CSCE 689 - Special Topics in NLP for Science

Lecture 14: Molecule Language Models

Yu Zhang

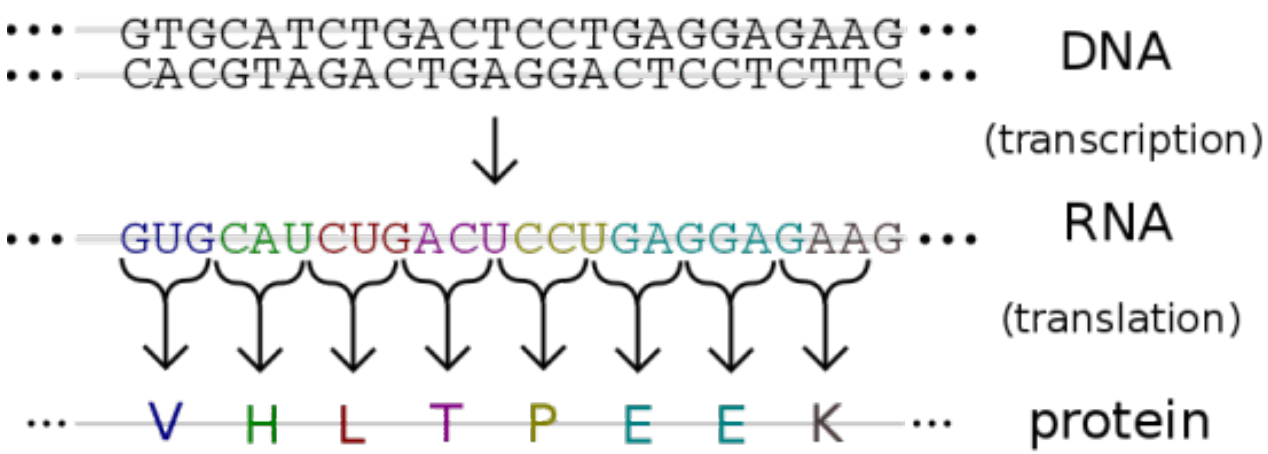
yuzhang@tamu.edu

March 4, 2025

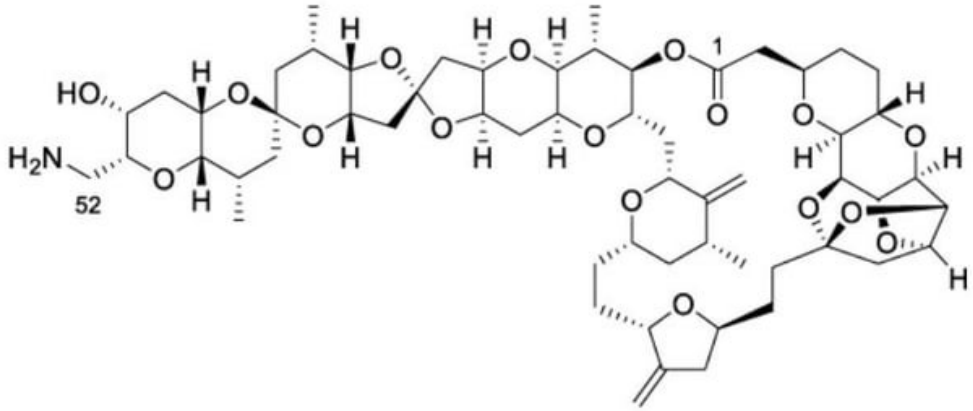
Course Website: https://yuzhang-teaching.github.io/CSCE689-S25.html

Differences between Molecules and Protein/DNA/RNA Sequences

- Protein, DNA, RNA: naturally sequential



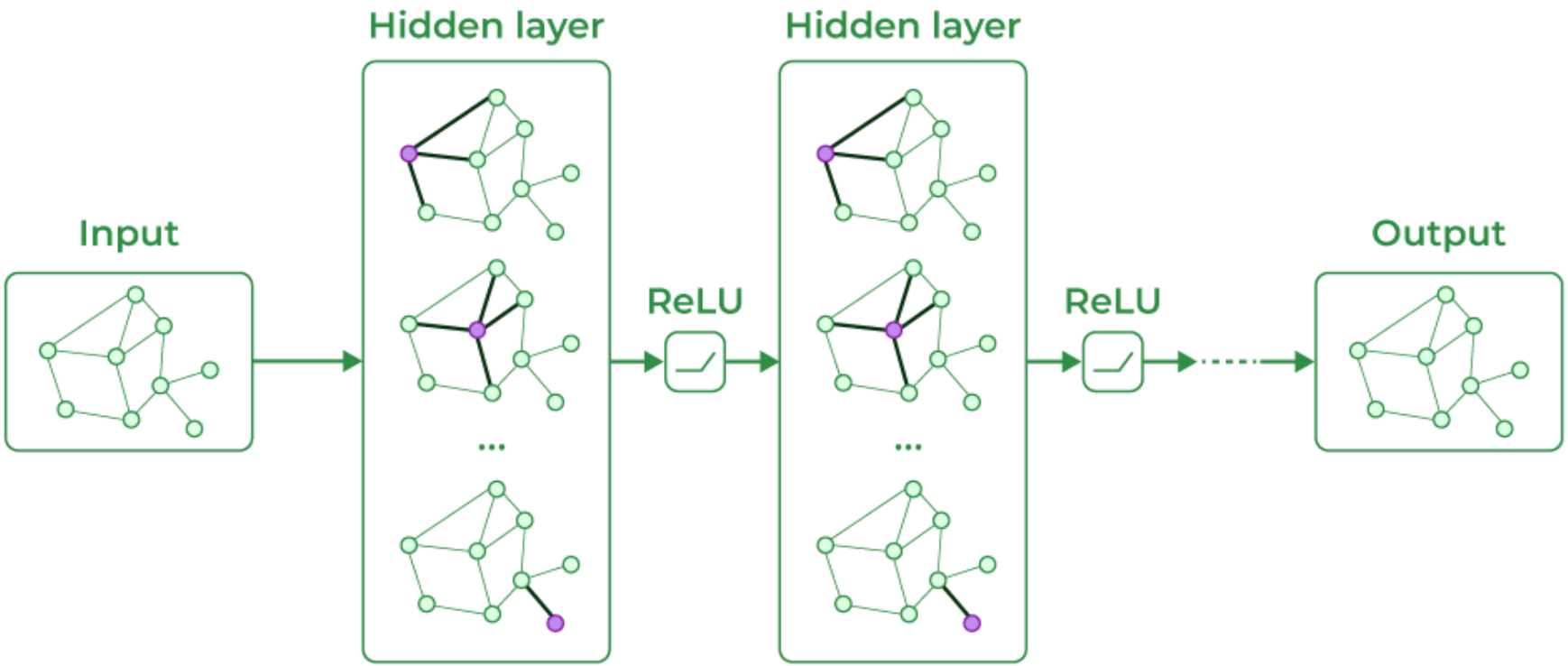
- Molecule: not naturally sequential
  - Strategy 1: using a graph encoder
  - Strategy 2: using a "sequential" language to describe molecules



C52-halichondrin-B amine (E7130)

Strategy 1: Using a Graph Encoder

- Graph Neural Networks (GNNs)
  - Each node in the graph has a representation vector at each layer.
  - The vector is obtained by aggregating information from its neighbors.



