Cross-modal representation alignment of molecular structure and perturbation-induced transcriptional profiles

Samuel G. Finlayson[1 2 †], Matthew B.A. McDermott[2], Alex V. Pickering[3], Scott L. Lipnick[3], Isaac S. Kohane[3]

[1]*Department of Systems, Synthetic, and Quantitative Biology, Harvard Medical School, Boston, MA*
[2]*Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA*
[3]*Department of Biomedical Informatics, Harvard Medical School, Boston, MA*
[†]*E-mail: samuel_finlayson@hms.harvard.edu*

Modeling the relationship between chemical structure and molecular activity is a key goal in drug development. Many benchmark tasks have been proposed for molecular property prediction, but these tasks are generally aimed at specific, isolated biomedical properties. In this work, we propose a new cross-modal small molecule retrieval task, designed to force a model to learn to associate the structure of a small molecule with the transcriptional change induced by perturbing a cell with this small molecule. We develop this task formally as multi-view alignment problem, and present a coordinated deep learning approach that jointly optimizes representations of both chemical structure and perturbational gene expression profile. We probe what aspects of generalizability are most critical for this task, and benchmark our results against oracle models and principled baselines.

Keywords: Representation Learning, Therapeutics, Gene Expression, Deep Learning, Information Retrieval

1. Introduction

Identifying molecules that are likely to have a specific biological effect is a cornerstone of drug discovery and a key component of efforts to achieve precision medicine. Classically, computational profiling of small molecules has centered on predicting affinities for specific biological targets, using tools ranging from biophysics-driven techniques such as molecular docking[1] to literature-mined annotations.[2]

Small molecule modeling has also recently become a major area of interest for the deep learning community, a trend catalyzed by graph neural networks[3] and the availability of benchmarking datasets.[4] Graph neural networks allow for end-to-end modeling of raw molecular graphs,[5–8] and have resulted in state-of-the-art performance on certain tasks.[9,10] To date, deep learning efforts in this space have generally focused on two extremes: first, a suite of highly local, biochemical prediction problems, which test the model's ability to predict specific chemical properties, and second, models are also tasked to infer more global, clinical properties, such as indication or side effect prediction. Missing from the field, however, are benchmark tasks between these two extremes, that test the ability of deep models to encode rich, general representations of a molecule's broad-spectrum effect on cellular biology.

In parallel to these developments, connectivity mapping has emerged as a alternative approach for drug development.[11] In connectivity mapping, compounds are foremost characterized not by individual chemical properties or downstream targets, but by the broad transcriptional effects they induce in cells. Connectivity mapping begins by first developing a large dataset of *perturbational signatures* of various molecules by physically treating cell lines with these molecules, then measuring the resultant changes in gene expression. These datasets are then compared to one or more *query signatures*, which are typically differential gene expression (GE) signatures representing disease states that the investigators hope to reverse or mimic. Various public datasets have been curated to enable these efforts,[12,13] and researchers have sought to use these for drug repurposing and precision medicine.[14–20]

Connectivity mapping is promising because it can be used to search for new indications of drugs without making any specific *a priori* assumptions about their mechanism of action. However, the typical framework for connectivity mapping is limited by the fact that it can only query against drugs that have already been profiled using the transcriptional assay. In other words, connectivity mapping is – in principle – very flexible with respect to the disease signatures they accept as a query, but is transductive rather than inductive with respect to the target small molecule signatures. (In practice, connectivity mapping relies upon the assumption that drug responses *in vitro* (typically in cancer cell lines) are representative of drug responses *in vivo*.) This is the perfect complement to structure-based computational chemistry, which is typically inductive to new drug structures but can only make predictions for diseases with known targets.

In this work, we combine these two fields, by using deep chemical embedders to learn the transcriptional space encoded by CMAP profiling. More specifically, we train coordinated networks to jointly embed chemical structures and perturbational gene expression profiles such that learned chemical representations are most similar to the encodings of the transcriptional patterns they induce.[*]. Note that this task naturally fills the gap in inductive molecular modelling discussed previously; by tasking the model to produce highly similar embeddings for chemical structures and the perturbational profiles they induce, we force the model to learn a *transcriptome-wide* reflection of the drug's action on the cell. We then evaluate these chemical representations by using gene expression signatures as queries into the embedding space and recovering their corresponding compounds. (See Figure 1). Crucially, the evaluation is set up such that the validation and test set compounds and cell lines are not used in training, which allows us to test the ability of the model to generalize to new drugs and cell lines.

In the rest of this work, we first offer some background on joint embedding alignment, then detail the methods used in this work. Finally, we walk through the results and discussion of these experiments, then close with concluding thoughts. A version of this work, with supplementary material present, can be found here: `https://github.com/sgfin/molecule_ge_coordinated_embeddings/blob/master/paper_08_2020.pdf`.

2. Background

Multi-view representation learning seeks to learn representations that relate information from multiple views of the same data, such as an image and a text description of a single scene. In

[*] Code available here: `https://github.com/sgfin/molecule_ge_coordinated_embeddings`

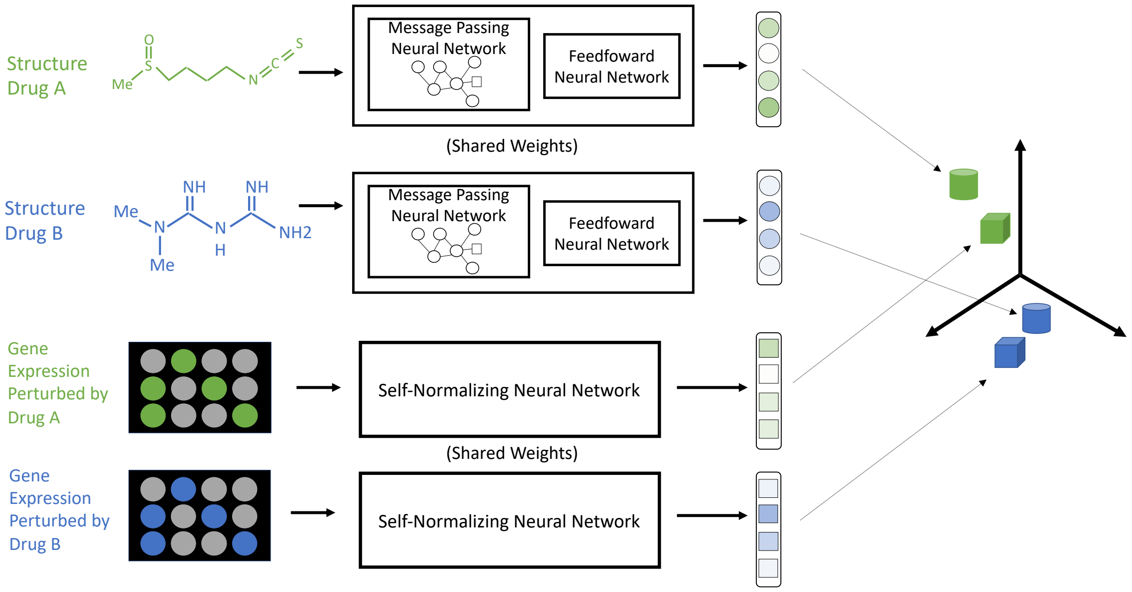
Structure Drug A
O
S
Me
N
C
S
Message Passing Neural Network
Feedfoward Neural Network
(Shared Weights)
Structure Drug B
NH
NH
Me
N
N
H
NH2
Me
Message Passing Neural Network
Feedfoward Neural Network
Gene Expression Perturbed by Drug A
Self-Normalizing Neural Network
(Shared Weights)
Gene Expression Perturbed by Drug B
Self-Normalizing Neural Network

