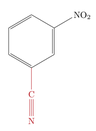
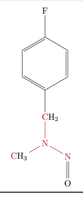
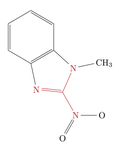
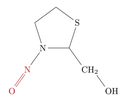
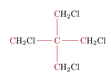
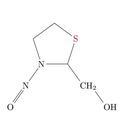
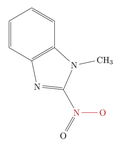
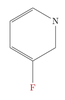
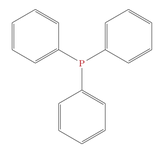
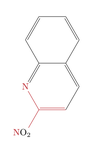
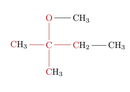
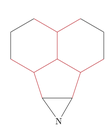
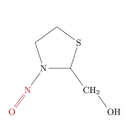
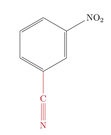
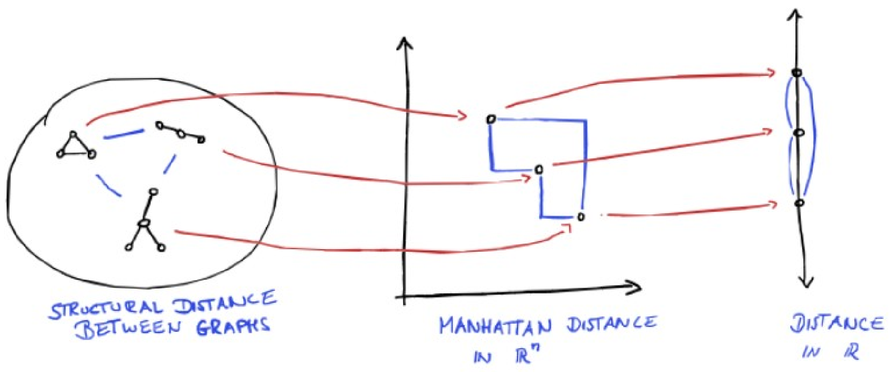
Graph Neural Networks focus on relevant subgraphs that chemist know for a long time

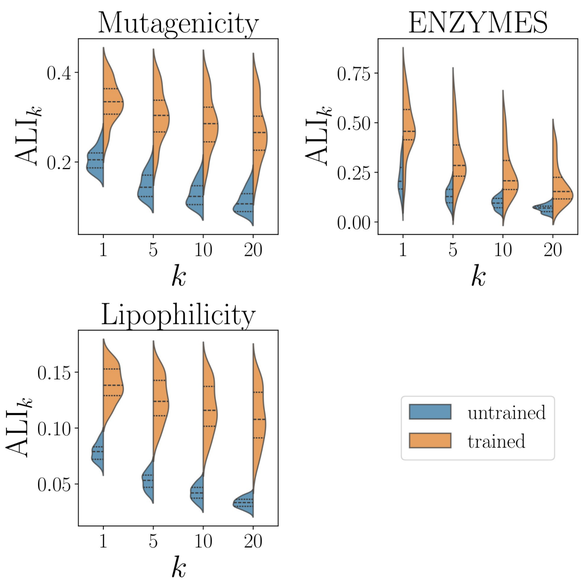
1) three-membered heterocycle (epoxide)	(2)	(3)	(4) alphatic halide	(5)
				
(6) nitroso	(7)	(8)	(9)	(10) alphatic halide
				

(1)	(2) three-membered heterocycle (epoxide)	(3) alphatic halide	(4)	(5)
				
(6)	(7)	(8) nitroso	(9)	(10)
				

- Mutagenicity dataset contains 4337 molecular graphs
  - 2401 mutagens and 1936 non-mutagens
- Graph Convolutional Neural Networks with
  - sum pooling (left)
  - mean pooling (right)
- Highlighted in red are substructures that influence the metric space on hidden embeddings the most
- Four (resp three) of the ten most influential structures are among seven toxicophores suspected to increase mutagenic effect of molecules (Kazius et al., 2005)

