Evaluating the Use of Foundational Chemical Language Models in Multimodal Graph Fusion

Collin Francel
Department of Computer Science
University of Alabama
Tuscaloosa, AL 35487
ctfrancel@crimson.ua.edu

Massimiliano Lupo-Pasini
Computational Sciences and Engineering Division
Oak Ridge National Laboratory
Oak Ridge, TN 37830
lupopasinim@ornl.gov

Zachary R Fox
Computational Sciences and Engineering Division
Oak Ridge National Laboratory
Oak Ridge, TN 37830
foxzr@ornl.gov

Abstract

Rapid and accurate prediction the physicochemical properties of molecules given their structures remains a key challenge in cheminformatics. Machine learning approaches offer high-throughput options, but the relevant inductive biases and data representations are up for debate. For example, masked language models (MLMs) can be trained in a self-supervised way on hundreds of millions to billions of readily available SMILES strings. However, this model will not be clued into the rich geometric information found in molecular structures. Another option is graph neural networks (GNNs), which can operate directly on molecular structures. Yet, generating these molecular graphs is computationally intensive, leading to a relative scarcity in data compared to SMILES strings. This makes it challenging to scale to large pre-trained GNNs, in contrast to the situation for LMs. As such, it becomes attractive to combine these two paradigms, benefiting from pre-training on a large corpus of SMILES strings and embedding these representation into a fine-tuning that uses geometric information. Despite the promise of such an approach, contrary to previous studies, we find mixed results with the combination of the LMs and GNNs on several molecule datasets. In particular, we found evidence for improvement on the FreeSolv and QM7 benchmarks, but degraded performance on the ESOL, LIPO and QM9 datasets compared to a GNN baseline.

1 Introduction

Accurately and efficiently predicting the physicochemical properties of molecules based on their structures is a fundamental challenge in the field of cheminformatics. While quantum and classical mechanical simulations provide valuable insights, advances in machine learning (ML) have introduced new possibilities for high-throughput predictions. However, several key challenges remain, particularly around the choice of inductive biases and the optimal representation of molecular data.

Inductive biases, which refer to the assumptions embedded within a model architecture, play a critical role in the success of machine learning models, particularly when data are limited. It is up for debate over the most appropriate inductive biases for molecular prediction tasks, with different models offering distinct advantages depending on the specific properties being predicted.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Equally important is the representation of molecular structures in a form that is both informative and efficient for machine learning algorithms to process. Traditional representations, such as SMILES (Simplified Molecular Input Line Entry System) strings [13] or molecular fingerprints, have been widely used [5, 12, 2, 1, 7], but they may not fully capture the intricate spatial and electronic features of molecules. Graph neural networks (GNNs), which treat molecules as graphs where atoms are nodes and bonds are edges, provide a promising inductive bias for processing molecular graphs. Several GNN architectures have been introduced to work with 3D molecule graphs, including DimeNet [6] and SchNet [10].

The key scientific question to address is (1) how can the extensive knowledge about chemical grammar within chemical language models be used, and (2) for what tasks (3) and under what data regimes are they useful. We address the first using a multimodal data fusion method that embeds representations from the pretrained chemical LM into GNN node embeddings. This method builds off of a similar work, MolPROP [9]. Our work mainly differs from MolPROP in that we use 3D graphs over 2D graphs, and use ChORBERT over ChemBERTa-2.The second point is addressed by exploring several tasks that may depend on different physical phenomena and benchmarks, including hydration free energy, aqueous solubility, lipophilicity, as well as QM7 and QM9 properties.

