nature portfolio

Peer Review File

sChemNET: A deep learning framework for predicting small molecules targeting microRNA function

Reviewers' comments:

Reviewer #1 (Remarks to the Author):

The authors describe sChemNET, a two-layer neural network model that predicts bioactivity (while not distinguishing between up- or down-regulation) against specific, predetermined miRNAs. In general the paper is well-written, the experiments are well-executed, and the results are both understandable and well-supported. However, a number of specific areas for improvement exist:

1. The authors state that only miRNA having 5 or more bioactive small molecules are selected for inclusion in the study, and these known activities are the source of training labels. What is not explicitly clear is whether those molecules also have known non-activity against other miRNA in the model (i.e. are true negatives, p(y_{iu}|x_i) == 0), or whether this model's labels are only the positive cases (i.e. true positives, p == 1). What is the extent of many-to-many molecule-to-miRNA label relationships in the training data?

2. The authors use a loss function that incorporates miRNA sequence similarity as a downweighting factor to minimize the discrepancy in predicted labels from other miRNAs that are less similar. This seems to imply that sChemNET model is not a global model learned once and over all miRNAs, but learned independently, locally and separately for each miRNA u. Or, that for those sets of miRNAs that have bioactivity against the same molecule that each prediction is counted in the loss function multiple times. Some additional clarity here would be appreciated.

3. The authors describe calculating "global similarity" using an alignment procedure from BioPython; this procedure should be described in greater detail (scoring matrix, gap penalty choices) and those parameter choices justified, given the very short sequences. Additionally, the "normalization" of these similarity scores between 0 and 1 seems to imply that a (randomly occurring) "worst" similarity would be rescaled to 0, but the procedure used for this normalization and rescaling was not provided.

4. In figures 2b and 3b, performance metrics are shown as barplots with error bars of some sort -- these are not described in the figure legend. While sChemNET with sequence similarity performs better, on average, than without sequence similarity, the error bars seem to indicate this difference isn't meaningful. In fact, most methods are only marginally better than chemical similarity; not surprising given the sparsity of training data available.

5. Figures 2b and 3b are easy to mis-interpret, given the authors choice of wording: among 100 "top" molecules, approx. 10% of bioactive molecules were identified by sChemNET -- thus, is 10% the recall (sensitivity), or the TPR of the top-100 selection? An ROC curve (with confidence intervals) would be a more informative and straightforward way to visualize performance.

6. The authors use a loss function that incorporates a regularization technique that penalizes for promiscuous (and presumably false positive) predictions across the unlabelled portion of the training set. This is meant to represent the chemical "space" in which the miRNA bioactivity prediction is optimized -- and thus the model should be relearned for any new screening library to be surveyed. Thus it remains unknown whether the various performance and learning behaviors of sChemNET presented here are specific to this particular choice of chemical space or would generalize to other chemical libraries of interest.

Reviewer #2 (Remarks to the Author):

This manuscript addresses a challenging topic, namely the interaction between small molecules and microRNA, by means of machine learning and validation. The authors are commended for doing experimental validation.

However, the number of concerns outweighs the benefits of the merits of this paper, which preclude publication at this time.

Here are some of the concerns:
1. data - the SM2miR database has not been updated since 2015; one would assume that in decade that has passed since the original publication, a lot more such molecules have been published. Some of these should have been used as a temporal split (external) validation set. More molecules here? PMC6546413
2. ML descriptors - sChemNet models are trained using the 127 features of the MACCS fingerprint. This is an unfortunate choice, given that MACCS keys in that set were built on the World Drug Index in the late 1980s, and are literally representative of "old chemistry". There is a significant number of more modern approaches to chemical features that can be used to address this question.
3. Model imbalance - there is a serious concern about the 10:1 ratio between inactive and active compounds. It is unclear if the 2400 unlabelled molecules are indeed inactive. Perhaps a more clear description would have helped; true inactives should incorporate compounds known to permeate mitochondria (e.g., metformin).
4. But the more serious problem is model validation. Calcitriol and its effect on miRNA has been disclosed in 2016 - PMC4714233. Given the extremely high similarity between calcediol and calcitriol, it is really difficult to imagine that a) this choice was accidental and b) that the authors had no knowledge of this result.

Last but not least, docetaxel is an antineoplastic taxane; if it modulates glucocorticoid receptors, please provide evidence.

Reviewer #3 (Remarks to the Author):

In this work Galeano et al. have made an attempt to develop a method to target microRNAs with small molecules by using deep learning approaches.

The key idea in this method is training a learning model by using the information of "unlabeled" structures. During the training phase each of these small molecules are assigned a prediction score to each miRNA.
There are several fundamental questions that authors need to provide a more detailed explanation and rationalization.

1. Authors claim that "Unlabeled small molecules have unknown biological activity against targeted miRNA" and then they use the Drug Repurposing Hub database for creating the unlabeled set of small molecules. It has been shown that there is a significant overlap between the chemical space of known approved drugs and RNA binders which bind to microRNA (J. Am. Chem. Soc. 2021, 143, 33, 13044–13055). Authors need to calculate the physicochemical properties of their labeled and unlabeled libraries to show the similarities and differences and then they need to map the chemical space of 6,433 small molecules to the bioactive molecules to show the unlabeled library is not already biased toward binding to microRNAs.

2. In line 155 authors use the term "unique" to describe the unlabeled small molecules. They need to provide more information on what determined this "uniqueness".

3. What is the reason behind selecting the MACCS fingerprint as it has been shown that extended-connectivity fingerprints of diameters 4 and 6 are among the best performing fingerprints in ranking diverse structures by similarity and fingerprints to avoid when measuring similarity include Daylight-type path-based fingerprints and MACCS keys (J Cheminform. 2016; 8: 36). As the scoring of the test

set is based on the extracted chemical similarity using this fingerprint it also raises another question whether the authors have benchmarked other fingerprints or not? This makes the whole scoring workflow highly questionable. Authors need to rationalize their choice of this fingerprint and then benchmark other fingerprints to validate their results.

Dear Reviewers,

We thank you for your thoughtful and productive comments, which have guided us to further improve the rigor and clarity of our manuscript. We believe the revised version of our manuscript is now much stronger due to the revisions we have done to address your comments.

The original comments of the reviewers are listed in black; our responses are written in blue.

On behalf of the authors,
Diego Galeano, Ph.D.

Responses to Reviewer #1

The authors describe sChemNET, a two-layer neural network model that predicts bioactivity (while not distinguishing between up- or down-regulation) against specific, predetermined miRNAs. In general, the paper is well-written, the experiments are well-executed, and the results are both understandable and well-supported. However, a number of specific areas for improvement exist:

We thank the reviewer for praising that our paper is "*well-written, the experiments are well-executed, and the results are both understandable and well-supported*". Nevertheless, we took very seriously the reviewers' comments and aimed at addressing them to the best of our abilities.

1. The authors state that only miRNA having 5 or more bioactive small molecules are selected for inclusion in the study, and these known activities are the source of training labels. What is not explicitly clear is whether those molecules also have known non-activity against other miRNA in the model (i.e. are true negatives, p(y_{iu}|x_i) == 0), or whether this model's labels are only the positive cases (i.e. true positives, p == 1). What is the extent of many-to-many molecule-to-miRNA label relationships in the training data?

The experimental small molecule - miRNA association data that we used in our study from the SM2miR database only contains associations that are true positive cases (i.e. small molecules that are known to experimentally alter the expression of miRNAs). Unfortunately, experimentally validated true negative associations (i.e. p == 0) are not reported on public datasets. Therefore, for instance, all the 1,102 associations between 131 small molecules and 126 miRNAs from *Homo sapiens* are a source only for positive labels. Negative labels, as such, are not explicitly provided in the database.

The problem of having databases containing only positive drug-target interactions is not unique to the SM2miR database, but it is a challenge for many researchers engaged in drug target predictions, drug side effect predictions, and protein-protein interactions. In such studies, typically, “unknown” or “missing” associations in the dataset are represented with zero values[1–3] ; this is also the representation we used in our modelling.