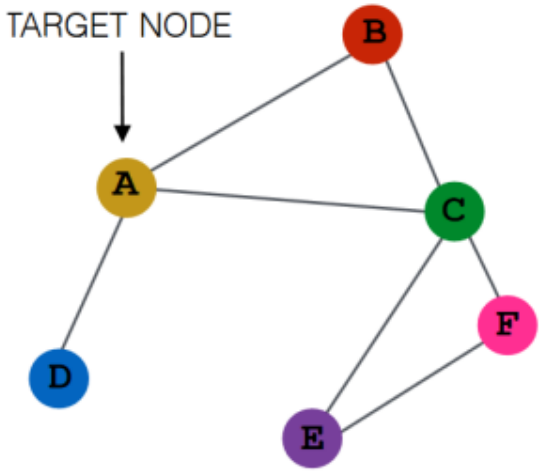
Strategy 1: Using a Graph Encoder

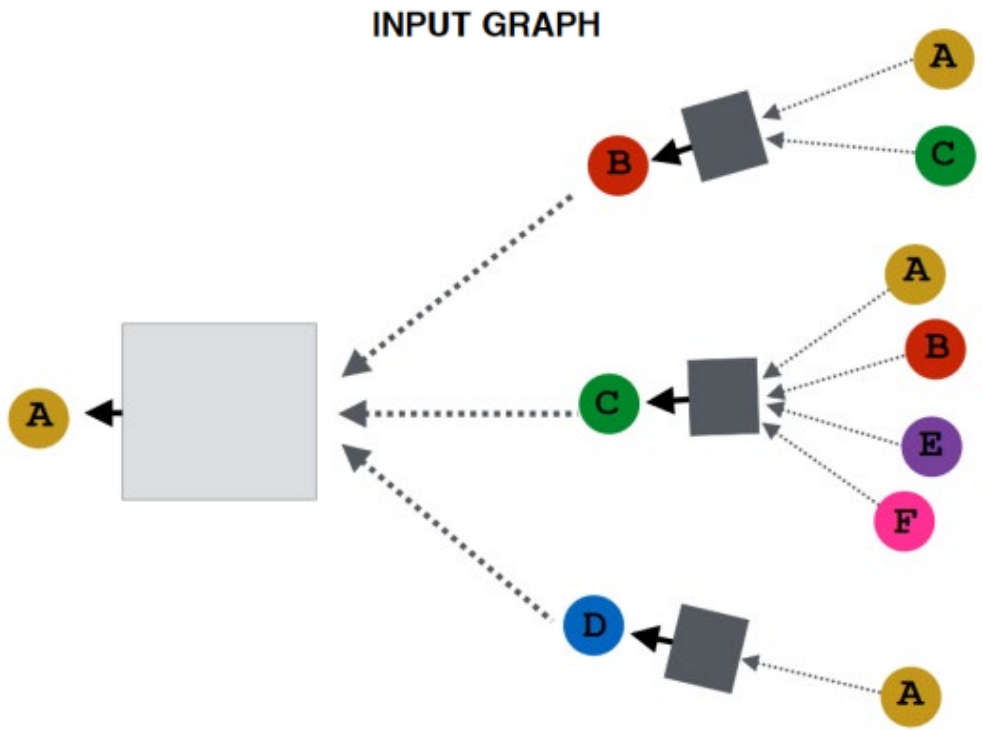
- Graph Neural Networks (GNNs)
  - Each node in the graph has a representation vector at each layer.
  - The vector is obtained by aggregating information from its neighbors.
  - For node v at layer t,

$$h_v^{(t)} = f\left(\underline{h_v^{(t-1)}}, \left\{\underline{h_u^{(t-1)} | u \in \mathcal{N}(v)}\right\}\right)$$

representation vector from previous layer for node v

representation vectors from previous layer for node v's neighbors





Strategy 1: Using a Graph Encoder

- Example 1: Graph Convolutional Networks (GCN) [1]

$$\mathbf{h}_v^k = \sigma\left(\mathbf{W}_k \sum_{u \in N(v) \cup v} \frac{\mathbf{h}_u^{k-1}}{\sqrt{|N(u)||N(v)|}}\right)$$

$\mathbf{W}_k$: weight matrix at layer k, shared across different nodes

- Example 2: Graph Sample and Aggregate (GraphSAGE) [2]

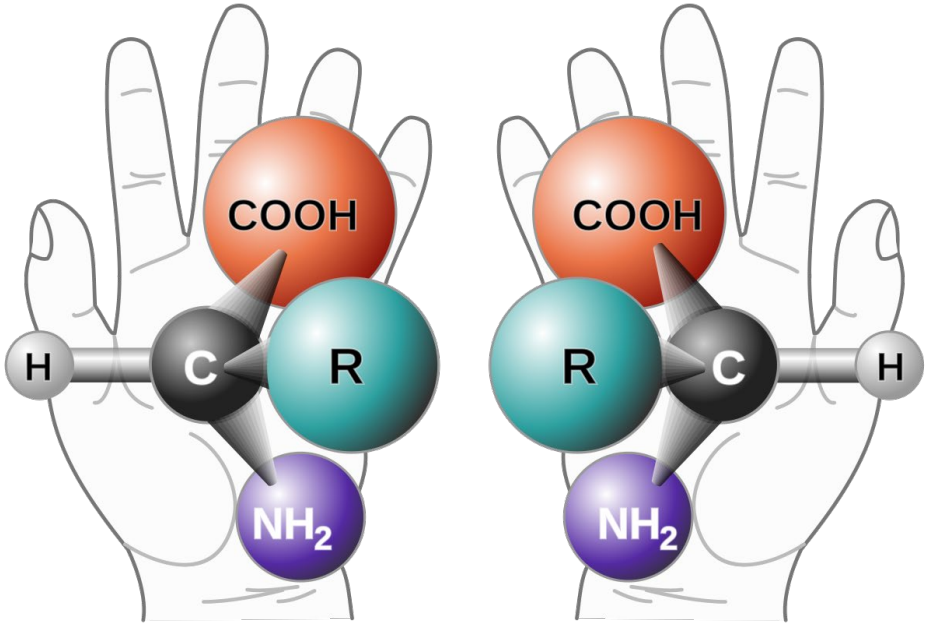
$$\mathbf{h}_{\mathcal{N}(v)}^k \leftarrow \text{AGGREGATE}_k(\{\mathbf{h}_u^{k-1}, \forall u \in \mathcal{N}(v)\})$$
$$\mathbf{h}_v^k \leftarrow \sigma\left(\mathbf{W}^k \cdot \text{CONCAT}(\mathbf{h}_v^{k-1}, \mathbf{h}_{\mathcal{N}(v)}^k)\right)$$

