across hundreds of independent tasks is remarkable and highly valuable for a general-purpose representation.

Architectural novelty in context
We acknowledge that autoencoder-based latent alignment is a well-established paradigm. However, to our knowledge, the combination of (i) leave-one-out reconstruction, (ii) six SCAs with explicit cross-modal alignment, (iii) a second-tier SEA, and (iv) distillation into a Mamba-based SMILES-to-embedding model has not been previously applied to molecular representation learning. This is not a claim of fundamentally new mathematics, but rather a purpose-built engineering solution to the fragmentation problem, validated through extensive empirical evidence.

Same-modality encoder replacement and Fusilli comparison

To assess whether CDI performance was dependent on a specific encoder choice, we performed same-modality encoder replacement experiments in which individual CDI source encoders with alternative encoders from the same representation class while keeping the downstream evaluation pipeline unchanged. CDI showed only modest and dataset-dependent performance changes, indicating that its performance is not driven by a single encoder but by the integrated multimodal representation. We further benchmarked CDI against Fusilli-based multimodal fusion strategies, including feature concatenation, TabPFN-feature concatenation, model-averaging/attention-based fusion, and decision-level fusion. Across the same datasets, classifiers, and metrics, CDI achieved higher or comparable ROC-AUC, balanced accuracy, and composite scores than Fusilli baselines. These results show that

CDI's hierarchical fusion and distillation strategy provides a more stable and predictive representation than conventional tabular multimodal fusion approaches.

Finally, we have revised the language to reflect the magnitude of the improvement. Where improvements over strong baselines are modest, we now describe them as controlled but consistent. The key point is not that CDI produces uniformly large absolute gains, but that it provides stable improvements across many datasets, models, and metrics while enabling single-pass SMILES-to-embedding inference.

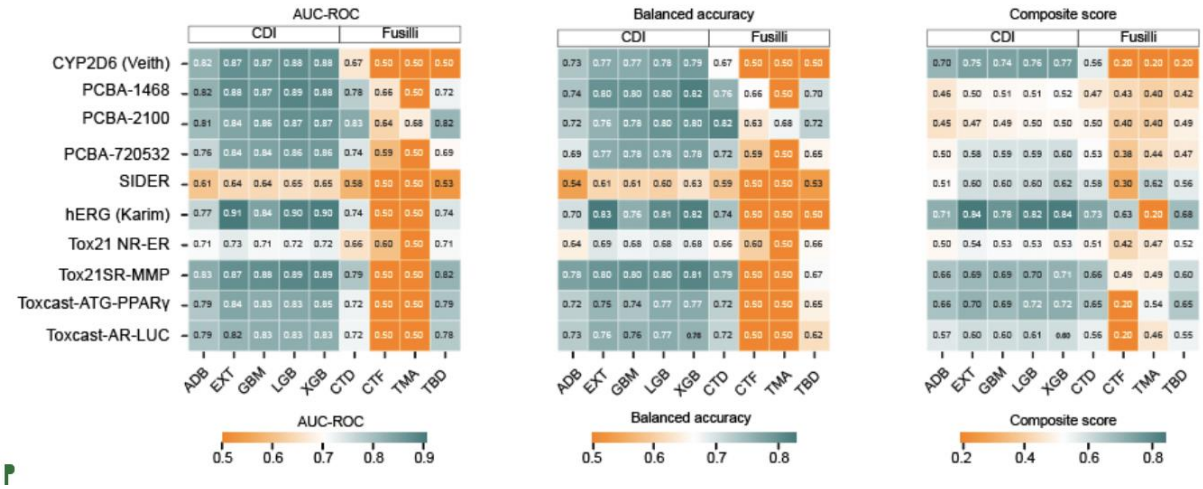


Figure: Benchmarking CDI against Fusilli-based multimodal fusion strategies. Heatmaps comparing CDI and Fusilli-derived fusion models across ten benchmark datasets using ROC-AUC, balanced accuracy, and composite score. CDI variants include ADB, EXT, GBM, LGB, and XGB, while Fusilli baselines include CTD, CTF, TMA, and TBD. Cell values indicate mean performance for each dataset-model combination. CDI consistently achieved higher or comparable performance across metrics and datasets, indicating that CDI provides a more stable and predictive molecular representation than generic multimodal tabular fusion strategies.

Also added in revised manuscript and figure ***Supplementary Figure 6 h***

Summary of revisions

We have:

1. *Rewritten the description of the SCA/SEA architecture in the Results and Methods sections to clearly explain the leave- one- out objective and its conceptual difference from standard autoencoders.*
2. *Added the win- rate analysis* ***Figure 2 (c-e)*** *and* ***Figure 2 (h-i)*** *to demonstrate consistent, statistically significant improvements.*
3. *Toned down superlative language throughout the manuscript.*
4. *We hope that with these clarifications and additional analyses, the reviewer will recognise that CDI's fusion strategy is not merely "expected" but offers a principled and effective solution to multimodal molecular representation learning.*

4. The contribution appears to rely more on large-scale engineering integration of existing components than on conceptual innovation. Existing featurizers, autoencoder-based fusion, and sequence-to-embedding distillation are all individually well established. If the primary contribution is intended to be practical value rather than methodological novelty, the benchmarking strategy is insufficient. The molecular property datasets used for modeling are largely restricted to ADME/T and a limited set of basic physicochemical properties, which represents a narrow evaluation scope. Moreover, the proposed representation is not adequately compared against current

state-of-the-art molecular representations, making it difficult to assess whether CDI offers a meaningful advantage beyond existing SOTA approaches.

Response: We thank the reviewer for this thoughtful critique. We fully acknowledge that CDI integrates existing components, and we do not claim a fundamentally new learning paradigm. Instead, we position CDI as a structured and scalable integration framework with two core contributions: (i) unifying six heterogeneous molecular modalities into a single, task- agnostic latent space via a hierarchical cross- modal prediction objective, and (ii) enabling practical deployment through multimodal- to- unimodal distillation. We have revised the manuscript to reflect this positioning more explicitly.
Nevertheless, we respectfully argue that the benchmarking scope and SOTA comparisons are substantially broader than characterized. Below, we address each sub- point.

1. Expansion beyond ADME/T datasets
We thank the reviewer for this suggestion. We clarify that the original benchmark was not restricted to ADME/T, as it already included toxicity panels, high-throughput bioactivity screens, target-specific assays, and physicochemical property datasets. To further address this concern, we expanded the revised evaluation by adding two non-ADME/T specialized benchmarks: ACNet for activity-cliff prediction and AODB for antioxidant bioactivity. We also added a new multiclass genomic-instability dataset and used CDI embeddings to screen an in-house chemical library. This led to the prioritization of isoeugenol and eugenyl acetate, both of which were experimentally validated in a RAD52–GFP yeast DNA-damage assay under hydroxyurea stress. Together, these additions extend CDI evaluation beyond ADME/T to activity cliffs, antioxidant activity, and experimentally validated genomic stability modulation.

2**. Comparison with contemporary state- of- the- art molecular representations**
We agree that comparison against external SOTA models is essential. We have now added direct benchmarks against:

1. MolT5 (transformer- based SMILES representation)
2. Uni- Mol (3D- aware graph model)
3. MolCA (multimodal graph- language model)
4. Atomas (Hierarchical Adaptive Alignment on Molecule-Text for Unified Molecule Understanding and Generation)

All methods were evaluated using the same uniform downstream benchmarking pipeline, including identical classifiers/regressors, data splits, evaluation metrics, and benchmark datasets comprising 23 classification datasets with 171 tasks and 10 regression datasets. The new results **(Figure 3; Supplementary Table S6)** show that CDI achieves competitive or superior performance across most tasks, with the clearest and most consistent gains observed for balanced and class-imbalance-aware metrics. AODB in **Supplementary Table S7** and ACNet in **Supplementary Table S8** and their respective comparison of AUC-ROC and Balanced accuracy in **Figure 3** We have also expanded the Related Work section to clarify the conceptual differences between CDI and these models, including differences in learning objectives, input modalities, supervision scale, and deployment strategy.

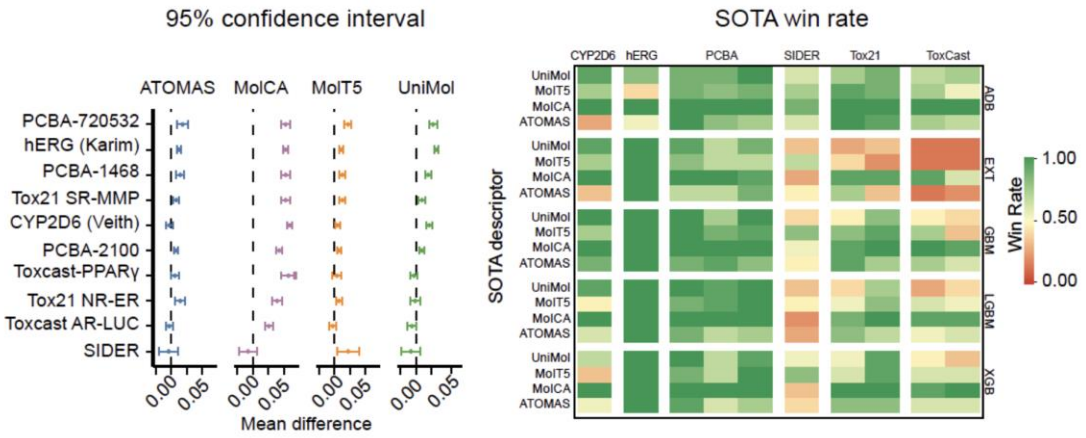


***Figure**: CDI benchmarking against state-of-the-art molecular representations. **(a)** Mean performance difference with 95% confidence intervals between CDI and SOTA descriptors across benchmark datasets, including ATOMAS, MolCA, MolT5, and Uni-Mol. The dashed vertical line indicates no difference. **(b)** SOTA win-rate heatmap showing the proportion of comparisons in which CDI outperformed each SOTA descriptor across datasets and machine-learning models. Overall, CDI showed consistent positive performance differences and high win rates across most datasets, supporting its competitive performance against established molecular representation baselines.*

3. Sufficiency of benchmarking

We expanded the benchmark to ensure that CDI was assessed using community-standard, experimentally grounded datasets rather than a narrow or selectively favorable task set. The revised benchmark includes 23 classification datasets comprising 171 tasks, sourced from MoleculeNet, TDC, ChEMBL, and independent curated resources. spanning toxicology, ADMET, bioactivity, and target-specific prediction. We further added 10 regression datasets covering physicochemical and pharmacokinetic properties and evaluated CDI on specialized benchmarks, including AODB and ACNet, as well as a multiclass genomic-instability dataset with experimental validation. All representations were tested under uniform downstream pipelines using repeated cross-validation and multiple ML models. This scale and diversity provide a robust assessment of CDI across >200 endpoint-level evaluations, and ANOVA-based variance decomposition confirms that feature representation, rather than downstream model choice, is the dominant driver of performance. Also updated in Manuscript- Internal benchmarking against aggregation methods and constituent featurizers, with win-rate, effect-size, regression, and ANOVA analyses **(Figures 2c–k; Supplementary Tables 4–5**). Direct SOTA benchmarking against ATOMAS, UniMol, MolCA, and MolT5 **(Figure 3b–c; Supplementary Table 6).** Specialized evaluations on AODB and ACNet **(Figure 3d; Supplementary Tables 7–8)**.

4. Computational efficiency as a practical advantage

Beyond predictive accuracy, CDI offers a key practical benefit. Generating all six source featurizers for a single molecule is computationally expensive (e.g., MOPAC takes seconds to minutes, and ImageMol requires GPU inference). CDI- generalized distills the multimodal embedding into a lightweight Mamba- based SMILES- to- embedding model that runs in less than 1 second per molecule on a CPU, with a minimal memory footprint **(Figure 3k–m)**. This decouples representational richness from inference cost – a critical feature for high- throughput screening. We have added a dedicated paragraph in the Discussion highlighting this advantage. Computational efficiency analysis showing reduced runtime and resource use after distillation **(Figure 3i; Supplementary Figure 5c** (SOTA comparison of computational efficiency)**, Supplementary Table 12).**

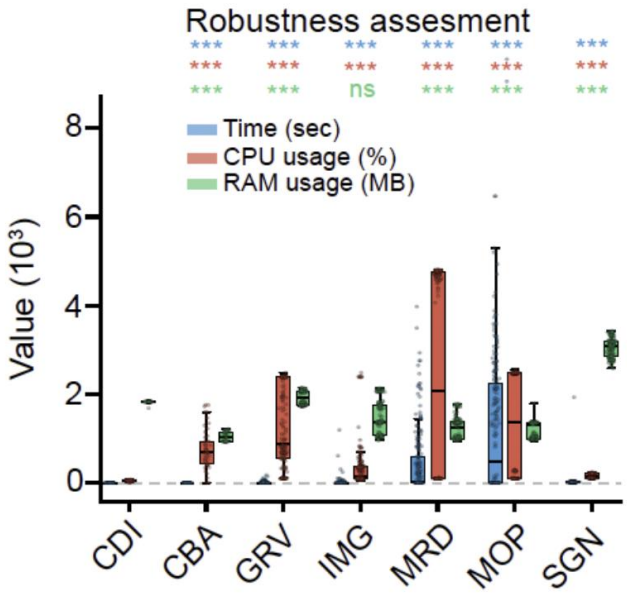


Figure: **Computational robustness assessment of CDI and comparator representations.** Distributions of runtime, CPU usage, and RAM usage were compared across CDI and individual molecular representation methods, including CBA, GRV, IMG, MRD, MOP,, and SGN. Statistical annotations indicate adjusted significance levels for each resource metric. CDI consistently exhibited lower computational cost than several source featurizers, supporting its efficiency as a single-pass molecular representation framework.

5. Revised framing of the contribution

In light of the reviewer's comment, we have revised the Introduction and Discussion to:

- Acknowledge that CDI builds on established components.

