

MolDualNet as a multi-modal architecture for small-data analog-space molecular property prediction

Corresponding Author: Dr Dan Wu

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

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The manuscript presents MolDualNet, a multimodal multitask architecture that integrates molecular graphs, string representations, expert-crafted descriptors, and bond-level geometric features via a gated bidirectional cross-attention mechanism. The model is evaluated across four endpoints: aqueous solubility, hydration free energy, lipophilicity, and BACE1 inhibitory activity, using scaffold-based splits and additional generalization settings. Overall, the study addresses an important and timely topic - the integration of heterogeneous molecular representations within a unified predictive framework. The proposed architecture is technically well-motivated, and the benchmarking against diverse baselines is a clear strength. However, several aspects of the study require further clarification and, in my opinion, additional experimental support before strong conclusions can be drawn.

The authors conclude that multimodal architectures provide robust and generally superior performance across molecular property prediction tasks. However, the evaluated endpoints span fundamentally different categories: physicochemical properties (e.g., solubility, hydration free energy, lipophilicity) and biological activity (BACE1 inhibition). These endpoints differ substantially in terms of underlying signal:

physicochemical properties are often governed by local or quasi-local interactions and solvation physics, bioactivity depends on complex, target-specific and often non-local interactions.

Therefore, it is unclear whether the observed performance gains stem from: true synergistic integration of modalities or simply increased representational capacity and exposure to broader chemical information. I recommend tempering the claims of universality or, ideally, providing a more fine-grained analysis separating these endpoint categories.

For example, it would be highly informative to compare MolDualNet against a strong single-modality model trained on substantially larger unlabeled or weakly labeled datasets (e.g., ZINC-scale data), or pretraining strategies that enrich chemical space coverage without multimodal fusion.

Given that only one of the four tasks reflects target-specific activity, the manuscript does not convincingly demonstrate the benefits of multimodality for drug discovery-relevant endpoints. I strongly encourage the authors to include experiments focusing specifically on activity-related tasks, for example:

polypharmacology settings for targets with varying degrees of similarity (e.g., closely related vs. distinct protein families).

Reviewer #2

(Remarks to the Author)

Major revision. This manuscript presents MolDualNet, a multimodal architecture that integrates molecular graphs, molecular string representations, expert descriptors, and lightweight bond-level geometry through a gated bidirectional cross-attention mechanism for molecular property prediction under distribution shift. The topic is timely and relevant, especially given the increasing interest in multimodal molecular representation learning and robustness under distribution shift. However, there

are several important concerns regarding the experimental design and baseline comparisons that need to be addressed before it can be considered for publication.

Major Concerns

1. The authors state that the model incorporates molecular string representations through a Transformer branch, but no string-based Transformer baseline (e.g., ChemBERTa, MolBERT, or other sequence-based models) is included.
2. The manuscript includes SchNet as the only 3D-aware baseline. However: SchNet (2017) is now significantly outdated relative to current 3D molecular learning methods. More recent and stronger models exist, such as: Uni-Mol, GEM. As a result, the comparison does not reflect the current state-of-the-art and weakens the validity of the conclusions.
3. A notable point of confusion is that MolDualNet adopts direct end-to-end supervised training for molecular property prediction without any pretraining, and the authors provide no explanation for this design choice. What is more curious is that the model achieves such good performance with a relatively small dataset (only 7,068 labeled molecules). Given that self-supervised pretraining has become a common strategy to enhance model performance, especially for molecular representation learning with limited data, it is recommended that the authors clarify why pretraining was not used. If necessary, please supplement pretraining experiments.
4. The manuscript includes a qualitative visualization of cross-attention maps; however, the interpretability analysis remains relatively limited and insufficiently rigorous. It remains unclear whether the learned attention patterns are consistently meaningful or merely anecdotal.
5. Figure 3 presents the same results as Table 3 just in a different for, making it somewhat redundant and unnecessarily affecting the readability of the manuscript. Please delete the duplicate figure and ensure that all figures (especially prediction-residual plots and cross-attention maps) have clear labels and legends, consistent with the journal's formatting requirements.
6. The study uses a relatively small dataset (only 7,068 labeled molecules) and only involves 4 molecular property tasks. Such limited data volume and task diversity are insufficient to fully verify the model's generalization ability, as the model's performance may not be robust when extended to more diverse molecular properties or larger, more complex datasets. It is recommended that the authors discuss this limitation and clarify the potential impact of insufficient data and task diversity on the model's practical applicability. If time is enough, please supplement relevant experiments.

The research topic of this manuscript is of great significance to the field of computational drug discovery, and the proposed MolDualNet model has certain innovation, which is consistent with the publication scope of Communications Chemistry. However, it is worth noting that the authors did not use pretraining and only used a relatively small dataset, but the model still achieved excellent performance. If this performance advantage is indeed attributed to the multi-representation fusion design of MolDualNet, it would be a remarkable achievement.

Reviewer #3

(Remarks to the Author)
Comments to the Authors

In this work, Zihan Zhang et al. introduce a MolDualNet approach that takes the molecular graph, molecular SMILES, molecular fingerprint, and some 3D-geometric representations as input and predicts the molecular property and activity. The innovation lies in combining (Fusing) these representations before doing the final prediction. The manuscript is generally clear and well-written. However, several shortcomings need to be addressed before I can recommend this manuscript for publication.