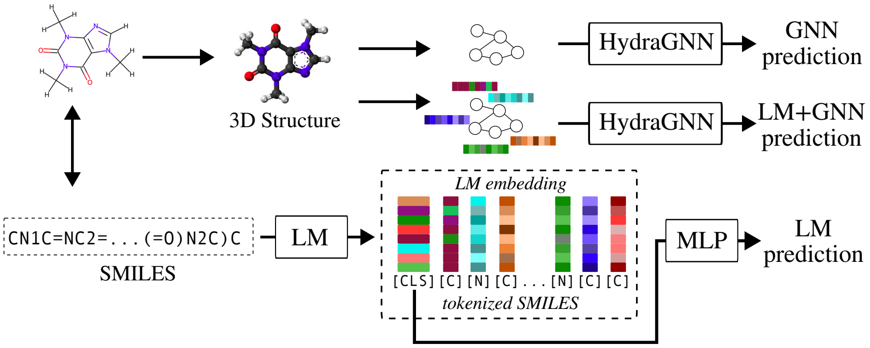


Figure 1: *Overview of multimodal graph fusion workflow.* Atom-wise embeddings of a molecule are obtained from a pretrained language model via its SMILES string, and concatenated to node features in a 3d molecular graph representation. We then train a GNN on this graph to predict downstream molecular properties.

2 Related Works

Integration of pre-trained protein language models into geometric deep learning networks by [14], trains a Protein Language Model (PLM) on amino acid sequences, creating amino acid embeddings that are used as features for a graph. They then train a Graph Geometric Neural Network (GGNN) on these enhanced graphs, finding improvements on most benchmarks compared to baselines and existing SOTA methods. They hypothesized that the relative scarcity in 3D geometric structures as compared to amino acid sequences allowed the PLM to supplement the GGNN approach, leading to the observed performance boost.

MolPROP: Molecular Property prediction with multimodal language and graph fusion by [9] mirrors the work done in [14], working with molecules instead of proteins. MolPROP generates 2D graphs from the SMILES string. MolPROP then uses ChemBERTa-2, a pre-trained chemical language model trained on SMILES strings, to create embeddings to concatenate to node features on the graph, similar to our approach. MolProp then trains a GAT or GCN based GNN on these augmented graphs, achieving improved performance on the FreeSolv and ESOL benchmarks and competitive results on Lipo and QM7 but poor results on the classification tasks examined - BACE, BBBP, and ClinTox.

3 Methods

Models

To evaluate the utility of fusing a pretrained chemical language model's embeddings into a graph neural network for molecular property prediction, we compare three modeling approach. First, we use the `[CLS]` token of the pretrained chemical language model [4], and add a feed-forward layer to predict the molecular property of interest. We refer to this approach as LM. The second baseline is a graph neural network DimeNet [6] implemented in [8], which receives the atomic positions and types as input, which we refer to as GNN. Finally in the multimodal graph fusion approach, we extract the per-atom embeddings from the tokenized SMILES strings and assign them as additional node features in the molecular graph, as shown in Fig. 1.

Datasets

We evaluate our approach on several benchmark datasets: FreeSolv, ESOL, Lipophilicity (Lipo) and QM7 [15]. FreeSolv is a dataset of experimental and calculated hydration-free energies for 642 small molecules in water. ESOL is a dataset of 1128 compounds and their solubility. Lipo is a dataset of 4200 molecules from the ChEMBL database with their corresponding experimentally dervied lipophilicities. Lastly, QM7 is a subset of the GDB-13 molecule database, consisting of 6830 molecules which have up to 7 heavy atoms (Carbon, Nitrogen, Oxygen, and Sulfur) along with their respective atomization energies, which are derived from Density Functional Theory (DFT) calculations.

To compare with [9], 80% of the data were used for training, with 10% reserved for testing and validation. We used the Bemis-Murcko [3] scaffold split to separate molecules of the same scaffold, making model performance on the test set more indicative of model performance on unseen molecules.

For each dataset, we needed to have both a 3D graph and corresponding SMILES string available. These datasets did not have 3D structural information available. We used the `RDKit` library to generate them by using the `EmbedMolecule` function to to position the atoms in 3D space, and then Merck Molecular Force Field (MMFF) [11] refined the 3D structure. Though we expect these generated strucutres to be lower quality compared to those computed with more complex DFT simulations, this allows us to get a basis for comparison with MolPROP that have been tested on FreeSolv, ESOL, Lipo and QM7.

