GeneLLM: Unveiling gene functions through literature-driven transformer embeddings

Ala Jaraweh*[1,2], Oladimeji Macaulay*[2], David A Arredondo*[2], Olufunmilola M Oyebamiji[2], Luis Tafoya[2], Kushal Virupakshappa[2], Avinash Sahu[1,2,3]

[1]Department of Computer Science, [2]UNM Comprehensive Cancer Center, [3]NM VA Medical Center

{KVirupakshappa, Asahu} @salud.unm.edu

avisahuai.com

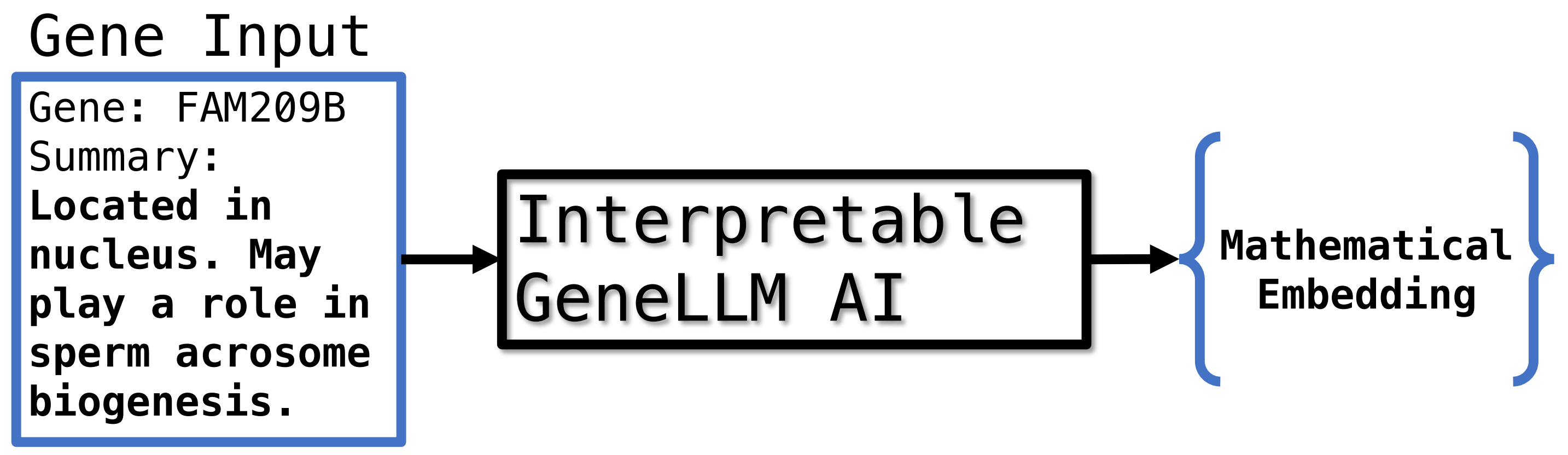
NIH R00CA248953

COMPREHENSIVE CANCER CENTER

TUMOR AI

Most bioinformatics models do not incorporate text data.

GeneLLM employs a Large Language Model trained on biomedical texts to understand summaries of gene functions and interactions, translating this knowledge into predictive models for gene properties.



```
Gene Input
Gene: FAM209B
Summary:
Located in
nucleus. May
play a role in
sperm acrosome
biogenesis.
```

Conclusion: GeneLLM is able to embed summaries in the context of high level concepts aggregated from vast studies, which themselves often include computational models utilizing quantitative data.