Major comments

1. Evaluation is limited by a restricted selection of datasets and comparative models.
The datasets employed for benchmarking are highly restricted (comprising only four relatively small datasets), and the comparative analysis with other ML models is similarly constrained (7 models in total, which are generally considered classic methods and don't incorporate newer, more powerful approaches).
For context, in the MolFormer paper, which was published in 2022, they have used the following datasets and models for reporting:
Datasets:
BBBP, Tox21, ClinTox, HIV, BACE, SIDER, QM9, QM8, ESOL, FreeSolv, Lipophilicity
Models:
RF, SVM, MolCLR, GEM, ChemBERTa, ...
I recommend adding more datasets and more modern models (like MolFormer) for comparison.
1- <https://doi.org/10.1038/s42256-022-00580-7>
2- <https://doi.org/10.1039/D5DD00348B>
3- <https://doi.org/10.1038/s42256-021-00438-4>
4- <https://doi.org/10.1038/s42256-022-00447-x>
5- <https://doi.org/10.48550/arXiv.2404.11568>

In this way, we can have a better understanding of how your model performs across diverse datasets (varying in size and characteristics), and also how it performs compared to the more advanced models. This provides a clearer picture to the reader and user of whether it is worthwhile to combine the modalities/representations or not.

2. The presented results are not convincing enough to use the proposed approach.

In Table 2, the authors have presented the results of applying the proposed approach to the four benchmarking datasets. For example, the reported RMSE for ESOL is 0.565. The reported RMSE in the MolFormer paper for ESOL is 0.2787. The same applies to FreeSolv data. So, in the literature, there are reported metrics that are better than your presented method without using multiple modalities and representations (by just using a pretrained SMILES-based model).

I believe that you need to carefully show what the real benefit of combining multiple modalities is compared to more advanced models that use pretrained models based on a single modality. If you can not prove (empirically) that your method has some advantages over them, I can not see that it adds much value for the readers and practitioners. So, addressing this point is very crucial for my final decision on this manuscript. This multimodal combination is the major novelty of this manuscript.

3. The advantages of a multi-task setting compared to a single-task setting are not well-studied and ablated.

If I understood correctly, the authors have used the shared encoder and specific heads for each task (dataset). Then, you have trained the model in the multi-task setting setup, where you are using a mask for a specific task. Can you quantify the clear advantage of this setting compared to training in a single-task setting? Often, multi-task learning can be helpful when the underlying tasks share some similarities. Also, it depends on the size of each task. Can you run some ablation to show which tasks, and under what conditions, this multi-task setup is advantageous compared to single-task learning? Since three of your tasks (Esol, FreeSolv, and Lipophilicity) are property predictions that share some similarities, I guess they benefit more from a multi-task setup, but in other scenarios, when you have more diverse tasks like binding affinity against different targets, I have doubts that a multi-task setup will be helpful.

P.S.: I found out you have done it for FreeSolv on Table S6 in the supplementary. I suggest having it for all tasks, and then highlighting which one benefits more.

4. The definition of the "strict scaffold-held-out" splitting method is missing.

Regarding the evaluation of generalization, the authors state: "We next evaluated generalization under two stronger settings than standard scaffold-split benchmarking: strict scaffold-held-out extrapolation for FreeSolv and Lipophilicity, and true external cross-dataset validation for ESOL and BACE."

However, the exact definition of the "strict scaffold-held-out" in the manuscript is missing. Please clarify the exact distinctions between this specific scaffold split and the conventional approach utilized elsewhere in the manuscript; furthermore, I recommend quantifying the chemical space divergence between the resulting test set and that of standard splitting methods. Can you put some figures (maybe a 2D visualization) to show how the chemical space produced by this split is different from the chemical space of the conventional scaffold splitting approach?

5. The comparison with the baseline is not completely fair.

In the Comparison with the established method section, the authors stated: "All baselines were trained per-task on identical Bemis–Murcko scaffold splits with three random seeds, using published default hyperparameters and early stopping".

So, for the baseline, authors have used per-task experiments, but for the proposed method, authors have used multitask training. Multitask training can be advantageous, specifically for the small-scale datasets like FreeSolv. So, I believe that comparing to the multitask baseline would be a fairer comparison. In that way, the effect of representation fusion, which is the main innovation, would be clearer.

6- For the generalization under the scaffold and external shift, the comparison with the baseline models is missing.

In the section on generalization under the scaffold and external shift, where the authors have introduced new test datasets, the comparison with previous baseline models is missing. I do not understand why the authors have compared their proposed method under scaffold split with previous ML models, but suddenly, in this section, they compare their model with the RDKit prediction. It is not clear how their model works in this "out-of-distribution" setting, compared to other ML models.

Minor comments

1- In Table S1, for example, for features like degree, the authors state "one-hot(0-10)," which clearly has 11 dimensions, but the "Dim" column lists 12. This is not just for the degree, but also for formal charge and others where the "Dim" value does not match the actual dimensionality.

2. You are calling the scaffold split "In-distribution scaffold-split". Why do you consider the scaffold split as the In-distribution split? How much is it different on the reported datasets from the random split?

3. When using external datasets from ChEMBL, how have you harmonized the data so that it would be comparable to the data that is coming from MoleculeNet?

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors addressed all the comments and updated the manuscript accordingly, which has now improved significantly. It can be published in the current form.