To estimate the extent of the many-to-many molecule-to-miRNA label relationship in training labeled data, we analyzed the *Homo sapiens* labeled data, which consists of 1,102 associations between 131 small molecules and 126 miRNAs. We first check the distribution of the number of shared bioactive small molecules (i.e., $y_{ij} = 1$) for pairs of miRNAs – see **Figure R1a**. The average number of shared bioactive small molecules for the 7,875 pairs of miRNAs is 1.77+-1.51 (mean and s.t.d.), indicating that most miRNAs in our dataset tend to share few bioactive small molecules. A similar trend is observed for the distribution of the number of shared miRNAs for each of the 8,515 pairs of bioactive small molecules (mean and s.t.d. of 0.63 +- 1.91) – see **Figure R1b**.

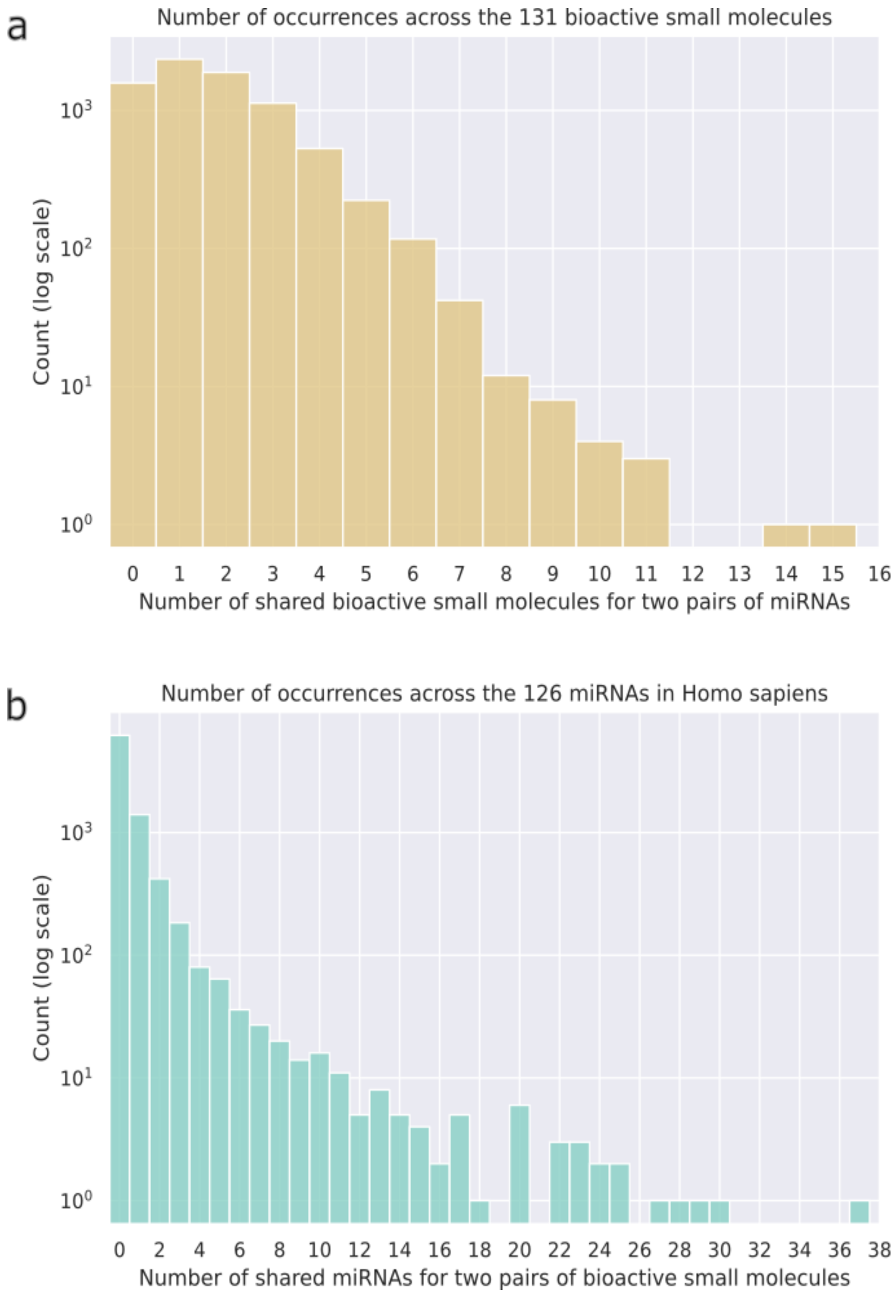

Figure R1 - Histograms of labeled relationships in the training dataset for 126 miRNAs from Homo sapiens. (a) The number of bioactive small molecules shared between 7,875 pairs of miRNAs; (b) The number of miRNAs shared between 8,515 pairs of bioactive small molecules.

Changes to the manuscript	**Changes to the Supplementary Materials**
New explanations in lines 147-149 and lines 154-156.	new **Supplementary Fig. 3** (**Fig. R1** in this document)

2. The authors use a loss function that incorporates miRNA sequence similarity as a down weighting factor to minimize the discrepancy in predicted labels from other miRNAs that are less similar. This seems to imply that sChemNET model is not a global model learned once and over all miRNAs, but learned independently, locally and separately for each miRNA u. Or, that for those sets of miRNAs that have bioactivity against the same molecule that each prediction is counted in the loss function multiple times. Some additional clarity here would be appreciated.

We apologize for the lack of detail that in turn affects the clarity of the manuscript. To clarify how sChemNET works, we have now added a new paragraph in the **Methods section** describing "**The sChemNET model training**".

Briefly, sChemNET is not a global model, and prediction scores for a given miRNA u are independently learned (or optimized). Still, sChemNET uses all the labeled information available from all other miRNAs in a way that their contribution to the learning is weighted according to how similar they are in sequence to miRNA u – the miRNA for which we are learning the prediction scores.

The down-weighting based on sequence similarities is indicated in our loss function shown in Eq. (1) in the main manuscript. The parameter s_{uv} allows us to downweigh each miRNA's labeled information's contribution based on their similarities in sequence. Although sChemNET requires a separate training for each miRNA u to incorporate sequence similarity information, we found that it only takes a few seconds to run sChemNET on a laptop, and thus does not represent an issue in practice.

In the particular case in which sChemNET is trained without sequence information (i.e. $s_{uv} = 1$ for all miRNA pairs), sChemNET could be trained only once to generate prediction scores for all miRNAs.

For the sets of miRNAs that have bioactivity against the same small molecule, their contribution would be counted multiple times but with different penalization weights depending on the similarity in sequence to miRNA u. This may add evidence during learning that improves the prediction score for a given small molecule.

Changes to the manuscript	**Changes to the Supplementary Materials**
additional paragraph in Methods section "**The sChemNET model training**". Lines 651-655.	-

3. The authors describe calculating "global similarity" using an alignment procedure from BioPython; this procedure should be described in greater detail (scoring matrix, gap penalty choices) and those parameter choices justified, given the very short sequences. Additionally, the "normalization" of these similarity scores between 0 and 1 seems to imply that a (randomly occurring) "worst" similarity would be rescaled to 0, but the procedure used for this normalization and rescaling was not provided.

Please excuse this oversight. In the revised version of the manuscript, we have provided the details of the alignment procedure from BioPython in the **Methods section "miRNA sequence similarity and linear re-scaling for sChemNET loss function".** Given that miRNA sequences are very short and of similar length, we followed previous computational modelling work[4,5] and used global alignment to build a sequence similarity matrix using the Needleman-Wunsch algorithm (with a match score of 1, and mismatch and gap scores of zero). In miRNA mature sequences, we typically do not have gaps; therefore, using gap scores of 0 is the appropriate choice.