We additionally looked at the QM9 dataset [15]. QM9 is a subset of 134 thousand molecules with up to 9 heavy atoms (Carbon, Nitrogen, Oxygen and Fluorine) from GDB-17. The molecules were modeled using DFT, producing high quality 3D structures, and thus 3D graphs, as well as 12 different molecular properties. Out of these 12, we focused on predicting dipole moment, as this is a complex property that depends on both the molecular geometry and electronic structure. For this dataset, we used a random split of 70/15/15.

Training

To ensure a fair evaluation across the different models, we optimized hyperparamters using `Optuna`. For each method (LM, GNN, LM+GNN), we ran 48 trials for 50 epochs each on FreeSolv, ESOL, Lipo and QM7, and 100 epochs for QM9. We used an early stopping patience of 10 for all trials. Parameters were chosen with the `TPE` sampler. All HPO and model training were performed on AMD MI250X GPUs on the Frontier supercomputer at Oak Ridge National Laboratory.

Evaluation

After HPO, we apply the best hyperparameters to an 8 fold cross validation for FreeSolv, ESOL, Lipo and QM7. We then report the average RMSE plus or minus the standard deviation for models trained on each of the 8 training sets. For the QM9 dataset, we only run one trial of the best hyperparameters on our split and report a single test validation.

Additionally, we trained models on various subsets of the combined training and validation dataset to determine whether each method's performance depended on dataset size. We ran Lipo and QM7 with subsets of size 250, 500, 750, 1,250, and QM9 with subsets of size 250, 500, 750, and 1,000. At each

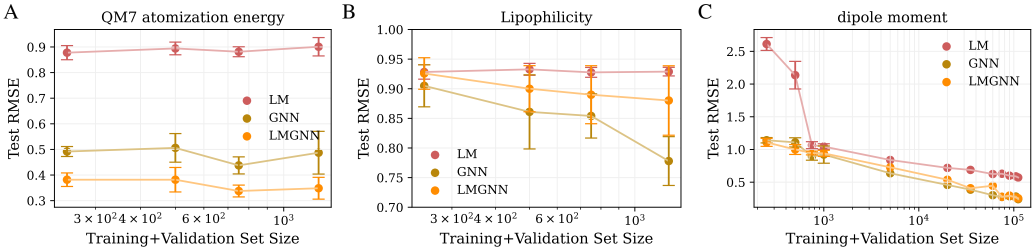


Figure 2: *Model performance scales with dataset size.* Performance of the different model types on random subsets of the (A) Lipophilicity, (B) QM7 atomization energy, and (C) QM9 dipole moment datasets. Each data point for a training set size less than 1,300 is averaged over 8 trials, error bars being one standard deviation. Results for datasets larger than 1,000 for (QM9 data, (C)) are from a single training run.

subset size, we run 8 trials with different subsets and report the average and standard deviation. For QM9, we additionally trained 1 trial for subset sizes 5,000, 20,000, 35,000, 60,000, 75,000, 90,000, 105,000, and 111,206 (the combined train and validation set).

On QM9, we use Principal Component Analysis (PCA) to examine how well the GNN and LM+GNN are representing the molecules. We select one of the 8 models (or the only model for larger subsets) that were trained on a given subset size, and use it to produce embeddings on the full test dataset. We then use PCA to visualize the embedding space. We expect better models will reveal more structured representations, with points in the test set having a smooth gradient from low dipole moment to high dipole moment and each molecule being close to others with similar dipole moments.