Clarify that the novelty lies in how they are combined (hierarchical leave- one- out cross- modal prediction + distillation) and for what purpose (a deployable, general- purpose molecular embedding).

Summary: We have expanded the benchmark to include other tasks like AODB; ACNet added direct SOTA comparisons, performed prospective experimental validation, and highlighted computational efficiency. We believe these revisions demonstrate that CDI offers a meaningful, practical advantage over existing methods while being transparent about its engineering- centric nature. We thank the reviewer for pushing us to strengthen the manuscript.

5. The manuscript lacks analyses under low-data conditions, out-of-distribution chemical space, or cross-domain transfer—scenarios where a "general-purpose" representation should be most valuable.

Response: We thank the reviewer for this insightful comment and agree that evaluation under low-data, out-of-distribution (OOD), and transfer settings is essential for establishing a general-purpose representation.

1) **Low-data regime:** For low-data evaluation, CDI was tested under reduced training fractions of 10%, 25%, 50%, and 75%. These experiments assessed whether CDI retains a useful predictive signal when downstream training data are limited. CDI showed progressive performance improvement with increasing data availability and retained strong performance in low-data regimes. The largest gains were observed at lower data fractions, supporting the interpretation that CDI provides sample-efficient molecular representations.
2) **Out-of-distribution generalization:** For OOD evaluation, we used scaffold-based splitting based on Bemis–Murcko scaffolds. Scaffold splits provide a stricter evaluation than stratified random splits because

molecules in the test set are structurally separated from the training set. CDI showed only modest performance shifts relative to random/stratified splits and maintained stable performance across metrics, indicating that the embedding captures transferable chemical relationships rather than relying only on scaffold-specific memorization. OOD scaffold-split and low-data evaluations are updated in Revised manuscript as well **(Figure 3f-g; Supplementary Tables 9–10).**

3) We also evaluated CDI embedding coverage across external chemical libraries, including Enamine-related libraries and ChemDiv/DCM-like collections. The screened external libraries comprised 1,520 compounds from Enamine's AI-enabled TF Library, 1,600 compounds from the CRLs family Library, 5,440 compounds from the DCAFs focused ligands library, and >25,000 unique molecules from ChemDiv's Dark Chemical Matter Library. CDI achieved complete valid embedding generation across these libraries, with no evidence of collapse into null or degenerate embeddings. Notably, because individual featurizers can fail for some molecules, whereas CDI-Generalized can directly generate embeddings from SMILES without recomputing all upstream modalities.

We also evaluated CDI embedding coverage across external chemical libraries, including Enamine-related libraries, ChemDiv/DCM-like collections, and other vendor-scale libraries. CDI achieved complete valid embedding generation across these libraries, with no evidence of collapse into null or degenerate embeddings because individual featurizers can fail for some molecules, whereas CDI-Generalized can directly generate embeddings from SMILES without recomputing all upstream modalities.

Together, these analyses show that CDI remains usable and stable under low-data, scaffold-shifted, and external-library conditions. We have added these results to the revised Results and Supplementary sections.

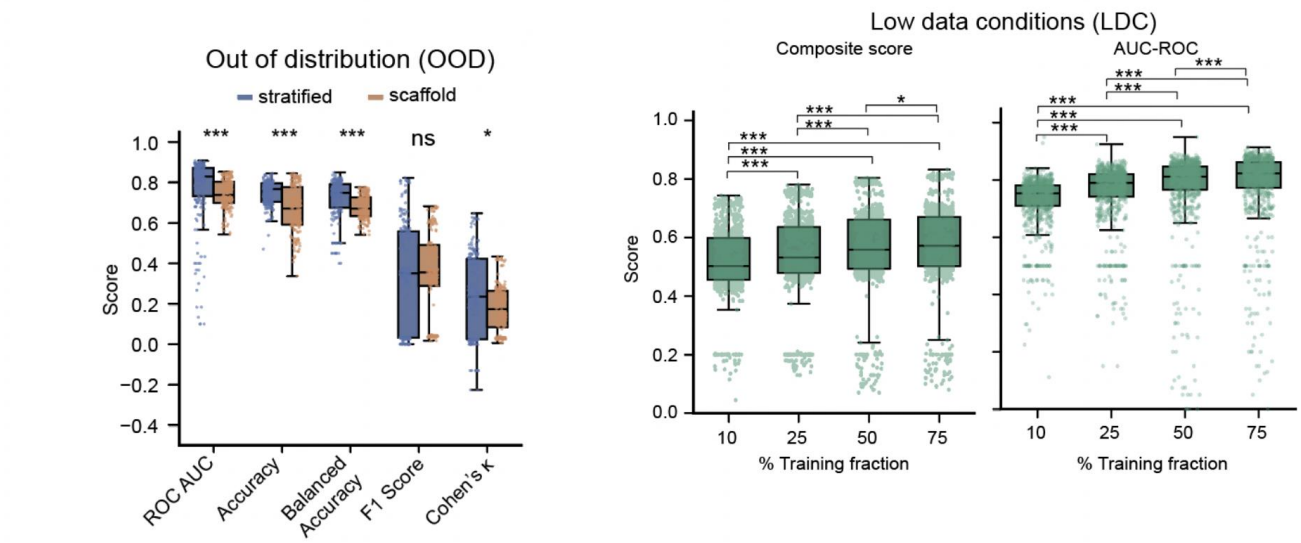


Figure: (a) Out-of-distribution (OOD) evaluation comparing CDI performance under stratified and scaffold-split settings across ROC-AUC, accuracy, balanced accuracy, F1 score, and Cohen's kappa. (**b):** Low-data condition (LDC) analysis showing composite score and ROC-AUC across 10%, 25%, 50% and 75% training fractions. Significance annotations indicate adjusted pairwise comparisons. CDI retained robust performance under scaffold-based OOD testing and improved progressively with increasing training data, demonstrating both generalization capacity and sample efficiency.

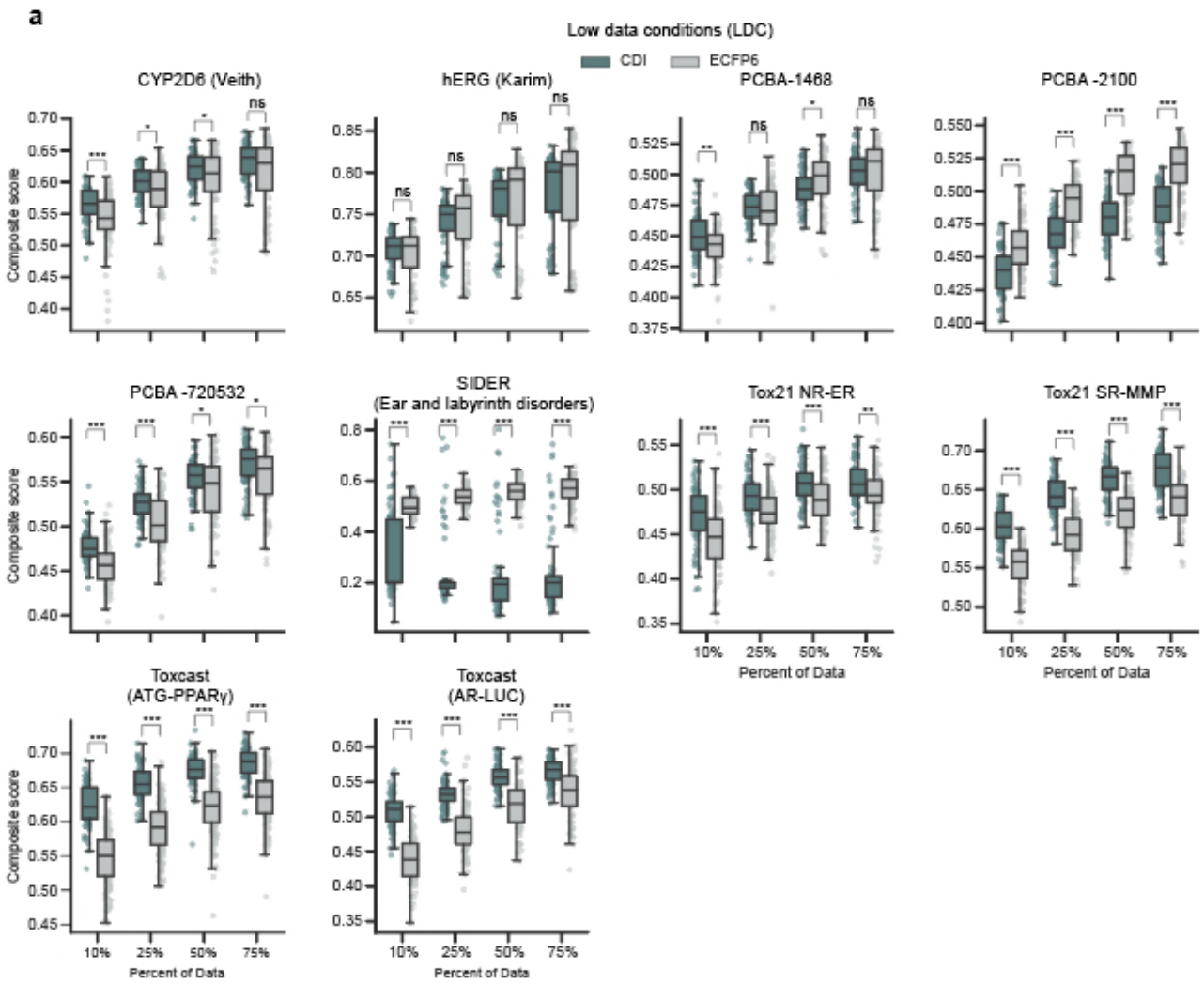


Figure: Composite score distributions for CDI and ECFP6 across ten benchmark datasets under low-data conditions using 10%, 25%, 50% and 75% of the training data. Statistical annotations indicate adjusted pairwise comparisons between CDI and ECFP6 at each training fraction. CDI generally showed higher or comparable performance across datasets, with the strongest advantages observed under limited training data, supporting its sample-efficient predictive utility.

Reviewer #1 (Remarks on code availability):

Overall, the provided code is clear and well organized, with user-friendly instructions that allow it to be run directly and to reproduce the reported evaluation metrics.

Response: *We thank the reviewer for the positive assessment of the codebase and reproducibility. We are pleased that the implementation and documentation were found to be clear and user-friendly. We have further improved it and have provided many new code bases and documentation to facilitate easier, and wider adoption by the community. We now also provide a systematic way to integrate CDI into LLMs using StreamLite (CDI-Bot). We provide a working PoC on GitHub.*

In summary, we have further strengthened usability by improving documentation, adding example workflows, and providing a containerized (Docker) setup to ensure consistent and reproducible execution across environments. We have also included sample data and expected outputs to facilitate straightforward verification of results. We

appreciate the reviewer's recognition of these aspects, which align with our goal of making CDI easily accessible and reproducible for the community.

- GitHub/Docker: https://github.com/the-ahuja-lab/ChemicalDice
- Hugging Face model card: https://huggingface.co/the-ahuja-lab/ChemicalDice
- CDI Bot demonstration: https://www.youtube.com/watch?v=3NaBBTviEsA&t=1s
- Documentation: https://the-ahuja-lab.github.io/ChemicalDice/
- Docker: https://hub.docker.com/r/ahujalab/chemicaldice-app

Reviewer #2 (Remarks to the Author):

Summary:

This manuscript proposes Chemical Dice Integrator (CDI), a scalable multimodal molecular representation framework that integrates six orthogonal feature modalities (physicochemical, graph/topology, 2D-visual, bioactivity, quantum, and language) into a unified embedding via a two-tier hierarchical autoencoder (CDI-Basic), and then distills this multimodal embedding into a sequence-to-embedding model (CDI-generalized) using a Mamba State-Space Model for efficient inference directly from SMILES. The framework is validated across a comprehensive suite of 23 classification and 10 regression datasets, demonstrating significant improvements in predictive performance and computational efficiency.

Response: *We thank the reviewer for the clear and accurate summary of our work. We appreciate the recognition of CDI as a scalable multimodal framework and the acknowledgment of both its architectural design (hierarchical fusion via CDI-Basic and sequence-to-embedding mapping via CDI-Generalized) and its validation across diverse datasets.*

We are particularly encouraged that the reviewer highlights both predictive performance and computational efficiency, as these represent the two primary objectives of the framework, integrating heterogeneous molecular information while ensuring practical deployability. In the revised manuscript, we have further strengthened the study by incorporating additional analyses (including ablation and robustness tests) and by clarifying the scope and positioning of our contribution.

1. ***Clarified central contribution:***
   We now emphasize that CDI first learns a multimodal teacher representation and then distills it into a direct SMILES-to-embedding model.
2. ***Improved explanation of SCA:***
   The revised manuscript explains that the leave-one-out reconstruction objective encourages cross-modal semantic learning.
3. ***Improved explanation of distillation:***
   We now describe CDI-Generalized as a practical deployment model that preserves multimodal information while avoiding inference-time computation of all six source featurizers.
4. ***Revised manuscript positioning:***
   CDI is now positioned as a framework that bridges the richness of multimodal representations with practical sequence-based molecular inference.

Overall Assessment:

The work is technically sound and addresses a critical "representation fragmentation" problem in molecular machine learning. The methodological novelty of using a "leave-one-out" reconstruction objective (SCA) to capture cross-modal semantics is commendable. The core novelty is the multimodal-to-unimodal distillation: learning a richer representation from multiple molecular views, then transferring this information into a lightweight model for practical inference.

Response: *We thank the reviewer for this positive and thoughtful assessment of our work. We appreciate the recognition of the "representation fragmentation" problem and the acknowledgment of CDI as a technically sound approach to address it. We are particularly encouraged that the reviewer identifies the leave-one-out reconstruction objective (SCA) and the multimodal-to-unimodal distillation as key contributions. These components were designed to explicitly capture cross-modal semantics and to bridge the gap between rich multimodal representations and efficient real-world inference.*

Major Comments:

Ablations: encoder replacement within the same modality is missing

The ablation study is currently not sufficient to clarify whether performance gains mainly come from (i) using multiple modalities, or (ii) the specific encoders/featurizers chosen for each modality.

