nature portfolio

Peer Review File

Impact of molecular multimodality on neural network models for prediction tasks related to drug discovery

Corresponding Author: Mr Marcos Martinez Galindo

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)
The paper presents a comprehensive study on the impact of integrating neural network models from multiple molecular modalities, including SMILES, SELFIES, 2D images, and 3D geometries. The experiments are performed on 2 major tasks in drug discovery, i.e., drug-target binding affinity prediction and molecular property prediction. The paper performs analysis on different model architectures (OK and SNN), fusion strategies (intermediate and late fusion), modality combinations, and dataset splits. The extensive results and key observations may potentially provide researchers with a roadmap about how to choose and integrate multiple modalities in real-world practices.

However, several critical issues have hindered my recommendation for the publication of the paper. While the work is aimed at studying “molecular multimodality”, detailed elaboration and deeper insights are lacking. See below:
(1) The paper seems to focus merely on multi-model ensembling. The inductive bias [1] of each modality, the intuitive connections between molecular modalities and downstream prediction tasks should be discussed.
(2) The key observation, i.e. the best ensemble strategy relies on benchmark, is rather vague and trivial. In-depth analysis on the experiment results for each modality and their combinations should be discussed, providing readers with clear guidance of which models and fusion strategies are recommended in practical scenarios.
(3) In Lines 716-717 and Lines 762-766 the authors discuss the complementarity of different modalities, and further justifications are required. Concerning model predictions (late fusion), it seems in Figure 4 (a)(d)(g) that the complementarity arise from the pre-trained models (possibly due to different pre-training data and strategies) instead of modalities. Concerning model representations (intermediate fusion), the authors should perform analysis on the cosine similarity between structurally and functionally similar molecules [2].
(4) Why do the authors choose 3 models for 1D SMILES strings and 1 model for the other modalities? Are they truly representative?
(5) 2D graphs is a significant molecular modality, and models such as GROVER [3] and GraphCL [4] have demonstrated great effectiveness in molecular property prediction. Why do authors overlook this modality?
(6) The OK architecture introduces a GNN to enhance the molecular representations with a KG. Can we consider KG as another modality and the GNN as an advanced fusion architecture? How is the performance compared with KGE-NFM [5] which performs intermediate fusion on molecular fingerprint features and knowledge graph embeddings? Can KG-enhanced embeddings improve molecular property prediction, as suggested by previous works [6]?
(7) Have authors tried other intermediate and late fusion strategies, such as mixture-of-experts [7] and majority voting?

Besides, some experiment results and technical details within the paper is confusing or questionable. See below:
(8) In Fig.1, the ELO scores for OK-LF and SNN-LF are exactly the same, and I’m not sure if it’s a typing error. Besides, the authors should explain how the model pool is created to calculate the ELO score, and why using 1 modality achieves better mean score but lower ELO score compared with using 3 and 4 modalities.
(9) The metrics reported in Table 1 and Table 2 are not clear.
(10) In Figure 2 and Figure 3, the authors group molecular property prediction tasks based on the evaluation metric, which compromises deeper insights. I recommend grouping the tasks based on their types and biological meaning.
(11) The contribution of each modality in Figure 4(b)(c) are mostly within [-0.01, 0.03]. Are the results statistically significant across different random seeds?
(12) Details of the OK architecture are confusing, and the term “modality” in section 4.2.1 seems to derive from [9], which differs from the definition of this paper. I recommend authors to provide an illustration of the network architectures and fusion

strategies.
(13) In Lines 196-197 the authors claims that “late fusion approach is computationally less expensive”, which contradicts Lines 710-712 where “late fusion significantly increases computational costs”.
(14) In Lines 859-861 the authors claim that they use Pearson correlation coefficient instead of MSE or MAE as the loss function for regression. Is the score computed within the batch, and if so, what’s the batch size? Why do authors choose this loss function?
(15) Lines 865-867 mention that the authors train their model for 25,000 epochs. Do they observe over-fitting issues?
Minor:
(16) The best result of the 3rd row in Table 1 is not bolded. The decimal places in Table 1, Table 2, and the Table in Figure 1 and Figure 2 are not consistent.
(17) Lines 190-198 are identical with Lines Lines 448-455.