Reviewer #2

(Remarks to the Author)

Recommendation: Accept

The authors have provided a thoughtful and substantive response to comments. Every major concern raised in the initial review has been fully addressed through rigorous experimental validation, expanded analyses, and careful revision of the manuscript. The revised manuscript now presents a well-validated, clearly bounded, and scientifically rigorous contribution to the field of multimodal molecular property prediction. It meets all the standards for publication in Communications Chemistry. I recommend acceptance without further revision.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Dear Reviewers,

We sincerely thank you for your detailed and valuable comments. We are grateful that all three reviewers recognized the timeliness and innovation of MolDualNet's multimodal architecture, and your constructive feedback has directly helped improve the rigor of this work. Individual comments are listed below in italics, with our responses directly following.

Reviewer 1: We thank Reviewer for the careful reading and for encouraging us to define the scope of the multimodal claim more precisely.

Comment 1: The evaluated endpoints are heterogeneous, and the manuscript should temper claims of universal multimodal superiority.

We thank the reviewer for raising this central point. We agree. The revised manuscript no longer claims that MolDualNet provides universal or generally superior performance across all molecular property prediction settings. Instead, we explicitly distinguish physicochemical endpoints from bioactivity endpoints and separate interpolation from extrapolation regimes.

In the revised manuscript, we now state that MolDualNet's strongest evidence is concentrated in random-split analog-space interpolation, especially for small-data molecular-property tasks. We also explicitly report that under scaffold-split, OOD, and broader bioactivity evaluations, the advantage becomes task-dependent and is reversed on several settings. For example, under matched scaffold-split evaluation, MolDualNet remains competitive on small-data physicochemical endpoints but trails MolFormer-XL on Lipophilicity and BACE. In the added bioactivity panel, MolDualNet is comparable to GIN and consistently better than MolFormer-XL. We added a limitations section and a method-selection guide to make this boundary explicit. The revised message is therefore not that multimodal fusion is universally better, but that it is most useful when the evaluation regime is analog-rich small-data interpolation and when task-dependent complementary information can offset the added model complexity.

Comment 2: The observed gains may reflect increased representational capacity rather than true multimodal synergy.

We thank the reviewer for highlighting this important distinction. We therefore expanded the ablation analysis and now interpret the model in a task-dependent way.

The revised manuscript reports a five-seed ablation study removing cross-attention, expert descriptors, and bond-level geometry, as well as a graph-only control. The ablation results show that the multimodal components do not help uniformly. Removing cross-attention produces the largest average drop, but the gain is concentrated mainly in Lipophilicity and partly in BACE. On FreeSolv, simpler variants are comparable to or slightly better than the full model, indicating that the added fusion capacity is not intrinsically beneficial on very small datasets.

We also report relative training cost: the graph-only model retains approximately 98% of the full-model average R2 at approximately 18% of the training cost. We therefore explicitly state that a cheaper graph-only baseline remains defensible when rapid iteration or OOD robustness is the priority.

Comment 3: Compare MolDualNet against strong single-modality models trained on large unlabeled or weakly labeled datasets.

We appreciate this important benchmarking suggestion. We addressed it by adding pretrained SMILES Transformer baselines, including MolFormer-XL and ChemBERTa-2, under matched protocols.

In the random-split benchmark, MolDualNet is compared with MolFormer-XL, which benefits from large-scale PubChem pretraining, and with ChemBERTa-2. MolDualNet achieves the lowest RMSE on Lipophilicity, is statistically competitive with the best methods on ESOL and FreeSolv, and remains competitive on BACE, where MolFormer-XL is best. Under scaffold-split comparison, MolDualNet remains competitive on ESOL and FreeSolv but trails MolFormer-XL on Lipophilicity and BACE. We report this transparently and no longer claim a universal advantage over large pretrained single-modality models.

Thus, the revised conclusion is narrower: MolDualNet can be competitive without large-scale pretraining under matched random-split analog-space interpolation, but pretrained Transformers remain preferable for some drug-like scaffold-shifted tasks.

Comment 4: Only one of the four original tasks is target-specific bioactivity; include activity-related or polypharmacology experiments.

We thank the reviewer for this suggestion, which motivated a substantial extension of the activity-related evaluation. We added a three-target bioactivity polypharmacology panel comprising EGFR (ChEMBL203, 10,252 molecules), HIV-1 protease (ChEMBL243, 2,278 molecules), and VEGFR2 (ChEMBL279, 9,410 molecules), evaluated under matched random 80/10/10 splits across three seeds.

First, we evaluated MolDualNet in a single-task setting against matched per-target bioactivity baselines:

Target	MolFormer-XL R2	GIN R2	MolDualNet single-task R2
EGFR	0.418 +/- 0.027	0.642 +/- 0.015	0.618 +/- 0.024
HIV-1 protease	0.491 +/- 0.038	0.592 +/- 0.077	0.600 +/- 0.031
VEGFR2	0.308 +/- 0.037	0.589 +/- 0.019	0.557 +/- 0.028

These results are deliberately not presented as broad bioactivity superiority. They show that MolDualNet is competitive with a strong classical-GNN baseline (GIN) and consistently better than the pretrained SMILES Transformer MolFormer-XL on these three ligand-only bioactivity tasks.

