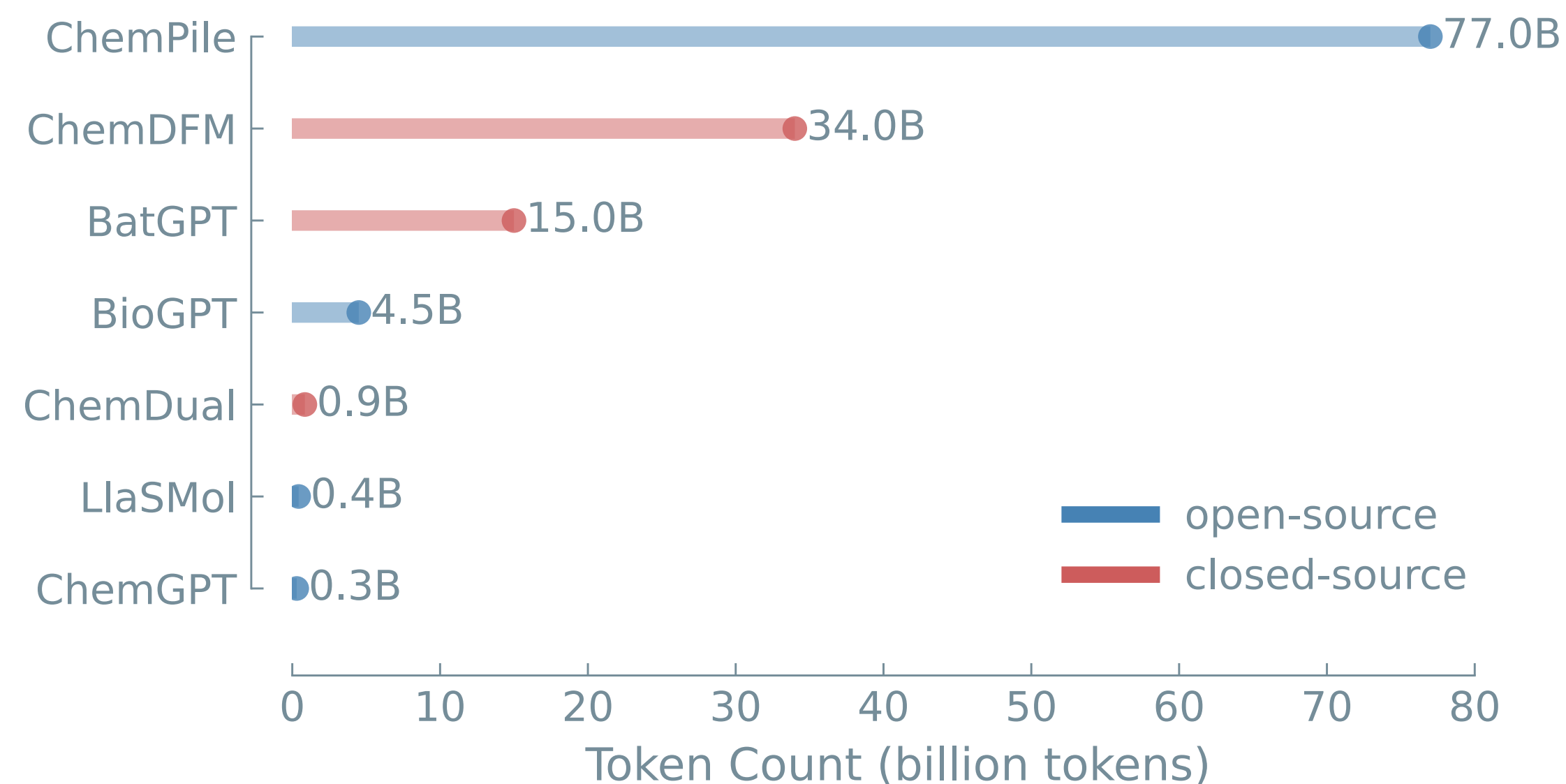


ChemPile — A 250GB Dataset for Chemical Foundation Models