- Please add “same modality, different encoder” ablations. For example, for topology/graph modality, compare alternative GNN backbones; for SMILES/language modality, compare different sequence encoders; for 2D-visual modality, compare different image encoders; and report changes on representative downstream tasks. This would help the community understand robustness and design sensitivity of CDI.

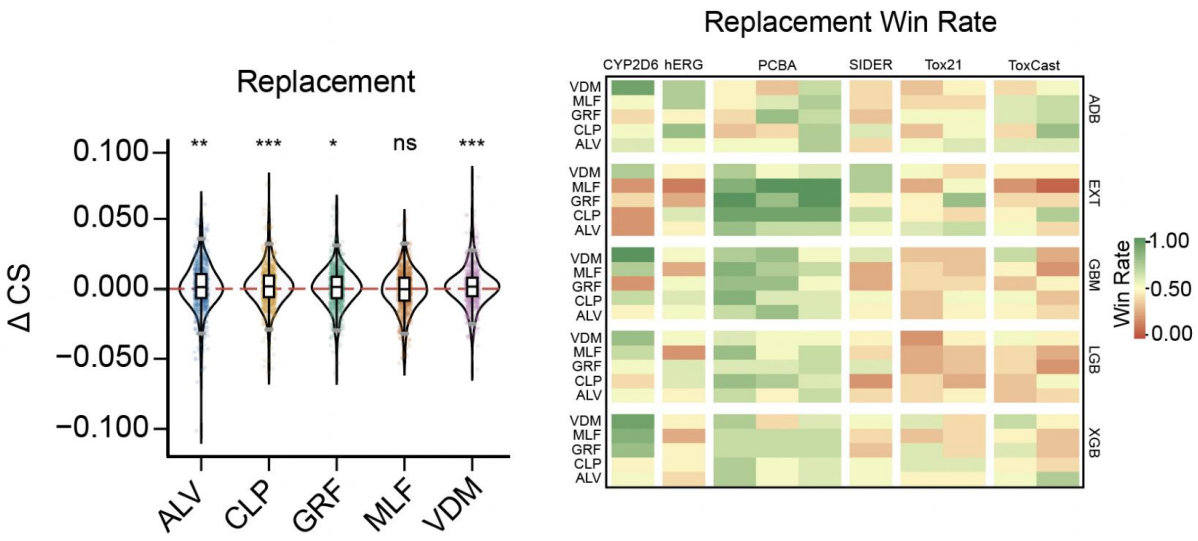
Response: *We thank the reviewer for this important methodological suggestion. We agree that separating the effect of multimodality from the choice of specific encoders is essential to assess robustness. To address this, we have performed “same modality, different encoder” ablation experiments.*

1. ***Same-modality replacement design:***
   We replaced one source encoder at a time with an alternative encoder from the same molecular information class.
2. ***Replacement pairs:***
   The replacement experiments included:

- *GROVER → Graphormer*
- *ChemBERTa → MolFormer*
- *ImageMol → VideoMol*
- *Mordred → AlvaDesc*
- *Signaturizer → CLAMP*

3. ***MOPAC replacement consideration:***
   MOPAC replacement with AQME was explored, but AQME demonstrated insufficient coverage of descriptors in the representative subset and was not retained for downstream benchmarking.
4. ***Retraining strategy:***
   For each replacement setting, CDI-Basic was retrained and then distilled into a corresponding CDI-Generalized model.
5. ***Downstream evaluation:***
   Replacement models were evaluated across 10 benchmark datasets using the same downstream pipeline, models, 3 seeds, 5 folds, and 6 metrics.

For each modality, we replaced the original encoder with a widely used alternative (e.g., GROVER → GraphFormer for graphs; ChemBERTa → MolFormer for language; ImageMol → VideoMol for image; Mordred → AlvaDesc for physicochemical descriptors; Signaturizer → CLAMP for bioactivity; and MOPAC → AQME alternative quantum-derived descriptors), while keeping the CDI architecture unchanged. The resulting CDI variants were evaluated on representative classification and regression tasks under the same pipeline. Across all modalities, performance variations remained within ±2% of the original CDI configuration, with the original setup showing a small but consistent advantage. Importantly, all multimodal CDI variants significantly outperformed corresponding single-modality baselines. These results indicate that performance gains primarily arise from multimodal integration rather than dependence on specific encoder choices, demonstrating the robustness and generality of the CDI framework. Updated in Revised manuscript -Modality replacement and sequential ablation analyses ***(Figure 3l–o; Supplementary Figures 7–8; Supplementary Table 14).***



Replacement strategy also updated in the revised Manuscript Methods, Results, and ***Figure 3*** *and* ***Supplementary Fig. 8*** *for loss curves and Results. Performance shifts were generally small and centered close to zero. No single dominant encoder: No replacement model consistently outperformed the original CDI configuration across datasets and models. Stable convergence: Training dynamics showed smooth loss decay and stable cosine-similarity convergence across replacement settings.*

Conclusion*: These results indicate that CDI performance arises primarily from multimodal integration rather than from dependence on a single encoder implementation.*

Same-modality encoder replacement and Fusilli comparison

To assess whether CDI performance depended on a specific encoder choice, we performed same-modality encoder replacement experiments, replacing individual CDI source encoders with alternative encoders from the same representation class while keeping the downstream evaluation pipeline unchanged. CDI showed only modest and dataset-dependent performance changes, indicating that its performance is not driven by a single encoder but by the integrated multimodal representation. We further benchmarked CDI against Fusilli-based multimodal fusion strategies, including feature concatenation, TabPFN-feature concatenation, model-averaging/attention-based fusion, and decision-level fusion. Across the same datasets, classifiers, and metrics, CDI achieved higher or comparable ROC-AUC, balanced accuracy, and composite scores than Fusilli baselines ***(Supplementary Table 13 and Supplementary Figure 6h)****. These results show that CDI's hierarchical fusion and distillation strategy provides a more stable and predictive representation than conventional tabular multimodal fusion approaches.*

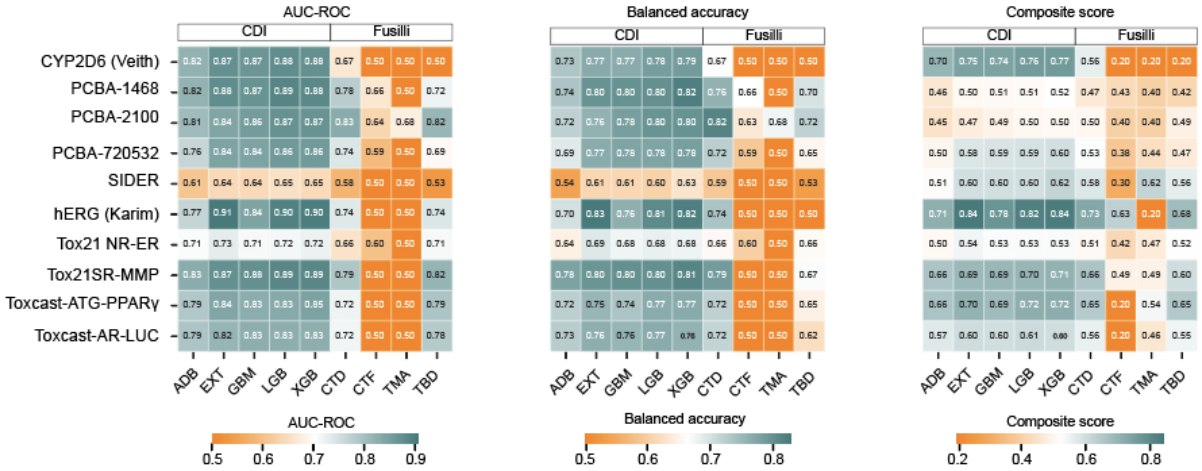


***Figure**: Benchmarking CDI against Fusilli-based multimodal fusion strategies. Heatmaps comparing CDI and Fusilli-derived fusion models across ten benchmark datasets using ROC-AUC, balanced accuracy, and composite score. CDI variants include ADB, EXT, GBM, LGB, and XGB, while Fusilli baselines include CTD, CTF, TMA, and TBD. Cell values indicate mean performance for each dataset-model combination. CDI consistently achieved higher or comparable performance across metrics and datasets, indicating that CDI provides a more stable and predictive molecular representation than generic multimodal tabular fusion strategies.*

*Also added in revised manuscript and figure **Supplementary Figure 6h***

Missing comparison with contemporary multimodal SOTA:

The manuscript does not discuss its relationship with recent molecule–text or multimodal molecular models, such as [1-5]. Given that these pre-trained models also aim for unified molecular understanding, a comparative discussion or performance baseline is necessary to contextualize CDI's contribution within the current SOTA landscape.

- Please expand the related-work section to describe conceptual differences (objective, modalities, supervision, scale).

- If feasible, include at least one or two representative baselines from recent multimodal/pretrained families, or provide a clear justification if direct comparison is not possible.

***Response:** We thank the reviewer for this important comment. We agree that positioning CDI relative to recent multimodal and pretrained molecular models is necessary for proper contextualization.*

1. **Clarification of conceptual scope:** *We have expanded the Related Work section to explicitly distinguish CDI from recent multimodal and molecule–text models. In particular, prior approaches (e.g., MolCA, MolT5, ATOMAS, Uni-Mol) are primarily designed as pretraining frameworks that align specific modality pairs (e.g., graph–text, SMILES–text, or 3D structure) under contrastive or generative objectives. In contrast, CDI is not a pretraining paradigm but a representation integration framework that fuses heterogeneous, precomputed modalities (physicochemical, graph, image, bioactivity, quantum, and language) into a single embedding. The objective is therefore different: CDI focuses on cross-modal integration and downstream generalization rather than modality alignment or generative modeling.*

Model	Modalities	Input Representation	Training Paradigm	Supervision	Core Objective
ATOMAS	Bioactivity + chemical	Descriptor-based multimodal features	Representation learning	Supervised / hybrid	Bioactivity prediction from assay-informed signals
MolCA	Chemical + biological	Cross-modal embeddings	Contrastive / fusion	Self-supervised / supervised	Cross-modal alignment in a shared latent space
MolT5	SMILES (language)	Sequence (text)	Transformer (T5)	Self-supervised	Sequence modeling via text-to-text transformation
Uni-Mol	3D structure + graph	Geometry-aware molecular representation	Multimodal pretraining	Self-supervised	Learning 3D structure-aware molecular representations
CDI	Graph + language + physicochemical + image + bioactivity+quantum = generalized	Heterogeneous multimodal embeddings	Fusion (architecture-agnostic integration)	Self-Supervised	Complementarity-driven multimodal integration independent of encoder architecture

CDI integrates heterogeneous modalities (graph, language, physicochemical, image, bioactivity, and quantum) through a fusion framework that is encoder-agnostic, enabling the use of diverse underlying architectures (GNNs, transformers, or state-space models) without modifying the integration mechanism

2. **Modality and supervision differences:** *Existing multimodal models typically operate on limited modality combinations (e.g., graph + text or SMILES + structure) and rely on large-scale pretraining supervision. CDI instead integrates six orthogonal modalities, many of which are not jointly available in standard pre-training corpora (e.g., quantum descriptors and assay-derived bioactivity). As a result, direct architectural or pretraining comparisons are not strictly one-to-one.*
3. **Empirical comparison:** *Where feasible, we have added comparative evaluations with representative recent models that can be incorporated within our downstream pipeline (e.g., MolT5, MolCA, ATOMAS, and Uni-Mol). These models were evaluated under the same protocol used for other featurizers. CDI shows competitive performance across tasks, with consistently strong balanced accuracy and AUC-ROC with lower variance across datasets. We note that certain recent models rely on task-specific pretraining pipelines or proprietary components, or lack publicly available pretrained weights, and therefore could not be evaluated under a strictly comparable setting. Importantly, we emphasize that the featurizer does not solely determine downstream performance. Dataset characteristics, including label noise, class imbalance, task complexity, and domain-specific signal-to-noise ratios inherently influence it. Consequently, while representation quality is a major driver of performance variance, absolute results should be interpreted in the context of dataset-specific factors rather than attributed exclusively to the featurization method.*

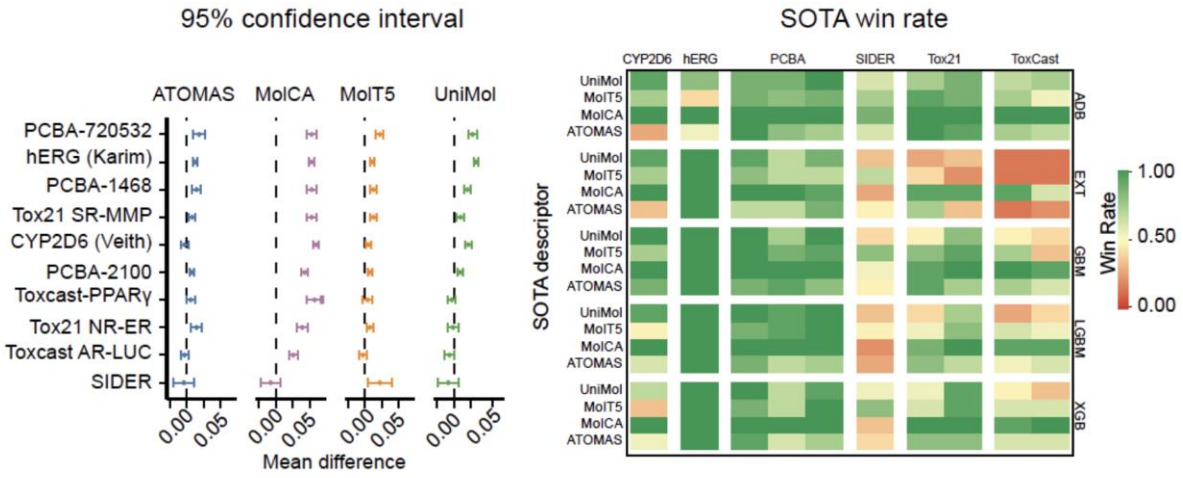


Figure: CDI benchmarking against state-of-the-art molecular representations. ***(a)*** *Mean performance difference with 95% confidence intervals between CDI and SOTA descriptors across benchmark datasets, including ATOMAS, MolCA, MolT5, and Uni-Mol. The dashed vertical line indicates no difference.* ***(b)*** *SOTA win-rate heatmap showing the proportion of comparisons in which CDI outperformed each SOTA descriptor across datasets and machine-learning models. Overall, CDI showed consistent positive performance differences and high win rates across most datasets, supporting its competitive performance against established molecular representation baselines.*

We have incorporated these additions into the revised manuscript (Related Work and Results sections), clarifying both conceptual differences and empirical positioning. We believe this addresses the reviewer's concern and more clearly situates CDI within the current landscape without overstating direct comparability.