Second, we added single-task versus multi-task comparisons for the same targets. The result is dataset-size and task-family dependent: HIV-1 protease, the smallest target, benefits substantially from joint EGFR+HIV-1 protease+VEGFR2 training (R2 improves from 0.600 +/- 0.031 to 0.711 +/- 0.021), whereas EGFR and VEGFR2 are statistically unchanged. We further tested a BACE+EGFR+HIV-1 protease companion configuration. Adding BACE reduced EGFR R2 from 0.625 +/- 0.006 to 0.571 +/- 0.006 and HIV-1 protease R2 from 0.711 +/- 0.021 to 0.672 +/- 0.057, while BACE itself remained similar across companion settings. This supports a bounded conclusion: multi-task sharing is useful when the companion tasks provide transferable signal, but it is not automatically beneficial for all bioactivity targets.

We explicitly limit the polypharmacology claim to the EGFR/HIV-1 protease/VEGFR2 panel and note that protein-aware modelling remains outside the scope of the present ligand-only architecture.

Reviewer 2: We thank Reviewer for the detailed technical comments, especially the requests for stronger baselines, clearer pretraining analysis, and more rigorous interpretability evidence.

Comment 1: *No string-based Transformer baseline was included.*

We thank the reviewer for identifying this important omission. We added two pretrained SMILES Transformer baselines: ChemBERTa-2 and MolFormer-XL. Both were evaluated under the same per-task random 80/10/10 split protocol as MolDualNet, across three random seeds, using the same data preparation and evaluation metric.

The revised Table 3 reports these baselines alongside RF, XGBoost, SVM, GIN, D-MPNN, SchNet, DimeNet++, and MolDualNet. MolFormer-XL is the best method on BACE and highly competitive on the other tasks, while MolDualNet remains competitive without large-scale pretraining and achieves the best RMSE on Lipophilicity. These results are now discussed as a regime-specific comparison rather than as a universal advantage of multimodal fusion.

Comment 2: *SchNet is outdated as the only 3D-aware baseline; newer models such as Uni-Mol or GEM should be considered.*

We thank the reviewer for this important benchmarking point. We agree that SchNet alone is insufficient as a 3D-aware comparison. We therefore added DimeNet++ as a modern directional message-passing 3D-aware baseline, evaluated under the same random-split protocol as the main benchmark and under the scaffold-split comparison.

In the random-split benchmark, DimeNet++ is highly competitive and achieves the best FreeSolv RMSE. In the scaffold-split comparison, DimeNet++ is also best on FreeSolv, whereas MolFormer-XL is best on Lipophilicity and BACE. This addition substantially changes the interpretation: MolDualNet is not presented as uniformly better than modern 3D-aware methods; instead, the revised manuscript reports complementary strengths across model families.

We cite Uni-Mol and GEM in the revised manuscript and discuss them as representative modern 3D molecular learning approaches. For the experiments, we selected DimeNet++ because it could be reproduced under our matched split and coordinate-generation protocol without requiring a pretrained checkpoint whose pretraining corpus and overlap with MoleculeNet-style benchmarks would complicate leakage control.

Comment 3: *MolDualNet is trained without pretraining; please clarify why and supplement pretraining experiments if necessary.*

We thank the reviewer for raising this important question. We added both an explanation and experiments.

First, the revised manuscript clarifies that the model is intentionally designed as a no-pretraining supervised architecture for settings where large-scale domain-matched molecular pretraining may not be available or may not transfer cleanly. Second, we added a self-supervised pretraining ablation using masked molecular-string modelling, masked atom prediction, and contrastive graph-string alignment before supervised fine-tuning.

The pretraining ablation did not yield a consistent improvement under the matched downstream protocol (reported in Supplementary Table S8). The SSL-initialized model was pretrained with the generic SSL checkpoint and then fine-tuned under the same multi-task scaffold-split protocol across 3 seeds:

Task	From-scratch MT RMSE	SSL-init MT RMSE	SSL-init R2
ESOL	0.858 +/- 0.037	0.942 +/- 0.073	0.776 +/- 0.031
FreeSolv	2.474 +/- 0.668	2.693 +/- 0.517	0.532 +/- 0.049
Lipophilicity	0.860 +/- 0.054	0.862 +/- 0.043	0.458 +/- 0.071
BACE	0.954 +/- 0.079	0.970 +/- 0.043	0.425 +/- 0.088

Thus, SSL initialization is similar to or slightly worse than from-scratch training on all four tasks in this matched scaffold-split setting. We therefore do not claim that pretraining is unnecessary in general; rather, we conclude that this particular generic SSL setup did not improve the present small-data benchmark family. We report this as a limitation and position MolDualNet as a no-pretraining alternative for analog-space interpolation, not as a replacement for large pretrained models in all regimes.

Comment 4: *The attention visualization is anecdotal; the interpretability analysis needs to be more rigorous.*

We appreciate this concern and expanded the interpretability analysis in the main text with quantitative attention diagnostics.

In addition to representative bidirectional cross-attention maps, the revised main text now includes an aggregate cross-attention entropy diagnostic computed on 200 held-out molecules per task after averaging over attention heads. Mean entropy ranges from 1.94 to 2.19 bits, corresponding to normalized entropy no higher than 0.27 relative to the 8.0-bit diffuse-attention reference. This indicates that the learned maps are concentrated rather than uniformly spread across tokens. The corresponding numerical values are also provided in Supplementary Table S10.

We have tempered the interpretation accordingly in the main text: the attention maps support chemical plausibility of graph-string alignment but are not presented as proof of mechanistic causality.