References
[1] Relational inductive biases, deep learning, and graph networks
[2] The Platonic Representation Hypothesis
[3] Self-Supervised Graph Transformer on Large-Scale Molecular Data
[4] Graph Contrastive Learning with Augmentations
[5] A unified drug–target interaction prediction framework based on knowledge graph and recommendation system
[6] Toward unified ai drug discovery with multimodal knowledge
[7] FuseMoE: Mixture-of-Experts Transformers for Fleximodal Fusion

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3

(Remarks to the Author)
Overall Summary:
Galindo et al. investigated two different ways of combining ("fusing") multiple molecular representations to improve the prediction of target binding affinities and molecular properties. They find that such fusion can improve the predictive performance of ML models on the investigated tasks. Yet, the optimal combination of representations and fusion varies depending on the benchmark.
The manuscript presents an extensive dataset, but the sheer volume of results obscures the central scientific narrative. While the thoroughness of the experimental work is commendable, readers struggle to identify the key findings and understand their broader significance among the numerous analyses presented.
The manuscript suffers from several structural issues that impede effective communication of the research. The organisation lacks a clear logical flow that would guide readers through the complex experimental design to the main conclusions. Furthermore, the tables and figures, while containing valuable data, require substantial revision in both content organisation and visual presentation to meet the standards expected for a high-impact journal.
From a scientific perspective, while the finding that optimal representation and fusion strategies vary across benchmarks is methodologically sound, it represents an incremental advance rather than a significant breakthrough that would warrant publication in Nature Communications. The study confirms what practitioners in the field might reasonably expect - that different approaches work better for different datasets - but does not provide sufficient novel insights or theoretical understanding to justify publication in a journal of this caliber.
The combination of presentation challenges and limited novelty leads me to recommend rejection. However, I believe this work could potentially be suitable for a more specialised journal, such as JCIM, RSC CompChem or MolInf, after substantial revision to clarify the narrative and highlight the most significant contributions.

Methodology and experimental design:
The addition of the Otter-Knowledge graph to the assessment does not help the clarity. Also, it's not discussed why the authors chose to do so.
There is no discussion around how and why the seven representation were selected. This should be mentioned, together with the dimensionalities of each representation.
For drug-target interaction prediction, the authors use a single method to represent proteins. Were other methods assessed or why was ESM chosen?
It is mentioned that hyperparameter optimisation was performed for ADMET benchmarks. And for the other? What was optimised exactly?
Simple neural network models: please do use a clearer description, such as fully-connected feed forward neural networks.

Results and interpretation:
The amount of results presented in this study is vast. Even though this showcases the thorough investigation of fusing molecular representations for prediction tasks, for me as a reader it is hard to follow the results. Some are presented in tables, some in figures and some even in bullet-point lists in the appendix. Together with the finding, that the optimal combination of representations and fusion varies depending on the benchmark and anyway needs to be investigated on a case by case basis, the reader is not really motivated to dive into the results in the current form.

Novelty and significance:
The main novelty of this study lies in the investigation of optimal fusion strategies for molecular representations.
Yet, many of the recent publications in the field do use graph or set-based representations of small molecules. However, the authors do not include any graph-based representation in their assessments. This should be changed, especially because some of them perform well in comparable benchmarks.

Clarity, organisation, writing and presentation:
In general, the overall experimental approach and workflow are difficult to follow throughout the manuscript and would benefit from clearer organisation. One can get lost in the details and can't see the forest for the trees, and several sections contain text that would belong in other sections. For example, the introduction already contains discussion of findings (page 4) that should be moved to the results / conclusion section. This also leads to repetitive writing in the results section, e.g. page 4 line 168++. Further, the results section describes many methods that are later repeated as well.
Also, the numerous abbreviations make it hard to follow which representation performs well. I suggest to at least either add some sort of classification to the results tables, or explain the abbreviations again in the table headers.

Figures and tables:
- I do miss a "Figure 1" that visualises the study approach. This would help readers to better understand what was done in this study.
- Figure 2 and 3 both contain a table and figures, these should be divided into a separate table and figures for clarity. The explanations of what is what in the table (page 5, line 219) should go into the caption. Also, it's not clear what the "Score" on the x-axis exactly is. Is higher or lower better?
Further, the text and box plots in the figure part is too small to read.
Figure 4, even though informative, is overwhelming. Text is way to small, and a color scale should be added.