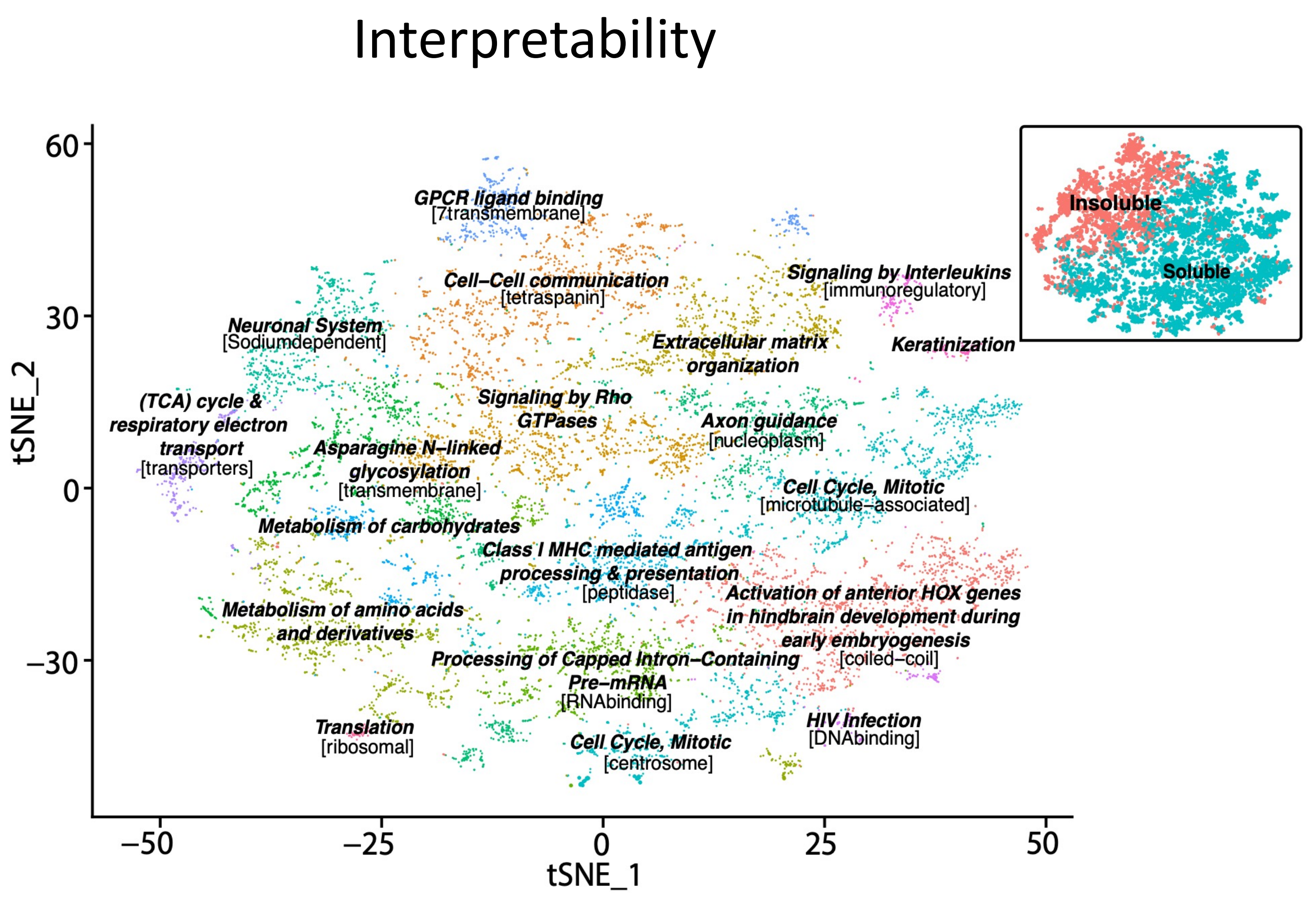
GeneLLM *is a foundation model that can be finetuned for downstream tasks and outperforms task-specific models.*