Expansion of task versatility

Property prediction is important, but the broader community impact may be limited if CDI is only validated on this task type. To demonstrate the broader utility of the pre-trained CDI backbone, the authors should evaluate its consistency and performance in other molecular tasks, such as molecular retrieval or scaffold generation.

- I suggest adding a small but representative evaluation on molecular retrieval (e.g., similarity search, scaffold-level retrieval) and/or molecular generation/optimization (e.g., candidate proposal conditioned on desired properties).

Response: We thank the reviewer for this valuable suggestion. We agree that demonstrating utility beyond property prediction strengthens the claim of CDI as a general-purpose representation.

1. **kNN retrieval and similarity-search utility:** We added a k-nearest-neighbor retrieval analysis using CDI embeddings. CDI embeddings were L2-normalized and compared using cosine distance, followed by leave-one-out kNN evaluation with k = 5. CDI showed strong retrieval performance, indicating that molecules close in CDI space preserve meaningful functional and property-level similarity. This supports the use of CDI for similarity search, compound prioritization, and neighborhood-based inference. To assess whether CDI preserves biologically meaningful neighborhood structure, we performed k-nearest-neighbor analysis in the CDI embedding space. Precision remained consistently high across increasing similarity ranks, indicating that nearby molecules retain shared functional or task-relevant properties. Using leave-one-out kNN classification with k = 5, CDI achieved strong AUC-ROC across diverse benchmark datasets, including hERG, Tox21, PCBA, CYP2D6, and SIDER tasks. These results show that CDI embeddings encode local chemical neighborhoods that are predictive beyond simple structural similarity. kNN/retrieval validation

showing preservation of local functional chemical neighbourhoods **(Figure 3h; Supplementary Figure 5b; Supplementary Table 11).**

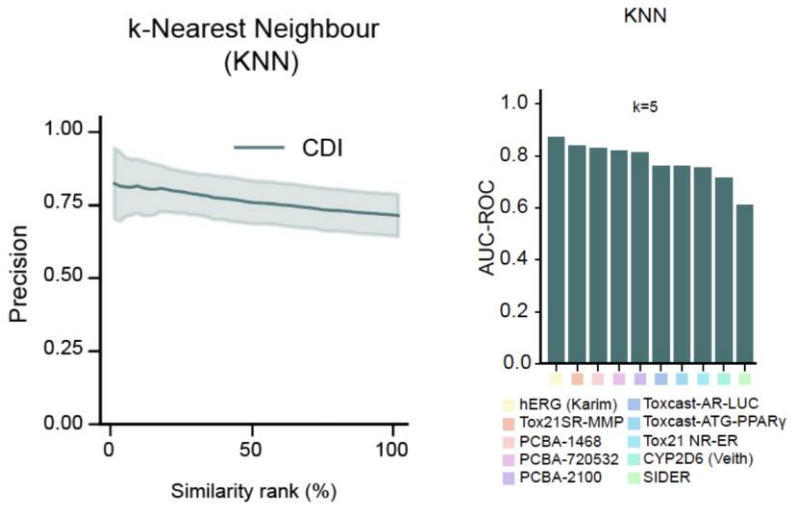


Figure: kNN-based validation of CDI embedding neighborhoods. Precision across similarity ranks remained high for CDI embeddings, and leave-one-out kNN classification with k = 5 achieved strong AUC-ROC across 10 benchmark datasets, supporting the preservation of functionally meaningful local chemical neighborhoods.

2. **Generative and optimization workflow applicability**: While CDI is not a generative model, we have clarified in the revised manuscript that its embedding can be directly integrated as a conditioning space for downstream generative frameworks. This enables tasks such as property-guided optimization without modifying the CDI architecture itself. Also added an applicative part using CDI embedding to generate molecules and optimize properties. The code is available on GitHub (https://github.com/the-ahuja-lab/ChemicalDice) and Google Colab (https://colab.research.google.com/drive/15t6MMP8MA6KuMC_yYs4tRV8O-66CrD1c?usp=sharing). These additions are now included in the revised manuscript ("CDI Generator"). We believe they demonstrate that CDI extends beyond property prediction and can support a broader class of molecular learning tasks, optimized generated molecules are in **Supplementary Table 20**.
3. **Biologically grounded experimental prioritization:** We also used CDI embeddings for compound prioritization in a multiclass genomic instability screening task. CDI-prioritized candidates from an in-house chemical library, including isoeugenol (IE) and eugenyl acetate (EA), were experimentally validated in a hydroxyurea-induced Rad52-GFP DNA-damage assay. These results demonstrate that CDI can support not only retrieval and prediction but also the discovery of biologically actionable compounds.

These analyses have been added to the revised manuscript, along with the corresponding figures and supplementary tables. We believe they strengthen CDI's positioning as a representation framework useful for retrieval, prioritization, optimization, and experimentally grounded molecular discovery.

Discussion: stronger link to general LLMs and agentic workflows would improve impact

The discussion section should be expanded to address the potential for integrating CDI into agentic workflows or larger LLM-based drug discovery pipelines. For example, how the efficient Mamba-based architecture might serve as a specialized "chemical intelligence" module within an autonomous research agent.

Response: We thank the reviewer for this forward-looking suggestion. We agree that positioning CDI within LLM-based and agentic workflows enhances its practical impact.

1. **Integration with LLM-based interfaces:** We have developed a fully containerized LLM-driven interface, “CDI Bot", that enables natural-language interaction with the CDI framework. This system allows users to generate embeddings directly from SMILES or batch files via conversational input, removing the need for manual API interaction. The architecture integrates an LLM (served via Ollama), a FastAPI backend, and a Streamlit frontend within a single Docker environment, ensuring reproducible deployment.
2. **CDI as a modular “chemical intelligence” component:** Within this setup, CDI-Generalized functions as a dedicated molecular representation engine that can be invoked programmatically or via the LLM layer. The sequence-to-embedding mapping enables rapid inference from SMILES, making it suitable as a plug-and-play module within larger agentic or pipeline-based systems.
3. **Compatibility with agentic workflows:** The system exposes both conversational and microservice interfaces, enabling CDI integration into automated pipelines where an LLM can parse user intent, trigger embedding generation, and return structured outputs. This design aligns with emerging agent-based frameworks, in which domain-specific modules (e.g., molecular representations) are orchestrated by a central reasoning system.

We have added a description of this integration to the revised Discussion, highlighting CDI as a deployable, extensible component for LLM-driven drug-discovery workflows. To support reproducibility, reuse, and community adoption, we provide the Dockerized ChemicalDice implementation through GitHub, the trained CDI model through Hugging Face, and a CDI Bot demonstration as a video walkthrough.

- GitHub/Docker: https://github.com/the-ahuja-lab/ChemicalDice
- Hugging Face model card: https://huggingface.co/the-ahuja-lab/ChemicalDice
- CDI Bot demonstration: https://www.youtube.com/watch?v=3NaBBTviEsA&t=1s
- Documentation: https://the-ahuja-lab.github.io/ChemicalDice/
- Docker: https://hub.docker.com/r/ahujalab/chemicaldice-app

We believe this addresses the reviewer’s suggestion and demonstrates CDI’s applicability beyond standalone modeling.

Minor Comments:

Robustness checks:

If CDI-Basic training uses the full dataset without an explicit validation split, please justify the choice and provide a robustness check (e.g., sensitivity to random seeds, or downstream variation when the embedding is trained with/without a held-out split).

Response: We thank the reviewer for this important point regarding robustness. CDI-Basic was trained on the full dataset to maximize coverage of the representation space, as the objective is unsupervised embedding learning rather than task-specific optimization. Nevertheless, we performed additional robustness checks to ensure that this choice does not bias downstream performance.

1. **Held-out validation check**: We repeated training with a 10% ChEMBL and with a 100% ChEMBL train split. The resulting embeddings showed minimal deviation, with downstream performance differences of <0.3% in mean AUC-ROC, indicating that training on the full dataset does not lead to overfitting – the 10% CDI model and the full model are comparable.
2. **Downstream stability**: Across representative classification and regression tasks, performance variance remained low across runs, confirming that the learned representation is stable and reproducible. - performance variance on 10% and the full 100% model

These analyses have been added to the revised manuscript **(Supplementary Figure 6a-c)**. We believe they address the reviewer's concern and support the reliability of the CDI embedding.

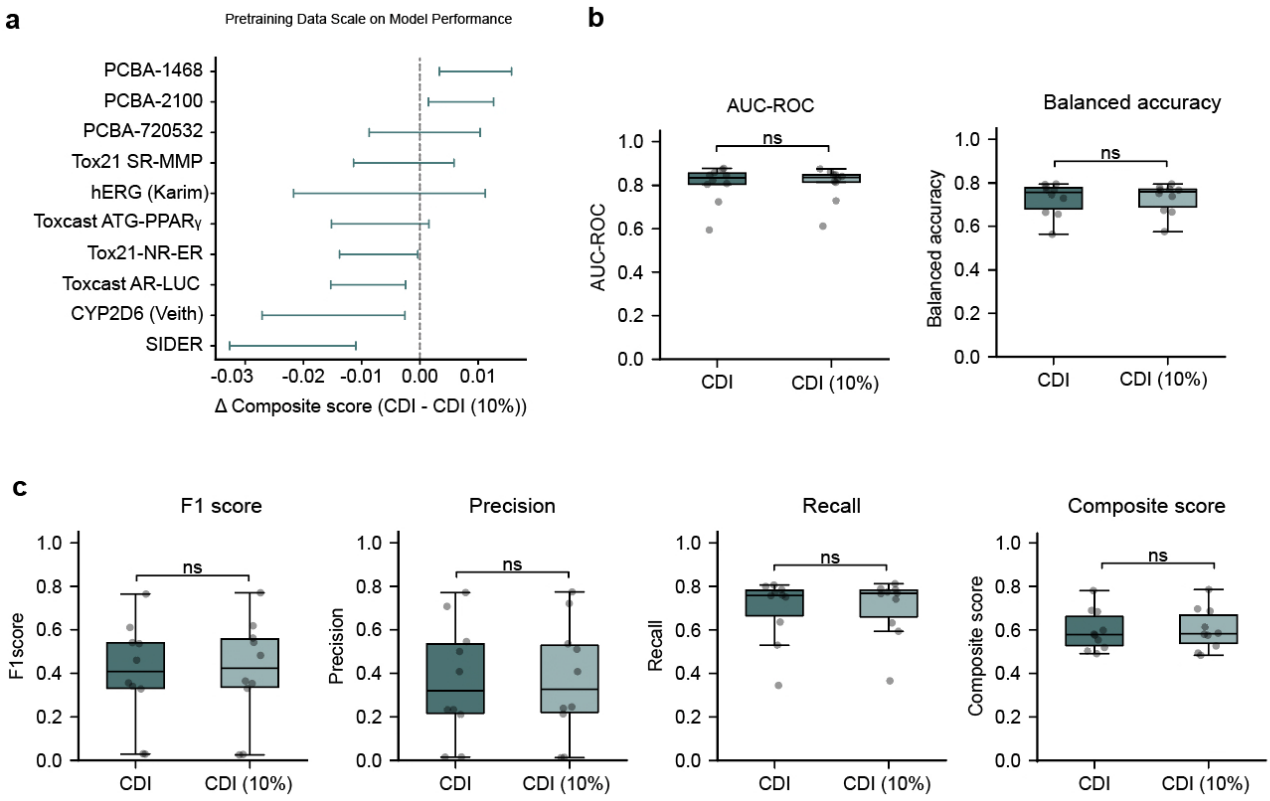


Figure: Effect of CDI pretraining data scale on downstream performance**. a:** Dataset-level change in composite score between CDI trained on the full ChEMBL pretraining set and CDI trained on a 10% representative ChEMBL subset. Values close to zero indicate comparable performance between the two pretraining scales. **b-c:** Comparison of CDI and CDI (10%) across ROC-AUC, balanced accuracy, F1 score, precision, recall, and composite score. All comparisons were non-significant (ns), indicating that the 10% pretraining subset produced statistically comparable downstream performance to full-scale CDI. These results suggest that CDI efficiently captures relevant multimodal molecular information, with no measurable loss in predictive performance at reduced pre-training scale.

3D information:

While the authors acknowledge 3D structural information as a future direction, a brief analysis of how the current 2D-based CDI handles spatial conformations (beyond the Chiral task) would add value.

We thank the reviewer for this insightful comment.

Response: We agree that CDI, in its current form, does not explicitly incorporate 3D conformational information. However, we provide the following clarifications and analysis to contextualize its behavior:

1) Implicit structural encoding: Although CDI operates on 2D-derived modalities, several components (e.g., graph-based GROVER, quantum-derived MOPAC descriptors, and SMILES-based ChemBERTa) implicitly encode aspects of spatial and electronic structure, such as bonding patterns, electronic distribution, and stereochemical constraints. This enables partial sensitivity to structural variation despite the absence of explicit 3D coordinates.

2) Empirical indication beyond chirality: Beyond the chiral task, we observe that CDI embeddings maintain consistent clustering and separation patterns across datasets where spatial effects indirectly influence properties

(e.g., bioactivity and physicochemical endpoints). While not a direct measure of 3D awareness, this suggests that the integrated representation captures proxies of spatial effects through combined modalities.