References:
In my opinion, some recent key publications of the field of molecular representation learning are missing, e.g.
- https://doi.org/10.1038/s41467-023-42328-w
- https://doi.org/10.1039/C8SC04175J
- https://doi.org/10.1038/s42256-024-00856-0

Further comments and questions:
- A drug is an approved molecule that does not need any further improvement or predictions. I recommend changing the word drug to molecule (e.g. page 2 line 100).
- A SotA representation in molecular property prediction is the combination of ECFP fingerprints with all calculable RDKit descriptors. I suggest adding the RDKit descriptors to the evaluation.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I would like to thank the authors for their thorough response and the significant revisions made to the manuscript. The updated draft is considerably stronger, and I appreciate the clarity provided for points (1), (6)–(9), and (11)–(17), which I now consider fully addressed.
However, several critical points remain partially resolved. To ensure the study meets the rigorous standards of Nature Communications and provides maximum utility to the community, I invite the authors to address the following follow-up queries:

(1) To provide a more definitive guide for practitioners, I suggest expanding Table D5 to include the results of the optimal fusion strategies alongside previous State-of-the-Art methods across all 25 benchmarks.
(2) The results presented in Figure 7 reinforce my concerns that the performance gains may stem from the diversity of models rather than the integration of distinct modalities.
I suggest the authors to perform a systematic pairwise comparison between different models. Specifically, compare a control group (e.g., combinations consistently incorporating Model A) against an experimental group (consistently incorporating Model B). This ablation is necessary to determine if adding a new modality offers a genuine advantage over simply adding a different model based on the same modality.
(3) The manuscript posits that information from 2D graphs is already subsumed within 2D spatial representations and 3D geometric structures. Please provide empirical evidence, which I think a targeted experiment comparing UM vs. GROVER [1] + UM, and IMC vs. GROVER + IMC should suffice.
(4) The authors should supplement their current reporting with relative rank distributions for each task group. Shifting the focus from raw values to ranking will provide a more statistically grounded view of the model's relative performance and consistency.

References
[1] Rong Y, Bian Y, Xu T, et al. Self-supervised graph transformer on large-scale molecular data. Advances in Neural Information Processing Systems, 2020, 33: 12559-12571.

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3

(Remarks to the Author)
I thank the authors for their answers to my questions. In my opinion, the manuscript has now improved in quality and readability, and I only have some minor comments on technical and formatting details:

- Lines 726 to ~736 are redundant
- Wherever possible, increase the font size / weight of the fonts in the Figures. Many of them are barely readable and hence not really publication ready.
- Some structural issues (overflowing pages, table in references) and naming convention mixups (Supplementary Information vs. Appendix)
- Several references are incomplete (issues / pages / journal missing)

Besides that, I do not have any further comments.
Good luck

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have addressed all of my concerns, and I believe the revised manuscript is now suitable for publication.

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Dear reviewers,

Thank you for your valuable feedback. We have tried to address all your problems, and we have reviewed the paper according to your comments. We have rewritten the paper to improve the clarity and the structure, the presentation of the results and the typos and comments suggested. Besides that, please find below an explanation to the other concerns.

The paper seems to focus merely on multi-model ensembling. The inductive bias [1] of each modality, the intuitive connections between molecular modalities and downstream prediction tasks should be discussed.

We provide and extensive work with thousands of experiments to address the research questions stated in the paper, beyond only multi-model ensembling. Each modality offers a distinct inductive bias, from symbolic topology to spatial geometry, and with our work we give provide valuable insights on this matter, helping readers to understand how different combinations, architectures and fusion approaches work together in different downstream tasks.

The key observation, i.e. the best ensemble strategy relies on benchmark, is rather vague and trivial. In-depth analysis on the experiment results for each modality and their combinations should be discussed, providing readers with clear guidance of which models and fusion strategies are recommended in practical scenarios.

We have added new insights and a quantitative breakdown of the relative contributions, that we think will serve valuably to future researchers.

In Lines 716-717 and Lines 762-766 the authors discuss the complementarity of different modalities, and further justifications are required. Concerning model predictions (late fusion), it seems in Figure 4 (a)(d)(g) that the complementarity arise from the pre-trained models (possibly due to different pre-training data and strategies) instead of modalities. Concerning model representations (intermediate fusion), the authors should perform analysis on the cosine similarity between structurally and functionally similar molecules [2].

Whether the complementarity arise from the pre-trained models instead of modalities is crucial and is something difficult to assess. By using 3 models based on SMILES we can have an insight, and we see that the 3 SMILES modalities behave similarly. We also performed a post-hoc meta-analysis using a surrogate model to systematically quantify

feature importance, for all modalities, architectures and fusion approaches. Although the analysis is aligned with the result, is something that needs more work.