Comment 5: *Figure 3 duplicates Table 3 and should be removed; figures should have clear labels and legends.*

We revised this part carefully. After expanding the comparison to ten methods across four tasks, the numerical table became dense. We therefore retained a compact bar-chart companion to Table 3 because it allows readers to see the cross-model pattern immediately, while Table 3 provides the exact numerical values. To address the reviewer's readability concern, we revised the caption to state explicitly that Figure 3 is a companion visualization to Table 3, improved the labels and legend, and ensured that the exact values remain in the table.

The purpose of the retained figure is not to duplicate the table mechanically, but to make the cross-model comparison easier to read. In this matched benchmark, MolDualNet remains competitive against strong fingerprint, graph, 3D-aware, and pretrained Transformer baselines across all four tasks: it achieves the best RMSE on Lipophilicity among the evaluated methods, is statistically close to the best methods on ESOL and FreeSolv, and remains close to the BACE leader, where MolFormer-XL performs best. This cross-task pattern is central to the revised

narrative that no single inductive bias dominates all endpoints. If the editor prefers a shorter main text, we would be willing to move this companion bar chart to the Supplementary Information.

Comment 6: *The study uses only 7,068 labelled molecules and four tasks; discuss this limitation and add experiments if possible.*

We thank the reviewer for pointing out this limitation. We addressed it in two ways.

First, we added a three-target ChEMBL bioactivity panel, expanding the activity-related evaluation beyond BACE to EGFR, HIV-1 protease, and VEGFR2. These targets range from 2,278 to 10,252 molecules and include both kinase and protease activities. We also added matched GIN and MolFormer-XL baselines on this bioactivity panel, so the added datasets are not only descriptive extensions but also comparative benchmarks. Second, we expanded the limitations section to state explicitly that the original four-task benchmark is not sufficient to claim broad molecular-property generalization.

The revised conclusion is therefore bounded. MolDualNet is not claimed to be a universal molecular-property predictor; it is a no-pretraining multimodal architecture whose strongest evidence is small-data analog-rich interpolation, with explicit limitations under label scarcity, scaffold shift, external OOD transfer, and ligand-only bioactivity modelling.

Reviewer 3: We thank Reviewer for the constructive comments on dataset scope, matched baselines, split definitions, and OOD evaluation. These comments directly shaped the revised experimental design and the more conservative manuscript framing.

Comment 1: *The datasets and comparative models are too restricted; add more datasets and modern models such as MolFormer.*

We thank the reviewer for this important suggestion and substantially expanded both the model comparisons and the datasets.

For the four original tasks, we added modern baselines including MolFormer-XL, ChemBERTa-2, and DimeNet++, in addition to classical fingerprint models and graph neural networks. The revised random-split Table 3 now reports ten methods under the same split protocol and three random seeds.

We also added three additional bioactivity datasets from ChEMBL: EGFR, HIV-1 protease, and VEGFR2. These were evaluated under matched random 80/10/10 splits and used for both single-task and multi-task comparisons. For the bioactivity panel we additionally trained GIN and MolFormer-XL under the same per-task protocol. The revised manuscript therefore evaluates MolDualNet on seven total endpoints: four MoleculeNet property/activity tasks plus three ChEMBL bioactivity targets.

Comment 2: *Published MolFormer results appear better than the reported MolDualNet results; carefully show the real benefit of multimodal fusion against advanced pretrained single-modality models.*

We thank the reviewer for raising what we agree is the central comparison. The revised manuscript avoids comparing numbers across incompatible protocols and instead evaluates MolFormer-XL directly on our datasets under matched splitting, preprocessing, and evaluation. The results are mixed and are reported transparently. Under matched random-split interpolation, MolDualNet achieves the best Lipophilicity RMSE, is statistically close to the best methods on ESOL and FreeSolv, and trails MolFormer-XL on BACE. Under matched scaffold-split evaluation, MolDualNet is competitive on small-data physicochemical endpoints but trails MolFormer-XL on Lipophilicity and BACE. Under the added bioactivity panel, MolDualNet single-task is consistently better than MolFormer-XL and comparable to GIN. Under OOD evaluation, MolDualNet remains task-dependent and does not uniformly lead the representative baseline panel.

Comment 3: *The advantage of multi-task training over single-task training is not sufficiently studied; evaluate all tasks and identify when multi-task helps.*

We added this analysis. For the four original tasks, Supplementary Table S7 compares single-task and multi-task MolDualNet under matched scaffold-split protocol:

Task	Single-task RMSE	Multi-task RMSE	Interpretation
ESOL	0.887 +/- 0.039	0.858 +/- 0.037	modest MT benefit
FreeSolv	2.938 +/- 0.734	2.474 +/- 0.668	largest MT benefit
Lipophilicity	0.814 +/- 0.009	0.860 +/- 0.054	ST slightly better
BACE	0.961 +/- 0.020	0.954 +/- 0.079	essentially unchanged

This pattern indicates that multi-task learning helps most on smaller tasks such as FreeSolv, while it can saturate or mildly hurt larger data-rich endpoints.