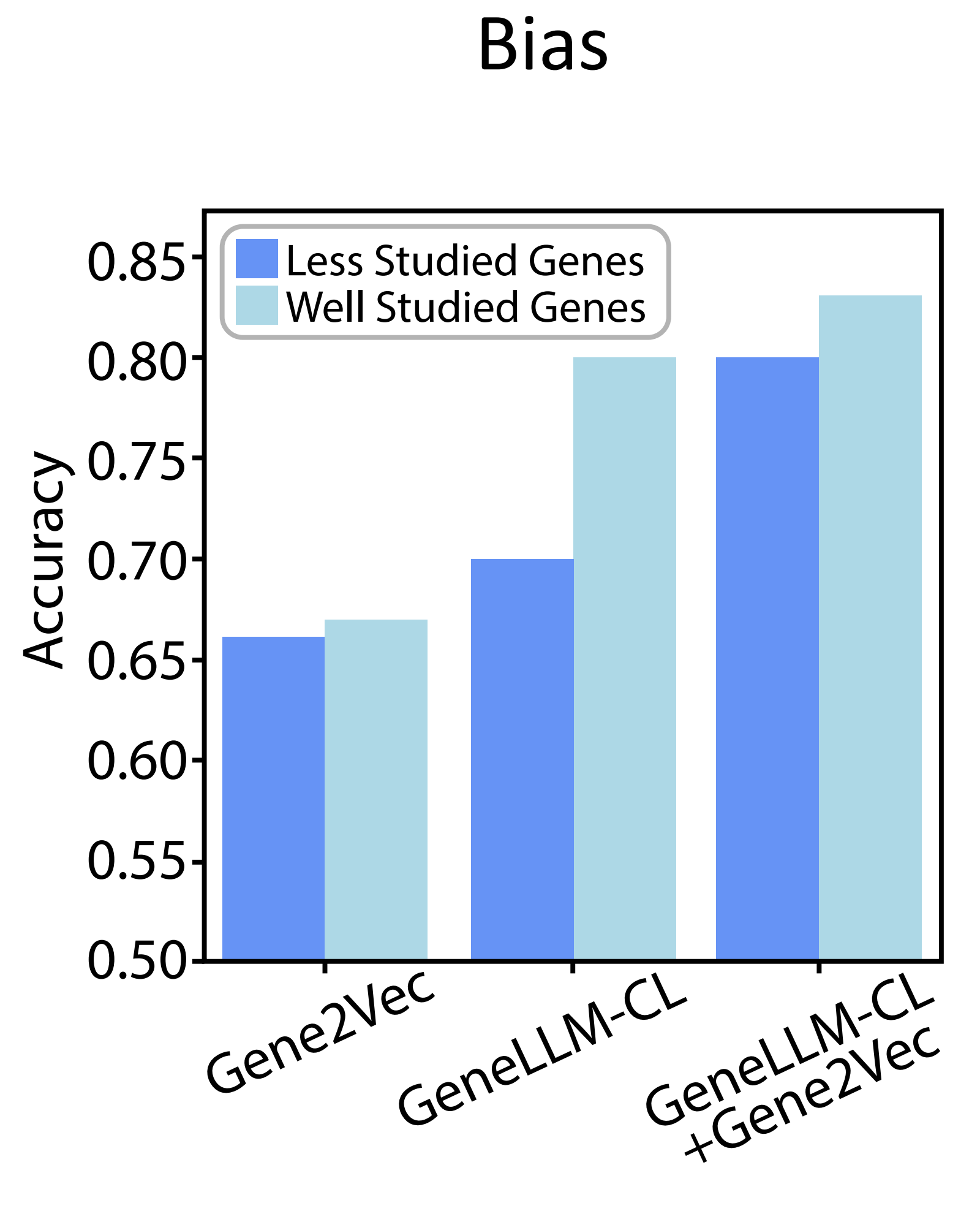
Model	Dosage Sensitivity	BivalentVs Lys4 Methylated	BivalentVs Non Methylated	Tf range	Tf target type	Solubility
Majority Classifier	0.73 ± —	0.58 ± —	0.75 ± —	0.73 ± —	0.41 ± —	0.52 ± —
GPT2	0.74 ± 0.04	**0.86 ± 0.04**	0.80 ± 0.11	0.71 ± 0.03	0.18 ± 0.02	0.80 ± 0.02
Doc2Vec	0.74 ± 0.04	0.84 ± 0.06	0.78 ± 0.05	0.66 ± 0.07	0.26 ± 0.01	0.71 ± 0.03
PMC-LLaMA	0.86 ± 0.05	0.77 ± 0.04	**0.84 ± 0.07**	0.64 ± 0.08	0.08 ± 0.01	0.78 ± 0.03
XLNet	0.74 ± 0.06	0.84 ± 0.06	0.83 ± 0.08	0.69 ± 0.05	0.12 ± 0.01	0.79 ± 0.02
Gene2Vec	0.84 ± 0.04	0.84 ± 0.06	0.75 ± 0.06	**0.75 ± 0.08**	0.21 ± 0.01	0.56 ± 0.02
BERT-Base	0.76 ± 0.09	0.83 ± 0.06	0.77 ± 0.10	0.68 ± 0.04	0.17 ± 0.01	0.77 ± 0.02
GeneLLM	**0.87 ± 0.06**	**0.86 ± 0.09**	0.82 ± 0.08	0.74 ± 0.07	**0.49 ± 0.04**	**0.89 ± 0.01**

Conclusion: These results suggest that textual data is particularly effective for tasks necessitating a comprehensive understanding of biological processes and molecular functions

Most AI models are black boxes, biased toward more well-studied genes.