Why do the authors choose 3 models for 1D SMILES strings and 1 model for the other modalities? Are they truly representative?

We included multiple models to (i) assess the impact of model choice within a single modality, and (ii) quantify inter- and intra-modality similarity, thereby evaluating whether combining complementary models or modalities provides additional benefits. SMILES are the most broadly used modality by researchers, and that's why we selected it as the modality to assess these tests with.

2D graphs is a significant molecular modality, and models such as GROVER [3] and GraphCL [4] have demonstrated great effectiveness in molecular property prediction. Why do authors overlook this modality?

The selection of the above modalities was guided by both diversity and representativeness. We included heterogeneous molecular modalities (including fingerprints, sequence-based (SMILES/SELFIES) or 1D, 2D spatial representations, and 3D geometric structure) to explore how distinct types of representations interact and complement one another. Sequence-based modalities (e.g., SMILES, protein sequences) primarily capture topological and symbolic relationships, reflecting the local chemical grammar and connectivity of molecules. Image-based and 3D structure modalities encode spatial and geometric information, which is crucial for capturing stereochemistry and binding-site complementarity in interaction tasks. Fingerprint-based representations (e.g., Morgan Fingerprint) incorporate handcrafted substructural patterns, acting as high-level summaries of chemical features. While there are several other modalities, due to the computational and time constraints, only a representative subset of modalities was analyzed. We didn't consider 2D graphs as we thought the information provided by them should be already encoded in the 2D spatial representations and the 3D geometric structure.

The OK architecture introduces a GNN to enhance the molecular representations with a KG. Can we consider KG as another modality and the GNN as an advanced fusion architecture? How is the performance compared with KGE-NFM [5] which performs intermediate fusion on molecular fingerprint features and knowledge graph embeddings? Can KG-enhanced embeddings improve molecular property prediction, as suggested by previous works [6]?

In our work, we consider a modality a way of representing a molecule, for example via SMILES or via 3D. The KG is not representing a single molecule but the domain knowledge,

and it includes multiple modalities, so we don't consider it as a modality. The GNN could be seen as an advanced fusion approach, when using in intermediate fusion, but not in the late fusion approach. We haven't compared against KGE-NFM, but according to the TDC leaderboard OK is the best model to date. Whether KG-enhanced embeddings can improve molecular property prediction is beyond the scope of this work, where we try to get an insight on the impact of the multimodality across architectures and fusion approaches.

Have authors tried other intermediate and late fusion strategies, such as mixture-of-experts [7] and majority voting?

The datasets we use in our work are regression problems, which makes majority voting unfeasible. Instead, we use mean fusion which is the most similar approach.

Again, due to the computational and time constraints we can't use all the existing methods. Specially, mixture-of-experts usually requires a lot of data, probably more than that provided by the datasets we use.

In Figure 2 and Figure 3, the authors group molecular property prediction tasks based on the evaluation metric, which compromises deeper insights. I recommend grouping the tasks based on their types and biological meaning.

The images show aggregation of experiments, and thus we can only aggregate experiments based on the metric they use, as we can't aggregate, for example, mean squared error (MSE) with area under the precision-recall curve (AUPRC).

The contribution of each modality in Figure 4(b)(c) are mostly within [-0.01, 0.03]. Are the results statistically significant across different random seeds?

As the image was overwhelming for readers, we have replaced it with easier-to-read images. Due to the computational and time constraints, we can't train different Otter-Knowledge models with different seeds.

Details of the OK architecture are confusing, and the term "modality" in section 4.2.1 seems to derive from [9], which differs from the definition of this paper. I recommend authors to provide an illustration of the network architectures and fusion strategies.

We have renamed the term "modality" to avoid confusions. We use the same architecture as [9] and we report the parameters we used. We have added an overview illustration of the approaches used.

In Lines 196-197 the authors claims that "late fusion approach is computationally less expensive", which contradicts Lines 710-712 where "late fusion significantly increases computational costs".

The added cost of moving from a single modality to a few additional ones under intermediate fusion is marginal, as seen in Figure 1, the number of parameters of the IF combination is slightly bigger than the models of the individual modalities. In contrast, late fusion requires training a separate model for each modality, substantially increasing computational demand, but offering greater flexibility: once individual models are trained, all possible combinations can be explored efficiently, and new modalities can be integrated without retraining.