We further added two bioactivity multi-task companion analyses:

Target	Single-task R2	EGFR+HIV+VEGFR2 MT R2	BACE+EGFR+HIV MT R2	Physical+BACE 4-task MT R2
EGFR	0.618 +/- 0.024	0.625 +/- 0.006	0.571 +/- 0.006	-
HIV-1 protease	0.600 +/- 0.031	0.711 +/- 0.021	0.672 +/- 0.057	-
VEGFR2	0.557 +/- 0.028	0.542 +/- 0.030	-	-
BACE	-	-	0.670 +/- 0.012	0.692 +/- 0.012

This directly supports the reviewer's intuition. Multi-task learning is task-pair specific rather than universally additive. HIV-1 protease, the smallest target, benefits substantially from the EGFR+HIV-1 protease+VEGFR2 bioactivity-only panel. In contrast, adding BACE to the bioactivity panel reduces EGFR and HIV-1 protease performance relative to the bioactivity-only companion set, while BACE itself is robust to companion choice. We therefore revised the text to state that multi-task learning is a secondary training affordance whose value depends on task relatedness and dataset size; it is not an automatic advantage over per-task models.

Comment 4: *The definition of strict scaffold-held-out splitting is missing; quantify chemical-space divergence and add a visualization.*

We appreciate this important request for clearer split definitions and chemical-space evidence. We added both the formal split description and a visual/quantitative chemical-space analysis.

The revised Methods now explicitly distinguish: (i) random 80/10/10 splitting, which probes analog-space interpolation; (ii) deterministic Bemis-Murcko scaffold splitting, used for scaffold-

split comparisons; and (iii) strict scaffold-held-out or external OOD evaluation, used for stronger distribution-shift testing. For FreeSolv and Lipophilicity, strict scaffold-held-out test molecules are selected so that their scaffolds are excluded from the training corpus. For ESOL and BACE, we use external AqSolDB and ChEMBL test sets after canonicalization and deduplication.

We also added a main-text chemical-space divergence figure. PCA projections of Morgan fingerprints show that random-split test molecules lie inside the training manifold, whereas scaffold-split test molecules occupy more peripheral or under-represented regions. Maximum Tanimoto similarity distributions quantify the same shift: for example, the median max-Tanimoto similarity drops from 0.54 to 0.28 for FreeSolv and from 0.70 to 0.53 for Lipophilicity when moving from random to scaffold splitting.

These analyses are now used to justify why random-split metrics should be interpreted as interpolation and scaffold-split metrics as a harder extrapolation regime.

Comment 5: *The comparison with baselines is unfair because baselines are trained per task whereas MolDualNet is trained in a multi-task setting.*

We agree that this needed to be separated. We now report matched single-task versus multi-task MolDualNet analyses, so that the effect of multi-task training can be distinguished from the effect of the multimodal architecture.

For the four original tasks, Supplementary Table S7 shows that multi-task training is not uniformly beneficial. It improves ESOL and FreeSolv, is essentially neutral on BACE, and slightly hurts Lipophilicity. This means the revised benchmark should not be interpreted as a simple multi-task advantage over per-task baselines. Instead, the benefit is task- and dataset-size dependent.

We also added per-target single-task MolDualNet results on the bioactivity panel and compare them directly with per-target GIN and MolFormer-XL baselines. This avoids attributing bioactivity performance to a multi-task advantage when the fair comparison is per-task. In Supplementary Table S16, single-task MolDualNet is comparable to GIN and better than MolFormer-XL on all three added bioactivity targets.

We restructured the manuscript to avoid implying that all improvements are due to multi-task learning. The revised discussion identifies the multimodal architecture as the primary contribution and the multi-task setup as a secondary training affordance that is useful in some small-data settings but not universally.

Comment 6: *The OOD section compares MolDualNet with RDKit only; comparisons with ML baselines are missing.*

We thank the reviewer for identifying this gap. We added a representative matched-protocol OOD comparison.

Supplementary Table S6 now compares MolDualNet with RF, XGBoost, SVM, GIN, D-MPNN, SchNet, ChemBERTa-2, and MolFormer-XL across four shifted settings: ESOL to AqSolDB, FreeSolv strict scaffold-held-out, Lipophilicity strict scaffold-held-out, and BACE to ChEMBL. All learning models are trained on the leakage-free merged training corpus with 815 OOD-test SMILES excluded.

The result is again mixed and reported honestly. MolDualNet is statistically close to the best methods on ESOL to AqSolDB, but it does not uniformly lead the representative baseline panel

on the other shifted settings. We therefore revised the OOD discussion to state explicitly that cross-scaffold and cross-dataset OOD generalization is not MolDualNet's strongest regime.

Minor comment 1: *Table S1 feature dimensionalities do not match one-hot descriptions.*

We corrected Table S1. The feature table now explicitly lists the additional "other" or padding categories that account for the dimensionality of degree, formal charge, number of hydrogens, hybridization, radical electrons, and chirality. The total atom-feature dimensionality is therefore internally consistent.

Minor comment 2: *Why call scaffold split "in-distribution"? How different is it from random split?*

We removed the ambiguous "in-distribution scaffold-split" wording and now explicitly separate random-split interpolation from scaffold-split extrapolation. The revised text states that random splits probe analog-space interpolation, whereas Bemis-Murcko scaffold splits and external datasets probe progressively stronger distribution shift.

We also added PCA and Tanimoto similarity analyses to quantify how random and scaffold splits differ. These analyses show that scaffold-split test molecules have lower nearest-neighbor similarity to the training set and occupy less represented regions of fingerprint space.