Fig. 1. Overview of our coordinated representation learning method. Neural networks are jointly trained to embed gene expression profiles and small molecule structures such that transcriptional signatures are brought into close proximity with their corresponding perturbagens. Given a cross-modal alignment, gene expression signatures can be used as queries to rank chemical structures by their likelihood to induce such a signature.

multi-view representation alignment, embeddings of each data modality are learned separately but in a coordinated manner, such that different views of the same scene are similar. In this work, we learn aligned representations such that differential gene expressions are embedded in close proximity to the small molecules that induced them.

Multi-view representation alignment can be achieved through a variety of methods, including distance-based, similarity-based, and correlation-based alignment. One classic method for representation alignment, which we leverage as a baseline in this paper, is *canonical correlation analysis (CCA)*,[21] which produces a linear mapping between two datasets that maximizes the element-wise correlation between the embeddings of their rows. A wide range of additional methods for multi-modal alignment have also been proposed, including methods to align embeddings using distance-based, similarity-based, correlation-based, and ranking-based penalties during training.[22] Ranking-based methods for multi-view representation alignment, such as that described by Deng at al,[23] allow the incorporation of ranking information into the training procedure, which may be important in tasks such as gene expression where perturbation signals may be small relative to baseline state. In addition, the field of rank-based embedding learning is intertwined with a broader literature of uni-modal embedding learning, which pioneered such architectures as Siamese[24,25] and Triplet networks.[26] These networks optimize embeddings to bring similar data together while driving dissimilar data apart. An analysis of best practices working with this architecture can be found in Wu et al.[27]

3. Methods

3.1. *Dataset & Tasks*

Data Acquisition and Subsetting All raw data in this study comes from the Next-Generation Connectivity Map (CMap) Level 3 L1000 data provided by the LINCS Consortium and the NIH.[13] This dataset features 978-dimensional L1000 gene expression profiles from a variety of human cell lines treated with chemical and genetic perturbations. In order to ensure we had sufficient support over possible drugs within our dataset, our primary cut of these data uses the most frequent 8 cell lines, split into a train, validation, and test set split by *both* drug and cell line (i.e., no cell line or drug in the train set appears in the validation or test sets). To mitigate non-random missingness, we included only drugs that had been assayed in all cell lines, and limited experiments to those incubated with small molecules for 24 hours at a dose of $10\mu m$. Final statistics of these data cuts are shown in Table A1. For drug structures, we used the SMILES[28] structures provided by LINCS, canonicalized using RDKIT.[29]

Preprocessing and Feature Engineering Gene expression intensity values from the training, validation, and test sets were centered and scaled at the gene-level based on the mean and standard deviation of each gene intensity across the training set. We also augmented each gene expression profile with three additional sets of feaatures: the corresponding gene expression intensities from a control signature on the same plate, the $\log_2$ fold-change between the perturbation and control signatures, and the difference between these gene expression signatures. For use in our baseline and oracle models, we also computed two numerical representations of each small molecule: Morgan extended-connectivity fingerprints, a standard representation used in computational chemistry,[30] and the output of the ChemProp network.[6]

Detailed Task Description Our goal is to learn embedders which map molecular structures and gene expression profiles into a vector space such that "matching" pairs of structures and gene expression profiles (meaning the gene expression profile results from a cell that has been perturbed by the molecule in question) are near one another and non-matching pairs are not. This is depicted graphically in Figure 1.

Formally, given a collection of gene expression signatures $\mathcal{G}$, chemical structures $\mathcal{M}$, and similarity function $\mathrm{Sim} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we seek to learn a gene expression embedder $E_g : \mathcal{G} \to \mathbb{R}^d$ and chemical embedder $E_m : \mathcal{M} \to \mathbb{R}^d$ to maximize $\mathrm{Sim}(E_g(g_i), E_m(m_j))$ while simultaneously minimizing $\mathrm{Sim}(E_g(g_i), E_m(m_{\neg j}))$, where gene expression g_i was induced by molecule m_j. Unless otherwise specified, the similarity function can be assumed to be Pearson Correlation in our experiments. Across our baseline and oracle methods, we realize many variants of E_g and E_m.

3.2. *Baseline and Oracle Methods*

Nearest Neighbor Baseline Nearest-neighbor-based methods have been previously shown to establish very strong baselines for machine learning tasks on the L1000 data.[31,32] In our cross-modal, information retreival (IR) context, traditional NN methods are not applicable, so we employ the following "double nearest neighbor" baseline: given a gene expression profile as a query, we first identify the nearest gene expression profile in the train set and look up its corresponding small molecule. We then take this small molecule (from the train set) as a query, and return the most structurally similar drug from the *test* set as our final prediction.

In particular, given a mapping $G2M : \mathcal{G} \rightarrow \mathcal{M}$ from gene expression profiles to the small molecule that induced them, and a molecular embedding E_m (which may include molecular fingerprints, Chemprop embeddings, or embeddings learned from other models), we define embedder $E_g : g_{\text{query}} \mapsto E_m(G2M(\arg\max_{g_{\text{tr}} \in \mathcal{G}_{\text{train}}} \text{Sim}(g_{\text{query}}, g_{\text{tr}})))$. Then, we perform information retrieval (IR) analyses with such embedders as usual.

Canonical Correlation Analysis Baseline Given training matrices of transcriptional $\mathcal{G}_{\text{train}}$ and molecular $\mathcal{M}_{\text{train}}$ encodings, we can learn a set of linear mappings $E_g : \mathcal{G}_{\text{train}} \rightarrow \mathbb{R}^d$ and $E_m : \mathcal{M}_{\text{train}} \rightarrow \mathbb{R}^d$ via d-dimensional CCA such that these mappings optimize the correlation between elements of $\mathcal{G}_{\text{train}}$ and $\mathcal{M}_{\text{train}}$.

Note that this procedure requires a default numerical representation for molecules, which, as with other methods, can be either fingerprints, ChemProp embeddings, or learned embeddings by our learning model (described below). Note, too, that CCA can be performed atop other embedding systems to further optimize embedding results. Canonical Correlation Analysis was performed using SciKit Learn,[33] using 1000 iterations and 50 components, both chosen on the basis of preliminary experiments.