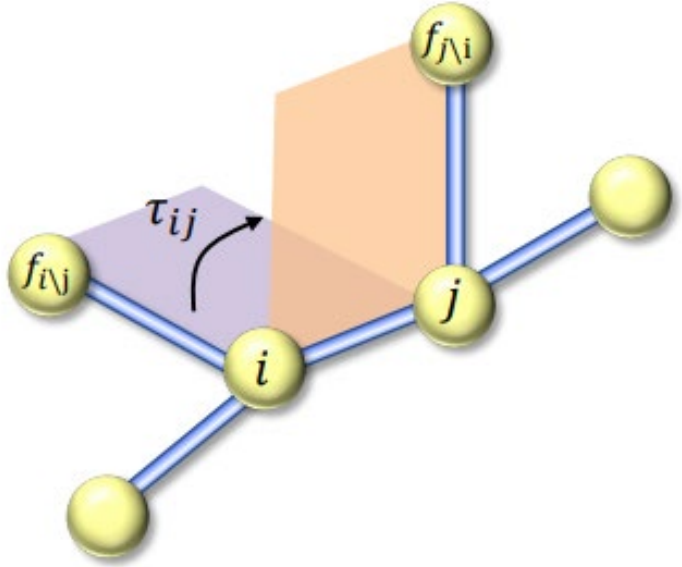
AGGREGATE_k: average, element-wise mean/max pooling, …

[1] *Semi-Supervised Classification with Graph Convolutional Networks*. ICLR 2017.
[2] *Inductive Representation Learning on Large Graphs*. NIPS 2017.

Strategy 1: Using a Graph Encoder

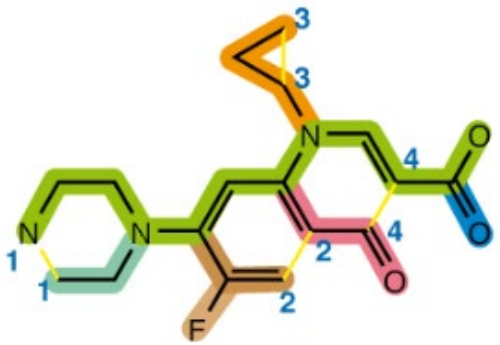
- Limitation: There are cases where graphs are not sufficient to describe a molecule.
  - Chirality
  - *ComENet: Towards Complete and Efficient Message Passing for 3D Molecular Graphs.* NeurIPS 2022.



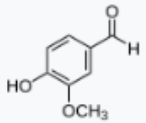
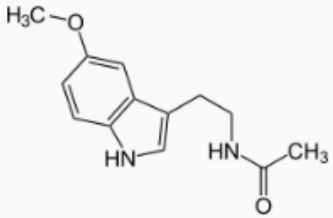


Strategy 2: Using a Sequential Language to Describe Molecules

- Simplified Molecular Input Line Entry System (SMILES)



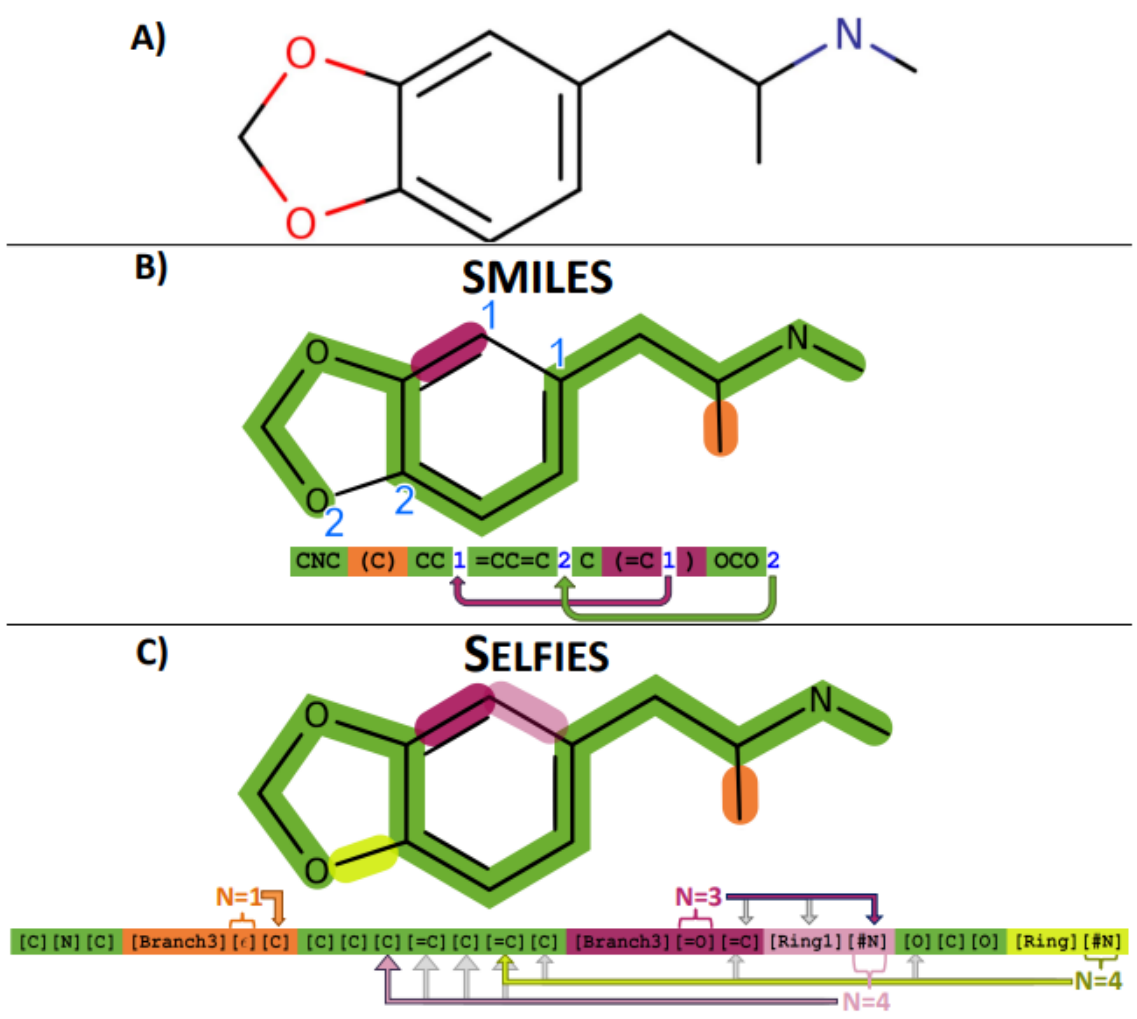
N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

Copper(II) sulfate	$Cu^{2+}SO_4^{2-}$	`[Cu+2].[O-]S(=O)(=O)[O-]`
Vanillin		`O=Cc1ccc(O)c(OC)c1` `COc1cc(C=O)ccc1O`
Melatonin ($C_{13}H_{16}N_2O_2$)		`CC(=O)NCCC1=CNc2c1cc(OC)cc2` `CC(=O)NCCc1c[nH]c2ccc(OC)cc12`

SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences 1988.