In computational modeling in machine learning, it is common to re-scale similarity values when used as penalization factors in a loss function – such as the one we developed in Eq. (1) for sChemNET – see examples in drug target prediction machine learning models[2]. This is done for practicality; larger penalization values can lead to issues in the optimization algorithms used for learning. Our re-scaling does not alter the similarity information because, in our cost function, what matters is that the relative importance of the information is preserved – which is our case because we used a linear re-scaling of the values. That is, for a given miRNA u, let $z_u = [z_{u1}, z_{u2}, \dots, z_{un}]$ be the sequence similarity score to all other miRNAs. We performed the following linear re-scaling of the sequence similarity values:

$$s_u = m \times z_u + (1 - m \times max(z_u))$$

where the $m = \frac{1-a}{max(z_u)-min(z_u)}$ is the slope and a is a constant value that needs to be defined to set the minimum value of the re-scaling. For the *Homo sapiens* chemical RNA dataset, we found that $a = 0$ works well ($0 \leq s_u \leq 1$), but for the model organisms, we used $a = 0.7$ ($0.7 \leq s_u \leq 1$). This difference might be due to the fact that for the model organisms the datasets available were much sparser than for *Homo sapiens*, and the model needed to rely more on the other labeled information to achieve optimal prediction performance.

As an example of the re-scaling performed on the sequence similarities, we showed in **Figure R2** the re-scaling performed for miR-let-7-5p on the *Homo sapiens* dataset.

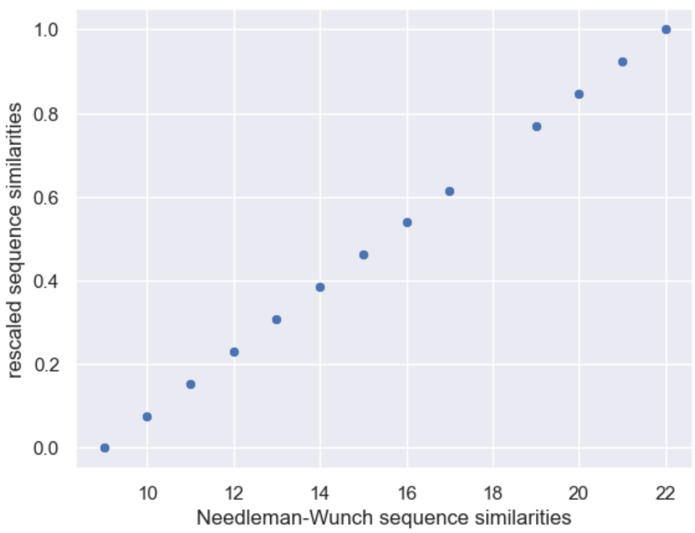

Figure R2. Example of re-scaling performed on the mature sequence similarities between miRNA hsa-let-7-5p and all other 125 miRNAs from Homo sapiens. The y-axis indicates the rescaled values that were used in the sChemNET loss function, and the x-axis indicates the sequence similarities scores obtained with the Needleman-Wunch alignment algorithm.

Changes to the manuscript	**Changes to the Supplementary Materials**
additional paragraph in Methods section “**miRNA sequence similarity and linear re-scaling for sChemNET loss function**” to explain the re-scaling procedure.	new Supplementary Figure 4 (Fig. R2)
We added details in the manuscript line 131	

4. In figures 2b and 3b, performance metrics are shown as barplots with error bars of some sort -- these are not described in the figure legend. While sChemNET with sequence similarity performs better, on average, than without sequence similarity, the error bars seem to indicate this difference isn't meaningful. In fact, most methods are only marginally better than chemical similarity; not surprising given the sparsity of training data available.

Prompted by the reviewer comment, in the revised version of the manuscript, we have improved the legend of **Figure 2** to explain that the error bars shown in **Figures 2b and 3b** represent the variation in recall at different top-K values (100, 300, 500 or 1000) across the 125 miRNAs from *Homo sapiens.*

Our evaluation to obtain the recall for each miRNA was as follows. For each miRNA, we removed one of its known bioactive small molecules from the training set, and we placed it on a test set together with other 3,999 randomly selected small molecules (see **Fig. 2a** in the main manuscript). Then, having trained each model using the available data in the training set, the goal now becomes to recover the bioactive small molecule from the pool of 4,000 previously unseen small molecules in the test set (see also **Fig. 2a**). To measure the prediction performance at recovering bioactive small molecules for a given miRNA i, we calculate the recall as follow:

$$\text{recall (miRNA } i\text{, top-}k\text{)} = \frac{\text{number of bioactive small molecules recovered at top}-\text{k}}{\text{number of instances assessed for miRNA i}}$$

Figures 2b and **3b** show this recall computed on different top-ks (100, 300, 500, and 1000) **-** x-axis **-** for each of the 125 miRNAs in *Homo sapiens* that were assessed. In the figures, the height of each rectangle indicates the median value, and the error bar indicates the standard deviation (s.t.d.).

The reviewer has also indicated that the differences in prediction performance between sChemNET with and without sequence similarity are not meaningful. To understand whether the differences in prediction performance between sChemNET with and without sequence similarity integration are statistically significant, we used the non-parametric one-sided Wilcoxon Sum Rank statistical test to assess whether the distribution of sChemNET recall across the different top-ks is significantly higher the distribution of recalls obtained by each of the other methods. We adjusted p-values using the Benjamini–Hochberg correction for multiple testing to keep the overall significance level below 0.05. **Figure R3** below shows the average recall improvement of sChemNET over each baseline method, including sChemNET without sequence information (or sChemNET with $s_{uv} = 1$).

Fig. R3 (Left) shows that, when considering all the evaluation instances (i.e., 1,097 associations between 131 small molecules and 125 miRNAs from *Homo sapiens*), sChemNET performs, on average, significantly better than sChemNET without sequence similarity by 2.73% (adjusted p < 0.042), XGBoost by 5.68% (adjusted p < 4.95e-05), logistic regression by 6.64% (adjusted p < 1.38e-06), random forest by 13.74% (adjusted p< 3.94e-20), FNN by 21.14% (adjusted p < 6.28e-48), chemical similarity baseline by 12.73% (adjusted p < 2.13e-17) and the random baseline by 23.07% (adjusted p < 1.17e-56). These results suggest that the average improvements in prediction performance obtained by sChemNET are statistically significant.

There are often chemically similar structures between the single bioactive small molecule in the test set and the set of bioactive small molecules in the training set. To understand whether sChemNET outperform the competitors at predicting bioactive small molecules chemically dissimilar from those available for training the model, we only kept the chemically dissimilar instances (i.e., 810 associations between 108 bioactive small molecules and 125 miRNAs from *Homo sapiens*) using a Tanimoto chemical similarity threshold of 0.6. **Fig. R3 (Right)** shows that, in the case of chemically dissimilar instances, sChemNET mean recall improvement is also significantly better than each of the baseline methods by: 5.56% (sChemNET without sequence similarity, adjusted p < 1.18e-03), 9.32% (XGBoost), 12.34% (logistic regression), 13.29% (random forest), 18.44% (FNN), 26.95% (chemical similarity baseline, adjusted p < 1.69e-67) and 19.65% (random baseline, adjusted p < 3.16e-34).