Oracle Models As described above, the central objective of our task is to learn small molecule embeddings that can stand in as surrogates for their corresponding gene expression signatures. To provide a rough upper-bound for expected performance on this task, we also implemented two "oracle" models, each of which queries test set GE signatures against pseudo-"chemical embeddings" that are in reality the average GE signatures from each test set drug when it was measured on either (1) the *train set cell lines*,[†] to simulate an embedder that perfectly associates all structures to perturbational profiles, but cannot generalize beyond the train set cell lines, or (2) the *test set cell lines*, which simulates a model of the same capabilities but now able to generalize perfectly to the test set as well. Note that these oracle models are still dependent on the representation of the underlying gene expression signatures, so further innovation there could offer improved upper bounds for this task. Formally, again given $G2M$ which maps gene expression profiles to their corresponding perturbing molecule, we define oracle embeddings $E_g^{\text{train}} : g_{\text{query}} \mapsto \text{Avg}(\{g_i \in \mathcal{G}_{\text{train}} | G2M(g_i) = G2M(g_{\text{query}})\})$, and $E_g^{\text{test}} : g_{\text{query}} \mapsto \text{Avg}(\{g_i \in \mathcal{G}_{\text{test}} | G2M(g_i) = G2M(g_{\text{query}})\})$.

3.3. *Deep Coordinated Metric Learning Approach*

For our learned model, we realize E_g as a self-normalizing neural network (of size dictated by hyperparameter search), and E_m as a directed message-passing neural network (D-MPNN), initialized by the Chemprop system, followed by a feed-forward output layer whose shape was dictated via hyperparameter search.[6] To train these architectures, we use a margin-based quadruplet loss, building on Wu et al's adaptive margin loss.[27] The base of the adapted margin loss is defined over two data points i and j as $\text{mar}_{\alpha,\beta} := (\alpha + y_{i,j}(D_{ij} - \beta))_+$, where D is distance

[†] Note that we can do this as we limited our choice of drugs to those that *were* measured in all 8 cell lines, even though our actual data split prohibits training on any drug that appears in the validation or test sets.

function (here euclidean distance), α defines a permissible margin of separation, β controls the boundary between positive and negative pairs, and $y_{i,j}$ is an indicator variable equal to 1 if i and j are of the same class and 0 otherwise. We allowed the system to tune α and β with other hyperparameters.

Given two pairs of matching gene expression and molecular structure embeddings, $(g_A, m_A), (g_B, m_B)$, our quadruplet loss is defined as the sum of the margin losses between all cross-modality pairs of embeddings: $\ell_{\text{quad}} = \text{mar}_{\alpha,\beta}(g_A, m_A) + \text{mar}_{\alpha,\beta}(g_A, m_B) + \text{mar}_{\alpha,\beta}(g_B, m_A) + \text{mar}_{\alpha,\beta}(g_B, m_B)$. The network is thus optimized to bring the positive embedding within the margin of the anchor and negative embedding outside the margin. For sampling these two pairs (an analog of negative sampling for a more traditional triplet network), we first sample one matching pair, choose the molecular structure for the other pair based on the distance-weighted negative sampling scheme described in Wu et al, which was successful with their margin-based approach,[27] then fill in the other gene expression profile to match the sampled molecular structure. To make this process computationally efficient we pre-computed the average distance in *average post-perturbational gene expression space* between every pair of small molecule structures in the dataset. We additionally tried other losses, including two varieties of more traditional triplet losses, and a quintuplet loss, but ultimately found the quadruplet loss to be most performant via our hyperparameter search.

Training and Hyperparameter Selection Each model was trained on a single Nvidia GeForce GTX 1080 GPU. Early stopping was used to select the model with the best mean reciprocal rank on the validation set. Hyperparameter tuning via the Bayesian Hyperopt library[34] was performed over a wide range of possible hyperparameters, including network depth and width (parametrized by the size of the first hidden layer and a growth rate), learning rate, total number of epochs, batch size, margin and β parameters, triplet model orientation (i.e., gene first or compound first), activation function/network type (e.g., SNNN vs. unconstrained fully connected network), and dropout, with no early stopping for hyperparameter search runs. The optimal hyperparameters from this search are shown in Appendix Section 7.2.

3.4. *Experiments*

We designed a range of experiments with two overarching purposes: First, we sought to evaluate if and how our deep coordinated representation learning method offers improvements over principled baselines. This entails a quantitative performance comparison against baseline methods and ablated versions of our model. Second, we introspect into the representations learned by training on this new task, to better understand the challenges and utility of the general framework. This entails a quantitative performance comparison against oracle models, a statistical analysis probing the ability of our models to generalize to new structures, and a qualitative exploration of the changes in chemical representations that are induced by our training scheme.

Quantitative performance analyses began by computing the embeddings of gene expression signatures and chemical structures in the test set, using the baseline, oracle, and deep coordinated methods defined in Sections 3.2 and 3.3. Using each embedded gene expression signature as a query, we ranked all chemical structures in the test set based on their proximity to that gene expression signature in the embedding space. These rankings were then used to compute standard information-retrieval metrics (precision-recall curves, MR, MRR, and Hits at 10 or 100 (H@10 or H@100, respectively)). For our ablation analyses, we repeated the above experiments using various combinations of raw and learned GE and chemical representations.

In addition to the information retrieval analyses, we probed the generalizability of our representations by analysing the statistical relationship between the average retrieval performance for each chemical structure and the structural similarity to the most similar chemical in the training set. Similarity was measured via the Tanimoto distance between all pairs of molecular fingerprints in the train and test set. We further examined performance vs. chemical specificity, using the number of genes that a molecule, on average, affected as a measure of specificity. Finally, we visualized the latent space of our chemical embedder (versus their pre-trained representations learned by ChemProp), and noted the relative position of drugs with the same mechanism of action (MOA) in each latent space.

4. Results & Discussion

4.1. *Quantitative IR Experiments*

Baseline and Proposed Method Information retrieval results from the baseline and proposed methods are reported in Figure 2. This figure shows that our model variants offer significant performance improvements over either of the CCA or NN baselines, and even approach the performance of the train-set only oracle model. All models fall dramatically short of the test-set generalizability oracle, which indicates that while our tasks offer significant improvement over baseline models here, there are still major gains possible, primarily by focusing on improving the generalizability to the novel cell-types of the test set.

In addition, we show various ablation studies over the baseline models in Table 1 to probe what gene or molecular representations would make them better or worse. Strikingly, we can note that uniformly using aligned representations (meaning representations based on our multi-view alignment neural network architecture) offers significant improvements over other representations, indicating that even with a baseline approach such as the double nearest neighbor (D-NN) model, improvements to the embedding quality translate to notable improvements to IR performance. Notably, this is true both for GE and Chemical embedders, with Aligned-Aligned representations yielding optimal performance for both D-NN and CCA query mechanisms. Additionally, it is also clear that CCA is the preferred query metric, over either raw correlation (Corr) based lookups or D-NN based lookups.