GeneLLM uses SHAP to elucidate the model's logic (interpretability) and combines embeddings of both text and traditional transcriptomic data to reduce bias.

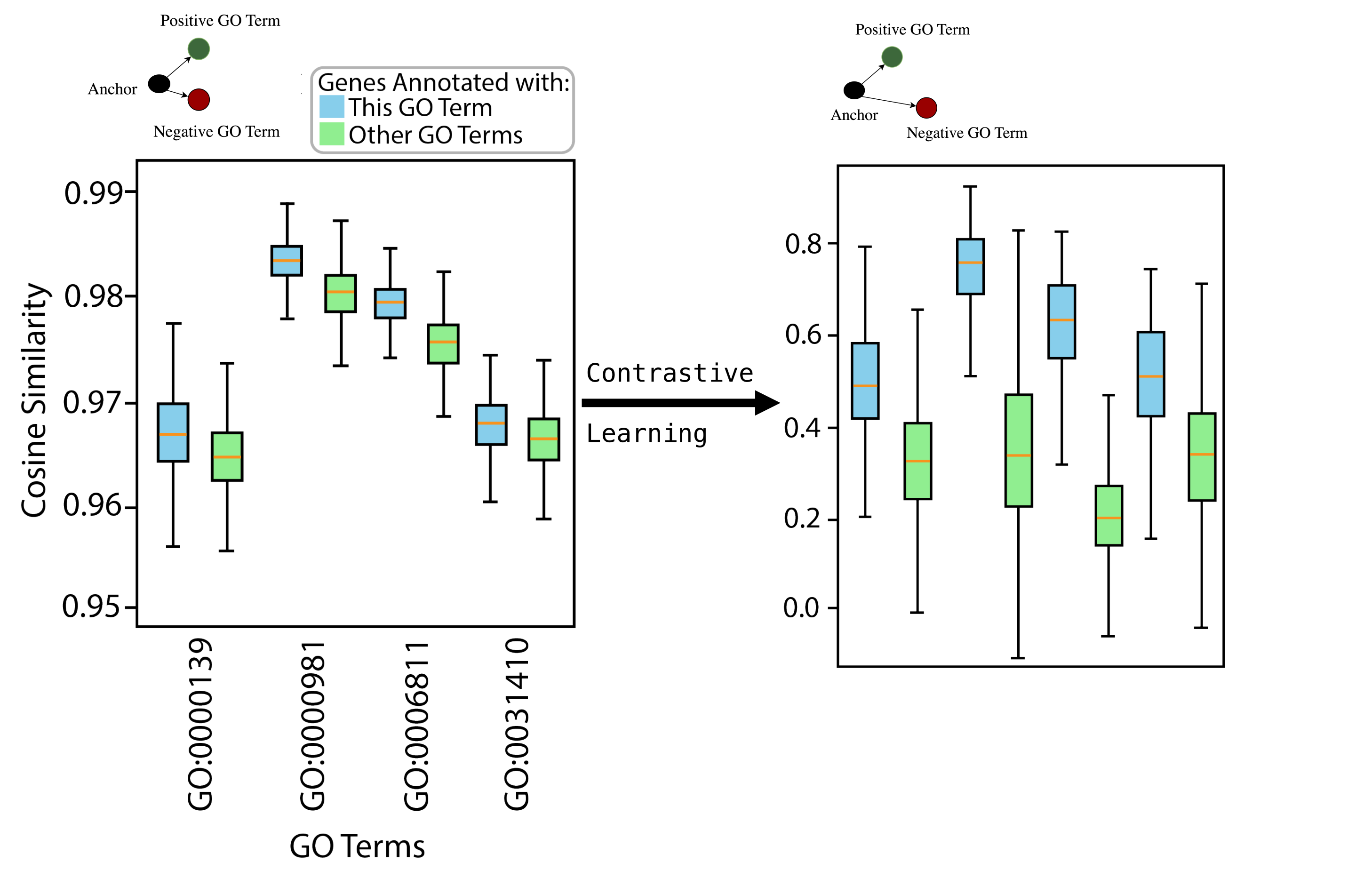


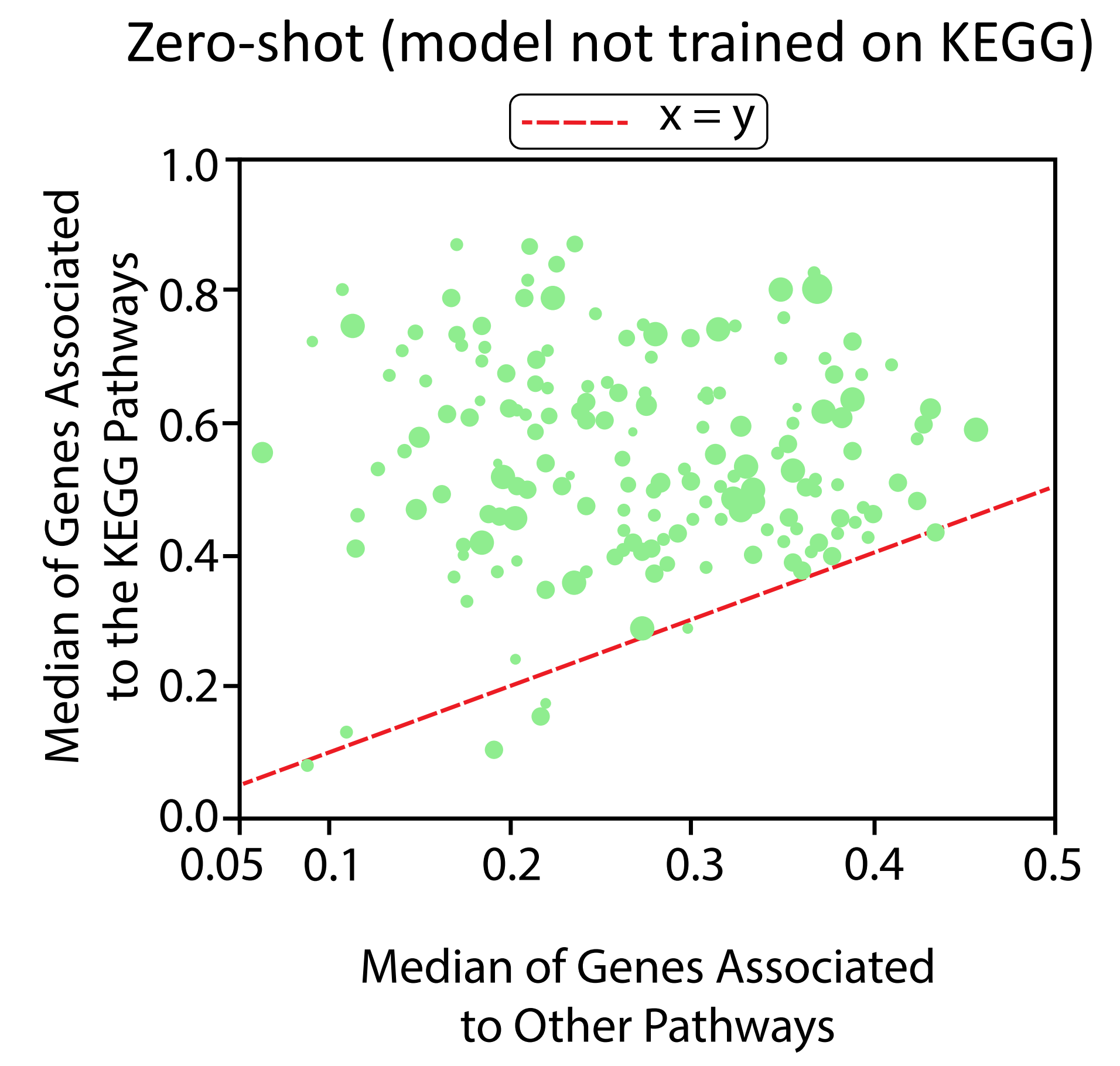


Conclusion: Interpretability enhances the overall value of AI technologies in biomedical applications and their potential for clinical adoption. This approach also provides evidence supporting the hypothesis that information contained in text is complementary to that found in structured databases.

Most predictive models depend on extensive labeled datasets.

GeneLLM advances predictions with unlabeled data, using contrastive learning to learn from gene text summaries in comparison to disease, pathway, and GO term summaries.





Conclusion: The use of contrastive learning enhances the ability of LLMs to predict a wide range of gene- and cell-related tasks and enables zero-shot predictions. By training the model on known GO terms (left), its ability to predict gene associations with unseen KEGG pathways was improved (right).