Strategy 2: Using a Sequential Language to Describe Molecules

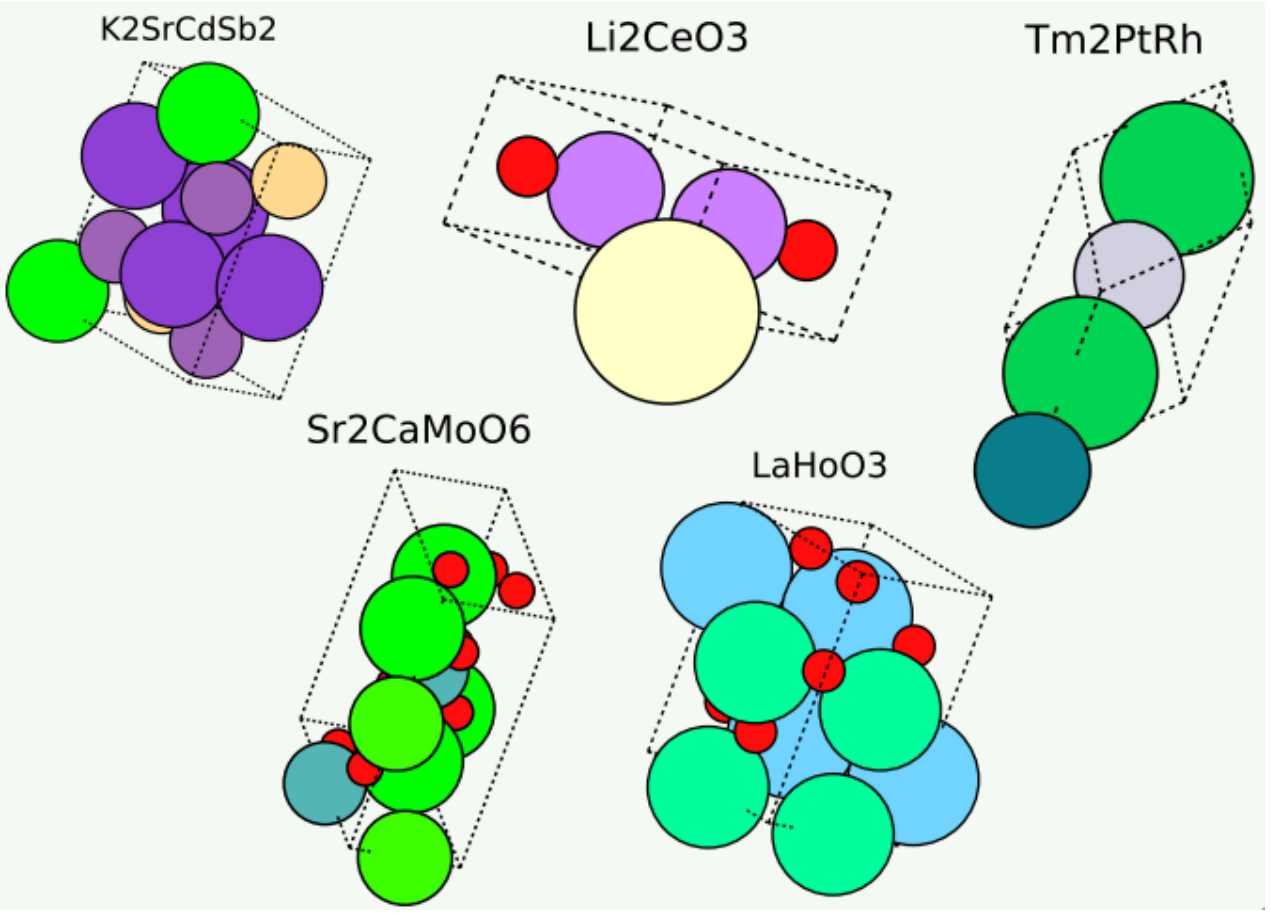
- Simplified Molecular Input Line Entry System (SMILES)
  - SMILES-BERT [1]: masked language modeling on SMILES
- Self-Referencing Embedded Strings (SELFIES) [2]
  - Used in BioT5



[1] *SMILES-BERT: Large Scale Unsupervised Pre-training for Molecular Property Prediction.* ACM BCB 2019.
[2] *Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation.* Machine Learning: Science and Technology 2020.

Strategy 2: Using a Sequential Language to Describe Molecules

- However, some structural information cannot be captured by the SMILES/SELFIES string.
  - Crystals (atom positions)
  - Can we use natural language to describe molecules?