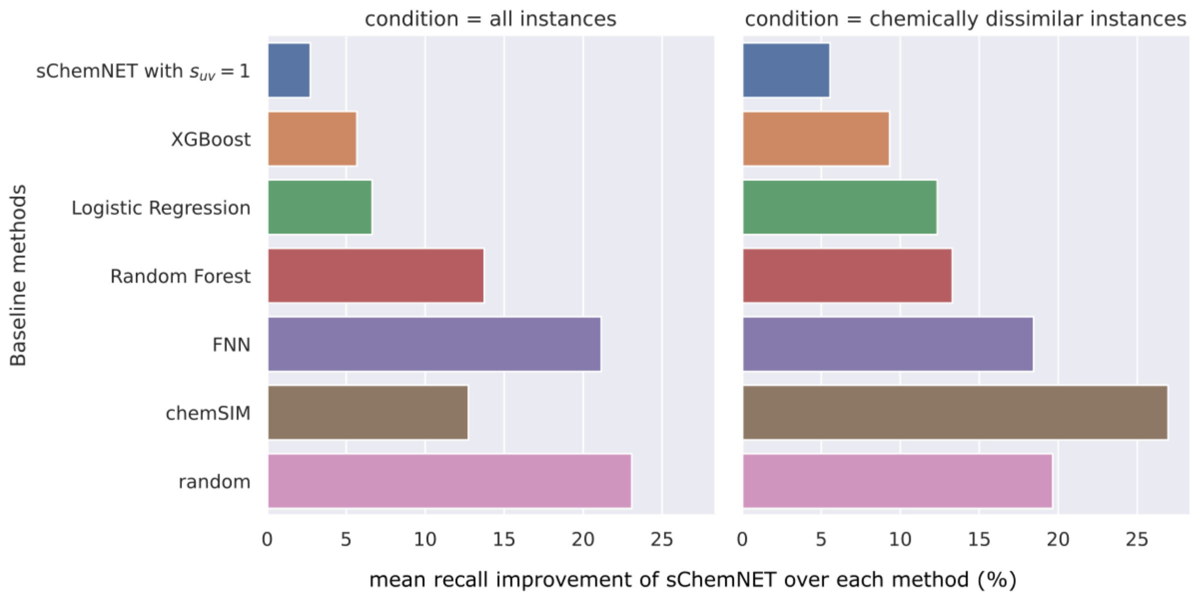

Figure R3. Mean recall improvement of sChemNET's recall performance over each baseline method computed over different top-ks (100, 300, 500 and 1000).

The reviewer pointed out that sChemNET's recall performance is only marginally better than the chemical similarity baseline. However, our experiments show that in the case in which we consider all the instances, sChemNET's mean improvement in recall over the chemical similarity baseline is **12.73%**; an improvement that is statistically significant (adjusted p < 2.13e-17). In the more challenging evaluation of chemically dissimilar instances between training and testing sets, the overall improvement is even higher (by **26.95%** in mean improvement in recall) with a statistically significant difference between the distribution of recall values (adjusted one-sided Wilcoxon Sum Rank Significance, p < 1.69e-67).

Changes to the manuscript	**Changes to the Supplementary Materials**
The results on Fig. R2 are mentioned in lines 199-201	new Supplementary Figure 5 (Fig. R3)
Figure 2 legend has been improved to explain the meaning of the error bars	
we have incorporated the explicit formula for the recall in the section "**In-silico LOOCV evaluation procedure**".	

5. Figures 2b and 3b are easy to mis-interpret, given the authors choice of wording: among 100 "top" molecules, approx. 10% of bioactive molecules were identified by sChemNET -- thus, is 10% the recall (sensitivity), or the TPR of the top-100 selection? An ROC curve (with confidence intervals) would be a more informative and straightforward way to visualize performance.

Thank you for indicating this point of confusion. We have now improved the clarity of our message with a new analysis.

Briefly, the percentages shown in **Figures 2b** and **3b** are the recall (sensitivity). Recall is typically used in leave-one-out-cross-validation (LOOCV) when assessing computational prediction models that deal with sparse data and in which the negative label is not well-defined[6]. In our evaluation, we used a LOOCV procedure for two main reasons:

1. **Because LOOCV allows us to accurately measure the expected prediction performance** (recall@top-k small molecules retrieved) for each miRNA separately, and when considering a realistic scenario in which a single bioactive small molecule needs to be found in a large pool of chemicals – which better reflects our practical application case, and,
2. **Because the sparsity in the dataset makes it difficult to generate a standard machine learning split** (e.g., 70-30%) that could allow us to use the ROC curve to measure the prediction performance for each miRNA accurately. Consider, for instance, the miRNAs with only five known labels for which an ROC curve could be misleading. Another important point to consider is that ROC curves are usually more informative when both positive and negative labels are well-defined, which is not the case for our problem, as negative labels represent unknown small molecule-miRNA associations rather than true negative associations.

To show how using a standard binary classification procedure becomes challenging in sparse and small datasets, we tried to run it for each miRNA. We used the *Homo sapiens* dataset which consisted of 1,102 associations between 6433 labeled and unlabeled small molecules and 126 miRNAs. Our evaluation procedure is described below:

For each miRNA, we split the labels as follows:

- **Positive labels** (i.e. known bioactive small molecules) were randomly split 70% into training and the remaining 30% into testing sets.
- **Negative labels** (i.e. yet unknown to be bioactive small molecules) were randomly split 38% into training and 62% into testing sets.

The reason for having different split ratios for negative and positive labels was to keep our original ratio between labels defined in our LOOCV so that each model's hyperparameters do not need to be re-tuned.

For the testing set obtained from the procedure described above, we calculated the Tanimoto 2D chemical similarity between "positive" small molecules between training and testing sets in order to remove from the testing set all the bioactive small molecules that were chemically similar (i.e. Tanimoto similarity > 0.6) to the bioactive small molecules in training. This step is crucial so that there is no bias and information leakage in our evaluation of the prediction performance.

Finally, we only kept instances in which we found at least five positive labels in the testing sets so that the AUROC measure is somewhat meaningful. The whole procedure for each miRNA was repeated five times to ensure different random splits. We then framed a binary classification problem in which we aimed to classify positive from negative labels and used the Area Under the Receiver Operating Characteristic Curve (AUROC) as a measure of the prediction performance.

We found that, out of the 126 miRNAs, only 3 miRNAs met the criteria to calculate the AUROC: (i) hsa-miR-21-5p with 35 known positive labels (the largest); (ii) hsa-miR-16-5p with 21 known positive labels; and (iii) hsa-let-7a-5p with 17 known positive labels. **Figure R4** below shows the summary distribution of the AUROCs obtained. We found that sChemNET outperforms the other approaches in terms of median AUROC (0.681 +- 0.05, median and s.t.d), Random Forest (0.592 +- 0.09), XGBoost (0.594 +- 0.06), Logistic Regression (0.655 +- 0.08) and FNN (0.56 +- 0.109).

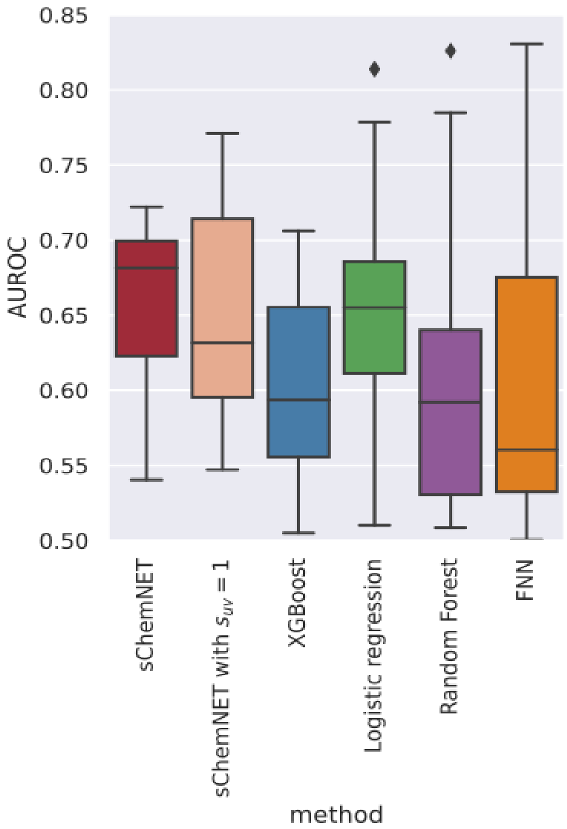

Figure R4. **Boxplots of the area under the receiver operating characteristic (AUROC) obtained for each method at predicting small molecules that are bioactive against three miRNAs from *Homo sapiens***: (i) hsa-miR-21-5p with 35 known positive labels (the largest); (ii) hsa-miR-16-5p with 21 known positive labels; and (iii) hsa-let-7a-5p with 17 known positive labels. A total of 6,433 small molecules were used.