Oracle Model Analysis Results from the oracle models are reported in Figure A1 and Table 2. As expected, oracle models using GE signatures from the test cell line greatly outperformed those using signatures from the train cell lines. In addition, aligned embeddings modestly improved the performance of test set oracles, and greatly improved the performance of train set oracles, consistent with the GE embedders learning slightly more generalizable representations. Of note, our proposed model achieved comparable results as the oracle model that leveraged raw GE signatures from the train cell lines. More specifically, our approach yielded slightly worse than the oracle on metrics (MRR, H@10) that emphasize early rankings, and slightly better on metrics (MR, H@100) that focus on more aggregate results. This is

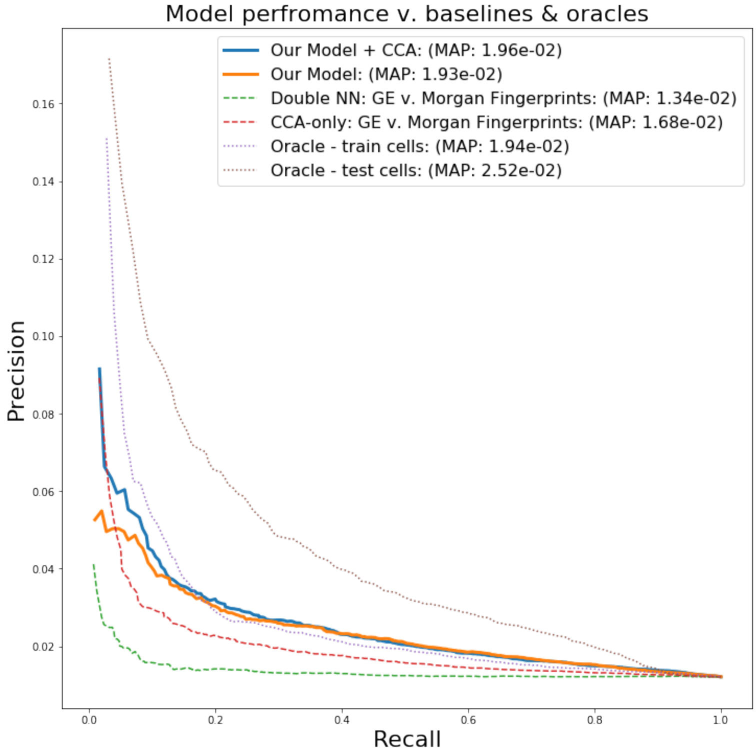

Fig. 2. Precision Recall curves for drug identification given gene expression signatures, across various baselines (dashed lines), oracles (dotted lines) and our model (solid lines).

Table 1. IR metrics across various configurations of the model/baselines. ‘Chemprop‘ refers to pretrained model from Yang et al.[6] ‘Aligned’ indicates representations learned from our method (see Section 3.3). MR=median rank, MRR=mean reciprocal rank, H@K=Hit/Recall at K.

GE	**Chemical**	**Method**	**MR**	**MRR**	**H@10**	**H@100**
Raw	Morgan FP	D-NN	206	0.025	0.037	0.240
Raw	Chemprop	D-NN	211	0.025	0.035	0.254
Raw	Aligned	D-NN	189	0.033	0.047	0.290
Aligned	Morgan FP	D-NN	214	0.025	0.041	0.256
Aligned	Chemprop	D-NN	196	0.022	0.037	0.278
Aligned	Aligned	D-NN	137	0.039	0.072	0.402
Raw	Morgan FP	CCA	180	0.027	0.045	0.303
Raw	Chemprop	CCA	184	0.024	0.040	0.294
Raw	Aligned	CCA	134	0.039	0.076	0.412
Aligned	Morgan FP	CCA	177	0.027	0.050	0.319
Aligned	Chemprop	CCA	163	0.028	0.051	0.334
Aligned	Aligned	CCA	130	**0.048**	**0.093**	0.425
Aligned	Aligned	Corr	**126**	0.042	0.085	**0.432**

also apparent in the precision-recall plot, which shows the aligned embeddings curve starting out slightly below that of the raw/train oracle curve but then moving rapidly above it as further results are considered.

The stark difference in performance between the train-cell-lines oracle and test-cell-lines oracle suggests that one of the largest barriers to performance here is the generalization gap between different cell lines. This further motivates for the curation of larger, cell-line heterogeneous datasets in the future.

Table 2. IR metrics for the various oracle methods.

Oracle Model	**MR**	**MRR**	**H@10**	**H@100**
Oracle - Train Cell Lines (Raw GE)	138	0.057	0.101	0.400
Oracle - Train Cell Lines (Embed)	110	0.064	0.128	0.466
Oracle - Test Cell Line (Raw GE)	82	0.076	0.147	0.565
Oracle - Test Cell Line (Embed)	79	0.093	0.160	0.568
Our Approach	126	0.042	0.085	0.432

4.2. *Introspection Analyses*

Figure 3A contains the results of our experiments comparing performance to distance from the training set. Regardless of the measure of chemical similarity, compound retrieval performance was inversely correlated with distance from the most similar compound in the training set. As can be seen in Supplementary Figure A2, the same trend held with learned gene expression embeddings, and was present but much weaker using raw gene expression profiles.

Figure 3B depicts the relationship between the transcriptional specificity of a compound and its ability to be retrieved using our analysis. As can be seen, there is a mild negative correlation, implying that molecules that affect the expression of many genes are easier to retrieve using this approach. Note that this observation is concordant with our findings on the difficulty of generalizing to new cell lines – drugs that affect a small, targeted set of genes are more likely to be cell line specific, and as our model is forced to surmount a significant generalization gap in evaluation, such cell-line specific signals are largely wiped out.

In addition, our analysis of the changes induced in the embedding space, shown in Supplementary Figure A4, reveal that our model's embeddings of molecules appear to better cluster shared MOAs than do the raw ChemProp embeddings, from which our model is initialized. This suggests that, as hypothesized regarding the nature of this task, our model is learning rich representations of the underlying molecules, though additional work remains to investigate this effect more thoroughly.