Agenda

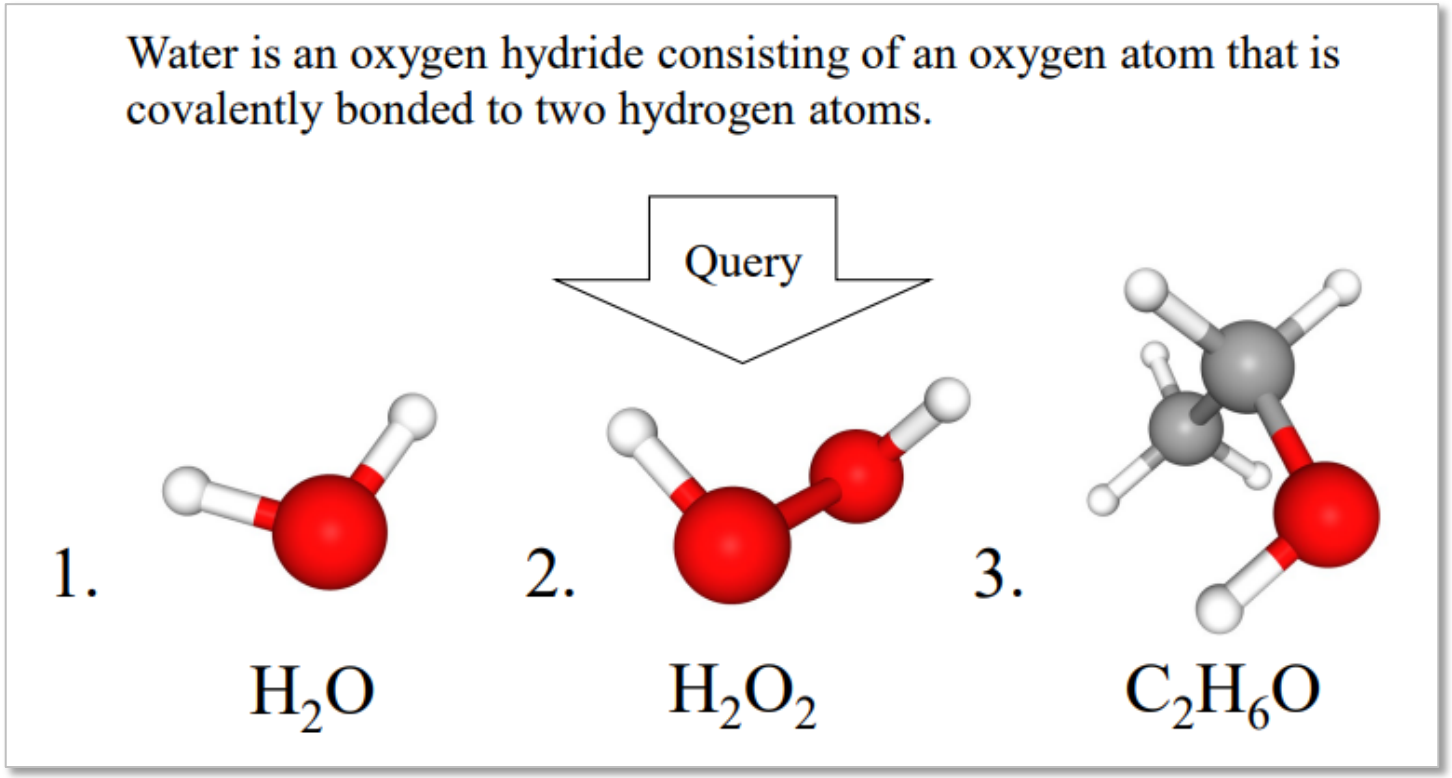
- Using a Graph Encoder
  - Text2Mol: CLIP
- Using SMILES/SELFIES to Describe Molecules
  - MolT5: Encoder-Decoder
  - LlaSMol: Decoder-Only + Instruction Tuning
- Using Natural Language to Describe Molecules
  - CrystalLLM: Decoder-Only + Instruction Tuning

Agenda

- Using a Graph Encoder
  - Text2Mol: CLIP
- Using SMILES/SELFIES to Describe Molecules
  - MolT5: Encoder-Decoder
  - LlaSMol: Decoder-Only + Instruction Tuning
- Using Natural Language to Describe Molecules
  - CrystalLLM: Decoder-Only + Instruction Tuning

Text-to-Molecule Retrieval

- Given a natural language description of a chemical, rank the corresponding molecule first among all the possible molecules.



Text2Mol: Cross-Modal Molecule Retrieval with Natural Language Queries. EMNLP 2021.

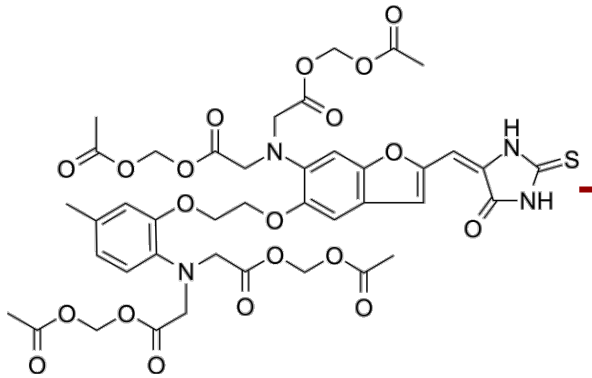
A Bi-Encoder Architecture

- Use GNN and BERT to encode molecules and text descriptions, respectively.

Fura red is a 1-benzofuran substituted at position 2 by a (5-oxo-2-thioxoimidazolidin-4-ylidene) methyl group, and at C-5 and C-6 by heavily substituted oxygen and nitrogen functionalities

SciBERT

[CLS]



GCN

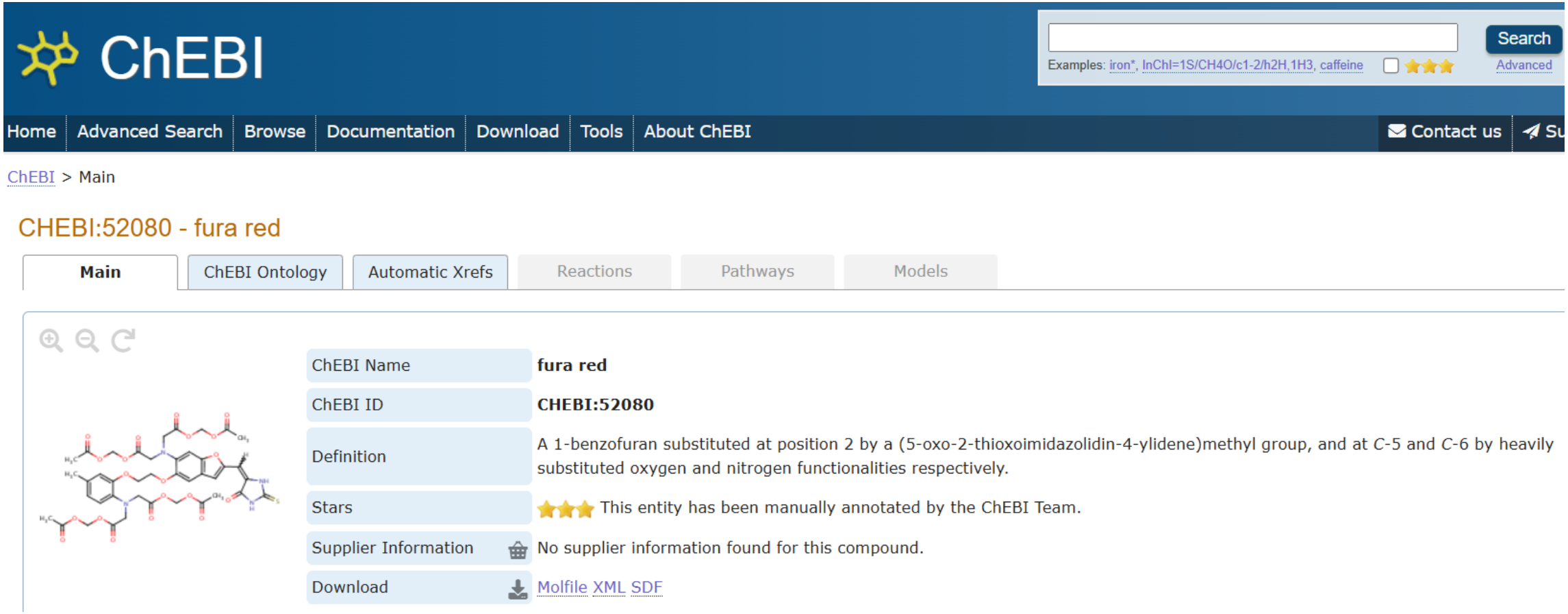
Average

Two vectors should be close if the molecule and the text are paired.