4 Results

Model	FreeSolv	ESOL	Lipo	QM7	QM9 μ_0
Size	642	1128	4191	6830	130831
Metric	RMSE ↓	RMSE ↓	RMSE ↓	RMSE ↓	MAE ↓
LM	4.136 ± 0.105	2.240 ± 0.008	1.114 ± 0.008	197.151 ± 2.283	0.572
GNN	2.069 ± 0.521	**0.746 ± 0.159**	**0.414 ± 0.0290**	72.968 ± 11.993	**0.241**
LM+GNN	**1.811 ± 0.454**	0.929 ± 0.147	0.539 ± 0.138	**70.291 ± 11.350**	0.265
ChemBERTa-2$_{\text{77M-MTR}}$	2.515 ± 0.00	1.025 ± 0.00	0.987 ± 0.00	147.9 ± 0.00	N/A
ChemBERTa-2$_{\text{77M-MLM}}$	2.047± 0.00	0.889 ± 0.00	0.798 ± 0.00	172.8 ± 0.00	N/A
MolPROP$_{\text{MTR+GAT}}$	2.05 ± 0.16	0.991 ± 0.11	0.799 ± 0.01	**131.8 ± 11.2**	N/A
MolPROP$_{\text{MLM+GAT}}$	**1.70 ± 0.09**	**0.777 ± 0.02**	**0.733 ± 0.02**	151.8 ± 10.0	N/A

Table 1: Performance comparison of different models across various datasets. Values represent mean ± standard deviation. 8-fold cross validation was performed on all datasets except QM9. Only 1 trial was run on the full QM9 dataset. We also report numbers from [9] from their two best models and their numbers for ChemBERTa-2 MLM and MTR, which are numbers from 10-fold cross validation evaluated on the same test set used in this paper.

We ran 8-fold cross validation for FreeSolv, ESOL, Lipo and QM7, obtaining a mean and standard deviation of test loss for each model. The results, along with numbers sourced from [9], are shown in Table 1. The LM+GNN model appears to perform competitively on FreeSolv and QM7, though the result is similar to the GNN baseline for QM7. We see that MolPROP similarly performs well on FreeSolv, but is not competitive to our GNN baseline on ESOL, Lipo, and QM7. We also found our LM approach to do quite poorly, suggesting that the ChORBERT model may not be well suited for predicting molecular properties on these datasets as compared to ChemBERTa-2 examined in MolPROP. We hypothesize this may be due to the difference in pre-training dataset (PubChem for ChemBERTa-2 and Enamine for ChORBERT), or possibly an issue with the regex tokenization approach used in this paper.

On QM9, we found that the performance of the performance of the LM+GNN and GNN were comparable, while the LM approach performed relatively poorly.

In our scaling experiments (Fig. 2), we see that the best model type is fairly consistently better or about the same as the next best model type for the dataset for different dataset sizes. There is no clear winner at a smaller dataset size. On the QM9 dataset, there is some evidence for the LM+GNN performing better at a lower dataset size, but this doesn't appear to be significantly different.

In Table 4, we see the result of the PCA algorithm reducing QM9 dipole moment model embeddings into 2D space using the first two principle components. We may then use these graphs to qualitatively assess model performance by how well the model appears to have structured its embedding space. For a fully trained model, we expect a smooth gradient from low dipole moment to high dipole moment. This can be seen in the full dataset (111,206 size train+validation set), where the GNN has a highly ordered structure going from low dipole moment to high dipole moment. In particular, we analyze dataset sizes where the LM+GNN may be doing better in the scaling graph. For the subset size of 500, where the LM+GNN appears to do somewhat better in the scaling graph, we see that the LM+GNN does have somewhat more structured embedding space, suggesting it does have a somewhat better representation compared to the GNN alone. On dataset sizes 750 and 1000, it is less clear which approach has a richer embedding space, mirroring the similar test MAE values in the scaling graph.

5 Discussion

In conclusion, we found that our combination of the LM+GNN did not conclusively show benefit compared to a simpler GNN approach for regression tasks. We found some evidence for improved performance on the FreeSolv and QM7 benchmarks, but evidence for degraded performance compared to a GNN baseline for ESOL, Lipo and QM9. In contrast, MolPROP finds state of the art performance on FreeSolv, surpassing our LM+GNN, as well as performance comparable to our GNN baseline for ESOL. Finally, on QM7, we find our method to be competitive against a GNN baseline, while MolPROP is significantly degraded on the benchmark.