4.3. *Future Work*

We see several opportunities for further work on this task. First, expanding our data coverage, across molecules, cell lines, dosages, and treatment durations will allow us to measure and improve generalizability here. Second, exploring additional strategies to use the over-sampled

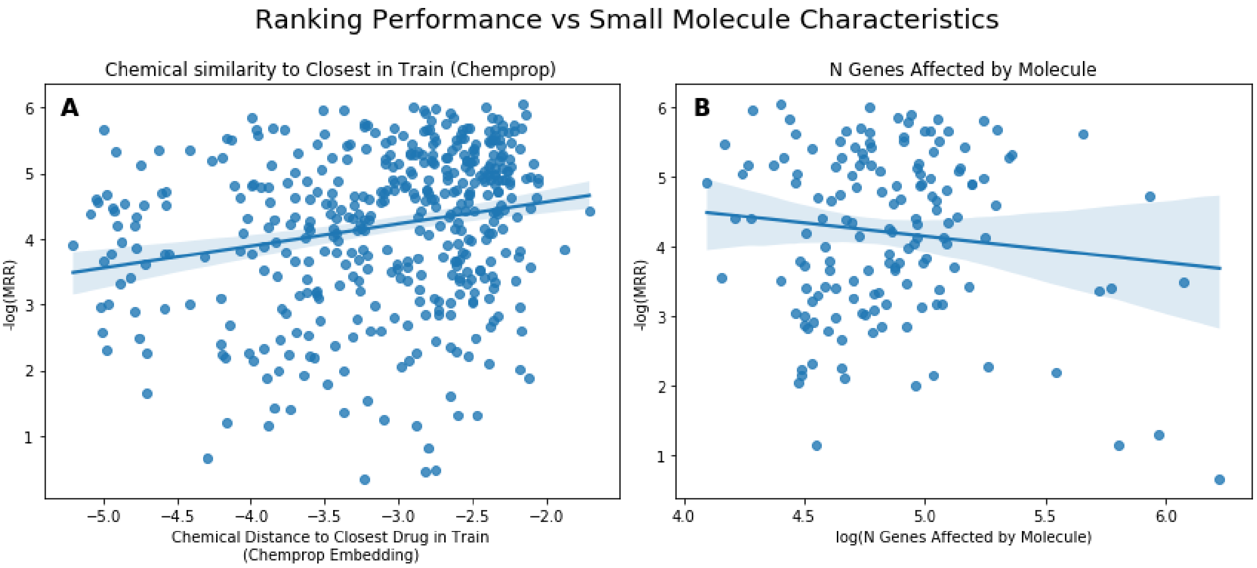

Fig. 3. Left: Performance (lower is better) vs. Structural distance to the nearest compound in the training set. See Appendix Figure A2 for analogous plots for four additional measures of distance from the training set. Right: Average Performance vs. # of Genes deferentially expression following treatment with the molecule of interest.

nature of these data (e.g., ensembling together control and perturbational signatures to reduce variance) could be beneficial. Third, a more robust exploration of model architectures and losses, could offer improvements here, as has been seen in other contexts.[22] Lastly, we recommend exploring methods to improve cell-line generalizability, e.g. incorporating information across many cell lines when forming predictions.

5. Conclusion

We present a new task: cross-modal multi-view alignment between drug structures and perturbational gene expression profiles. We argue this task fills a notable gap in the current suite of cheminformatics tasks; namely, it links molecular structure to an objective, functional readout of drugs with very broad biomedical relevance. We profile state-of-the-art representation learning methods on this new task, and inspect the learned chemical embeddings. We find that this modeling task induces an embedding space reflective of drug mechanism of action – which is not explicitly included in the training regime – and see modest generalization to both new structures and a new biological environment.

As more scientists consider using connectivity mapping techniques, we hope that this task will prove helpful in enabling a rapid, in silico evaluation of which compounds under consideration should be profiled further via gene expression or other assays. That being said, our oracle experiments highlight potential obstacles to deploying even classical connectivity mapping techniques in the (almost universal) scenario in which the results are intended to generalize to tissues other than those in which the perturbational assays were performed.

6. Acknowledgements

The authors thank Connor Coley and Kyle Swanson for graciously providing the pretrained chemical embedder from Yang et al.[6] S.G.F. was supported by training grant T32GM007753 from the National Institute of General Medical Science. M.B.A.M. was funded in part by National Institutes of Health: National Institutes of Mental Health grant P50-MH106933 as well as a Mitacs Globalink Research Award. The content is solely the responsibility of the authors.

References

1. T. Hansson, C. Oostenbrink and W. van Gunsteren, Molecular dynamics simulations, *Current opinion in structural biology* **12**, 190 (2002).
2. M. Krallinger, O. Rabal, A. Lourenco, J. Oyarzabal and A. Valencia, Information retrieval and text mining technologies for chemistry, *Chemical reviews* **117**, 7673 (2017).
3. P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner *et al.*, Relational inductive biases, deep learning, and graph networks, *arXiv preprint arXiv:1806.01261* (2018).
4. Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing and V. Pande, Moleculenet: a benchmark for molecular machine learning, *Chemical science* **9**, 513 (2018).
5. S. Kearnes, K. McCloskey, M. Berndl, V. Pande and P. Riley, Molecular graph convolutions: moving beyond fingerprints, *Journal of computer-aided molecular design* **30**, 595 (2016).
6. K. Yang, K. Swanson, W. Jin, C. W. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea *et al.*, Analyzing learned molecular representations for property prediction, *Journal of chemical information and modeling* (2019).
7. Y.-C. Lo, S. E. Rensi, W. Torng and R. B. Altman, Machine learning in chemoinformatics and drug discovery, *Drug discovery today* **23**, 1538 (2018).
8. B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding and V. Pande, Massively multitask networks for drug discovery, *arXiv preprint arXiv:1502.02072* (2015).
9. F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley and O. A. Von Lilienfeld, Prediction errors of molecular machine learning models lower than hybrid dft error, *Journal of chemical theory and computation* **13**, 5255 (2017).
10. J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals and G. E. Dahl, Neural message passing for quantum chemistry, in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017.
11. A. Musa, L. S. Ghoraie, S.-D. Zhang, G. Glazko, O. Yli-Harja, M. Dehmer, B. Haibe-Kains and F. Emmert-Streib, A review of connectivity map and computational approaches in pharmacogenomics, *Briefings in bioinformatics* **19**, 506 (2017).
12. J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross *et al.*, The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease, *science* **313**, 1929 (2006).
13. A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli, J. K. Asiedu *et al.*, A next generation connectivity map: L1000 platform and the first 1,000,000 profiles, *Cell* **171**, 1437 (2017).
14. A. Musa, S. Tripathi, M. Kandhavelu, M. Dehmer and F. Emmert-Streib, Harnessing the biological complexity of big data from lincs gene expression signatures, *PloS one* **13**, p. e0201937 (2018).
15. T.-P. Liu, Y.-Y. Hsieh, C.-J. Chou and P.-M. Yang, Systematic polypharmacology and drug

repurposing via an integrated l1000-based connectivity map database mining, *Royal Society open science* **5**, p. 181321 (2018).