Minor comment 3: *How were ChEMBL external data harmonized with MoleculeNet data?*

We clarified the external-data processing. ChEMBL BACE1 activity values were processed by target ChEMBL identifier, converted from IC50 to pIC50, canonicalized by RDKit, deduplicated by canonical SMILES/InChIKey, and checked against the MoleculeNet training union. Overlapping OOD-test molecules were excluded from the leakage-free training corpus. The same canonicalization and deduplication logic was applied before external evaluation, and the excluded SMILES list and prediction files are included in the supplementary data package.

Review Comments on Manuscript COMMSCHEM-26-0279-T

Recommendation

Major revision. This manuscript presents MolDualNet, a multimodal architecture that integrates molecular graphs, molecular string representations, expert descriptors, and lightweight bond-level geometry through a gated bidirectional cross-attention mechanism for molecular property prediction under distribution shift. The topic is timely and relevant, especially given the increasing interest in multimodal molecular representation learning and robustness under distribution shift. However, there are several important concerns regarding the experimental design and baseline comparisons that need to be addressed before it can be considered for publication.

Major Concerns

1. The authors state that the model incorporates molecular string representations through a Transformer branch, but no string-based Transformer baseline (e.g., ChemBERTa, MolBERT, or other sequence-based models) is included.
2. The manuscript includes SchNet as the only 3D-aware baseline. However: SchNet (2017) is now significantly outdated relative to current 3D molecular learning methods. More recent and stronger models exist, such as: Uni-Mol, GEM. As a result, the comparison does not reflect the current state-of-the-art and weakens the validity of the conclusions.
3. A notable point of confusion is that MolDualNet adopts direct end-to-end supervised training for molecular property prediction without any pretraining, and the authors provide no explanation for this design choice. What is more curious is that the model achieves such good performance with a relatively small dataset (only 7,068 labeled molecules). Given that self-supervised pretraining has become a common strategy to enhance model performance, especially for molecular representation learning with limited data, it is recommended that the authors clarify why pretraining was not used. If necessary, please supplement pretraining experiments.
4. The manuscript includes a qualitative visualization of cross-attention maps; however, the interpretability analysis remains relatively limited and insufficiently rigorous. It remains unclear whether the learned attention patterns are consistently meaningful or merely anecdotal.
5. Figure 3 presents the same results as Table 3 just in a different for, making it somewhat redundant and unnecessarily affecting the readability of the

manuscript. Please delete the duplicate figure and ensure that all figures (especially prediction-residual plots and cross-attention maps) have clear labels and legends, consistent with the journal's formatting requirements.

6. The study uses a relatively small dataset (only 7,068 labeled molecules) and only involves 4 molecular property tasks. Such limited data volume and task diversity are insufficient to fully verify the model's generalization ability, as the model's performance may not be robust when extended to more diverse molecular properties or larger, more complex datasets. It is recommended that the authors discuss this limitation and clarify the potential impact of insufficient data and task diversity on the model's practical applicability. If time is enough, please supplement relevant experiments.

The research topic of this manuscript is of great significance to the field of computational drug discovery, and the proposed MolDualNet model has certain innovation, which is consistent with the publication scope of Communications Chemistry. However, it is worth noting that the authors did not use pretraining and only used a relatively small dataset, but the model still achieved excellent performance. If this performance advantage is indeed attributed to the multi-representation fusion design of MolDualNet, it would be a remarkable achievement.

Comments to the Authors

In this work, Zihan Zhang et al. introduce a MolDualNet approach that takes the molecular graph, molecular SMILES, molecular fingerprint, and some 3D-geometric representations as input and predicts the molecular property and activity. The innovation lies in combining (Fusing) these representations before doing the final prediction. The manuscript is generally clear and well-written. However, several shortcomings need to be addressed before I can recommend this manuscript for publication.

Major comments

1. Evaluation is limited by a restricted selection of datasets and comparative models.

The datasets employed for benchmarking are highly restricted (comprising only four relatively small datasets), and the comparative analysis with other ML models is similarly constrained (7 models in total, which are generally considered classic methods and don't incorporate newer, more powerful approaches).

For context, in the MolFormer paper, which was published in 2022, they have used the following datasets and models for reporting:

Datasets:

BBBP, Tox21, ClinTox, HIV, BACE, SIDER, QM9, QM8, ESOL, FreeSolv, Lipophilicity

Models:

RF, SVM, MolCLR, GEM, ChemBERTa, ...

I recommend adding more datasets and more modern models (like MolFormer) for comparison.

- 1- <https://doi.org/10.1038/s42256-022-00580-7>
- 2- <https://doi.org/10.1039/D5DD00348B>
- 3- <https://doi.org/10.1038/s42256-021-00438-4>
- 4- <https://doi.org/10.1038/s42256-022-00447-x>
- 5- <https://doi.org/10.48550/arXiv.2404.11568>
- 6- <https://doi.org/10.1038/s41467-025-66685-w>

In this way, we can have a better understanding of how your model performs across diverse datasets (varying in size and characteristics), and also how it performs compared to the more advanced models. This provides a clearer picture to the reader and user of whether it is worthwhile to combine the modalities/representations or not.

2. The presented results are not convincing enough to use the proposed approach.

In Table 2, the authors have presented the results of applying the proposed approach to the four benchmarking datasets.

For example, the reported RMSE for ESOL is 0.565. The reported RMSE in the MolFormer paper for ESOL is 0.2787. The same applies to FreeSolv data. So, in the literature, there are reported metrics that are better than your presented method without using multiple modalities and representations (by just using a pretrained SMILES-based model).

