# Mechanistic Interpretability of Chemical Language Models: Molecular Geometry and Representations in MoLMFormer and SMI-TED

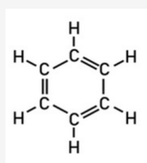
Wanqing Hao Zetong Xu Zhiyuan Han Xiaoyan Bai Chenhao Tan
University of Chicago

## Background:

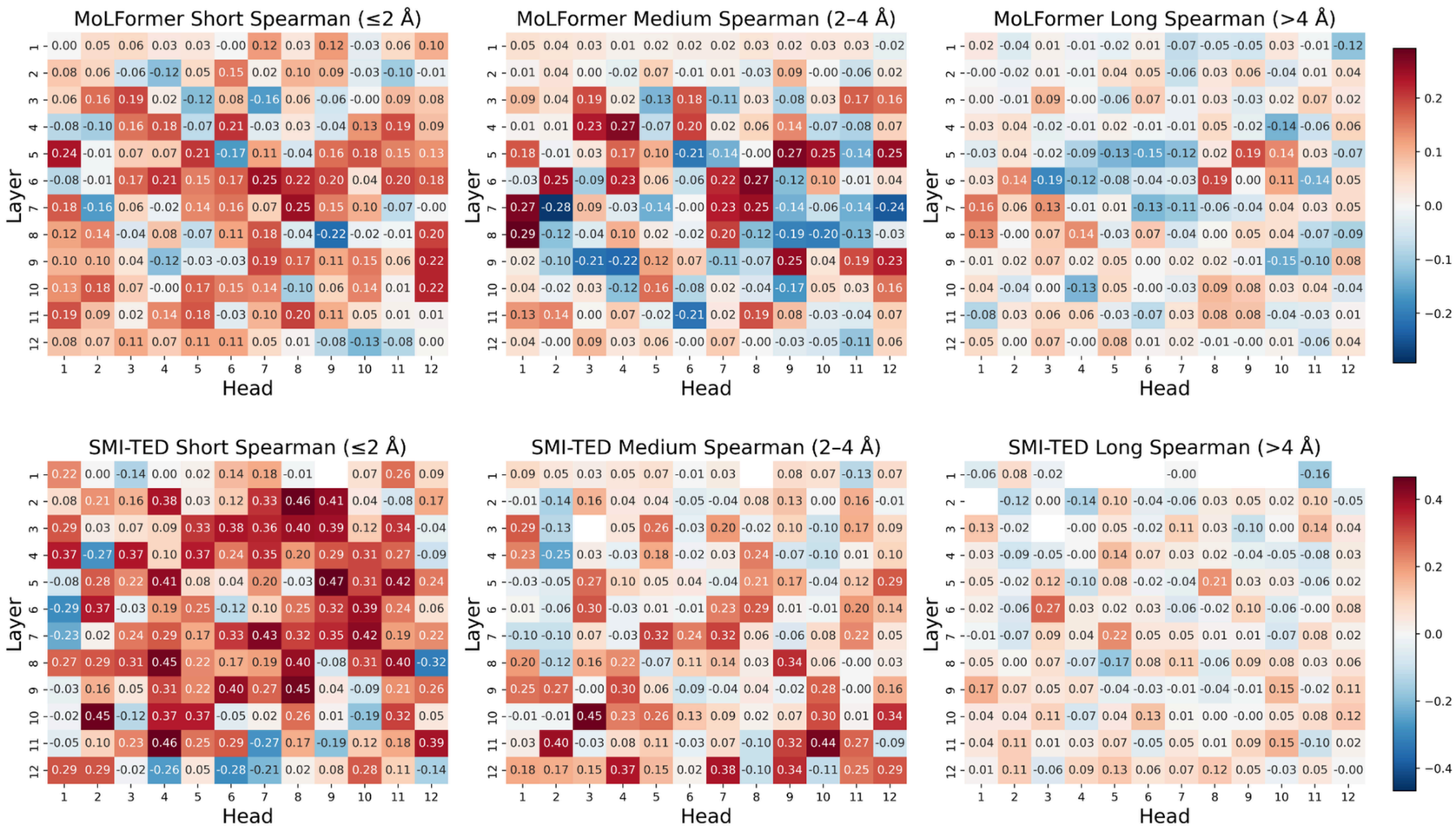
**MoLMFormer** (encoder-only): Masked language modeling
**SMI-TED** (encoder-decoder): SMILES string reconstruction

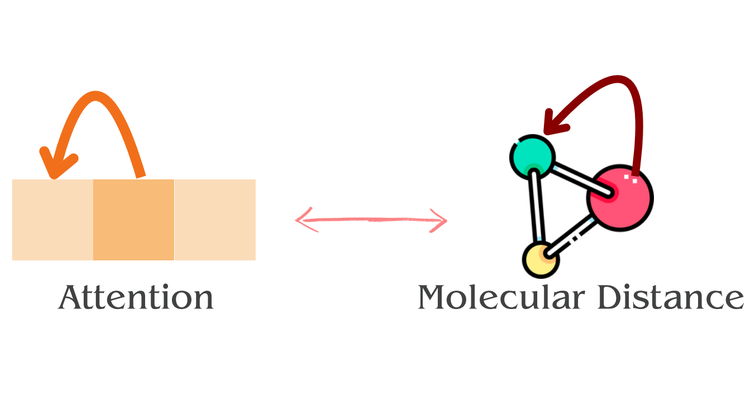
} No explicit geometry supervision during pretrain. Surface-level features are readable from tokens. Contextual features requires geometry context.