16. N. R. Clark, K. S. Hu, A. S. Feldmann, Y. Kou, E. Y. Chen, Q. Duan and A. Maayan, The characteristic direction: a geometrical approach to identify differentially expressed genes, *BMC bioinformatics* **15**, p. 79 (2014).
17. Y. Donner, S. Kazmierczak and K. Fortney, Drug repurposing using deep embeddings of gene expression profiles, *Molecular pharmaceutics* **15**, 4314 (2018).
18. A. B. Dincer, S. Celik, N. Hiranuma and S.-I. Lee, Deepprofile: Deep learning of cancer molecular profiles for precision medicine, *bioRxiv* , p. 278739 (2018).
19. L. Rampášek, D. Hidru, P. Smirnov, B. Haibe-Kains and A. Goldenberg, Dr. vae: improving drug response prediction via modeling of drug perturbation effects, *Bioinformatics* (2019).
20. J. Cheng, Q. Xie, V. Kumar, M. Hurle, J. M. Freudenberg, L. Yang and P. Agarwal, Evaluation of analytical methods for connectivity map data, in *Biocomputing 2013*, (World Scientific, 2013) pp. 5–16.
21. H. Hotelling, Relations between two sets of variates, in *Breakthroughs in statistics*, (Springer, 1992) pp. 162–190.
22. Y. Li, M. Yang and Z. M. Zhang, A survey of multi-view representation learning, *IEEE Transactions on Knowledge and Data Engineering* (2018).
23. C. Deng, Z. Chen, X. Liu, X. Gao and D. Tao, Triplet-based deep hashing network for cross-modal retrieval, *IEEE Transactions on Image Processing* **27**, 3893 (2018).
24. G. Koch, R. Zemel and R. Salakhutdinov, Siamese neural networks for one-shot image recognition, in *ICML deep learning workshop*, 2015.
25. T. M. Filzen, P. S. Kutchukian, J. D. Hermes, J. Li and M. Tudor, Representing high throughput expression profiles via perturbation barcodes reveals compound targets, *PLoS computational biology* **13**, p. e1005335 (2017).
26. E. Hoffer and N. Ailon, Deep metric learning using triplet network, in *International Workshop on Similarity-Based Pattern Recognition*, 2015.
27. C.-Y. Wu, R. Manmatha, A. J. Smola and P. Krahenbuhl, Sampling matters in deep embedding learning, in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.
28. D. Weininger, A. Weininger and J. L. Weininger, Smiles. 2. algorithm for generation of unique smiles notation, *Journal of chemical information and computer sciences* **29**, 97 (1989).
29. G. Landrum *et al.*, Rdkit: Open-source cheminformatics (2006).
30. D. Rogers and M. Hahn, Extended-connectivity fingerprints, *Journal of chemical information and modeling* **50**, 742 (2010).
31. R. Hodos, P. Zhang, H.-C. Lee, Q. Duan, Z. Wang, N. R. Clark, A. Maayan, F. Wang, B. Kidd, J. Hu *et al.*, Cell-specific prediction and application of drug-induced gene expression profiles, in *Pacific Symposium on Biocomputing*, 2017.
32. M. McDermott, J. Wang, W. N. Zhao, S. D. Sheridan, P. Szolovits, I. Kohane, S. J. Haggarty and R. H. Perlis, Deep learning benchmarks on l1000 gene expression data, *IEEE/ACM transactions on computational biology and bioinformatics* (2019).
33. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, Scikit-learn: Machine learning in python, *Journal of machine learning research* **12**, 2825 (2011).
34. J. Bergstra, D. Yamins and D. D. Cox, Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures, in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML13 (JMLR.org, 2013).

Appendix A.

7. Appendix

7.1. *Data Access*

Raw LINCS data are available for download at `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70138` and `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92742`. Our preprocessed versions of these files will be distributed with our Github repository.

7.2. *Final Optimal Parameters*

Table A1. Dataset cell lines, number of samples, and number of drugs per split.

	Train	Validation	Test
Cell Lines	HEPG2, HA1E, HCC515, VCAP, A375, PC3, MCF7	A549	HT29
# Samples	97210	1255	2317
# Drugs	3490	436	437

Final, optimal hyperparameters were selected according to the median rank of the parameter setting. Final values are shown in Table A2. We see lots of consistency between the two settings, despite the fact that both searches were independent. For example, both prefer very similar gene expression architectures, both run for 12 epochs, and both have relatively low margins.

Table A2. Final optimal hyperparameters for the deep neural network architecture.

Parameter	Drug Split
# Epochs	12
LR	5.6e-4
Batch Size	417
Rank Transform	No
Neg. Samp.	N/A
Gene Exp. Emb. Arch.	677, 1159
Embed Size	51
Dropout Prob	0.02
Activation	SELU
Linear Bias	False
Beta	1.3
Margin	0.31
RdKit Ft.	True