Data

- 33,010 pairs of (molecule, text) from ChEBI, where the length of text is 20+ words.
  - 80%/10%/10% train-validation-test split

https://www.ebi.ac.uk/chebi/



Rank Ensemble

- Combine multiple weaker ranking models to a stronger ranking model
  - These models can either share the same architecture or be totally different.
- Given M ranking models, we use each of them to ranks all candidates. Let $\mathrm{rank}_i(x)$ denote the rank of candidate x according to model i ($i = 1,2, \dots, M$).
- How to combine these ranks?
  - Mean rank: $\min \sum_{i=1}^{M} \mathrm{rank}_i(x)$
  - Mean reciprocal rank: $\max \sum_{i=1}^{M} \frac{1}{\mathrm{rank}_i(x)}$

- This work
  - Train the same model multiple times with different parameter initialization
  - Use mean rank to combine these models

Performance of Text2Mol

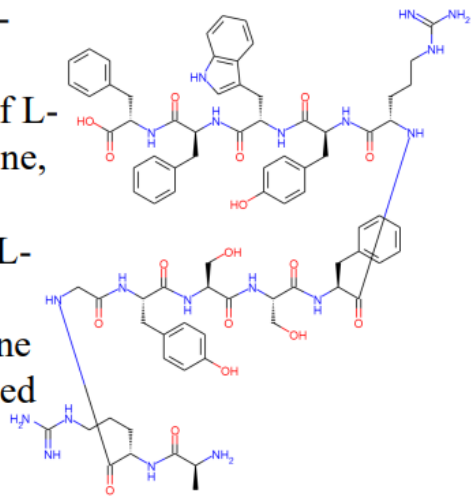
	Training				Test			
Model	Mean Rank	MRR	Hits@1	Hits@10	Mean Rank	MRR	Hits@1	Hits@10
MLP1	9.55	0.428	26.5%	77.5%	30.38	0.372	22.4%	68.6%
MLP2	9.82	0.425	26.4%	77.1%	30.72	0.369	22.3%	68.9%
MLP3	9.53	0.431	26.9%	77.8%	36.30	0.372	22.3%	67.9%
GCN1	10.22	0.432	27.2%	76.5%	42.28	0.366	21.7%	68.2%
GCN2	9.67	0.423	26.7%	77.4%	41.90	0.371	22.3%	68.9%
GCN3	10.12	0.420	25.8%	76.7%	39.11	0.366	22.3%	67.9%
MLP-Ensemble	5.81	0.520	35.1%	86.4%	20.78	0.452	29.4%	77.6%
GCN-Ensemble	6.09	0.516	35.0%	86.1%	28.77	0.447	29.4%	77.1%
All-Ensemble	**4.67**	**0.568**	**40.2%**	**89.8%**	**20.21**	**0.499**	**34.4%**	**81.1%**

Association Rules from Token to Chemical Substructure

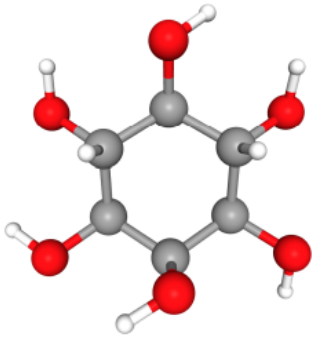
Token	Substructure	Supp	Conf
Titanium	Ti=O	1.29	0.65
Aluminium	Al^{3+}	4.31	0.23
Manganese	Mn^{2+}	10.08	0.30
Toluene	C – C=C	12.93	0.231
Toluene	C_7H_8	23.79	0.425
##chloro	Cl – C	18.81	0.207
pollutant	F – C	3.097	0.208
chromatography	C – Si	2.976	0.271
acid	C – O – H	2398.7	0.078
crown	C – C – O	4.18	0.325

Case Study: Correct Predictions

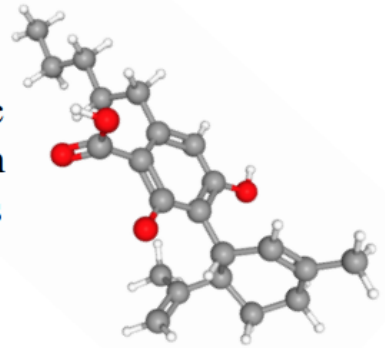
Argyssfrywff: Ala-Arg-Gly-Tyr-Ser-Ser-Phe-Arg-Tyr-Trp-Phe-Phe is an oligopeptide composed of L-alanine, L-arginine, glycine, L-tyrosine, L-serine, L-serine, L-phenylalanine, L-arginine, L-tyrosine, L-trytophan, L-phenylalanine and L-phenylalanine joined in sequence by peptide linkages.



Inositol: Myo-inositol is an inositol having myo-configuration. It has a role as a member of compatible osmolytes, a nutrient, an EC 3.1.4.11 (phosphoinositide phospholipase C) inhibitor, a human metabolite, a Daphnia magna metabolite, […]

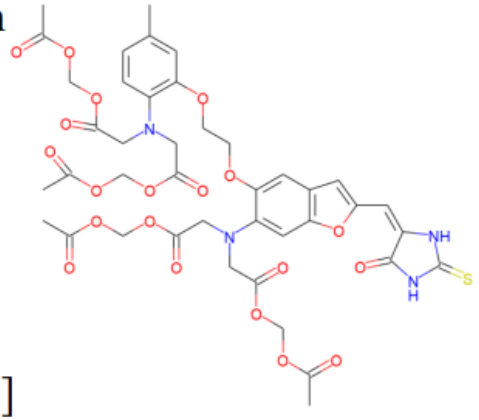


Cannabidiolate is a dihydroxybenzoate that is the conjugate base of cannabidiolic acid, obtained by deprotonation of the carboxy group. It derives from an olivetolate. It is a conjugate base of a cannabidiolic acid.