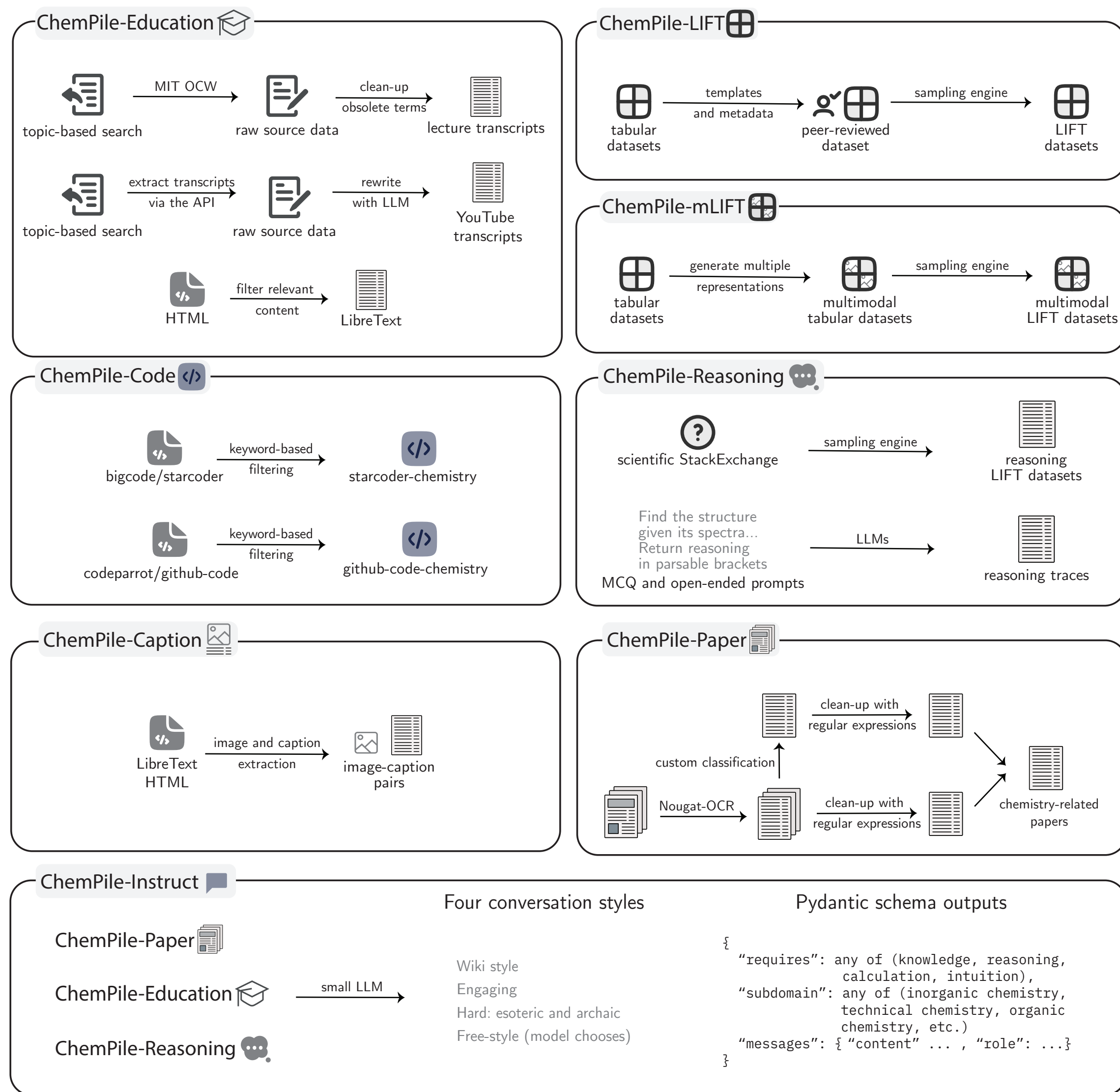
NeurIPS 2025 D&B Track | Adrian Mirza | 18.10.2025

What Were the Reasons for Creating ChemPile?