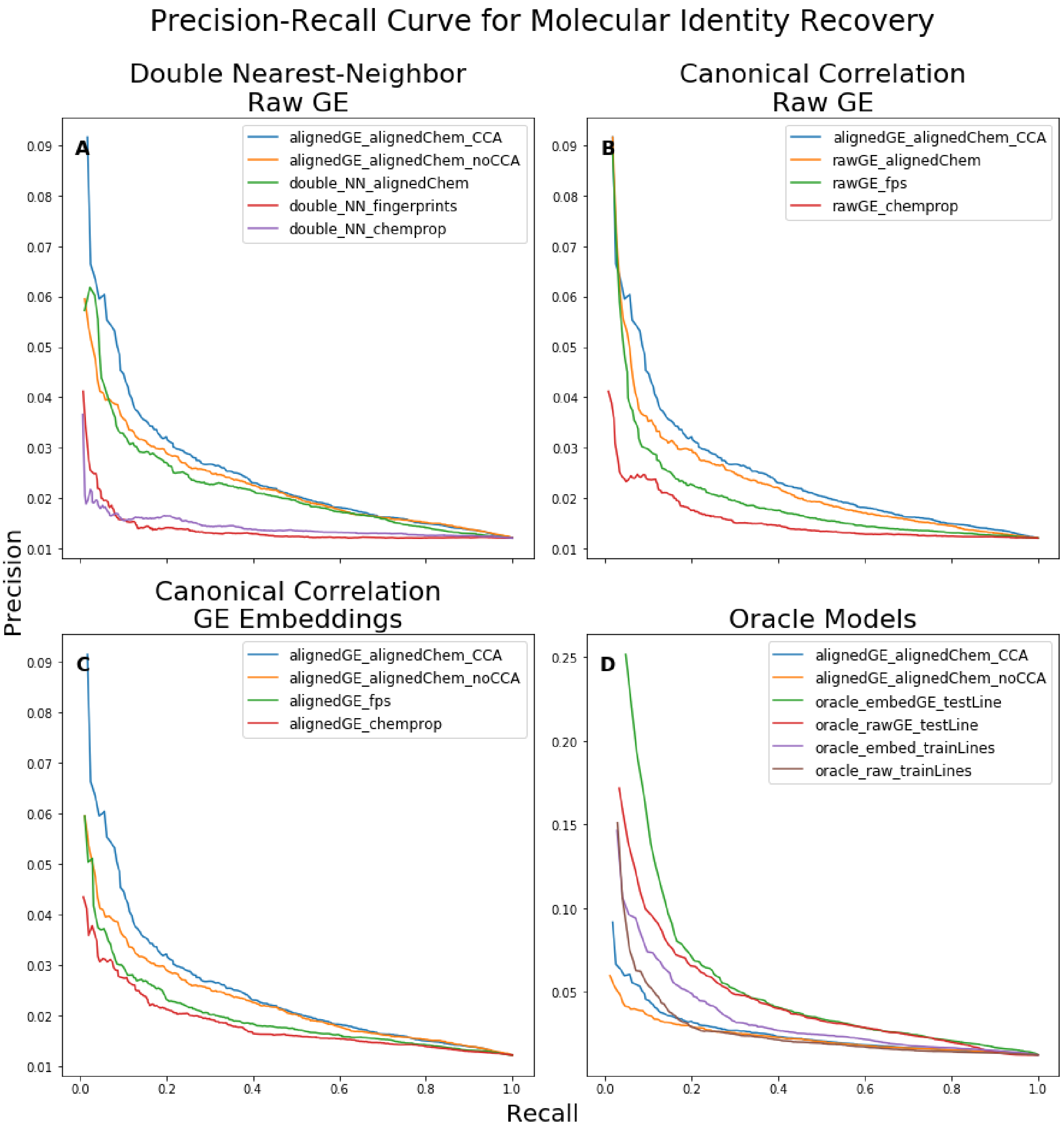

Fig. A1. Precision Recall curves for drug identification given gene expression signatures. The "alignedGE_alignedChem_CCA" and "alignedGE_alignedChem_noCCA" curves, plotted in blue and orange, respectively, in all plots, represents our approach as described in Section 3.3 with and without CCA post-processing. All other methods represent baseline or oracle methods as described in Section 3.2.

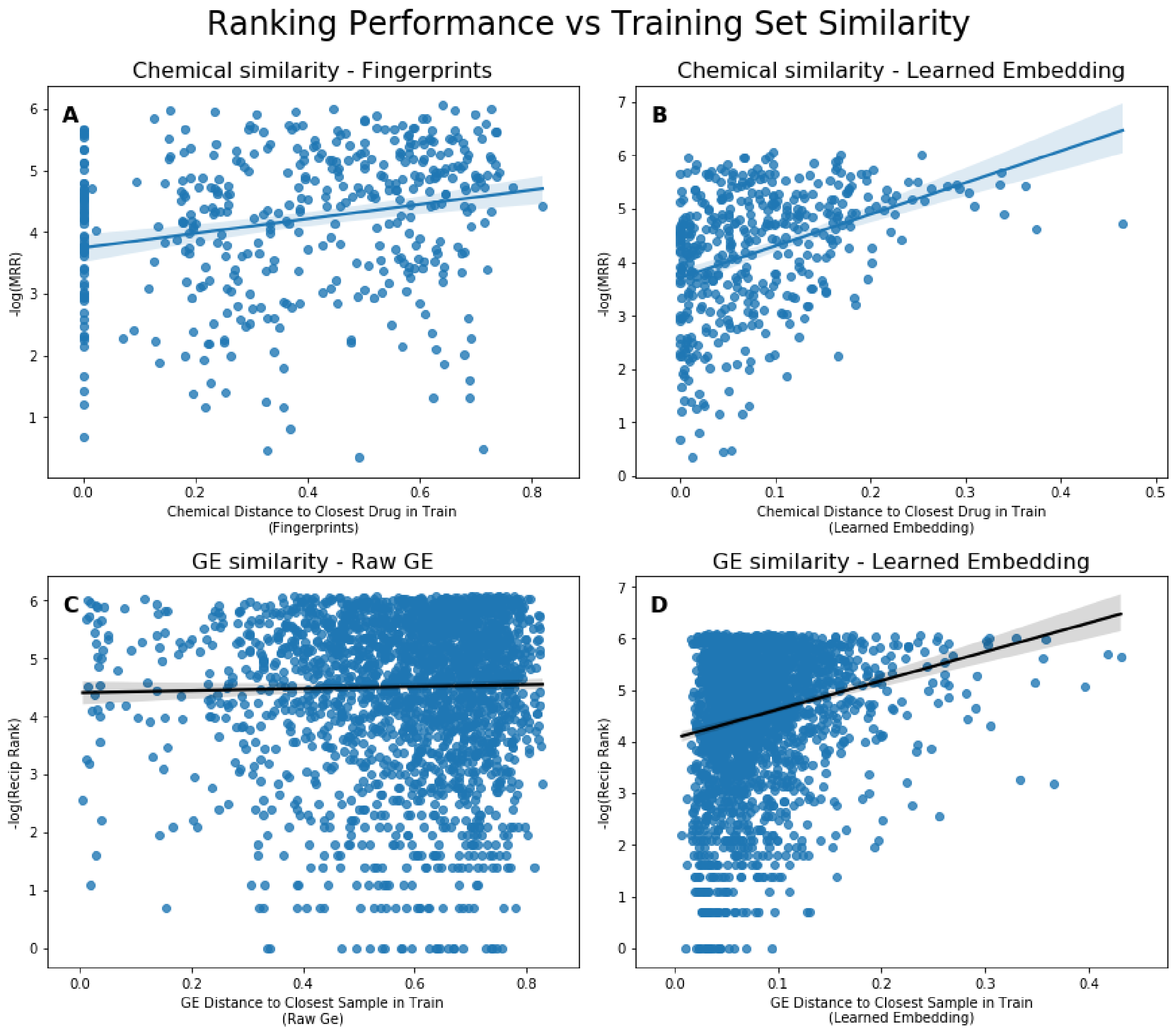

Fig. A2. Alternative measures of performance as a function of distance from the training set. Accompanies Figure 3. Note that while some test-set molecules have tanimoto distance of 0 to the train set, this does not imply they were shared between the two sets; rather, there are cases of two distinct molecules with identical molecular fingerprints. See Figure A3 for several examples.

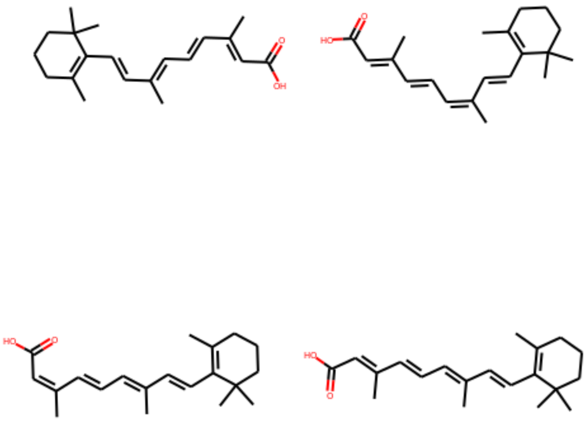

Fig. A3. Examples of four molecules that have slightly distinct structures, but have identical Morgan Circular Fingerprints.