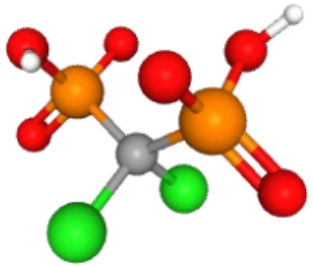


Case Study: Incorrect Predictions

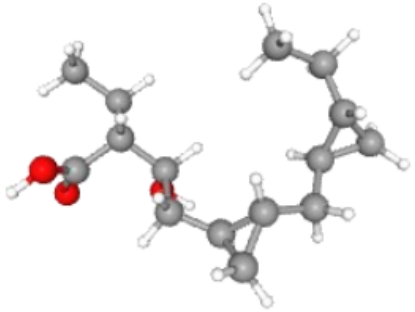
Fura red is a 1-benzofuran substituted at position 2 by a (5-oxo-2-thioxoimidazolidin-4-ylidene)methyl group, and at C-5 and C-6 by heavily substituted oxygen and nitrogen functionalities […]



Clondronate(2-) is the dianion resulting from the removal of two protons from clondronic acid. It is a conjugate base of a clodronic acid.



An alpha-mycolic acid is a class of mycolic acids characterized by the presence of two cis cyclopropyl groups in the meromycolic chain. It is an organic molecular entity and a mycolic acid. […]



Take-Away Messages

- The CLIP architecture can be extended from citation-enhanced LLMs, vision-language models, and protein language models to molecule language models. GNNs can be used as the molecule encoder.
- Rank ensemble is an effective way to combine multiple weaker ranking models to a stronger ranking model.

- Limitations
  - Molecules are heterogeneous graphs. Nodes have types (carbon, oxygen, …). Edges also have types (single bond, double bond, …). How to consider these signals in GNNs?
  - Because SMILES-BERT can handle molecules as sequences, can we build a CLIP model by just combining SMILES-BERT and SciBERT?
  - The CLIP model still relies on massive paired (molecule, text) data.

Agenda

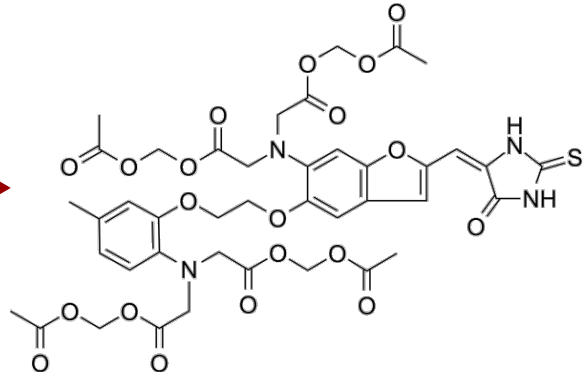
- Using a Graph Encoder
  - Text2Mol: CLIP
- Using SMILES/SELFIES to Describe Molecules
  - MolT5: Encoder-Decoder
  - LlaSMol: Decoder-Only + Instruction Tuning
- Using Natural Language to Describe Molecules
  - CrystalLLM: Decoder-Only + Instruction Tuning

Molecule Generation and Molecule Captioning

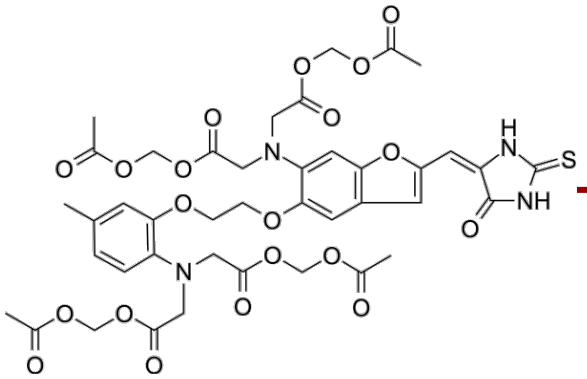
- Molecule Generation

Fura red is a 1-benzofuran substituted at position 2 by a (5-oxo-2-thioxoimidazolidin-4-ylidene) methyl group, and at C-5 and C-6 by heavily substituted oxygen and nitrogen functionalities

→ Text-to-SMILES →



- Molecule Captioning

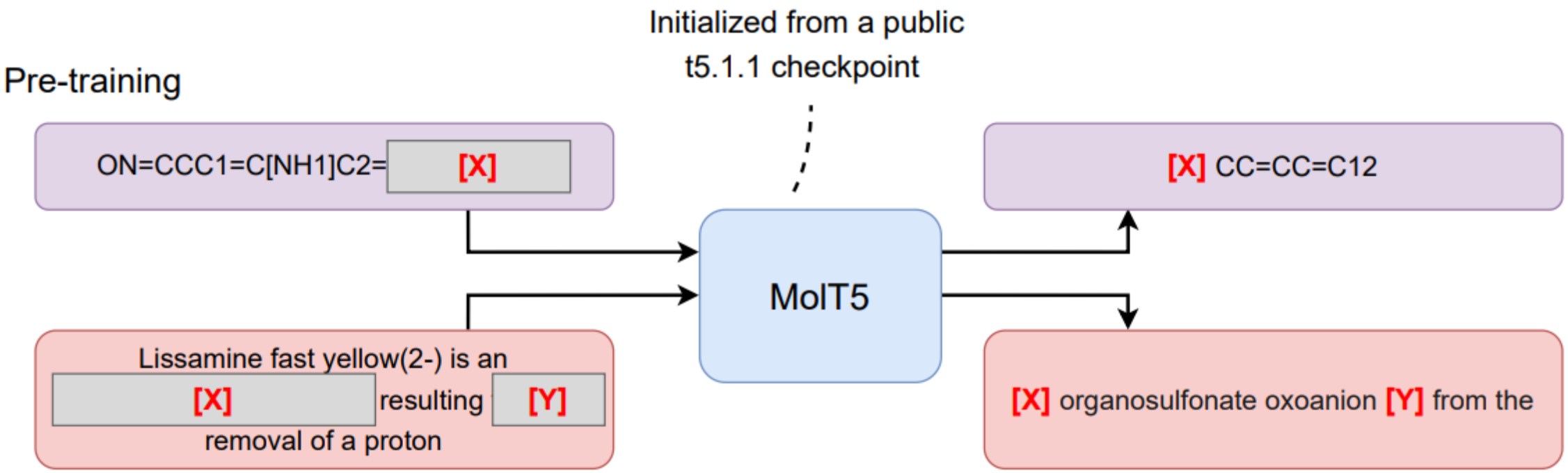


→ SMILES-to-Text →

Fura red is a 1-benzofuran substituted at position 2 by a (5-oxo-2-thioxoimidazolidin-4-ylidene) methyl group, and at C-5 and C-6 by heavily substituted oxygen and nitrogen functionalities

What if we do not have massive paired (molecule, text) data?