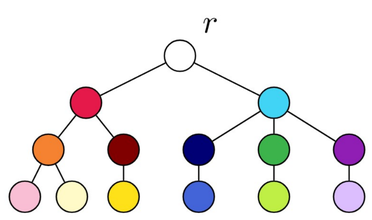
Graph Neural Networks align their hidden embeddings with the metric in the target space

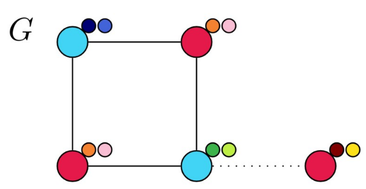




- Training
  - aligns the distances between hidden embeddings with distances between targets
  - disaligns structural metrics between graphs and distances between hidden embeddings

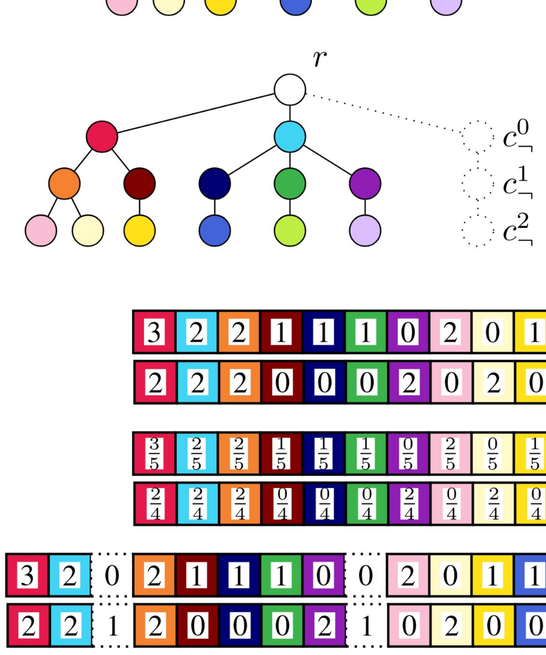
We present a tunable, interpretable metric that aligns with Graph Neural Network embeddings





	d_{WWL}	d_{WLOA}	$\dot{d}_{\mathrm{WILT}}$	$\bar{d}_{\mathrm{WILT}}$
Mutagenicity				
mean	9.25±0.87	18.74±3.36	1.74±0.52	3.34±1.01
sum	12.25±0.54	5.98±1.60	1.22±0.31	0.82±0.17
ENZYMES				
mean	12.18±0.23	16.79±2.33	2.71±0.38	4.64±0.67
sum	11.28±0.65	6.83±0.41	9.15±0.47	1.43±0.10
Lipophilicity				
mean	10.92±0.42	13.97±0.97	3.11±0.54	6.35±1.22
sum	10.83±0.73	10.00±1.34	2.50±0.67	2.64±0.74

RMSE between
- Structural graph metrics WWL, WLOA
- WILTing metric variants



$$d_{\mathrm{WILT}}(G, H; w) \coloneqq \min_{P \in \Gamma} \sum_{v_i \in V_G} \sum_{u_j \in V_H} P_{i,j} d_{\mathrm{path}}(c_{v_i}^{(L)}, c_{u_j}^{(L)})$$

where $\Gamma \coloneqq \{P \in \mathbb{R}^{|V_G| \times |V_H|} \mid P_{i,j} \geq 0, P\mathbf{1} = \mathbf{1}, P^T\mathbf{1} = \mathbf{1}\}$.

$$d_{WILT}(G, H; w) = \sum_{c \in V(T_{\mathcal{D}}) \setminus \{r\}} w\left(e_{\{c,p(c)\}}\right) \left|\nu_c^G - \nu_c^H\right|,$$

where $e_{\{c,p(c)\}}$ is the edge connecting c and its parent $p(c)$ in $T_{\mathcal{D}}$.

Masahiro Negishi, Thomas Gärtner, Pascal Welke

WILTing Trees: Interpreting the Distance Between MPNN Embeddings
International Conference on Machine Learning (2025), to appear