- Lack of other pretraining scale datasets for the chemical foundation models
- Lack of multimodal data
- Lack of modularity



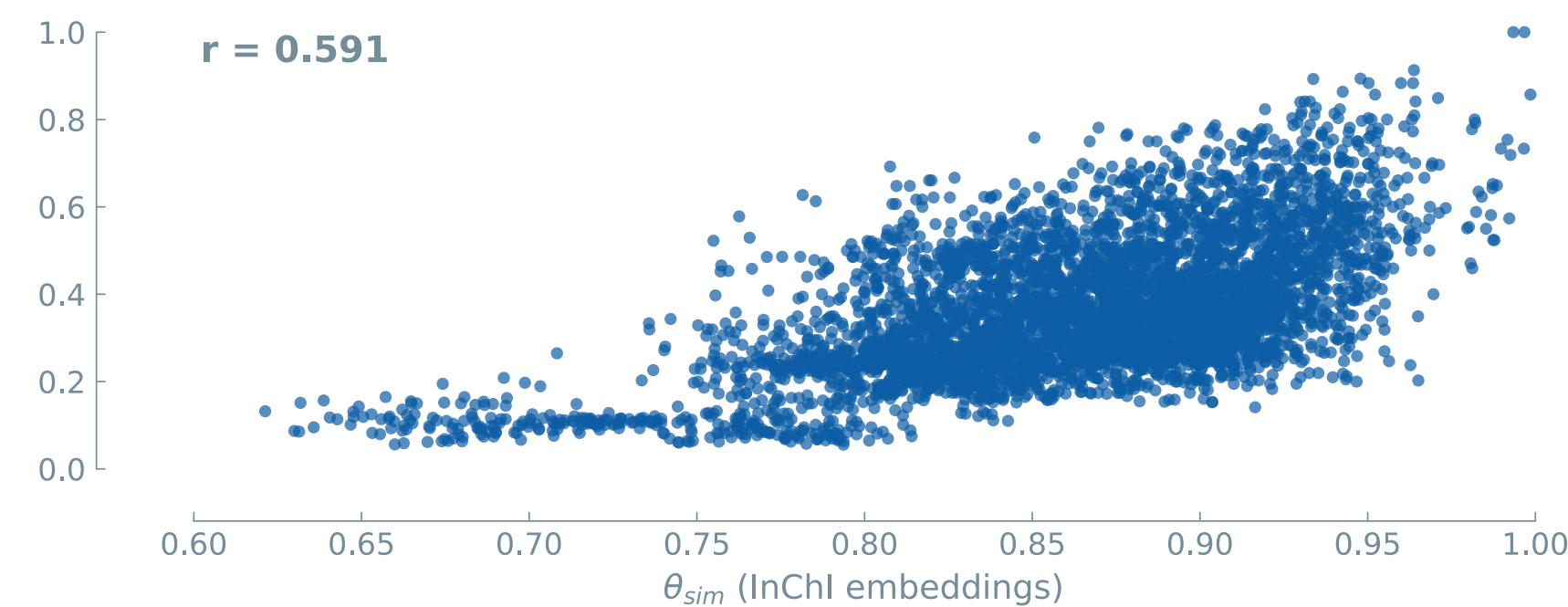
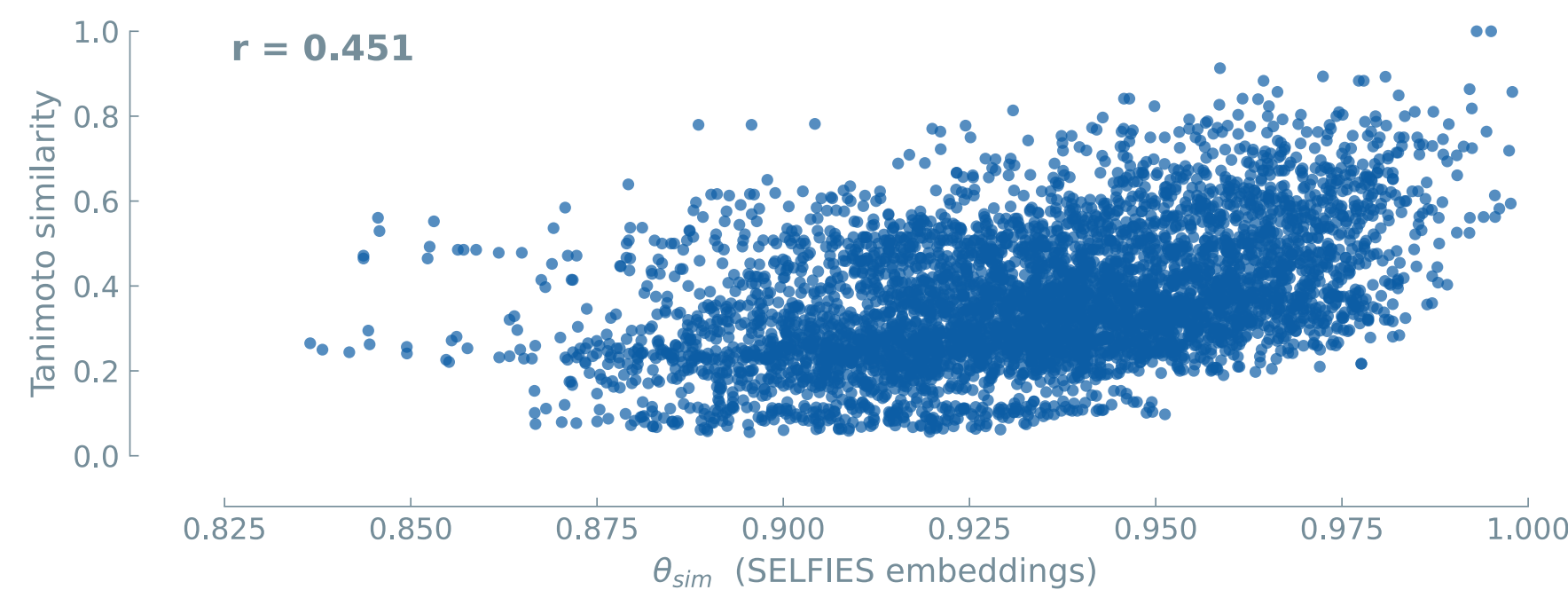