I believe that you need to carefully show what the real benefit of combining multiple modalities is compared to more advanced models that use pretrained models based on a single modality. If you can not prove (empirically) that your method has some

advantages over them, I can not see that it adds much value for the readers and practitioners. So, addressing this point is very crucial for my final decision on this manuscript. This multimodal combination is the major novelty of this manuscript.

Table 2 | Performance of fine-tuned MOLFORMER and other supervised GNN baselines on QM9, QM8, ESOL, FreeSolv and Lipophilicity regression benchmarks

	QM9	QM8	ESOL	FreeSolv	Lipophilicity
GC	4.3536	0.0148	0.970	1.40	0.655
A-FP	2.6355	0.0282	0.5030	0.736	0.578
MPNN	3.1898	0.0143	0.58	1.150	0.7190
MOLFORMER-XL	1.5894	0.0102	0.2787	0.2308	0.5289

For QM9 and QM8, we report average MAE, while root-mean-square error is reported for the remaining tasks. Baseline performances are taken from refs. [28](#),[40](#). Bold indicates the top-performing model.

3. The advantages of a multi-task setting compared to a single-task setting are not well-studied and ablated.

If I understood correctly, the authors have used the shared encoder and specific heads for each task (dataset). Then, you have trained the model in the multi-task setting setup, where you are using a mask for a specific task. Can you quantify the clear advantage of this setting compared to training in a single-task setting? Often, multi-task learning can be helpful when the underlying tasks share some similarities. Also, it depends on the size of each task. Can you run some ablation to show which tasks, and under what conditions, this multi-task setup is advantageous compared to single-task learning? Since three of your tasks (Esol, FreeSolv, and Lipophilicity) are property predictions that share some similarities, I guess they benefit more from a multi-task setup, but in other scenarios, when you have more diverse tasks like binding affinity against different targets, I have doubts that a multi-task setup will be helpful.

P.S.: I found out you have done it for FreeSolv on Table S6 in the supplementary. I suggest having it for all tasks, and then highlighting which one benefits more.

4. The definition of the “strict scaffold-held-out” splitting method is missing.

Regarding the evaluation of generalization, the authors state: "We next evaluated generalization under two stronger settings than standard scaffold-split benchmarking: strict scaffold-held-out extrapolation for FreeSolv and Lipophilicity, and true external cross-dataset validation for ESOL and BACE."

However, the exact definition of the “strict scaffold-held-out” in the manuscript is missing. Please clarify the exact distinctions between this specific scaffold split and the conventional approach utilized elsewhere in the manuscript; furthermore, I recommend quantifying the chemical space divergence between the resulting test set and that of standard splitting methods. Can you put some figures (maybe a 2D visualization) to show how the chemical space produced by this split is different from the chemical space of the conventional scaffold splitting approach?

5. The comparison with the baseline is not completely fair.

In the Comparison with the established method section, the authors stated: “All baselines were trained per-task on identical Bemis–Murcko scaffold splits with three random seeds, using published default hyperparameters and early stopping”.

So, for the baseline, authors have used per-task experiments, but for the proposed method, authors have used multitask training. Multitask training can be advantageous, specifically for the small-scale datasets like FreeSolv. So, I believe that comparing to the multitask baseline would be a fairer comparison. In that way, the effect of representation fusion, which is the main innovation, would be clearer.

6- For the generalization under the scaffold and external shift, the comparison with the baseline models is missing.

In the section on generalization under the scaffold and external shift, where the authors have introduced new test datasets, the comparison with previous baseline models is missing. I do not understand why the authors have compared their proposed method under scaffold split with previous ML models, but suddenly, in this section, they compare their model with the RDKit prediction. It is not clear how their model works in this "out-of-distribution" setting, compared to other ML models.

Minor comments

1- In Table S1, for example, for features like degree, the authors state "one-hot(0-10)," which clearly has 11 dimensions, but the "Dim" column lists 12. This is not just for the degree, but also for formal charge and others where the "Dim" value does not match the actual dimensionality.

Table S1: Atom feature encoding used by the graph branch.

Feature	Encoding	Dim.
Atom type	one-hot (C, N, O, S, F, Cl, Br, I, P, . . . , other)	22
Degree	one-hot (0–10)	12
Formal charge	one-hot (−5 to +5)	12
Number of H	one-hot (0–8)	10
Hybridization	one-hot (SP, SP ² , SP ³ , SP ³ D, SP ³ D ² , other)	7
Aromatic	binary	1
In ring	binary	1
Radical electrons	one-hot (0–4)	6
Chirality	one-hot (5 tags)	6
Atomic mass	normalized ($m/100$)	1
Ring sizes	multi-hot (3, 4, 5, 6+)	4
Total		82

2. You are calling the scaffold split “In-distribution scaffold-split”. Why do you consider the scaffold split as the In-distribution split? How much is it different on the reported datasets from the random split?
3. When using external datasets from ChEMBL, how have you harmonized the data so that it would be comparable to the data that is coming from MoleculeNet?

Review Comments on Manuscript COMMSCHEM-26-0279A

Recommendation: Accept

The authors have provided a thoughtful and substantive response to comments. Every major concern raised in the initial review has been fully addressed through rigorous experimental validation, expanded analyses, and careful revision of the manuscript. The revised manuscript now presents a well-validated, clearly bounded, and scientifically rigorous contribution to the field of multimodal molecular property prediction. It meets all the standards for publication in *Communications Chemistry*. I recommend acceptance without further revision.