Changes to the manuscript	**Changes to the Supplementary Materials**
Improved explanation of Fig. 2 on lines 185-187.	
New AUROC analyses are explained in lines 203-205.	new Supplementary Figure 7 (Fig. R4)

6. The authors use a loss function that incorporates a regularization technique that penalizes for promiscuous (and presumably false positive) predictions across the unlabelled portion of the training set. This is meant to represent the chemical "space" in which the miRNA bioactivity prediction is optimized -- and thus the model should be relearned for any new screening library to be surveyed. Thus it remains unknown whether the various performance and

learning behaviors of sChemNET presented here are specific to this particular choice of chemical space or would generalize to other chemical libraries of interest.

The reviewer makes a valid point. In our study, we focused on the chemical space defined by the Drug Repurposing Hub database. We believe using this dataset offers several advantages: (1) Most of the compounds were experimentally confirmed in purity and identified and annotated with literature-reported protein targets, indications, and disease areas and even provided with vendor information; (2) The chemical library contains FDA-approved drugs or small molecules that have reached clinical development and hence are easy to acquire for experimental testing on bioactivity against miRNAs; (3) Corsello et al.[7] showed that the Drug Repositioning Hub Library contains small molecules that are chemically and therapeutically diverse thus covering an important part of the chemical space for drug repositioning purposes.

It is out of the scope of our current study to test sChemNET on chemical libraries that are chemically distant from the drug repurposing hub. But it is an opportunity for future studies.

Changes to the manuscript	**Changes to the Supplementary Materials**
we have incorporated a sentence in the **Discussion** about this potential future work (line 486-487).	

Responses to Reviewer #2

This manuscript addresses a challenging topic, namely the interaction between small molecules and microRNA, by means of machine learning and validation. The authors are commended for doing experimental validation.

We thank the reviewer for the positive feedback on our manuscript, especially for praising our efforts to combine computational predictions with experimental validations.

However, the number of concerns outweighs the benefits of the merits of this paper, which preclude publication at this time.

Here are some of the concerns:

1. data - the SM2miR database has not been updated since 2015; one would assume that in decade that has passed since the original publication, a lot more such molecules have been published. Some of these should have been used as a temporal split (external) validation set. More molecules here? PMC6546413

The reviewer makes a good point. Unfortunately, we cannot use the small molecules from the paper mentioned by the reviewer[8] because the paper does not provide the SMILES information for their chemical structures.

However, we searched and found another recently published dataset called RNAInter[9] (*http://www.rnainter.org/*) that also contains associations between small molecules and microRNAs for *Homo sapiens*. We checked whether there were prospective associations in this dataset from 2022 that were not present in our dataset from 2015. We found 1,180 novel prospective associations between 123 miRNAs and 120 small molecules.

We used these prospective associations as a test set for each miRNA. That is, for each miRNA, we used our 2015 dataset (SM2miR database) to train the models, and the 2022 dataset (RNAInter database) as a test set. To avoid information leakage from similar chemical structures, we only kept in the test set chemically dissimilar compounds from those in training (Tanimoto chemical similarity < 0.6). We only considered cases in which we had at least five associations in the test set. In the test set, we also incorporated 4,000 randomly selected small molecules that were unknown to be bioactive against the miRNA under evaluation. The remaining unlabeled small molecules were used for training. We then framed a binary classification performance and used the area under the receiver operating characteristic curve (AUROC) to calculate the model's prediction performance for each miRNA.

Figure R5 shows that sChemNET outperforms all the baseline methods, with an average AUROC of 0.582 +- 0.096. The average prediction performance of the other methods was 0.562+-0.104 (sChemNET without sequence information), 0.541 +- 0.106 (FNN), 0.533+-0.07 (XGBoost), 0.475+-0.087 (Logistic Regression), 0.457+- 0.095 (Random Forest). Overall, these results indicate that sChemNET can also be helpful at predicting novel small molecules without known bioactivity against miRNAs.

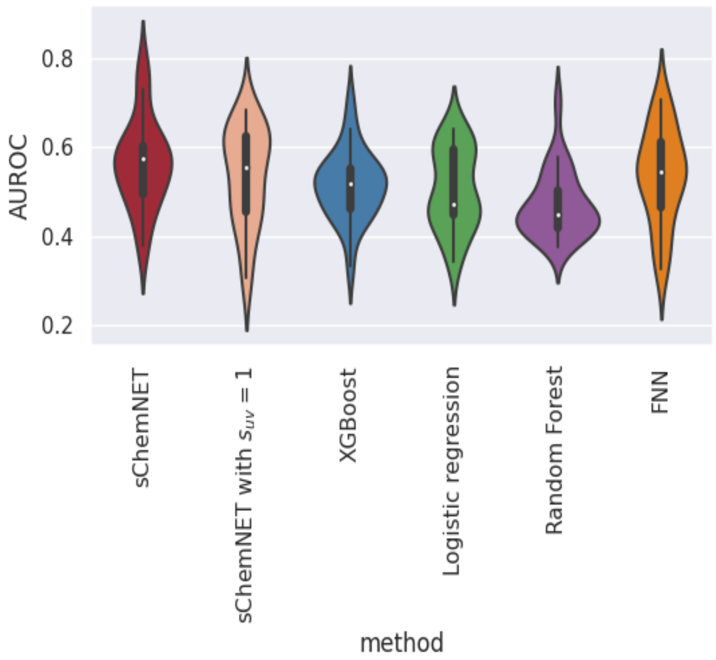

Figure R5. **Prospective evaluation.** Violin plots of the area under the receiver operating characteristic (AUROC) were obtained for 123 miRNAs from *Homo sapiens*. The positive labels in the training set consisted of associations obtained from the SM2miR 2015 database version. The positive labels in the test set consisted of associations obtained from the RNAInter 2022 database. Unlabeled small molecules were obtained from the Drug Repositioning Hub database.

Changes to the manuscript	**Changes to the Supplementary Materials**
We incorporated information about the new dataset in the Methods section **"Chemical Datasets"**	
We incorporated a new section in Methods "**Prospective evaluation**" to explain the new procedure	
new paragraph in lines 221-226 to explain the new results	new Supplementary Figure 9 (same as Fig. R5)

2. ML descriptors - sChemNet models are trained using the 127 features of the MACCS fingerprint. This is an unfortunate choice, given that MACCS keys in that set were built on the World Drug Index in the late 1980s, and are literally representative of "old chemistry". There is a significant number of more modern approaches to chemical features that can be used to address this question.

Our use of MACCS chemical fingerprints was motivated by the low dimensionality of this chemical fingerprint, which makes it suitable for building a machine-learning model with very sparse datasets, such as ours. In machine learning literature, it is well known that as the number of input feature sizes increases, the number of samples needs to increase to prevent the curse of dimensionality[10]. Other chemical fingerprints, such as Daylight-type or Morgan fingerprints, require much larger dimensionality (> 1,000), which in turn, are likely to cause overfitting and poor generalization of the model to new structures.

To understand whether other fingerprints perform better than MACCS, we assessed the prediction performance of sChemNET and all the competitors using the RDKit, ECFP4, and ECFP6 chemical fingerprints. The RDKit-specific fingerprint is inspired by public descriptions of the well-known Daylight fingerprint. The RDKit-specific algorithm is based on hashing molecular subgraphs. To compute the chemical fingerprint of each small molecule, we used the default set of parameters: minimum path size: 1 bond, maximum path size: 7 bonds, fingerprint size: 2,048 bits, number of bits set per hash: 2, minimum fingerprint size: 64 bits, target on-bit density 0.0. The Extended-Connectivity Fingerprints (ECFPs) of diameters 4 and 6 are topological fingerprints that have been shown to outperform other fingerprints in ranking diverse structures[11]. We also used the RDKit python library to calculate ECFP4 and ECFP6 from the SMILES representation of the small molecules. For ECFP4, we used maximum radius of 4 for the substructure and fingerprint length: 1,024 bits, standard Daylight atom features, and we ignore chirality. For ECFP6, we used the same parameters, but the maximum radius was set to 6. We then benchmarked the RDKit, ECFP4, and ECFP6 fingerprints using the LOOCV procedure we presented in our manuscript. Notice that while MACCS has only 127 dimensions, RDKit has 2,048 dimensions and ECFP4 and ECFP6, 1,024 dimensions.