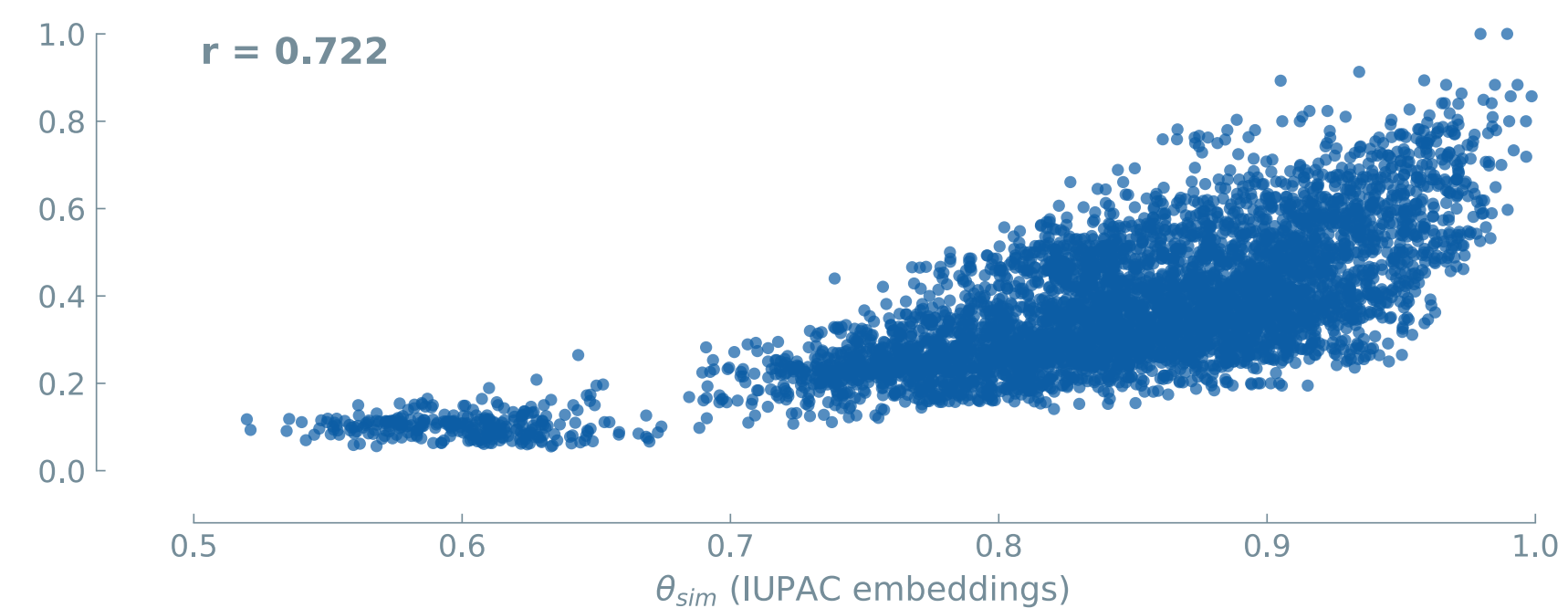
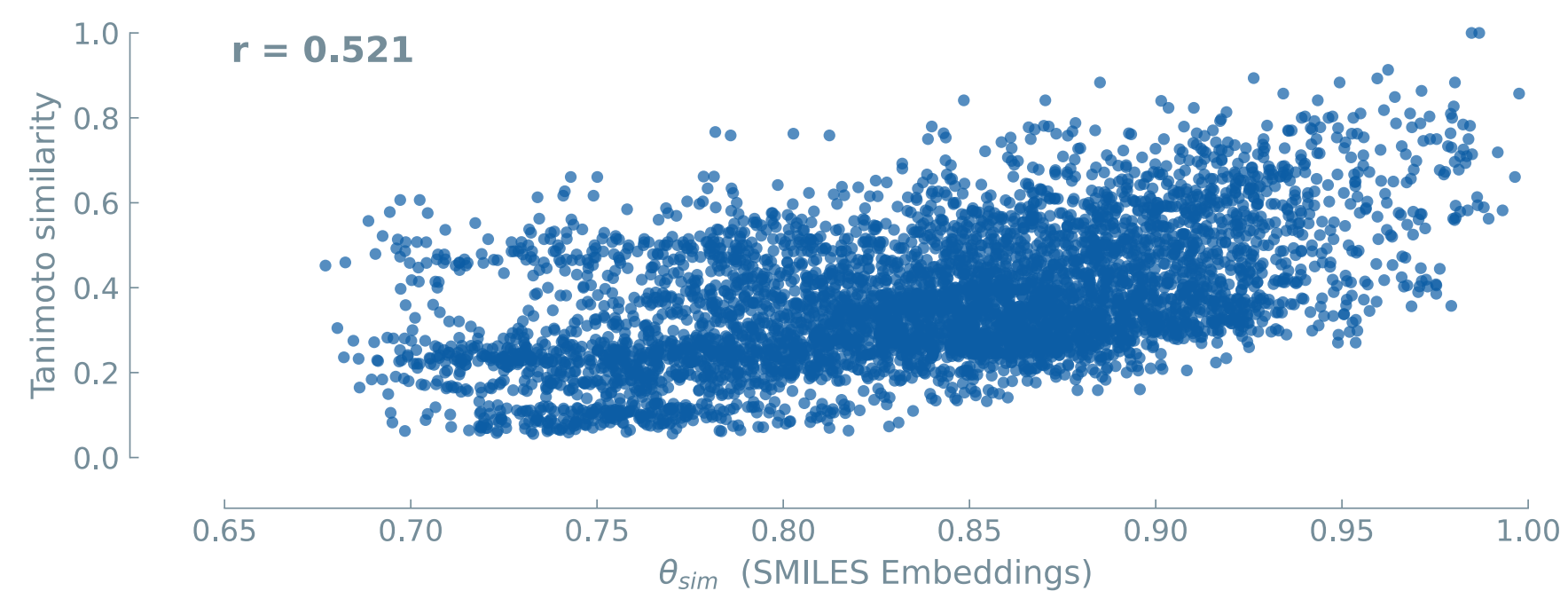
The Modular Structure of Our Dataset