So, if the trained models or the actual predictions for the single modalities are available, the cost of using late fusion is insignificant. However, if you have to train all the models for doing late fusion is significantly more expensive. In Lines 196-197 we are talking about combining all modalities, and not only the ones of different "type". In the experiments, we already had the values for the single modalities as we were doing some combinations, so combine 6 modalities instead of just 4 didn't mean train new models and thus was not expensive and doable.

In Lines 859-861 the authors claim that they use Pearson correlation coefficient instead of MSE or MAE as the loss function for regression. Is the score computed within the batch, and if so, what's the batch size? Why do authors choose this loss function?

We use MSE as the main loss function, but we use validation score (Pearson correlation coefficient) to select the best model. We have rewritten this part to make it more clear.

Lines 865-867 mention that the authors train their model for 25,000 epochs. Do they observe over-fitting issues?

We saw over-fitting issues, which led us to use model selection based on the validation score.

The addition of the Otter-Knowledge graph to the assessment does not help the clarity. Also, it's not discussed why the authors chose to do so.

Otter-Knowledge is (to date) the best performing model in the TDC Leaderboard for drug-target interaction (https://tdcommons.ai/benchmark/dti_dg_group/bindingdb_patent/) and it provides a different architecture than the Feed-Forward Neural Network, thus allowing to see if the results and the discussion applies across architectures.

There is no discussion around how and why the seven representation were selected. This should be mentioned, together with the dimensionalities of each representation.

We include discussion on how the modalities were selected in the Introduction, in the Methods section (Section 4.1) and detail of the modalities in the Appendix A.

For drug-target interaction prediction, the authors use a single method to represent proteins. Were other methods assessed or why was ESM chosen?

We focus solely on the impact of molecular modalities. The impact of protein modalities remains as future work. ESM was chosen as it is the most used method for representing protein sequences.

We want to thank again the reviewers for their valuable feedback. We hope this answers their concerns, and if not we are happy to answer any other question.

Best regards, the authors.

Table 1: Impact of adding sequence-based modalities (SMILES/SELFIES) across different experimental configurations in the DG dataset using the FFNN-IF architecture. Columns correspond to different sequence-based models.

	Original	MOLF	SMI	CB	MG
MF	0.559	0.572	0.560	0.572	0.570
IMC – MF	0.554	0.553	0.547	0.551	0.551
MF – UM	0.542	0.565	0.558	0.568	0.570
IMC – MF – UM	0.548	0.559	0.557	0.557	0.554

We thank the reviewers for their thorough evaluation of our manuscript and for their constructive and insightful comments. Their feedback has been invaluable in improving the clarity, rigor, and completeness of this work. In the revised version, we have carefully addressed all comments and incorporated additional experiments, analyses, and clarifications where appropriate. Below, we provide a detailed, point-by-point response to each comment.

Reviewer 1 - Comment (1) *To provide a more definitive guide for practitioners, I suggest expanding Table D5 to include the results of the optimal fusion strategies alongside previous State-of-the-Art methods across all 25 benchmarks.*

Response: As suggested by the reviewer, we have extended the table with all the benchmarks.

Reviewer 1 - Comment (2) *The results presented in Figure 7 reinforce my concerns that the performance gains may stem from the diversity of models rather than the integration of distinct modalities. I suggest the authors perform a systematic pairwise comparison between different models to determine whether adding a new modality provides a genuine advantage over simply adding a different model based on the same modality.*

Response: To investigate whether the improvements observed in the multimodal setting arise from the integration of distinct modalities rather than from the diversity of individual models, we conducted a controlled comparison in which models belonging to the same modality type were systematically substituted within identical experimental configurations.

Specifically, we focused on sequence-based modalities (SMILES/SELFIES representations), considering multiple encoders (MOLF, SMI, MG, CB). We used MF as the reference modality and analyzed the effect of adding other modalities (e.g., IMC, UM) while varying only the sequence encoder. This setup allows us to isolate whether performance differences are driven by the modality itself or by the specific model used to encode it.

The comparison was performed on the DG dataset, the largest dataset considered in this study, providing a reliable setting for controlled evaluation. We further restricted the analysis to the FFNN-IF architecture to minimize confounding effects from more complex components.