Figure R6 below shows the average recall of the methods across different tops when considering all the associations (**Fig. R6 *left panel***) and when considering only the chemically dissimilar associations between training and testing (**Fig. R6 *right panel***). The results show that, on average, the MACCS fingerprint outperforms all the other fingerprints when used with sChemNET. Our result indicates that although MACCS has limitations in the chemical representation, it works well with our sparse dataset of small molecule-miRNA associations.

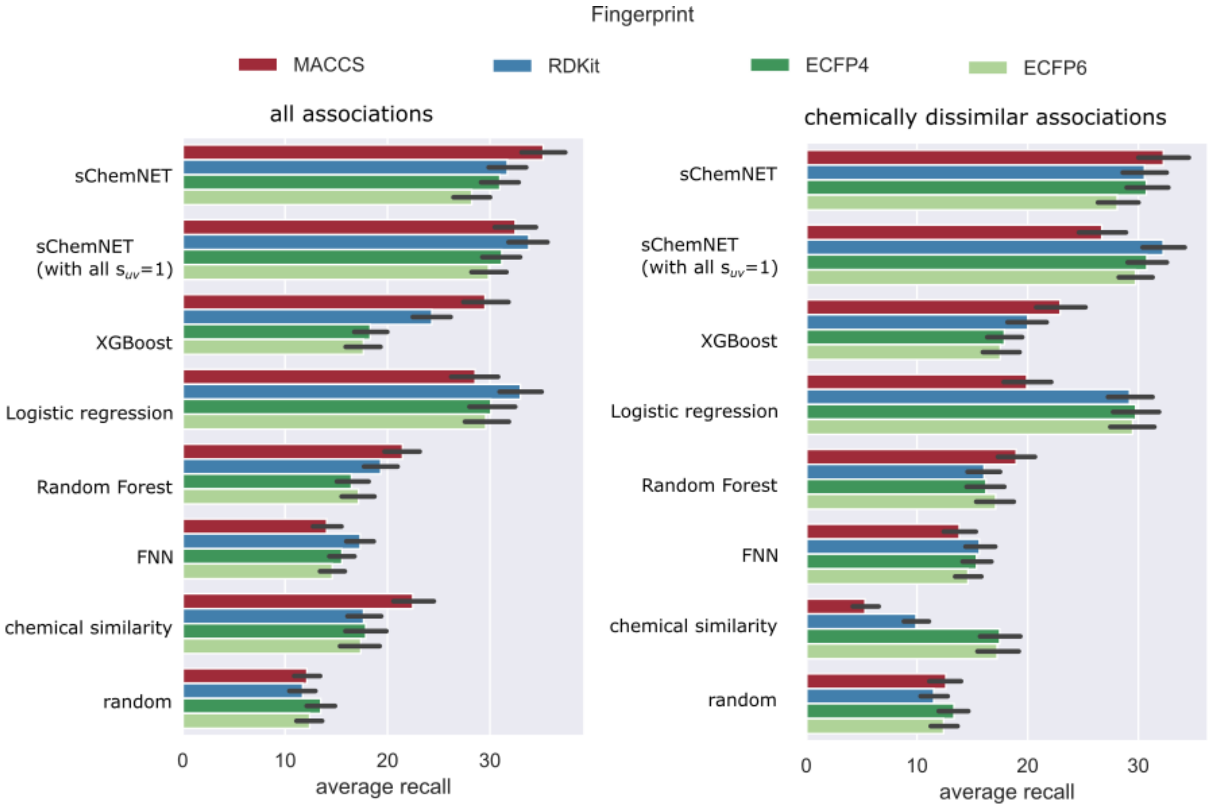

Figure R6. Prediction performance of different methods using different chemical fingerprints (MACCS, RDKit, ECFP4, and ECFP6). The x-axis shows the average recall across top-K (100,300,500, and 1000) predictions in a Leave-one-Out cross-validation procedure in which a single bioactive small molecule known to affect a miRNA is placed on a test set together with other 3,999 randomly selected small molecules yet unknown to affect the miRNA. Each method ranks 4,000 small molecules on the test set, and the recall was calculated for all the 1,102 associations between bioactive small molecules and 125 miRNAs from Homo sapiens. (*Left panel*) All the associations; (*Right panel*). Only chemically dissimilar associations between training and testing are considered (Tanimoto chemical similarity < 0.6).

Changes to the manuscript	**Changes to the Supplementary Materials**
New paragraph in the Discussion, lines 493-501.	new Supplementary Fig. 10 (same as Fig. R6)

3. Model imbalance - there is a serious concern about the 10:1 ratio between inactive and active compounds. It is unclear if the 2400 unlabelled molecules are indeed inactive. Perhaps a more clear description would have helped; true inactives should incorporate compounds known to permeate mitochondria (e.g., metformin).

We agree with the reviewer that model imbalance is a common problem in the drug target prediction research[2]. This is because, on one side, small molecule are known to affect a small number of biological targets. On the other, there is a long-tailed distribution in the number of biological targets targeted by small molecules[12], which we have shown also prevails for the distribution of small molecule-miRNA associations (see **Supplementary Figure 2**).

In our manuscript, we do not claim that our set of unlabeled small molecules is inactive against miRNAs. In fact, our hypothesis is that their activity against miRNAs is yet unknown. We believe that sChemNET can assist in the discovery of the true bioactive small molecules from the set of unlabeled small molecules. We have shown that sChemNET is able to achieve this goal both with *in-silico* simulations (see **Figures 2 and 3**), and with different wet-lab experiments and publicly available experimental data that confirmed sChemNET's predictions (**Figures 5 and 6**).

Including experimentally assessed inactive small molecules against miRNAs in our dataset, i.e. true negatives, would require extensive experimental validation that is beyond the scope of the current manuscript and would be better suited for future work.

Changes to the manuscript	**Changes to the Supplementary Materials**
Sentence in the Discussion about potential future work, lines 497-500.	

4. But the more serious problem is model validation. Calcitriol and its effect on miRNA has been disclosed in 2016 - PMC4714233. Given the extremely high similarity between calcediol and calcitriol, it is really difficult to imagine that a) this choice was accidental and b) that the authors had no knowledge of this result.

Thank you for identifying an excellent reference demonstrating that calcitriol upregulates miRNA, which we have added to our references. We were unaware of this study because the keyword "miRNA-Seq" was not mentioned in the abstract. This information adds to our rationale for the validation of the model. We want to emphasize that the first step of our process was that the sChemNET model predicted this relationship (see heatmap generated with sChemNET's predictions in **Figure 4**) using small molecule-miRNA associations from the SM2miR database release 2015 that incorporates information from scientific publications up to the year 2015. Therefore, it is not possible for our sChemNET model to have used the information mentioned by the reviewer which was published in 2016.

For the experimental validations we performed for the manuscript, we generated each hypothesis based on sChemNET predictions in an agnostic manner. Either by using direct ranking of small molecules for each miRNA or by using the heatmap in **Figure 4b-c** that contains statistically significant associations between sChemNET's predicted drug mode of action and drug indications and miRNAs from *Homo sapiens*. The purpose of these heatmaps was to condense the predictions of sChemNET for generating biological hypotheses that could be more meaningful to pursue for experimental validation. For instance, the statistically significant association between miR-451 and vitamin D receptor agonists inspired us to experimentally test α-Calcidol on the zebrafish embryos (see **Figure 5**); whose effect we have confirmed experimentally in our manuscript.