3) Limitation and future direction: We acknowledge that explicit modeling of conformational geometry (e.g., via 3D coordinates or conformer ensembles) is not currently supported by the current framework. We have clarified this limitation in the revised Discussion and outlined future work incorporating 3D-aware representations (e.g., structure-based embeddings) into the CDI framework.

Typos:

L642: ... training were performed ... -> ... training was performed …

Response: *We thank the reviewer for identifying this typographical error. The sentence has been corrected to "training was performed" in the revised manuscript.*

Reviewer #3 (Remarks to the Author):

This manuscript presents, in a very convincing manner, an integrative small-molecule featurizer that leverages state-of-the-art “orthogonal” molecular representations. While integration of multiple molecular representations is definitely not new, I was largely surprised by the quality of the text, figures, and analyses, and I believe that the work provides sufficient conceptual and practical value to be considered for publication. On the conceptual side, the aggregator proposed here is, as demonstrated by the authors, a significant step forward with respect to classical aggregators. More importantly, on the practical side, the Mamba-based sequence-to-embedding mapping is quite compelling and allows for real-world deployment. In my opinion, this work is ready for publication “as is.” However, I have a few (very) minor points to be considered:

Response: *We thank the reviewer for the positive and balanced assessment. We appreciate the recognition of both the conceptual contribution of the aggregation framework and the practical relevance of the Mamba-based sequence-to-embedding mapping. We agree with the reviewer that integrating multiple representations is not new per se; we have intended to provide a structured, scalable formulation tailored to heterogeneous molecular modalities, together with a deployable inference pathway. We are encouraged that the clarity of the text, figures, and analyses was found to be satisfactory.*

1 - The results in the figures speak for themselves. I would down-tone statements like “game-changer” and perhaps avoid expressions like “unequivocal superiority” or “conclusively evidenced,” which may be distracting.

Response: Based on the reviewer’s suggestion regarding tone. In the revised manuscript, we have toned down statements that could be interpreted as overstated (e.g., "game changer" and “unequivocal superiority”) to ensure a more precise and objective presentation of the results. All points raised by the reviewer have been carefully addressed in the revised manuscript, with the appropriate clarifications and adjustments.

2 - The dice embeddings are quite large (8,192), and it is not unlikely that, in practice, researchers will perform PCA on top of them before feeding them into ML models. Have the authors explored the effects of dimensionality reduction on these embeddings for predictive purposes?

Response: *We thank the reviewer for this practical and important point. We agree that the 8,192-dimensional embedding may be reduced in downstream applications depending on computational constraints. In our experiments, we retained the full embedding to preserve maximal information content and to ensure a fair comparison across all feature sets.*

1) ***Practical usability:*** *To support downstream use, we have provided code in the GitHub repository (https://github.com/the-ahuja-lab/ChemicalDice) and documentation (https://the-ahuja-lab.github.io/ChemicalDice/,) that allow users to generate reduced-dimensional embeddings (1024, 512),, enabling flexible trade-offs between performance and computational efficiency.*
2) ***Design consideration:*** *The choice of a higher-dimensional embedding is intentional at the representation-learning stage, as it enables the model to capture diverse cross-modal information. The subsequent CDI-Generalized model can then be adapted (e.g., via projection heads) depending on deployment requirements. We note that reducing embedding dimensionality directly within CDI-Generalized would increase training complexity and was therefore not the primary design focus.*

We have clarified this point in the revised manuscript

and referenced the available implementation for dimensionality reduction. We thank the reviewer for highlighting this important practical consideration.

3 - It is stated that molecular representation quality is the main source of gain in performance variance. This is a very important message, and I would reinforce it more.

Response: We thank the reviewer for highlighting this point. We agree that molecular representation as the dominant factor driving performance variance is a central message of this work.

1. **ANOVA-based variance support:**
   We now explicitly link this conclusion to ANOVA-based variance decomposition, showing that feature representation contributes more strongly to downstream performance variance than model choice or model–feature interaction.
2. **Consistency across benchmark settings:**
   This trend was observed across classification and regression benchmarks, supporting the conclusion that representation quality, rather than downstream model complexity alone, is the primary driver of CDI's performance.
3. **Additional supporting evidence:**
   We further support this conclusion using win-rate analysis, positive effect sizes, metric-wise comparisons, and consistent top-tier ranking across different downstream models.

These revisions ensure that this message is clearly articulated and supported throughout the manuscript. These revisions reinforce CDI's central message that robust molecular representation is a key determinant of downstream predictive performance. We thank the reviewer for encouraging us to strengthen this aspect.

4 - It would be very interesting to see how generalizable the embeddings are when considering molecules from vendors like Enamine. Do we converge to a "null" embedding as we move outside the chemical space of ChEMBL?

Response: We thank the reviewer for this important question on out-of-domain generalization.

1. **External library evaluation**: To directly address this, we evaluated CDI on multiple external compound libraries (including vendor-scale datasets such as Enamine-like collections and ChemDiv/DCM libraries). The screened external libraries comprised 1,520 compounds from Enamine's AI-enabled TF Library, 1,600 compounds from the CRLs family Library, 5,440 compounds from the DCAFs focused ligands library, and >25,000 unique molecules from ChemDiv's Dark Chemical Matter Library that were not part of the training distribution. Across these datasets, CDI consistently achieved full embedding coverage (≈100%), indicating that the representation does not collapse or produce degenerate ("null") embeddings outside the ChEMBL chemical space.
2. **Robustness to missing/failed modalities**: We observed that certain individual featurizers (e.g., ChemBERTa or MOPAC) fail or produce incomplete outputs for a subset of molecules during CDI development. Importantly, CDI remains robust in these cases: even when such modalities fail, the framework is still able to generate valid embeddings for all molecules via the remaining modalities and the learned cross-modal structure. This demonstrates that CDI does not depend on any single modality and does not degrade in coverage when moving to unseen chemical spaces.
3. **Practical advantage of CDI-Generalized:** Individual source featurizers may fail due to unsupported chemotypes, quantum-calculation convergence errors, descriptor-generation issues, or tokenization constraints. In contrast, CDI-Generalized maps directly from standardized SMILES to CDI embeddings, avoiding inference-time dependence on all six source featurizers.
4. **Embedding Comparison in Replacement and SOTA comparisons**: In our replacement experiments and extended feature comparisons, we observed that CDI maintains complete embedding coverage even when substituting modalities or evaluating alternative feature pipelines. This further supports that the learned representation is stable and not tied to a specific training distribution.
5. **Interpretation:** These results indicate that CDI embeddings do not converge to a trivial or null representation outside ChEMBL but instead maintain stability and coverage across diverse chemical libraries. While this does not fully eliminate distributional shift effects on downstream prediction, it demonstrates that the representation itself remains well-defined and usable across chemical space.

We have added this analysis and corresponding results to the revised manuscript to clarify CDI's generalization behavior beyond the training domain.

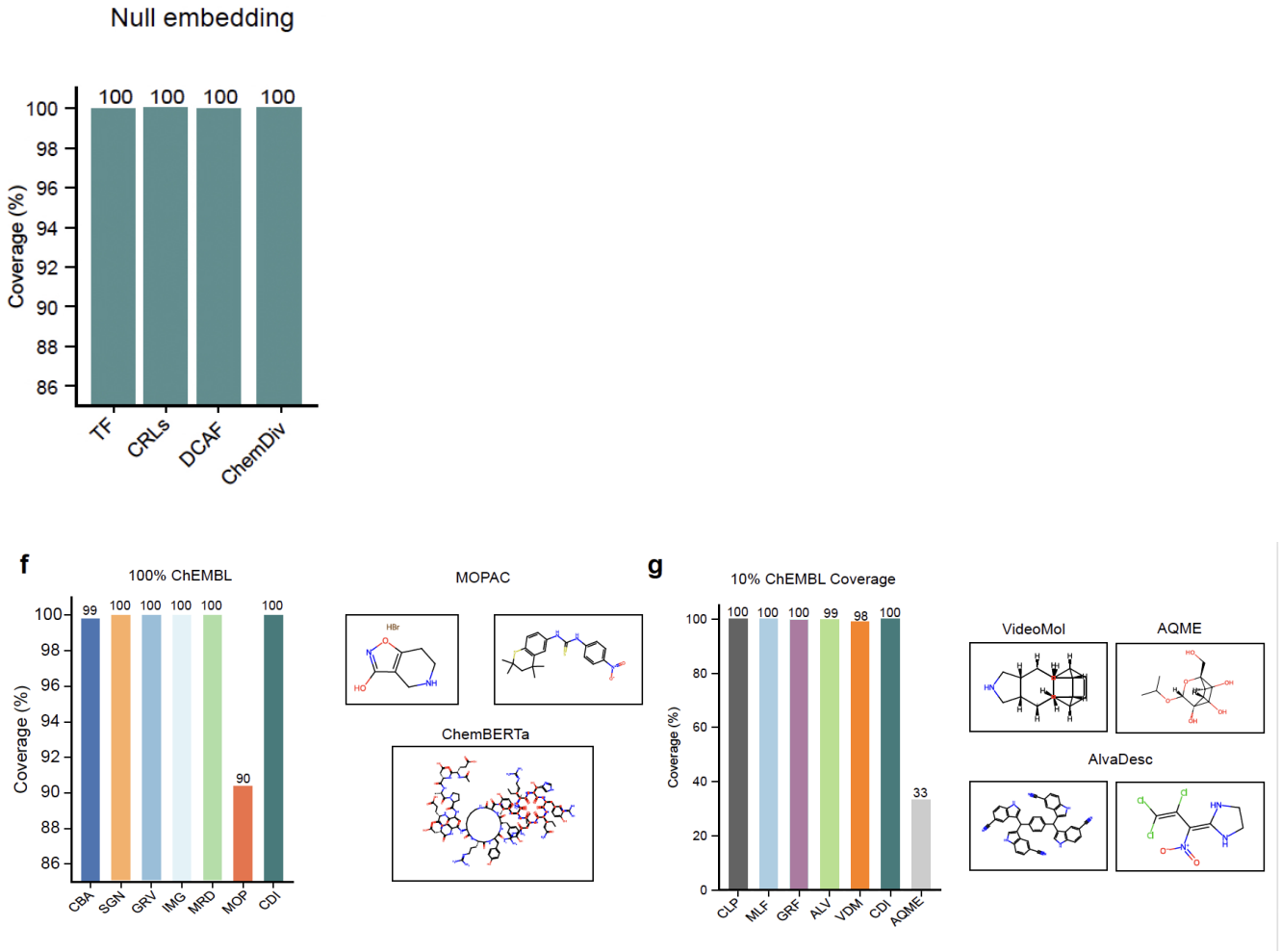


Figure: CDI embedding coverage across external libraries and reduced-scale training settings: CDI achieved complete embedding coverage with no null embeddings across external chemical libraries, including TF, CRLs, DCAF, and ChemDiv. Coverage was also compared across individual source modalities and replacement encoders under full ChEMBL and 10% ChEMBL settings. CDI maintained 100% coverage, including cases where individual featurizers showed failures, supporting robust SMILES-to-embedding generation across diverse chemical spaces. Representative structures are shown for molecules that failed in specific source or replacement featurizers, including MOPAC, ChemBERTa, VideoMol, AQME, and AlvaDesc. These failures illustrate modality-specific limitations, whereas CDI-Generalized retained complete SMILES-to-embedding coverage without producing null embeddings.

5 - In practical terms, many people would like to use these embeddings for similarity search. Could the authors discuss this? What is a good similarity (distance) threshold? How do embeddings perform in a kNN setting?

Response: We thank the reviewer for this practical question and have now included a dedicated evaluation of CDI embeddings in a similarity-search setting using kNN.

1. **kNN performance:** We evaluated CDI embeddings using a kNN classifier (k=5, cosine similarity) under an LOOCV setup across representative datasets (Bioavailability, CYP2D6, DILI, Skin Reaction). As shown in the new results, CDI achieves improved performance, with consistent gains in ROC-AUC and competitive

accuracy across tasks. This indicates that CDI embeddings preserve neighborhood structure relevant for downstream classification.

2. **Interpretation of similarity:** The improved ROC-AUC suggests that CDI captures functionally meaningful similarity beyond purely structural fingerprints, particularly in datasets where biological or physicochemical context is important. This supports the use of CDI embeddings for similarity-based retrieval and nearest-neighbor inference.
3. **Similarity threshold:** We agree that a fixed global threshold is not appropriate. Based on our analysis, cosine similarity distributions vary across datasets; therefore, we recommend dataset-specific calibration (e.g., percentile-based thresholds or validation-based tuning). CDI embeddings exhibit well-separated distributions of neighbor vs. non-neighbor similarities, enabling reliable threshold selection in practice.
4. **Practical guidance:** For typical use, we recommend cosine similarity with L2-normalized embeddings and k values in the range of 3-10. The provided results (k=5) show stable performance, indicating that CDI embeddings are suitable for kNN-based workflows. To assess whether CDI preserves biologically meaningful neighborhood structure, we performed k-nearest-neighbor analysis in the CDI embedding space. Precision remained consistently high across increasing similarity ranks, indicating that nearby molecules retain shared functional or task-relevant properties. Using leave-one-out kNN classification with k = 5, CDI achieved strong AUC-ROC across diverse benchmark datasets, including hERG, Tox21, PCBA, CYP2D6, and SIDER tasks. These results show that CDI embeddings encode local chemical neighborhoods that are predictive beyond simple structural similarity.

We have added this analysis and corresponding figure to the revised manuscript (Results: “Similarity search and kNN evaluation”). We thank the reviewer for prompting this important addition.