SMILES INPUT
**c1ccccc1**
*Benzene*



## Attention weakly reflects molecular geometry.




## Attention-distance correlation:

✓ Short- and Medium - range atomic interaction

✗ Long-distance interaction

Figure 1. Attention-distance correlation matrices per layer, head pair for MOLFORMER and SMI-TED

## Both models recover surface-level properties, but **SMI-TED** encodes contextual properties better.

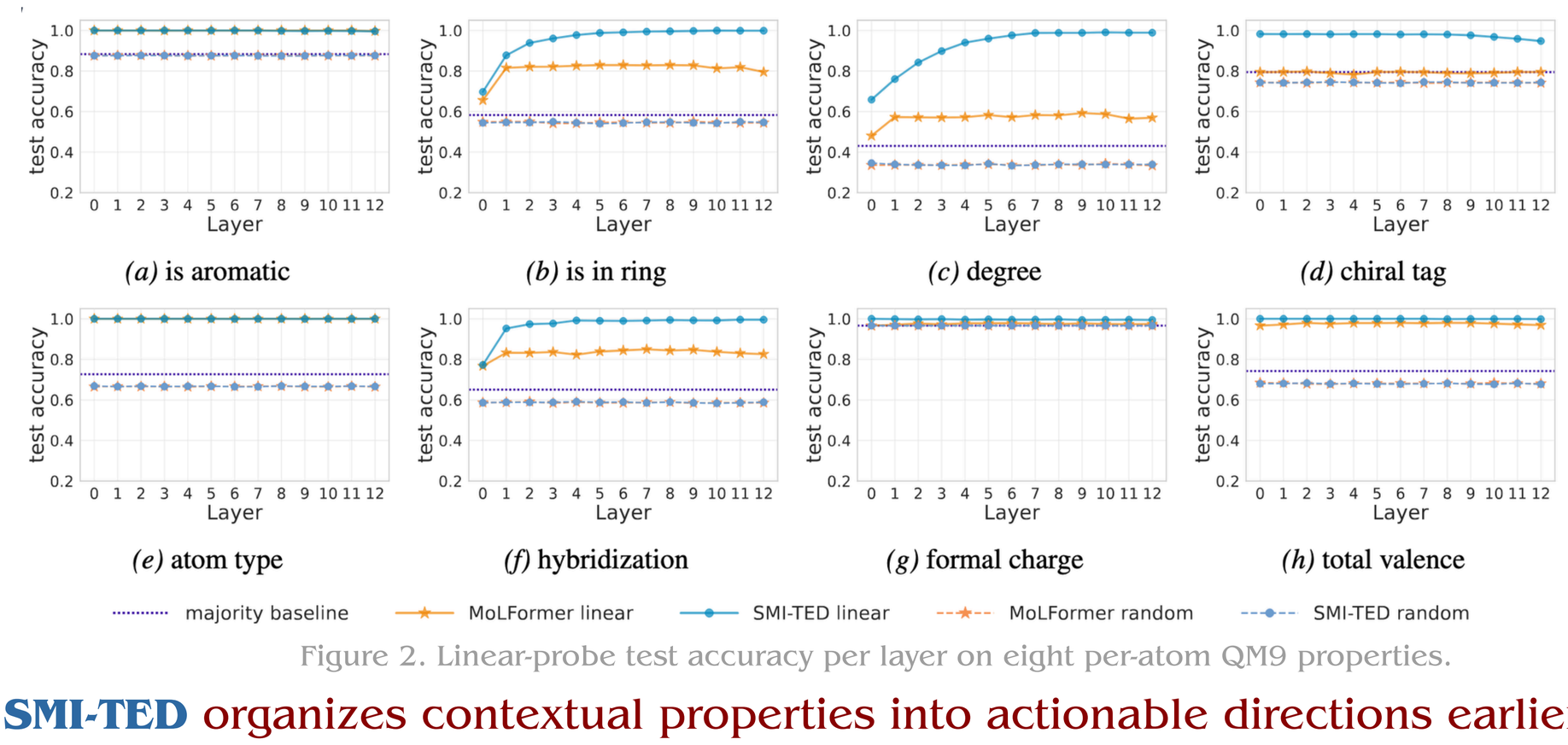



Figure 2. Linear-probe test accuracy per layer on eight per-atom QM9 properties.

## **SMI-TED** organizes contextual properties into actionable directions earlier.




Figure 3. Intervention success rate per steer layer on eight per-atom QM9 properties

## Limitation & Discussion

### What is encoded in science model?

Downstream performance ≠ human-like or correct scientific representations.

### Better interpretability tools for science model

Probing for scientific model can be easily confounded by tokenizer leakage, class imbalance, dataset bias.

Chemistry models need tailored interpretability tools.

Paper

Github