We speculate differences could be due to our use of the ChORBERT model as compared to ChemBERTa. ChORBERT is trained on an augmented Enamine dataset, while ChemBERTa-2 is trained on molecules from PubChem, possibly leading to differences in performances on molecules closer to their respective training datasets. Further work is needed to determine if our results depend on the language model used. An additional difference is our regex tokenization strategy. We chose this approach to simplify mapping embeddings from to the molecular graph, but it is possible that this choice of tokenization leads to worse performance for the language model. Future work will examine using a different tokenization strategy, such as the WordPiece tokenizer or digits-based tokenizer mentioned in [4].

Another area of difference is our use of DimeNet, a GNN architecture better suited for 3D molecular data, over MolPROP's use of GCN and GAT architectures. Further examination might explore using our LM+GNN method with these different architectures.

We hypothesize that the language model does provide some structure for training, especially on smaller datasets, but ultimately becomes much less useful than the 3D graph data, explaining the inconsistent performance improvements. This can be seen in the LM+GNN's more clear performance improvement on the smallest dataset, FreeSolv, as well as its competitive performance for small subsets of QM9 on dipole moment prediction.

Since we expect the 3D graph data to be much more useful for learning, we hypothesize that the language model embeddings cause the model to lose performance by utilizing them even when enough higher quality 3D graph data is available. We expect that the LM+GNN model eventually learns to avoid attending to the language model embeddings, leading to a convergence in performance between the GNN and LM+GNN on large datasets. This can be seen in the QM9 scaling curve, where the test MAE of the two models becomes similar at the four largest subset sizes.

We speculate that fine-tuning a model pre-trained on 3D graph data will lead to better performance than using a language model, as it can make use of the much higher quality graph data. Future work should focus on producing such a pre-trained graph model and comparing to existing baselines.

Acknowledgments and Disclosure of Funding

This work was supported by a non-employee stipend (CF) from the Oak Ridge Associated Universities (ORAU) through the Research Student Internship (RSI) program. This research was sponsored by the Artificial Intelligence Initiative (ZF, MLP) within Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy. This work used resources of the Oak Ridge Leadership Computing Facility, which is supported by the Office of Science of the U.S. Department of Energy, under INCITE award CPH161.

References

[1] Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models, 2022.

[2] Suryanarayanan Balaji, Rishikesh Magar, Yayati Jadhav, et al. Gpt-molberta: Gpt molecular features language model for molecular property prediction. *arXiv preprint arXiv:2310.03030*, 2023.

[3] Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, January 1996.

[4] Andrew E Blanchard, Mayanka Chandra Shekar, Shang Gao, John Gounley, Isaac Lyngaas, Jens Glaser, and Debsindhu Bhowmik. Automating genetic algorithm mutations for molecules using a masked language model. *IEEE Trans. Evol. Comput.*, 26(4):793–799, August 2022.

[5] Manajit Das, Ankit Ghosh, and Raghavan B Sunoj. Advances in machine learning with chemical language models in molecular property and reaction outcome predictions. *Journal of Computational Chemistry*, 45(14):1160–1176, 2024.

[6] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs, 2020.

[7] Ross Irwin, Spyridon Dimitriadis, Jiazhen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, January 2022.

[8] Massimiliano Lupo Pasini, Samuel Reeve, Pei Zhang, and Jong Youl Choi. HydraGNN, 2021.

[9] Zachary A Rollins, Alan C Cheng, and Essam Metwally. MolPROP: Molecular property prediction with multimodal language and graph fusion. *J. Cheminform.*, 16(1):56, May 2024.

[10] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller. Schnet – a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24), March 2018.

[11] Paolo Tosco, Nikolaus Stiefl, and Gregory Landrum. Bringing the mmff force field to the rdkit: implementation and validation. *Journal of Cheminformatics*, 6(1), July 2014.

[12] Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, pages 429–436, 2019.

[13] David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36, February 1988.

[14] Fang Wu, Lirong Wu, Dragomir Radev, Jinbo Xu, and Stan Z Li. Integration of pre-trained protein language models into geometric deep learning networks. *Commun. Biol.*, 6(1):876, August 2023.