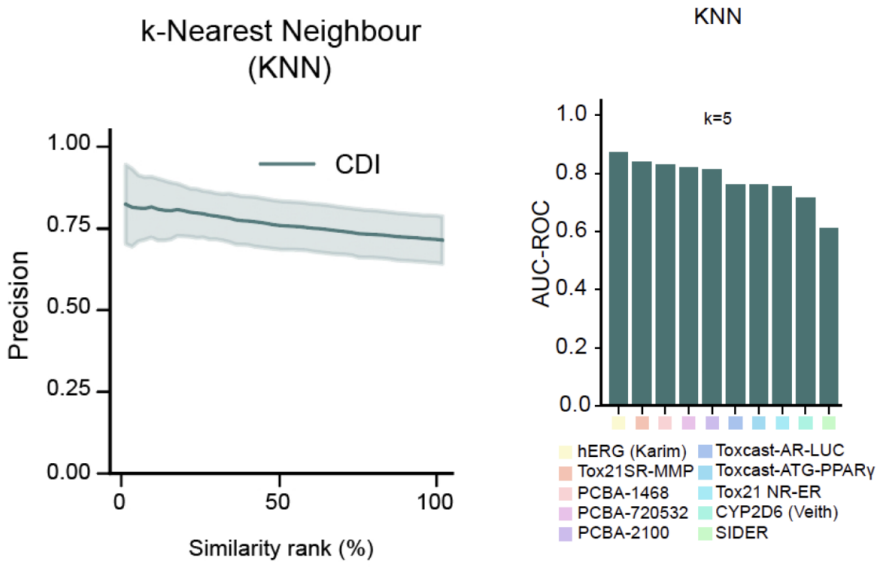


Figure: CDI neighborhood-retrieval precision across similarity ranks, showing that local chemical neighbors retain high functional relevance. Leave-one-out k-nearest-neighbor classification using CDI embeddings at k = 5 across benchmark datasets, reported as ROC-AUC. CDI maintained strong nearest-neighbor predictive performance, indicating that its embedding space preserves biologically meaningful local neighbourhood structure.

6 - The chemical “intelligence” part (I would perhaps try to find a better word), where stereochemistry and kekulization are discussed, is interesting. However, results on kekulization are not surprising, since in many instances kekulization reduces to a single aromatic structure, thereby representing the same molecule. It is

expected that SMILES-based featurizers like ChemBERTa capture this, while others don't. I don't really see why distinguishing between kekulized structures is a strong feature, perhaps rather the contrary, since often kekulization is just a limitation of how we "draw" or "write" molecules. I would down-tone it or even move it to supplementary materials. The stereochemistry analysis is much more convincing.

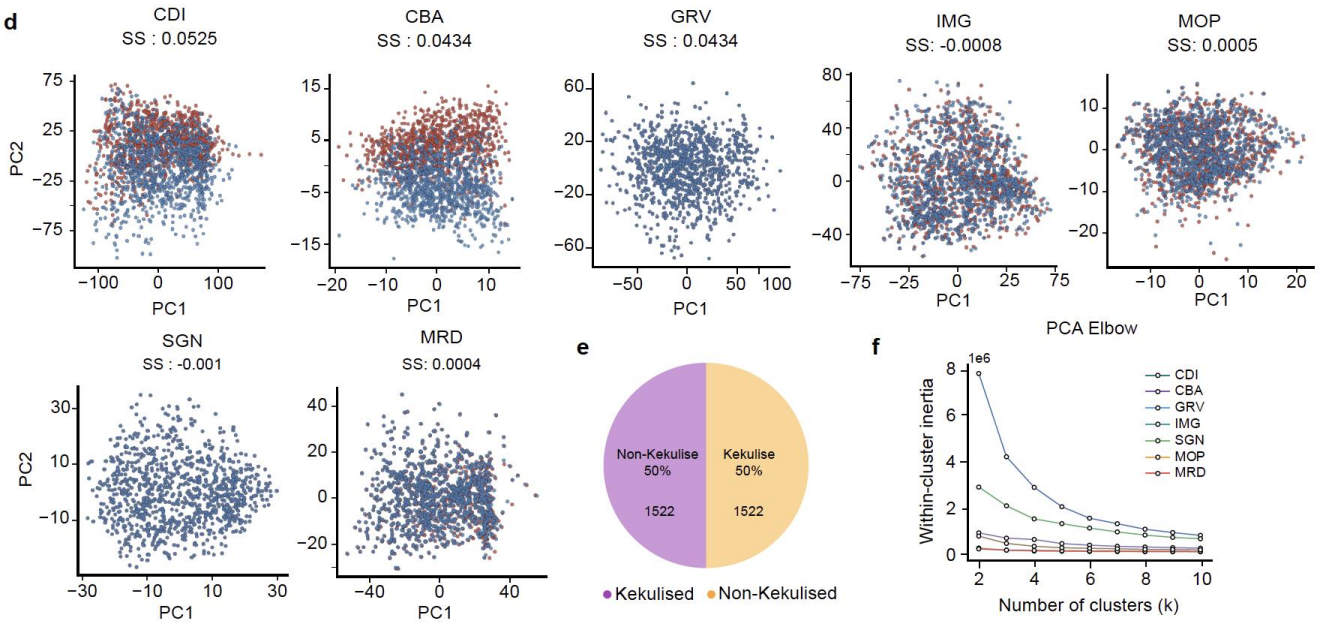
Response: *We thank the reviewer for this thoughtful and nuanced comment. We agree with the assessment regarding kekulization and have revised the manuscript accordingly.*

1) ***Positioning of kekulization results****: We acknowledge that kekulized and non-kekulized representations often correspond to the same underlying aromatic structure and that distinguishing them is not necessarily indicative of deeper chemical understanding. We have therefore toned down the interpretation of this result and clarified that it primarily reflects sensitivity to representational differences rather than a strong indicator of chemical intelligence. In line with the reviewer's suggestion, we have moved the kekulization analysis to the Supplementary section and reduced emphasis in the main text. The discussion has been revised to avoid overstating its significance.*
2) ***Emphasis on stereochemistry****: We agree that the stereochemistry (chiral) analysis provides a more meaningful evaluation of chemical awareness. Accordingly, we have strengthened its presentation and interpretation in the main manuscript, as this better reflects CDI's ability to capture non-trivial structural distinctions.*

We thank the reviewer for this constructive suggestion, which has helped improve the clarity and balance of our presentation.

7 - On a related note, coloring in Figure 4d is difficult to distinguish.

Response: *We thank the reviewer for pointing this out. We have revised the color scheme in Figure 4d* ***(Supplementary 9d)*** *to improve visual clarity and distinguishability between classes. Specifically, we increased contrast and ensured sufficient separation between hues for clear interpretation. The updated figure has been included in the revised manuscript.*



8 - It is not unlikely that researchers will try to use this codebase as a basis for integrating their own set of descriptors. Perhaps a note in the discussion on how to do this would help.

Response: *We thank the reviewer for this practical suggestion.*

1. ***Extensibility of the framework****: We clarify that CDI is modular by design, and additional modalities ("dice") can be incorporated by extending the input feature set and adding corresponding SCA blocks within the existing architecture. We have also provided the code to add or remove modules on GitHub.*
2. ***Implementation guidance****: In the revised manuscript and documentation, we have added a brief guideline outlining the integration process: (i) generate the new descriptor/embedding, (ii) align it with existing modalities at the molecule level, (iii) include it as an additional input branch in the SCA stage, and (iv) retrain the CDI-Basic model to incorporate the new modality.*
3. ***Practical note:*** *We also highlight that the CDI-Generalized model would require retraining to reflect the updated embedding space, as it learns a direct mapping to the fused representation.*

These additions have been included in the Discussion and documentation to facilitate adoption and extension by the community.

9 - Python and R implementations are provided, which is great. I also appreciate the efforts in containerization. I tried the code from PyPi, and it works perfectly. Thanks for working towards FAIRness.

Response: *We thank the reviewer for this positive feedback and for testing the PyPI package. We are pleased that the implementation worked as expected and that the Python and R support, along with containerization, was useful. Ensuring reproducibility and ease of adoption were key objectives of this work. We will continue to maintain and improve the codebase in line with FAIR principles.*

Reviewer #3 (Remarks on code availability):

I have used successfully the code on PyPi. The README file on GitHub is clear and, overall, code is of high quality.

Response: *We thank the reviewer for testing the PyPI package and for the positive assessment of the code quality and documentation. We are pleased that the README and overall implementation were found to be clear and effective. Ensuring usability, clarity, and reproducibility was a key focus in developing the codebase. We will continue to maintain and improve the repository to support the community. We have added extensive documentation for all applicable aspects of the CDI.*

1. **Code and documentation strengthened:**
   We have further improved the Data and Code Availability section to make the implementation easier to locate and reuse.

- GitHub: https://github.com/the-ahuja-lab/ChemicalDice
- Hugging Face model card: https://huggingface.co/the-ahuja-lab/ChemicalDice
- CDI Bot demonstration: https://www.youtube.com/watch?v=3NaBBTviEsA&t=1s
- Documentation: https://the-ahuja-lab.github.io/ChemicalDice/
- Docker: https://hub.docker.com/r/ahujalab/chemicaldice-app

2. **Resources provided:**
   The revised manuscript now includes links to the Python package, R implementation, GitHub repository, Docker image, Hugging Face model card, documentation, and CDI Bot demonstration. We will continue maintaining the repository and documentation to support community adoption.

Reviewer #4 (Remarks to the Author):

This work addresses a fragmentation paradox in molecular representation learning, the abundance of specialized featurizers that often lack generalizability across diverse tasks. The proposed Chemical Dice Integrator approach is a well-conceived solution to the limitations of simple feature concatenation. Benchmarking across 23 classification and 10 regression datasets demonstrates that CDI's claimed general-purpose nature yields improved performance across many tasks, including toxicity, ADMET, and physicochemical properties. However, the following specific points need to be improved.

Response: *We thank the reviewer for this thoughtful assessment. We appreciate the recognition of the "fragmentation paradox" in molecular representation learning and the acknowledgment of CDI as a structured solution beyond simple feature concatenation. We are also encouraged that the breadth of benchmarking across classification and regression tasks is viewed as supporting the general-purpose nature. We carefully address the specific points raised below and have revised the manuscript accordingly to strengthen clarity, evaluation, and positioning of the work.*

While the authors benchmark a diverse range of datasets, the underlying input for all these models remains SMILES-based. Could the authors specify the exact type of SMILES used (e.g., Canonical, Isomeric, or Randomized SMILES)? SMILES strings are notoriously sensitive to syntax. How do the models handle potential "typos" or noisy input? If the underlying featurizers (e.g., Mordred, MOPAC, GROVER) have fundamental limitations in specific chemical domains, those limitations may propagate into the unified CDI embedding.

Response: We thank the reviewer for this important clarification.

1. **Type of SMILES used:** We use canonical SMILES as the primary input representation, with stereochemical information preserved where available (i.e., isomeric SMILES). This ensures consistency across all featurizers and avoids variability introduced by randomized SMILES.
2. **Handling syntax sensitivity and noisy input:** We agree that SMILES are sensitive to syntax. All inputs are pre-validated using RDKit to ensure chemical correctness before feature generation. Invalid or malformed SMILES are filtered out at preprocessing, preventing propagation of erroneous inputs through the pipeline.
3. **Robustness to featurizer limitations:** We acknowledge that individual featurizers may have domain-specific limitations (e.g., failure cases in quantum calculations or descriptor generation). However, CDI mitigates this by integrating across modalities; no single modality is solely responsible for the final representation. Additionally, the cross-modal reconstruction objective (SCA) enables the model to learn redundancy across modalities, reducing sensitivity to failures in any one component.

We have clarified these points in the revised manuscript to better describe input handling and robustness considerations.

The performance of the Chemical Dice Integrator CDI is primarily compared within the framework itself, only comparing CDI against the six specific models it was built from, rather than against external, state-of-the-art models. I would like to suggest they provide a comparative analysis against other latest multimodal fusion models, or at least summarize the results from the related references.

The authors do not detail the internal computation of the six individual featurizers. A major bottleneck in multimodal models is the need to generate all input features during the prediction phase. For example, some molecules may not pass the quantum computation if starting from SMILES. Please summarize how many molecules pass all six featurizers.

Existing multimodal techniques include early and late fusion, hybrid fusion, and model ensembles. In a recent study, a machine learning-based fusion strategy has been explored (doi: 10.1016/j.csbj.2024.04.030). Please find other related review references.

Chemical Dice Integrator Basic will process the six type inputs. What are the dimensions and shapes of each hidden input? Could the author provide more details?