We then explored the literature and found that calcitriol regulation of multiple miRNAs was established, so we tested our predictive hypothesis in a cell model to see if the specific miRNA predicted by the sChemNET model was regulated by calcitriol.

In our manuscript, we do not claim that we are the first to demonstrate that calcitriol affects miRNAs, but rather to show a new ML model, sChemNET, that can use small-size bioactive compounds from available RNA chemical datasets, to predict bioactive compounds in a larger chemical library such as the Drug Repositioning Hub chemical database.

Changes to the manuscript	**Changes to the Supplementary Materials**
We included the reference mentioned by the reviewer	
In the Discussion, lines 459-461, we have added a sentence describing that α-calcidiol is converted to calcitriol, and thus functions in this system equivalently to calcitriol.	

5. Last but not least, docetaxel is an antineoplastic taxane; if it modulates glucocorticoid receptors, please provide evidence.

We thank the reviewer for pointing this out. Docetaxel, predicted in top- 3 by sChemNET when ranking over 6,400 small molecules for miR-451, is a tubulin polymerization inhibitor, and it is known to target BCL2 (also targeted by miR-451) according to the Drug Repositioning Hub database. We have corrected the known mode of action of docetaxel in the revised version of our manuscript.

Changes to the manuscript	**Changes to the Supplementary Materials**
In lines 279-281, we have corrected the known mode of action of docetaxel in the main manuscript.	

Responses to Reviewer #3

In this work Galeano et al. have made an attempt to develop a method to target microRNAs with small molecules by using deep learning approaches. The key idea in this method is training a learning model by using the information of “unlabeled” structures. During the training phase each of these small molecules are assigned a prediction score to each miRNA.

There are several fundamental questions that authors need to provide a more detailed explanation and rationalization.

1. Authors claim that “Unlabeled small molecules have unknown biological activity against targeted miRNA” and then they use the Drug Repurposing Hub database for creating the unlabeled set of small molecules. It has been shown that there is a significant overlap between the chemical space of known approved drugs and RNA binders which bind to microRNA (J. Am. Chem. Soc. 2021, 143, 33, 13044–13055). Authors need to calculate the physicochemical properties of their labeled and unlabeled libraries to show the similarities and differences and then they need to map the chemical space of 6,433 small

molecules to the bioactive molecules to show the unlabeled library is not already biased toward binding to microRNAs.

We thank the reviewer for the comment. We want to insist that the SM2miR dataset that we used in our study is not based on direct interactions, i.e. binding between small molecules and microRNAs. Our dataset indicates whether a small molecule elicits a transcriptional response that mimics the miRNA-mediated regulation, but the exact mechanism of action is not provided in the database.

The analysis by Zhang et al.[13] mentioned by the reviewer, is a different dataset and represents a different application problem that focuses entirely on binding interactions between RNA folds and small molecules.

Nevertheless, to understand the similarities/differences between the labeled and unlabeled libraries in terms of their physicochemical properties, we calculated 46 different physicochemical properties for each of the 6,300 small molecules in our dataset using SwissADME[14]. To compare small molecules' properties, we used the Euclidian distance between their physicochemical property vectors, in which each element of the vector corresponds to a specific property. We then calculated the Welch's t-test Significance between the vector of distances among the labeled small molecules (*intra-group*) and between the labeled and unlabeled small molecules (*inter-group*) to see whether there two distributions have equal mean. We found that the mean of the distribution of physicochemical distances underlying the inter-group of small molecules is significantly greater than the mean of the distribution of the intra-group (One-sided p-value < 1.10e-30). This suggests that, in terms of physicochemical properties distances, there are statistically significant differences between the labeled and unlabeled small molecules.

Changes to the manuscript	**Changes to the Supplementary Materials**
We added the reference mentioned by the reviewer	
We added the new analysis in a paragraph in the **Discussion** (lines 413-426).	
In the **Discussion**, we have mentioned that combining datasets/knowledge about direct binding and regulation between miRNAs	

and small molecules is an important avenue of future research (Lines 500-501).	

2. In line 155 authors use the term “unique” to describe the unlabeled small molecules. They need to provide more information on what determined this “uniqueness”.

The word unique in this context simply refers to small molecules with unique PubChem Identifiers (provided as CIDs).

Changes to the manuscript	**Changes to the Supplementary Materials**
We have clarified that “unique” refers to unique PubChem IDs in the manuscript in line 160.	

3. What is the reason behind selecting the MACCS fingerprint as it has been shown that extended-connectivity fingerprints of diameters 4 and 6 are among the best performing fingerprints in ranking diverse structures by similarity and fingerprints to avoid when measuring similarity include Daylight-type path-based fingerprints and MACCS keys (J Cheminform. 2016; 8: 36). As the scoring of the test set is based on the extracted chemical similarity using this fingerprint it also raises another question whether the authors have benchmarked other fingerprints or not? This makes the whole scoring workflow highly questionable. Authors need to rationalize their choice of this fingerprint and then benchmark other fingerprints to validate their results.

We thank the reviewer for this comment.**REV#2** had a similar concern, and we have addressed it with further analysis. It is presented in in our response to **REV#2 question 2**, but we also placed it below for your convenience.

Our use of MACCS chemical fingerprints was motivated by the low dimensionality of this chemical fingerprint, which makes it suitable for building a machine-learning model with very sparse datasets, such as ours. In machine learning literature, it is well known that as the number of input feature sizes increases, the number of samples needs to increase to avoid the curse of dimensionality. Other chemical fingerprints, such as Daylight-type or Morgan fingerprints, require much larger dimensionality (> 1000), which in turn, causes overfitting and poor generalization of the model to new samples.

To understand whether other fingerprints perform better than MACCS, we assessed the prediction performance of sChemNET and all the competitors using the RDKit, ECFP4, and ECFP6 chemical fingerprints. The RDKit-specific fingerprint is inspired by public descriptions of the well-known Daylight fingerprint. The RDKit-specific algorithm is based on hashing molecular subgraphs. To compute the chemical fingerprint of each small molecule, we used the default set of parameters: minimum path size: 1 bond, maximum path size: 7 bonds, fingerprint size: 2,048 bits, number of bits set per hash: 2, minimum fingerprint size: 64 bits, target on-bit density 0.0. The Extended-Connectivity Fingerprints (ECFPs) of diameters 4 and 6 are topological fingerprints that have been shown to outperform other fingerprints in ranking diverse structures[11]. We also used the RDKit python library to calculate ECFP4 and ECFP6 from the SMILES representation of the small molecules. For ECFP4, we used maximum radius of 4 for the substructure and fingerprint length: 1,024 bits, standard Daylight atom features, and we ignore chirality. For ECFP6, we used the same parameters, but the maximum radius was set to 6. We then benchmarked the RDKit, ECFP4, and ECFP6 fingerprints using the LOOCV procedure we presented in our manuscript.

Figure R6 below shows the average recall of the methods across different tops when considering all the associations (**Fig. R6 *left panel***) and when considering only the chemically dissimilar associations between training and testing (**Fig. R6 *right panel***). The results show that, on average, the MACCS fingerprint outperforms all the other fingerprints when used with sChemNET. Our result indicates that although MACCS has limitations in the chemical representation, it works well with our sparse dataset of small molecule-miRNA associations.