Table 1 reports the results. Across all configurations, sequence-based models exhibit highly consistent trends: replacing one sequence encoder with another leads to only minor and non-systematic variations in performance. In contrast, changing the modality composition (e.g., introducing IMC or UM) produces larger and more structured shifts in performance, although not all modalities contribute equally. To facilitate interpretation, these results are also visualized in Figure 1. The figure highlights that sequence-based encoders follow a consistent performance trend across model variants, whereas the addition of spatial or geometric modalities produces different effects. This behavior suggests that the dominant factor influencing the performance trend is the modality type rather than the specific encoder used to implement that modality.

To further quantify this observation, we conducted a pairwise sign analysis. When comparing models within the same modality, the results are mixed (e.g., 2–2 and 3–1 splits across pairs), with no model consistently outperforming others. This lack of transitivity indicates that performance differences are configuration-dependent and do not reflect a systematic advantage of any specific model.

In contrast, comparisons across configurations with different modality compositions exhibit a fully consistent pattern: all pairwise comparisons result in unanimous outcomes (4–0), indicating that modality changes induce systematic performance differences across all model variants.

The detailed pairwise comparisons are as follows:

- MOLF vs SMI: 4 - 0
- MOLF vs CB: 2 - 2
- MOLF vs MG: 3 - 1
- SMI vs CB: 1 - 3
- SMI vs MG: 1 - 3
- CB vs MG: 3 - 1
- MF vs IMC - MF: 4 - 0
- MF vs MF - UM: 4 - 0
- MF vs IMC - MF - UM: 4 - 0
- IMC - MF vs MF - UM: 0 - 4
- IMC - MF vs IMC - MF - UM: 0 - 4
- MF - UM vs IMC - MF - UM: 4 - 0

Taken together, these results demonstrate that the primary driver of performance differences is the modality being introduced rather than the specific encoder used to implement it. While individual models may induce small fluctuations, the dominant trends observed in our experiments are governed by

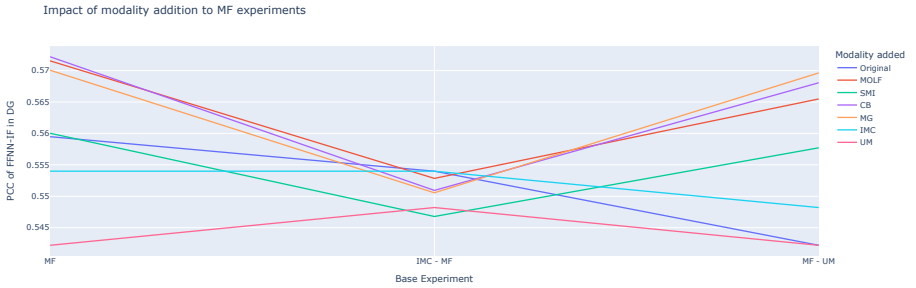

Figure 1: **Impact of sequence-based model substitution on multimodal performance.** Performance obtained adding modalities to experiments containing Morgan Fingerprint (MF) in the DG dataset using the Feed-Forward Neural Network architecture with Intermediate Fusion (FFNN-IF). Each curve corresponds to a different modality added, while the x-axis represents the experimental configuration. Models belonging to the same modality family exhibit consistent performance trends across configurations, whereas configurations involving spatial (Imagemol, IMC) and geometric (Unimol, UM) modalities display different behaviors. This suggests that the dominant performance trend is primarily associated with the modality type rather than the specific model used to implement that modality.

modality-level complementarity, supporting the interpretation that the reported gains arise from the integration of distinct molecular representations rather than from model diversity alone.

Reviewer 1 - Comment (3) *The manuscript posits that information from 2D graphs is already subsumed within 2D spatial representations and 3D geometric structures. Please provide empirical evidence, which I think a targeted experiment comparing UM vs. GROVER + UM, and IMC vs. GROVER + IMC should suffice.*

Response:
As suggested by the reviewer, and to empirically assess the contribution of 2D molecular graph representations, we conducted additional experiments incorporating a graph-based modality. As an alternative to GROVER, which checkpoint was not publicly accessible at the time of experimentation, we employed the MHG-GED model as a representative 2D graph-based approach, [1]. Due to the computational cost of the experiments, we have experimented only using the Feed-Forward Neural Network with Late Fusion (FFNN-LF) over the TDC DTI datasets, [2].

