Supporting Information for Site-level bioactivity of small-molecules from deep-learned representations of quantum chemistry

Kathryn Sarullo[†], Matthew K. Matlock[†], and S. Joshua Swamidass*

Washington University in St. Louis, School of Medicine, Department of Pathology and Immunology, Saint Louis, MO, USA

E-mail: swamidass@wustl.edu

List of Figures

List of Tables

[†]Authors contributed equally to this paper.

Table S1: Quantum domains used to train transfer models. For each domain, a graph-based deep learning model was trained to predict a selected subset of the available quantum chemical properties. Graph-based encodings were extracted from each trained model for each molecule and used for transfer learning. TE: total energy, HOMO: highest occupied molecular orbital, LUMO: lowest unoccupied molecular orbital, Gap: HOMO-LUMO gap energy, Tot: total valence, Bond: bonded valence, Chrg: Mulliken charge, Pop: Mulliken population, Order: Mayer bond order, Length: bond length

	Molecule			Atom				Bond	
		Orbital Energies		Valence		Mulliken			
D	TE	HOMO	Gap	Tot	Bond	Chrg	Pop	Order	Length
1								y	y
2				y	y	y	y		
3				y	y	y	y	y	y
4		y	y						
5		y	y					y	y
6		y	y	y	y	y	y		
7		y	y	y	y	y	y	y	y
8	y								
9	y							y	y
10	y			y	y	y	y		
11	y			y	y	y	y	y	y
12	y	y	y						
13	y	y	y					y	y
14	y	y	y	y	y	y	y		
QC	y	y	y	y	y	y	y	y	y

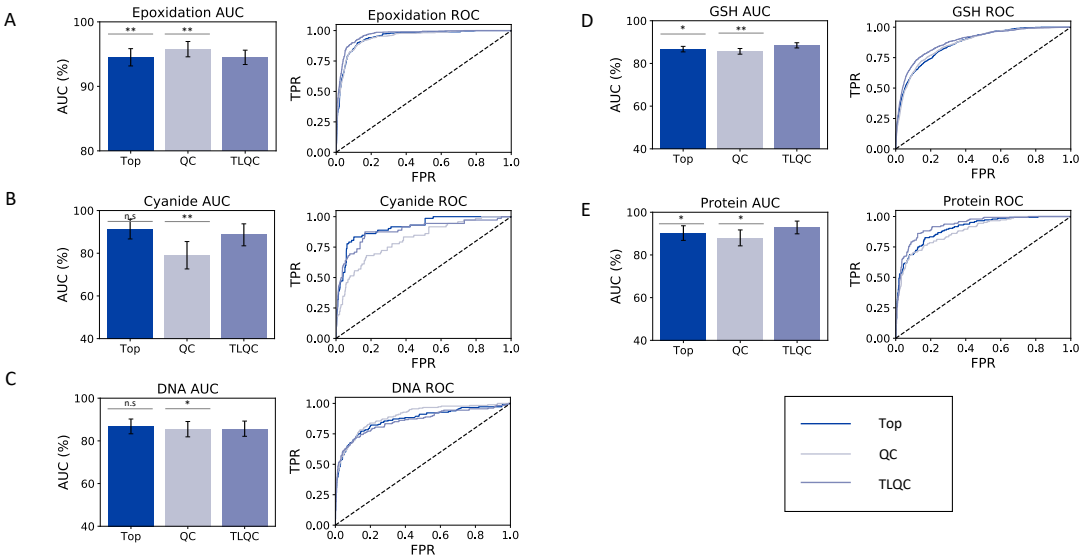

Figure S1: AUCs We computed task specific (Epoxidation, Cyanide, DNA, GSH, and Protein) AUC values and ROC curves for each model (Top, QC, and TLQC). *p-value $<$ 0.05, **p-value $<$ 0.001

Table S2: Top-2 accuracy performance of alternate topological representations We varied the architectures for the Top model. We chose the highest top-2 accuracy on average as a comparison for the TLQC model. The highest top-2 accuracy for the Top model was Top N3.