[15] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

A Appendix / supplemental material

Hyperparameter optimization

The tables below contain the hyperparameters which were optimized as the adapter layers for the LM (Table 2) and the LM+GNN approach (Table 3.

Hyperparameter	Search Range	Data Type
Dropout Rate	$[0.1, 0.5]$	Float
Learning Rate	$[10^{-6}, 10^{-1}]$	Float
Number of Layers	$[1, 5]$	Integer
ExponentialLR Gamma	$[0.9, 0.999]$	Float

Table 2: LM Adapter search space.

Hyperparameter	Search Range	Data Type
Hidden Dimension	$[50, 150]$ or $[50, 300]$ (QM9)	Integer
Number of Convolutional Layers	$[1, 5]$	Integer
Number of Head Layers	$[1, 2]$ or $[1, 3]$ (QM9)	Integer
Number of Shared Layers	$[1, 3]$ or $[1, 5]$ (QM9)	Integer
Dimension of Shared Layers	$[32, 100]$	Integer
Dimension of Head Layers	$[50, 100]$ (per layer)	Integer (multiple)
ExponentialLR Gamma	$[0.9, 0.999]$	Float
Learning Rate	$[10^{-6}, 10^{-2}]$ (log scale)	Float

Table 3: GNN and LM+GNN search space. On columns with another search range, the dataset which used the alternative search range is specified in parentheses.

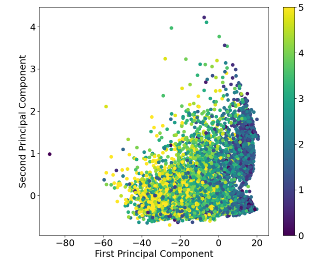
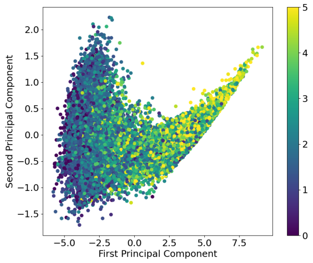
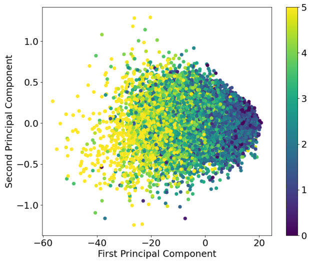
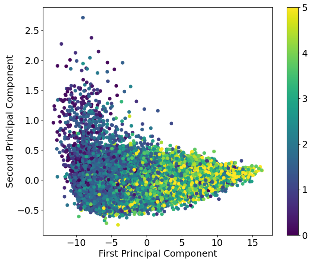
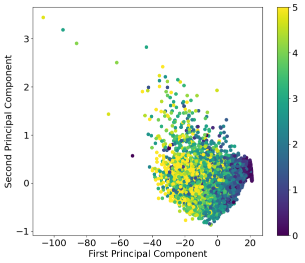
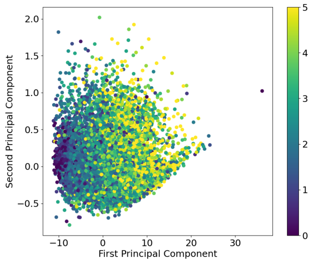
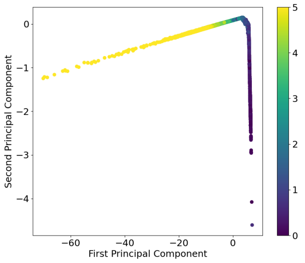
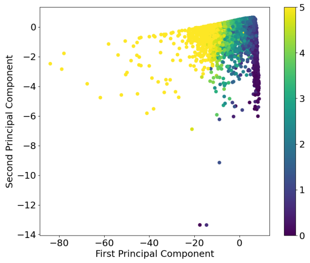
Subset size	GNN	LM+GNN
500		
750		
1000		
111206		

Table 4: Dimensionality reduction of embeddings for GNN and LM+GNN on the QM9 dipole moment task. Uses a randomly selected model from the 8 trials used in the scaling tests and runs PCA on test set embeddings.