Response: *We thank the reviewer for these detailed and constructive comments. We address each point below.*

1. Comparison with external SOTA models: *We agree that comparison beyond constituent featurizers is necessary. In the revised manuscript, we have included benchmarking against recent molecular representation models (e.g., MolT5, ATOMAS, MolCA, and Uni-Mol) under a unified downstream pipeline. These models span language-based, atom-level, contrastive, and lightweight paradigms. CDI demonstrates competitive performance with consistently stronger balanced metrics and lower variance. We have also expanded the Related Work section to summarize conceptual differences (fusion vs. pretraining vs. contrastive alignment) and to contextualize CDI within the broader multimodal landscape.*

Featurizer	Dimension	Representation Type
MolT5	768	SMILES language model
ATOMAS	1024	Atom-level structural encoding
MolCA	768	Contrastive molecular representation
CDI	8192	Multimodal fused embedding
MiniMol	512	Lightweight molecular embedding
AQME		Lightweight molecular embedding

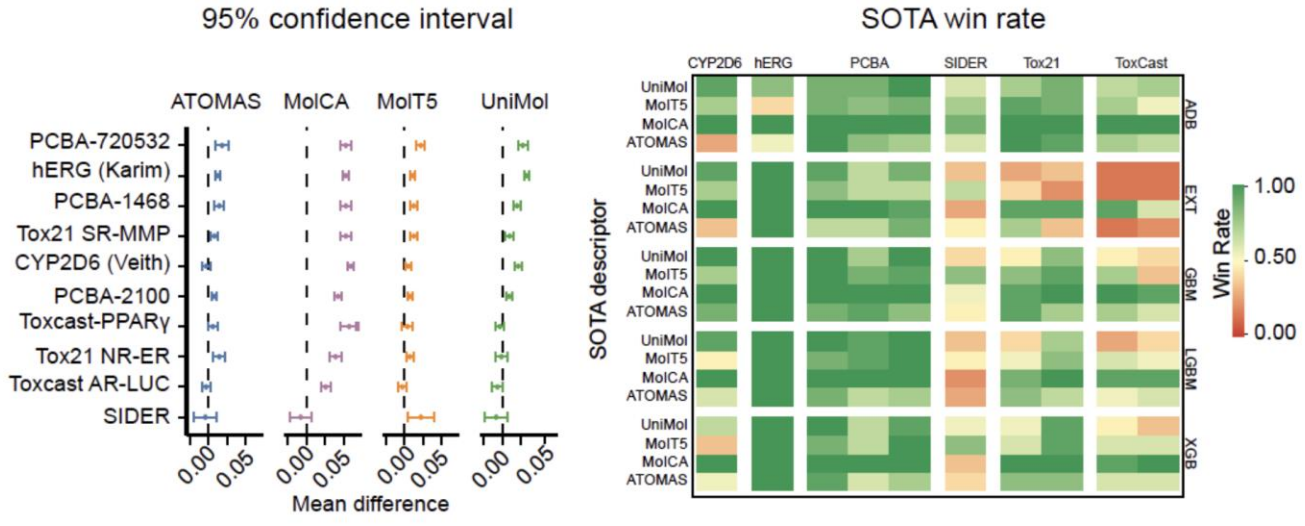


Figure: *Mean performance difference between CDI and SOTA descriptors, including ATOMAS, MolCA, MolT5, and Uni-Mol, across benchmark datasets with 95% confidence intervals. The dashed vertical line indicates no difference.*

Win-rate heatmap showing the proportion of comparisons in which CDI outperformed each SOTA descriptor across datasets and downstream models. CDI showed consistently positive mean differences and high win rates, supporting robust performance against contemporary molecular representation baselines.

2. Bottleneck of generating all modalities & coverage across featureizers: *We agree this is a critical limitation in multimodal systems. We now explicitly report featurizer coverage: while individual modalities exhibit failure cases (e.g., ChemBERTa fails on ~4,986 large SMILES; MOPAC fails on ~237k molecules), CDI achieves ~100% embedding coverage for valid SMILES. This is enabled by two factors: (i) training-time requirement of all modalities ensures a complete representation space, and (ii) inference-time decoupling via CDI-generalization removes dependency on modality generation. This directly addresses the reviewer's bottleneck.*

3. Relation to existing multimodal fusion strategies: *We have expanded the Related Work section to include classical and modern multimodal fusion paradigms, including early fusion, late fusion, hybrid approaches, and recent ML-based fusion strategies (including the cited CSBJ study and additional survey literature). We clarify that CDI differs in explicitly modeling cross-modal reconstruction (leave-one-out SCA) and hierarchical integration, rather than direct concatenation or alignment-only objectives.*

4. Details of input dimensions and architecture: *We agree that architectural clarity is important. We now explicitly report input dimensions of all modalities: MOPAC (212), ChemBERTa (384), Mordred (1421), Signaturizer (3200), ImageMol (10000), and GROVER (4800). Each SCA takes the concatenation of five modalities (excluding one) as input and learns a compressed latent representation aligned to the target modality. The six latent outputs are then concatenated and passed to the SEA to generate the final 8192-dimensional CDI embedding. All intermediate layer dimensions and encoder–decoder mappings are now clearly illustrated in Figure 1D and described in the Methods.*

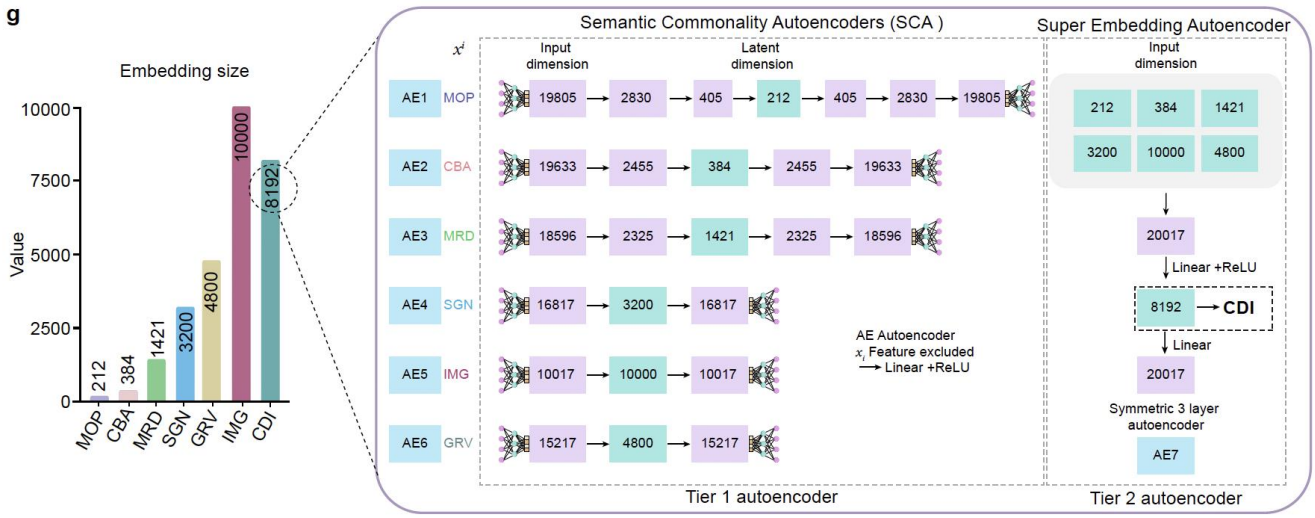


Figure: *Each SCA takes the concatenation of five modalities (excluding one) as input and learns to reconstruct the excluded modality via a compressed latent representation. The latent outputs from all six SCAs are concatenated and passed to the SEA, which generates the final unified 8192-dimensional CDI embedding.*

5. Summary of contribution relative to fusion vs. representation models: *We clarify that CDI is not a competing pretraining architecture but a multimodal integration framework. Unlike models such as MolT5 or MolCA, which learn representations from a single modality (or paired modalities), CDI fuses heterogeneous precomputed representations. The contribution is therefore complementary, focused on integration and deployment rather than modality-specific representation learning.*

We have incorporated all clarifications, additional benchmarks, and architectural details into the revised manuscript. We thank the reviewer for highlighting these important aspects, which have improved both clarity and positioning of the work.

The authors are not aware of the latest research on multimodal fusion, as they list only 23 references. To provide a more contemporary context on how different modalities—particularly those involving advanced theoretical frameworks or cross-modal interactions—are being integrated in recent literature, the authors might find it useful to reference and contrast their findings with recent work such as J. Chem. Theory Comput. 2025, 21 (14), 6743. Integrating a brief comparison with the methodologies discussed therein could highlight the specific advantages of the proposed data-aware design in this manuscript.

Response: We thank the reviewer for this important suggestion and for pointing out recent work in J. Chem. Theory Comput. 2025, 21, 6743. We agree that situating CDI within the most recent multimodal fusion literature strengthens the manuscript.

1. **Positioning relative to recent multimodal fusion work:** We have now incorporated a discussion of this study and related recent approaches. Recent frameworks increasingly emphasize data-aware, adaptive fusion, in which modality contributions are dynamically weighted or aligned via attention or cross-modal conditioning mechanisms. These methods typically rely on jointly learned encoders and require tightly coupled multimodal inputs.
2. **Conceptual distinction of CDI:** In contrast, CDI adopts a different design philosophy. Rather than learning fusion jointly from raw modalities, CDI operates on heterogeneous, precomputed representations and enforces cross-modal consistency via a leave-one-out reconstruction objective. This allows CDI to: (i) integrate modalities that are not naturally co-trained (e.g., quantum descriptors and bioactivity profiles), and (ii) remain robust when modalities are partially unavailable. This distinction aligns CDI more closely with a representation integration framework than with end-to-end multimodal pretraining models.
3. **Advantage over recent data-aware fusion approaches:** Recent works highlight key limitations of conventional fusion methods, including dependence on complete modality availability, sensitivity to modality imbalance, and computational overhead of attention-based fusion. CDI addresses these issues in a complementary manner: (i) cross-modal reconstruction reduces reliance on any single modality, (ii) hierarchical fusion stabilizes integration without requiring explicit attention weighting, and (iii) the CDI-Generalized model removes inference-time dependency on multimodal inputs entirely.
4. ***Same-modality encoder replacement and Fusilli comparison:*** *To assess whether CDI performance depends on a specific encoder choice, we performed same-modality encoder replacement experiments in which individual CDI source encoders were replaced with alternative encoders from the same representation class, while keeping the downstream evaluation pipeline unchanged. CDI showed only modest and dataset-dependent performance changes, indicating that its performance is not driven by a single encoder but by the integrated multimodal representation. We further benchmarked CDI against Fusilli-based multimodal fusion strategies, including feature concatenation, TabPFN-feature concatenation, model-averaging/attention-based fusion, and decision-level fusion. Across the same datasets, classifiers, and metrics, CDI achieved higher or comparable ROC-AUC, balanced accuracy, and composite scores than Fusilli baselines. These results show that CDI's hierarchical fusion and distillation strategy provides a more stable and predictive representation than conventional tabular multimodal fusion approaches.*

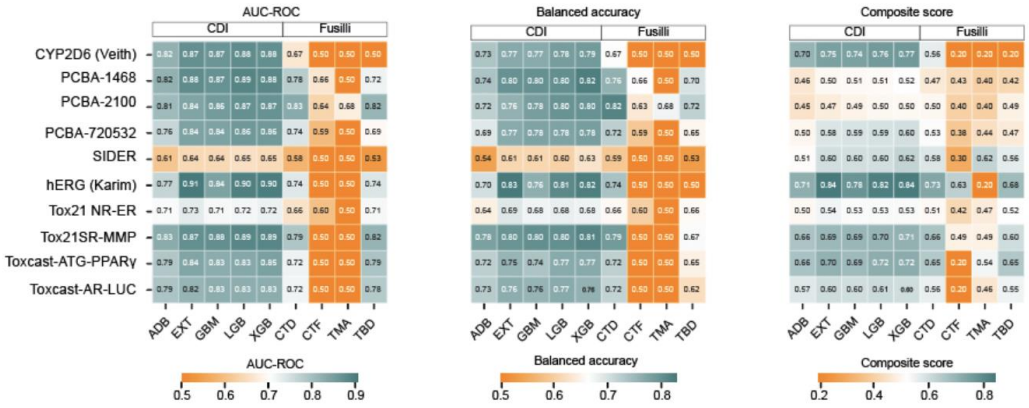


Figure**: Benchmarking CDI against Fusilli-based multimodal fusion strategies. Heatmaps comparing CDI and Fusilli-derived fusion models across ten benchmark datasets using ROC-AUC, balanced accuracy, and composite score. CDI variants include ADB, EXT, GBM, LGB, and XGB, while Fusilli baselines include CTD, CTF, TMA, and TBD. Cell values indicate mean performance for each dataset-model combination. CDI consistently achieved higher or comparable performance across metrics and datasets, indicating that CDI provides a more stable and predictive molecular representation than generic multimodal tabular fusion strategies. Also added in revised manuscript and figure **Supplementary Figure 6 h.

5. **Manuscript revision:** We have expanded the Related Work and Discussion sections to include these comparisons, clarifying that CDI complements recent advances in multimodal fusion by focusing on robust integration and deployability rather than adaptive fusion alone.

Fusion strategy	Key idea	Main limitation	CDI advantage
Feature concatenation	Directly combines all modality features	High-dimensional, redundant, noisy	Learns compact cross-modal embedding
Dimensionality reduction	Projects concatenated features using PCA/ICA/kPCA, etc.	Generic projection, weak semantic learning	Learns modality-aware semantic structure
Decision-level fusion	Combines predictions from separate modality models	Fusion happens after prediction	Integrates information before downstream learning
Attention/adaptive fusion	Learns modality weights dynamically	Often task-specific; may need all modalities at inference	Distills multimodal knowledge into SMILES-only inference

CDI	Learns and distills six-modality molecular information	Requires multimodal teacher training initially	Robust, scalable, deployable embedding for prediction, retrieval, and screening

We thank the reviewer for highlighting this relevant work, which has helped us better position CDI within the evolving multimodal molecular learning

Significant weakness of this work is its exclusive reliance on retrospective analysis of public datasets, lacking any form of prospective wet-lab validation. The authors should provide at least one case study involving experimental validation to confirm the model's ability.

Response: We thank the reviewer for this important concern regarding prospective validation. We agree that, although CDI is primarily a representation-learning and benchmarking framework, demonstrating that CDI-derived predictions can translate into experimentally measurable phenotypes is essential for establishing biological relevance. We have therefore added a dedicated proof-of-concept validation workflow linking CDI-based compound prioritization to wet-lab testing.