Architecure	Epoxidation	Cyanide	DNA	GSH	Protein	Average
N2	78.7%	78.4%	76.3%	58.0%	70.7%	66.1%
N3	**81.1%**	**62.7%**	**77.7%**	**65.4%**	**79.6%**	**71.1%**
N4	75.7%	68.6%	74.3%	62.1%	79.6%	67.8%
N5	77.3%	64.7%	76.3%	63.5%	75.2%	68.7%

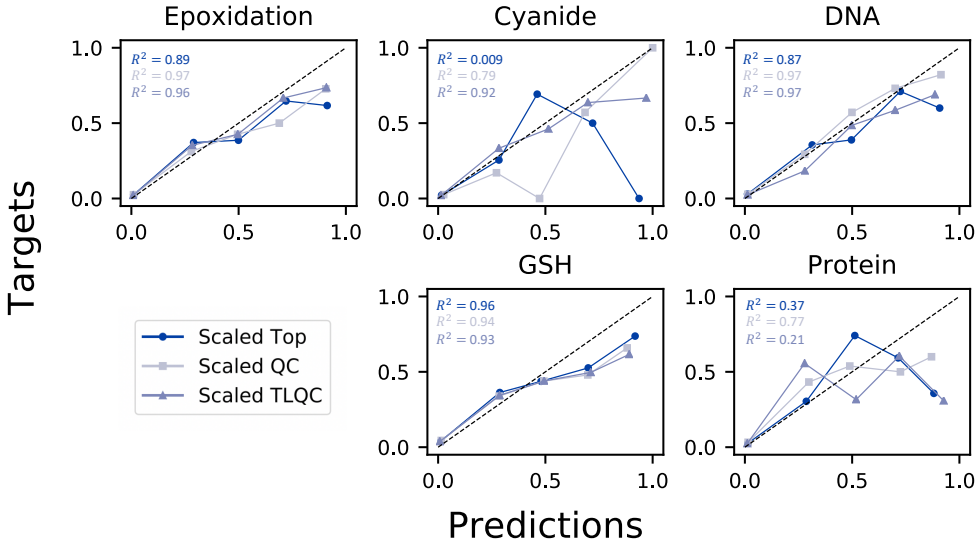

Figure S2: Reliability We performed platt scaling on the each of the task specific (Epoxidation, Cyanide, DNA, GSH, and Protein) predictions for each model (Top, QC, and TLQC) to get the scaled predictions. The reliabilty plots tell us if each tasks predictions are trustworthy. A perfect predictor would be aligned with the linear, black dashed line on each plot.

Table S3: Top-2 accuracy performance of alternate QC representations We varied the architectures for the QC model. We chose the highest top-2 accuracy on average as a comparison for the TLQC model. The highest top-2 accuracy for the QC model was QC N5.

Architecure	Epoxidation	Cyanide	DNA	GSH	Protein	Average
N2	78.9%	45.0%	68.0%	56.9%	76.9%	64.1%
N3	81.9%	56.8%	70.1%	55.8%	73.4%	64.6%
N4	81.4%	49.0%	69.4%	57.8%	73.4%	65.3%
N5	**79.2%**	**56.8%**	**74.3%**	**58.7%**	**70.7%**	**65.7%**

Table S4: Accuracy vs state size We trained several models to predict QC properties with decreasing state sizes; 32,48,64,96, and 128. Here we show the absolute error between the state size and each of the 9 QC properties we predicted. To minimize the number of parameters of TLQC models, used 64 as the state size of the quantum representation.

	Molecule			Atom				Bond	
		Orbital Energies		Valence		Mulliken			
State Sizes	TE	HOMO	Gap	Tot	Bond	Chrg	Pop	Order	Length
32	0.011	0.015	0.017	0.011	0.015	0.013	0.012	0.012	0.004
48	0.009	0.011	0.013	0.010	0.012	0.010	0.010	0.010	0.004
64	**0.007**	**0.010**	**0.011**	**0.010**	**0.012**	**0.010**	**0.010**	**0.010**	**0.003**
96	0.006	0.009	0.010	0.009	0.010	0.009	0.009	0.009	0.003
128	0.006	0.008	0.010	0.009	0.010	0.009	0.009	0.009	0.003