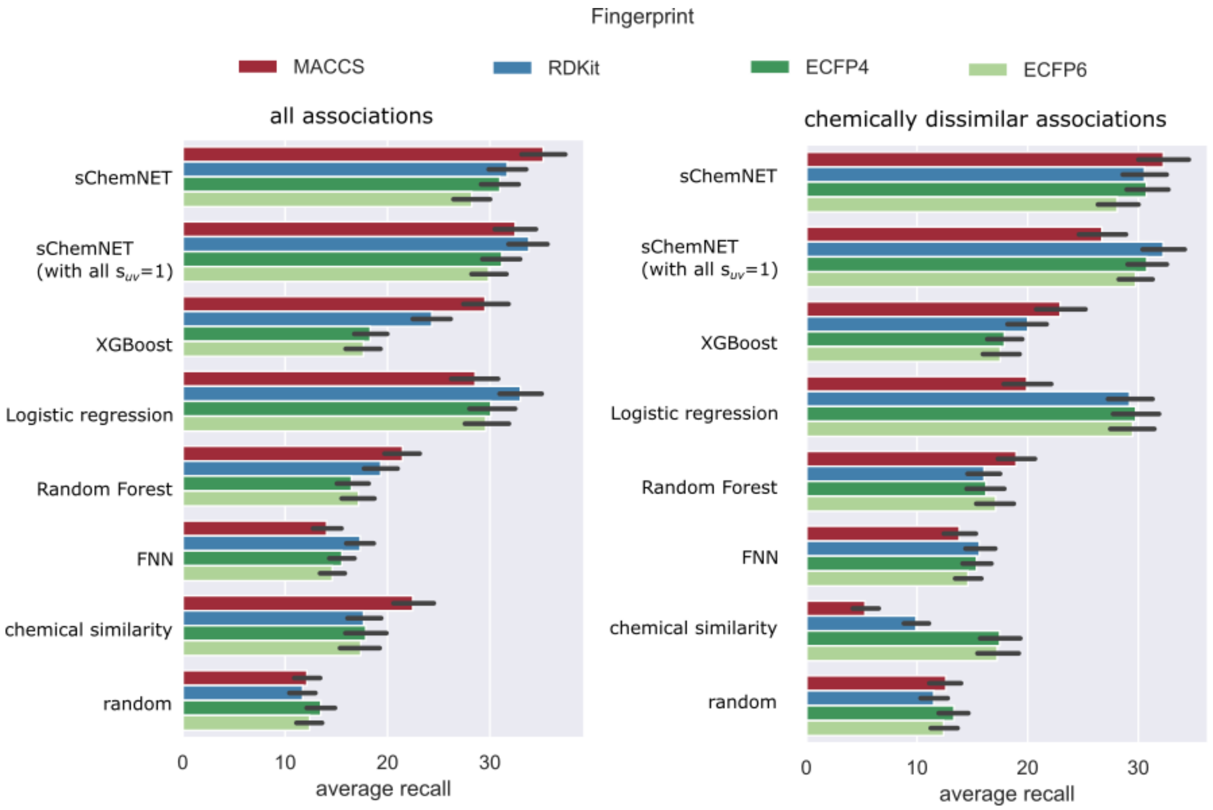

Figure R6. Prediction performance of different methods using different chemical fingerprints (MACCS, RDKit, ECFP4, and ECFP6). The x-axis shows the average recall across top-K (100,300,500, and 1000) predictions in a Leave-one-Out cross-validation procedure in which a single bioactive small molecule known to affect a miRNA is placed on a test set together with other 3,999 randomly selected small molecules yet unknown to affect the miRNA. Each method ranks 4,000 small molecules on the test set, and the recall was calculated for all the 1,102 associations between bioactive small molecules and 125 miRNAs from Homo sapiens. (*Left panel*) All the associations; (*Right panel*). Only chemically dissimilar associations between training and testing are considered (Tanimoto chemical similarity < 0.6).

Changes to the manuscript	**Changes to the Supplementary Materials**
New paragraph in the Discussion, lines 493-501.	new Supplementary Figure 10 (same as Fig. R6)

References

1. Galeano, D., Li, S., Gerstein, M. & Paccanaro, A. Predicting the frequencies of drug side effects. *Nat. Commun.* **11**, 4575 (2020).
2. Luo, Y. *et al.* A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* **8**, 1–13 (2017).
3. Santos, S. de S. *et al.* Machine learning and network medicine approaches for drug repositioning for COVID-19. *Patterns* 100396 (2021) doi:10.1016/j.patter.2021.100396.
4. Jiang, L., Ding, Y., Tang, J. & Guo, F. MDA-SKF: Similarity Kernel Fusion for Accurately Discovering miRNA-Disease Association. *Front. Genet.* **9**, 618 (2018).
5. Li, L. *et al.* SCMFMDA: Predicting microRNA-disease associations based on similarity constrained matrix factorization. *PLoS Comput. Biol.* **17**, e1009165 (2021).
6. Cáceres, J. J. & Paccanaro, A. Disease gene prediction for molecularly uncharacterized diseases. *PLoS Comput. Biol.* **15**, e1007078 (2019).
7. Corsello, S. M. *et al.* The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.* **23**, 405–408 (2017).
8. Fan, R. *et al.* Small molecules with big roles in microRNA chemical biology and microRNA-targeted therapeutics. *RNA Biol.* **16**, 707 (2019).
9. RNAInter v4.0: RNA interactome repository with redefined confidence scoring system and improved accessibility | Nucleic Acids Research | Oxford Academic. https://academic.oup.com/nar/article/50/D1/D326/6414580?login=false.
10. Bishop, C. M. *Pattern Recognition and Machine Learning*. (Springer, New York, 2006).

11. O'Boyle, N. M. & Sayle, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. *J. Cheminformatics* **8**, 36 (2016).

12. Yildirim, M. A., Goh, K.-I., Cusick, M. E., Barabási, A.-L. & Vidal, M. Drug-target network. *Nat. Biotechnol.* **25**, 1119–1126 (2007).

13. Zhang, P. *et al.* Reprogramming of Protein-Targeted Small-Molecule Medicines to RNA by Ribonuclease Recruitment. *J. Am. Chem. Soc.* **143**, 13044–13055 (2021).

14. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules | Scientific Reports. https://www.nature.com/articles/srep42717.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

No further comments; my concerns have been adequately addressed, for the most part.

Reviewer #2 (Remarks to the Author):

The authors have convincingly addressed most points raised (as Reviewer #2). They are indeed commended for using the 2015 data to predict 2022 data (temporal validation as it should be done).

The only one that remains somewhat contentious is the inference that MACCS fingerprints are superior to other chemical descriptors such as Morgan fingerprints ECFP path 6. From an ML perspective, the authors are correct. What's really happening is that "lower dimensional" FPs compress information, which is further compressed by the 2-layer NN that is sChemNet. This is not a technical disagreement per se, but the authors are strongly encouraged to point this out. Again:
200 chemical features == data compression of chemical structures
2000 chemical features, learned from the data set == better chemical representation
in the long run, the 2000-feature model becomes more explainable and usable for predicting bioactivity against miRNAs.

Reviewer #2 (Remarks on code availability):

I did not run the code.

Reply to the Reviewers

Reviewer #1 (Remarks to the Author):

No further comments; my concerns have been adequately addressed, for the most part.

Reviewer #2 (Remarks to the Author):

The authors have convincingly addressed most points raised (as Reviewer #2). They are indeed commended for using the 2015 data to predict 2022 data (temporal validation as it should be done).

We thank the reviewer for the positive feedback on our revisions.

The only one that remains somewhat contentious is the inference that MACCS fingerprints are superior to other chemical descriptors such as Morgan fingerprints ECFP path 6. From an ML perspective, the authors are correct. What's really happening is that "lower dimensional" FPs compress information, which is further compressed by the 2-layer NN that is sChemNet. This is not a technical disagreement per se, but the authors are strongly encouraged to point this out. Again:
200 chemical features == data compression of chemical structures
2000 chemical features, learned from the data set == better chemical representation in the long run, the 2000-feature model becomes more explainable and usable for predicting bioactivity against miRNAs.

As suggested by the reviewer, we will point this out in the Discussion of the revised version of our manuscript. (lines 500-501). The sentences that we
have incorporated are copied below.

Although ECFP-based offers better chemical representation, it is likely that MACCS outperform it due to overfitting in the presence of our small and sparse labeled dataset.