1. **Extension beyond retrospective benchmarking through a genomic-instability task:** To move beyond retrospective ADMET/property-prediction benchmarks, we trained a multiclass genomic-instability classifier using curated compounds labeled as genotoxic, non-genotoxic, or pro-genomic-stability. The CDI-based classifier showed strong cross-validated performance, with ROC curves supporting discrimination among the three classes. We then applied the trained model to an independent in-house chemical library of 584 compounds for external screening and candidate prioritization. To avoid data leakage, the only two compounds overlapping between the training set and screening library, β-sitosterol and L-alanine, were removed before downstream analysis, ensuring that prioritization was performed on an unseen chemical set. In addition, we used this CDI-based genomic-instability model to guide embedding-optimized molecule generation, linking the learned CDI representation to downstream molecular design. These additions have been incorporated into the Methods, Results, and Supplementary Table 20.
2. **CDI-guided prioritization of experimentally testable candidates:** From the screened in-house library, CDI identified a small pro-genomic-stability candidate subset. Based on model prediction, CDI latent-space positioning, and chemical rationale, we selected isoeugenol (IE) and eugenyl acetate (EA) for wet-lab validation. IE is a phenylpropanoid/eugenol isomer with reported antioxidant activity, while EA is an acetylated eugenol derivative related to redox-modulating phenolic scaffolds. In the CDI embedding space, IE and EA localized within the genomic-stability-associated chemical landscape, whereas individual single-modality featurizers produced more diffuse class organization. This supported the use of CDI for biologically meaningful candidate prioritization rather than only benchmark-level prediction.
3. **Wet-lab validation using a HU-induced Rad52-GFP DNA-damage assay:** We validated the prioritized compounds using the Saccharomyces cerevisiae BY4880 Rad52-GFP reporter strain, where Rad52-GFP foci serve as a marker of DNA damage-associated repair events. Hydroxyurea (HU) was used to induce replication stress because it inhibits ribonucleotide reductase, depletes dNTP pools, stalls DNA replication, and increases Rad52-GFP foci. We first standardized the HU challenge across 0.25, 0.50, 0.75, and 1.0 M HU, and selected 0.75 M HU because it produced the most robust and reproducible DNA-damage response across biological replicates. We then tested IE and EA in two complementary formats: co-treatment, where HU and compound were added simultaneously to assess protection during stress exposure, and sequential treatment, where cells were compound-treated before HU exposure to assess whether pre-conditioning

reduces subsequent HU-induced damage. A growth pre-screen identified 150 µg/ml as the working concentration for both IE and EA.

4. **Experimental outcome and interpretation:** Under optimized conditions, HU-treated cells showed prominent Rad52-GFP foci, whereas IE- and EA-treated cells showed visibly reduced foci burden. Quantification confirmed that both compounds significantly reduced HU-associated DNA-damage phenotypes relative to HU-only controls. In the sequential-treatment assay, both IE and EA significantly reduced normalized damage, supporting a rescue/preconditioning effect. In the co-treatment assay, IE produced the stronger reduction by 60 min, while EA showed a consistent but more moderate protective effect. Statistical comparisons were performed using two-sided Mann–Whitney U tests, with image-level damage proportion used as the unit of analysis.
5. **Positioning of the validation:** We present this wet-lab component as a proof-of-concept prospective validation, not as a complete mechanistic characterization of IE or EA. The primary contribution of the manuscript remains CDI's development as a general-purpose molecular representation framework. However, these experiments demonstrate that CDI embeddings can support biologically grounded candidate prioritization and that CDI-selected molecules can produce experimentally observable rescue of HU-induced genomic-instability phenotypes.

These results and the corresponding workflow have been added to the revised manuscript in **Figure 4**, **Supplementary Figure 10**, and **Supplementary Tables 16–18**. We believe this addition directly addresses the reviewer's concern and strengthens the translational relevance of CDI. We believe this addresses the reviewer's concern and strengthens the practical relevance of CDI.

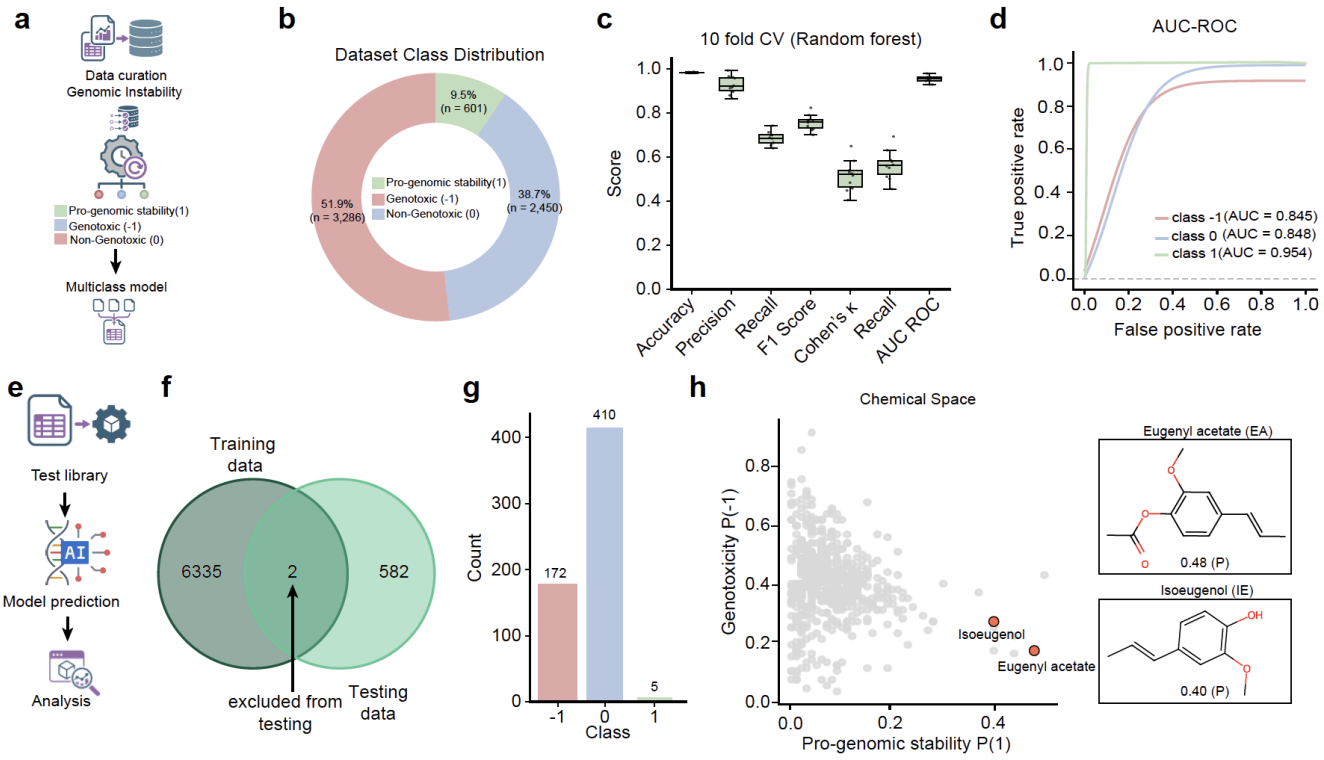


Figure: CDI-guided prediction and prioritization of genomic-instability modulators.

a: Workflow for curating genomic-instability labels and training a CDI-based multiclass classifier. **b:** Class distribution of the curated dataset comprising genotoxic, non-genotoxic, and pro-genomic-stability molecules. **c–d:** Ten-fold cross-validation performance of the Random Forest classifier and one-vs-rest ROC curves for each class. **e:** Screening workflow for applying the trained model to an in-house chemical library. **f:** Overlap analysis between

training and testing libraries, with shared compounds excluded before screening. **g:** Predicted class distribution of the screened library. **h:** CDI chemical-space projection highlighting prioritized pro-genomic-stability candidates, eugenyl acetate and isoeugenol, with predicted pro-stability probabilities and chemical structures.

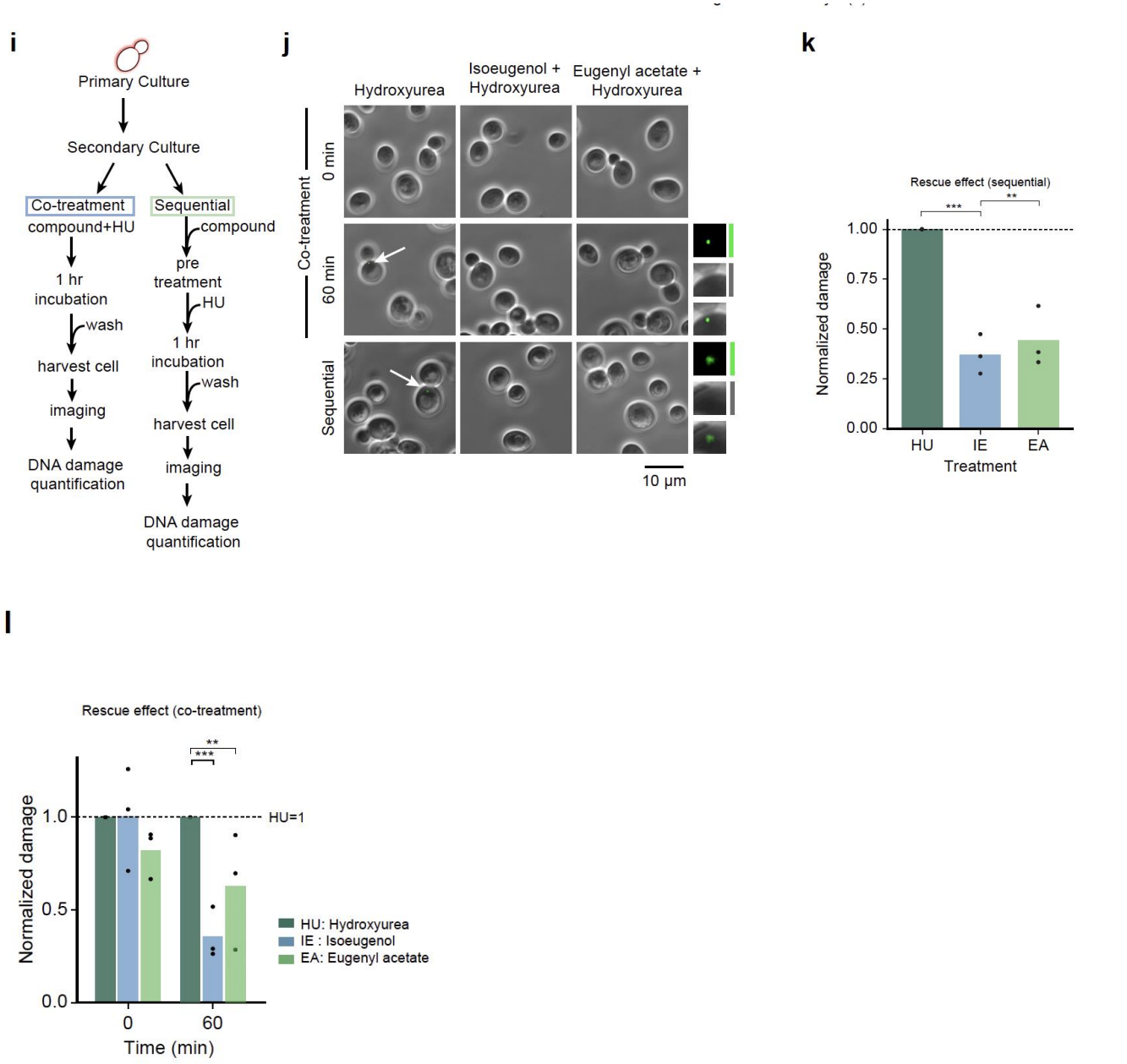


Figure : **(i)** Workflow for co-treatment and sequential-treatment assays using hydroxyurea (HU) and candidate compounds. **(j)** Representative microscopy images of HU-treated cells and cells treated with isoeugenol (IE) or eugenyl acetate (EA) under co-treatment and sequential-treatment conditions; arrows indicate RAD52–GFP DNA-damage foci. (**k**) Quantification of the sequential-treatment rescue assay showing reduced normalized DNA damage after IE and EA treatment compared with HU-only control. Statistical significance is indicated above the comparisons. **(l)** Quantification of the co-treatment rescue assay showing normalized DNA damage at 0 and 60 min after simultaneous HU and compound exposure. Isoeugenol (IE) and eugenyl acetate (EA) reduced HU-induced DNA damage at 60 min, with IE showing the stronger rescue effect**.**

Reviewer Response

Reviewer 4: Remarks to the Author: The authors have made the required revisions and have responded to my previous comments. However, I note that in their response letter, they refer to a "Related Work section" or "Related Work and Results sections". After carefully re-examining the revised manuscript, I was unable to locate such a section. I would kindly ask the authors to verify whether this section was inadvertently omitted from the submitted file or placed under a different heading.

Response: We sincerely thank Reviewer #4 for carefully re-examining the revised manuscript and for drawing our attention to this inconsistency. The reviewer is correct: the revised manuscript does not contain a separate section entitled **"Related Work,"** nor was such a section inadvertently omitted from the submitted manuscript. The references to a "Related Work section" and, in one instance, to "Related Work and Discussion sections" in our previous response letter were imprecise cross-references on our part, for which we sincerely apologize.

We have carefully verified the revised manuscript and confirm that material referred to in those responses is distributed across the Introduction, Results, Methods, and Discussion sections, rather than being presented under a standalone "Related Work" heading. Specifically, the Introduction provides conceptual background and positions CDI relative to existing molecular representations and conventional integration strategies. The Results present corresponding empirical comparisons, principally under the subsections "Distillation validation through internal benchmarking of aggregators and featurizers" and "Systematic evaluation and stress-testing of CDI representations across regimes," including comparisons with aggregation methods, constituent featurizers, contemporary molecular representation models such as ATOMAS, MolCA, MolT5, and Uni-Mol, and multimodal-fusion baselines. The corresponding methodological details are provided in the Methods under "Performance Benchmarking: Aggregator Module", "Performance Benchmarking: Featurizer Module," "Performance Benchmarking: State-of-the-Art (SOTA) Methods," and "Same-modality encoder replacement and Fusilli comparison". These subsections describe the benchmarking datasets, downstream models, cross-validation design, evaluation metrics, statistical analyses, win-rate calculations, and comparisons with both contemporary molecular representations and conventional multimodal-fusion strategies. The Discussion provides the broader conceptual interpretation and positioning of CDI, including its distinction from single-modality representation paradigms, its hierarchical cross-modal fusion and distillation strategy, and the deployment advantage of directly generating the unified embedding from SMILES without recomputing all constituent modalities at inference.

Accordingly, no manuscript section was omitted or submitted under an incorrect heading. The issue arose solely from imprecise terminology in our previous response letter. We have now revised the response letter throughout to remove references to a standalone "Related Work section" and replaced them with precise references to the relevant Introduction, Results, Methods, and Discussion sections and subsections. We sincerely thank the reviewer for identifying this issue and enabling us to improve the accuracy and clarity of our response.