We evaluated the impact of adding this modality to configurations already including 2D spatial (IMC) and 3D geometric (UM) representations. The results show that incorporating the graph-based modality (IMC–MHG–UM) does

Modalities	DG	DAVIS			KIBA		
		Drug	Rand	Target	Drug	Rand	Target
CB - IMC - UM	0.584	0.234	0.621	0.810	0.552	0.672	0.858
IMC - MF - UM	0.595	0.248	0.618	0.817	0.557	0.684	0.863
IMC - MG - UM	0.595	0.298	0.622	0.813	0.561	0.680	0.858
IMC - MOLF - UM	0.590	0.222	0.621	0.815	0.589	0.679	0.860
IMC - SMI - UM	0.575	0.190	0.607	0.815	0.574	0.660	0.841
IMC - UM	0.565	0.194	0.617	0.803	0.503	0.667	0.848
IMC - MHG - UM	0.541	0.126	0.612	0.749	0.378	0.620	0.816

Table 2: **Impact of incorporating a 2D graph-based modality on multimodal performance.** Performance of multimodal configurations including the graph-based modality (MHG-GED) compared to baseline combinations using 2D spatial (IMC) and 3D geometric (UM) representations across DTI benchmarks. The addition of the graph-based modality does not improve performance and, in most cases, leads to a decrease in predictive accuracy, suggesting limited complementary information in this setting.

not improve performance compared to the corresponding multimodal baselines without it (e.g., IMC–UM or IMC–UM combined with other modalities). In fact, we observe a consistent degradation in performance across all evaluated benchmarks when MHG is included, Table 2.

To further understand this behavior, we analyzed the similarity between modality embeddings, Figure 2. The graph-based modality (MHG) exhibits relatively high cosine similarity with other modalities, particularly those derived from SMILES representations (e.g., MF, SMI, MOLF), indicating a substantial degree of shared information. In contrast, UM remains the least similar modality, supporting its role in providing complementary geometric information.

Taken together, these results suggest that the information captured by the 2D graph-based modality is largely redundant with that already encoded by existing representations, particularly sequence-based and 2D spatial modalities. This empirical evidence supports our original claim that 2D graph information is, to a large extent, subsumed by the combination of 2D spatial and 3D geometric representations, and therefore does not provide additional complementary signal in this multimodal setting.

Reviewer 1 - Comment (4) *The authors should supplement their current reporting with relative rank distributions for each task group. Shifting the focus from raw values to ranking will provide a more statistically grounded view of the model's relative performance and consistency.*

Response: To provide a more statistically grounded comparison across heterogeneous benchmarks, we have supplemented our analysis with a rank-based evaluation of experimental configurations. For each task, experiments are ranked