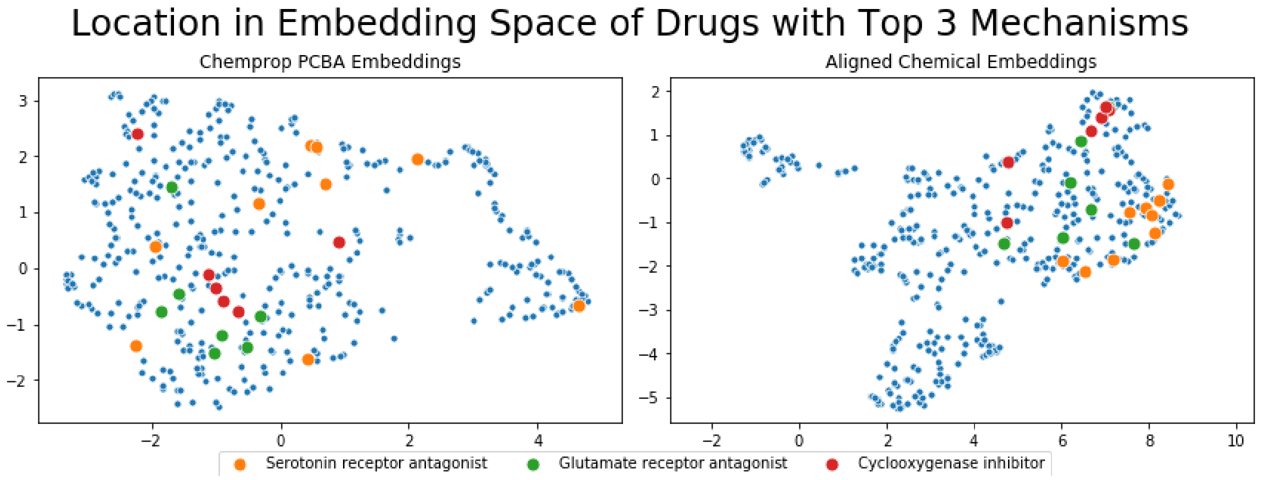

Fig. A4. This figure shows our induced embeddings (right) compared to established chemical embeddings (left), projected to 2-dimensions using UMAP, colored according to MOA. (Blue dots represent small molecules with different or unknown MOAs.)

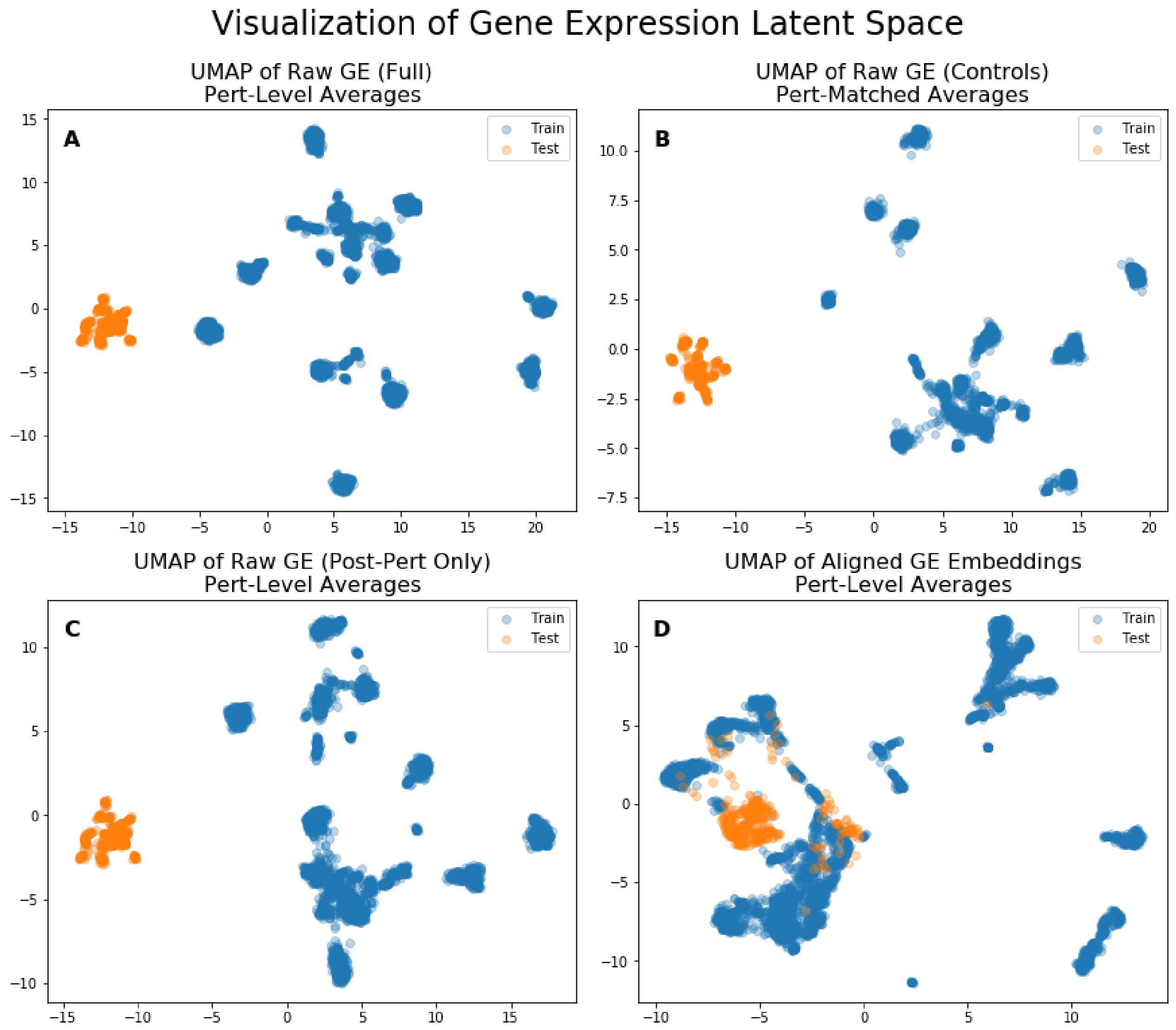

Fig. A5. UMAP latent space of gene expression signatures, colored by train vs test set. A: Full raw gene expression signatures (post-perturbational signature, plate control signature, log-fold change). B: Raw gene expression from control signatures. C: Raw gene expression values, only the post-perturbational measurments. D: Aligned embedding representations.