Dataset	Size (GB)	Number of text tokens	Number of documents
ChemPile-Education	0,25	130M	66,9K
ChemPile-Paper	31,6	14,1B	11,7M
ChemPile-LIFT	49.1	29,4B	185M
ChemPile-mLIFT	155	15,0B	61.6M
ChemPile-Code	15,6	18,0B	2,27M
ChemPile-Reasoning	0,10	20,0M	72,9K
ChemPile-Caption	3,23	10,3M	100K
ChemPile-Instruction	2,85	396M	410K
ChemPile	257	77,0B	260M


Multi-representation Tabular Datasets

- SMILES
- Images
- SELFIES
- IUPAC names
- InChI



Curation Process

- Review app — manual review of over 600 random samples from ChemPile to ensure good quality
- GitHub peer-review process for tabular data
- Verifiable generations
 - IUPAC name validation with OPSIN
 - SMILES validation for spectra elucidation traces

 **Data Annotation Tool**

A new set of random questions from multiple sources is loaded for you automatically.

Your Name

Enter your name (required for saving annotations)

Guidelines:

Check if there is something strange in the formatting.
Check if there seems to be a factual mistake.
Check if the style of the entry seems unnatural.
Describe the problem in the comment box (bottom right).

Thank you for your contribution towards making ChemPile better!

Progress

Entry 1 of 99

Text Entry to Validate

Sound pressure levels generated at risk volume steps of portable listening devices: types of smartphone and genres of music. Sound pressure levels generated at risk volume steps of portable listening devices: types of smartphone and genres of music. <h4>Background</h4>The present study estimated the sound pressure levels of various music genres at the volume steps that contemporary smartphones deliver, because these levels put the listener at potential risk for hearing loss.<h4>Methods</h4>Using six different smartphones (Galaxy S6, Galaxy Note 3, iPhone 5S, iPhone 6, LG G2, and LG G3), the sound pressure levels for three genres of K-pop music (dance-pop, hip-hop, and pop-ballad) and a Billboard pop chart of assorted genres were measured through an earbud for the first risk volume that was at the risk sign proposed by the smartphones, as well as consecutive higher volumes using a sound level meter and artificial mastoid.<h4>Results</h4>The first risk volume step of the Galaxy S6 and the LG G2, among the six smartphones, had the significantly lowest (84.1 dBA) and highest output levels (92.4 dBA), respectively. As the volume step increased, so did the sound pressure levels. The iPhone 6 was loudest (113.1 dBA) at the maximum volume step. Of the music genres, dance-pop showed the highest output level (91.1 dBA) for all smartphones. Within the frequency range of 20~ 20,000 Hz, the sound pressure level peaked at 2000 Hz for all the smartphones.<h4>Conclusions</h4>The results showed that the sound pressure levels of either the first volume step or the maximum volume step were not the same for the different smartphone models and genres of music, which means that the risk volume sign and its output levels should be unified across the devices for their users. In addition, the risk volume steps proposed by the latest smartphone models are high enough to cause noise-induced hearing loss if their users habitually listen to music at those levels.</endofcontext>

Your Validation

is this entry fine?

☐ Correct

☐ Wrong

Comments (optional)

Add context...

Previous (No Save)

Save & Next

Using ChemPile to Improve Models on Chemistry Tasks

- Data quality was assessed by training different combinations of LoRA adaptors
 - LoRA—Merge
 - LoRA—Ensemble