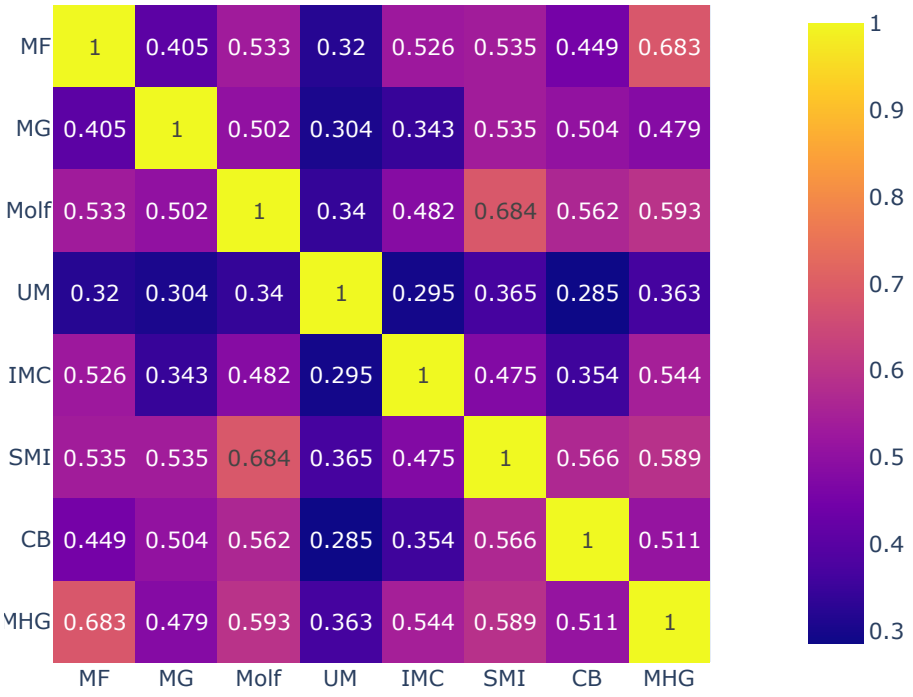

Figure 2: **Cosine similarity between modality embeddings.** Pairwise cosine similarity between learned modality representations. The graph-based modality (MHG) exhibits relatively high similarity with other representations (e.g., MF, SMI, MOLF), and specifically MHG is the most similar modality to UM and IMC, indicating overlapping encoded information. In contrast, the 3D modality (UM) remains the most dissimilar to other modalities, supporting its role as a complementary source of geometric information.

according to their performance, allowing comparisons that are independent of the scale of the evaluation metric. We then analyze the distribution of ranks across tasks for the architecture, fusion strategy, and number of modalities.

For drug–target interaction (DTI) benchmarks, the analysis (Appendix C) shows that the OK architecture and late fusion (LF) strategy consistently achieve the best ranks, and that incorporating more modalities systematically improves performance. Specifically, the mean rank decreases monotonically as the number of modalities increases, indicating that multimodal integration drives consistent improvements across heterogeneous benchmarks.

For ADMET property prediction, a single architecture is used, so the analysis focuses on fusion strategy and number of modalities. The results confirm the same trends: LF outperforms intermediate fusion (IF), and the mean rank decreases from 124.56 for unimodal experiments to 49.95 for six modalities. These findings demonstrate that the benefits of multimodal integration are robust across diverse molecular prediction tasks, and are not solely due to the choice of architecture or isolated model effects.

Due to space and figure limitations, these ranking analyses have been included in the appendix (Appendix C). We believe this additional material strengthens the statistical interpretation of our results and provides a clearer picture of the consistency and generality of the design choices investigated in our study.

Reviewer 2 - Comment (1) *Lines 726 to 736 are redundant.* **Response:**

The statements in this section refer to conclusions drawn from two complementary analyses: the first describes the similarity structure observed between modalities, while the second interprets these relationships in the context of the shared embedding space obtained after multimodal projection. Although both analyses lead to similar conclusions regarding the redundancy among SMILES-based modalities and the complementary nature of the UM representation, they originate from different experimental observations.

To improve clarity and avoid unnecessary repetition, we have revised this paragraph in the manuscript to streamline the explanation and better emphasize the distinct role of each analysis.

Reviewer 2 - Comment (2)

Wherever possible, increase the font size / weight of the fonts in the Figures. Many of them are barely readable and hence not really publication ready.

Response:

We have revised the figures to improve readability where possible, increasing font sizes and line weights in several plots. The figures are included as high-resolution vector graphics to ensure that labels remain clear when zoomed.

Some of the apparent size limitations arise from the formatting constraints of the current manuscript template, which uses a reduced page layout during the review stage. These constraints will be addressed during the production stage, where figures will be typeset according to the journal's final layout requirements. Nevertheless, we have made adjustments to improve the clarity of the figures in the revised manuscript.

Reviewer 2 - Comment (3)
Some structural issues (overflowing pages, table in references) and naming convention mixups (Supplementary Information vs. Appendix)

Response:
In the revised manuscript we have corrected the tables appearing in the reference section and standardized the terminology used throughout the manuscript to consistently refer to the additional material as the Appendix.

Some minor layout issues, such as occasional page overflows, arise from the constraints of the review template currently used for submission. These formatting aspects are typically adjusted during the production stage when the manuscript is typeset using the journal's final layout. Nevertheless, we have made minor adjustments in the revised version to improve the document structure where possible.

Reviewer 2 - Comment (4)
Several references are incomplete (issues / pages / journal missing).

Response:
We have reviewed the reference list and completed the missing bibliographic information where available (including journal names, volume numbers, and page ranges). In cases where the cited work corresponds to preprints or online repositories, we have verified that the provided citation information is consistent with the official source. The reference list has been updated accordingly in the revised manuscript.

References

[1] Kishimoto, A. *et al.* Mhg-gnn: Combination of molecular hypergraph grammar with graph neural network (2023). URL `https://arxiv.org/abs/2309.16374`. `2309.16374`.

[2] Drug-Target Interaction — tdcommons.ai. `https://tdcommons.ai/multi_pred_tasks/dti/`. [Accessed 14-05-2025].