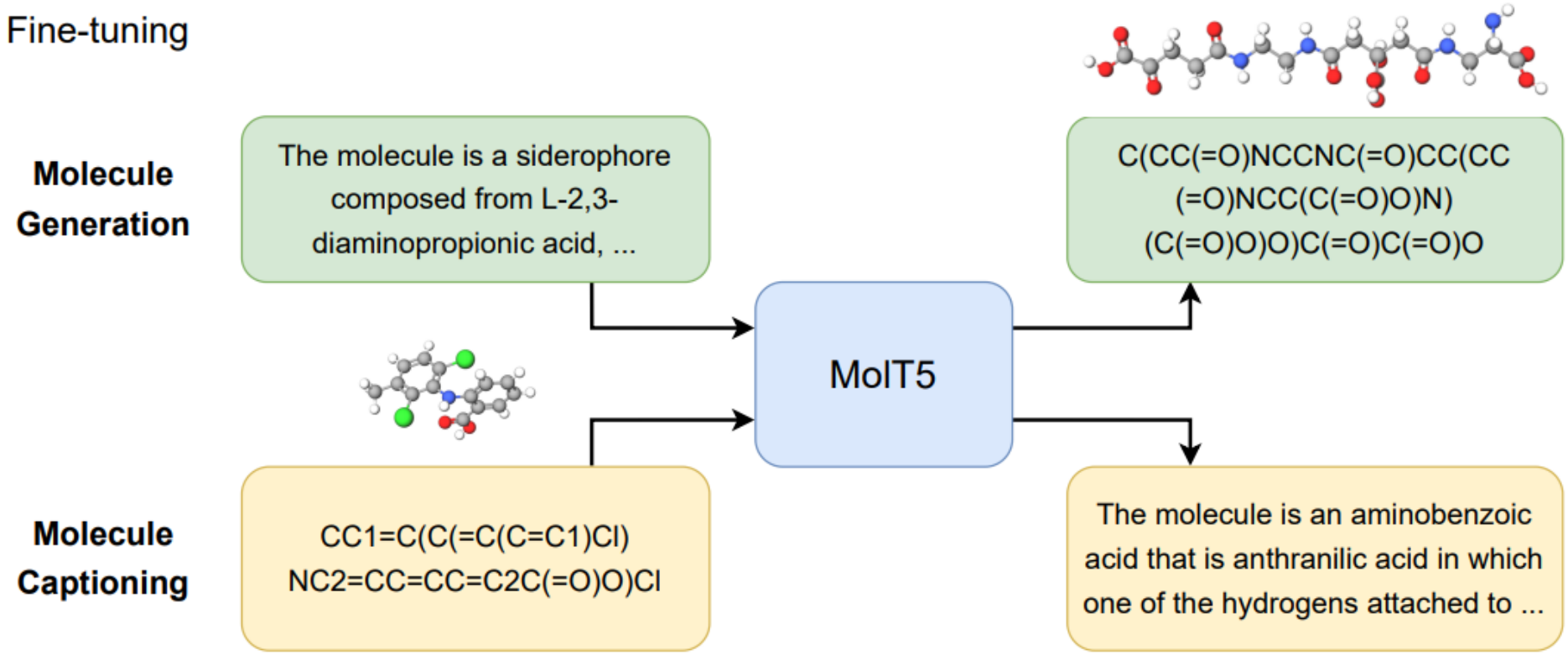
- Pre-training the model within the molecule modality and the text modality only
  - The "input modality = output modality" case in BioT5



Translation between Molecules and Natural Language. EMNLP 2022.

What if we do not have massive paired (molecule, text) data?

- Fine-tuning the model with a small number of paired (molecule, text) samples
  - The "input modality ≠ output modality" case in BioT5

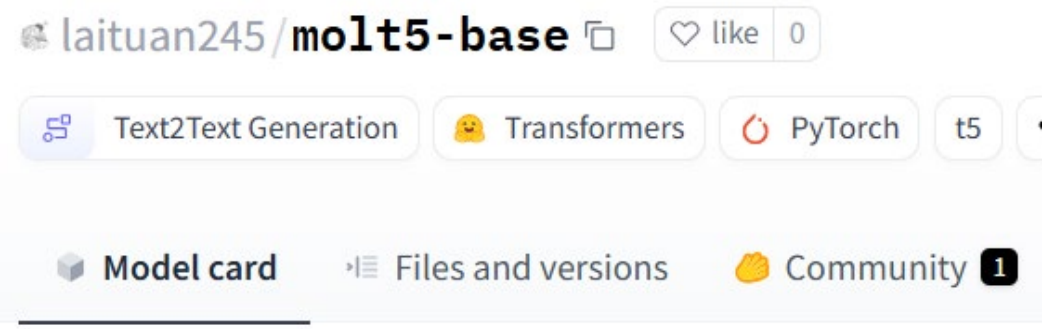


Translation between Molecules and Natural Language. EMNLP 2022.

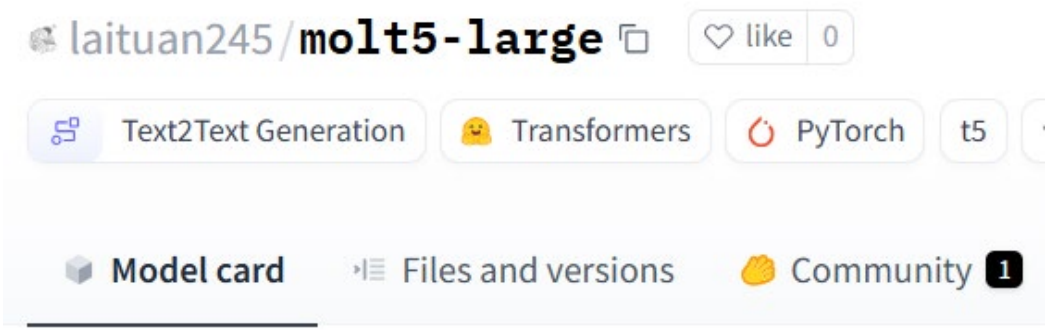
More Details of MolT5

- Initialized from T5 (small: 60M parameters, base: 220M parameters, large: 770M parameters)
- Unpaired text data: Colossal Clean Crawled Corpus (https://github.com/google-research/text-to-text-transfer-transformer#c4)
- Unpaired molecule data: 100M SMLIES strings selected from ZINC-15 (https://zinc15.docking.org)
- Paired (molecule, text) data: ChEBI

https://huggingface.co/laituan245/molt5-base



https://huggingface.co/laituan245/molt5-large



Performance of MolT5

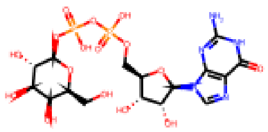
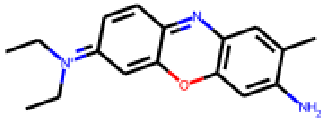
Model	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	Text2Mol
Ground Truth							0.609
RNN	0.251	0.176	0.450	0.278	0.394	0.363	0.426
Transformer	0.061	0.027	0.204	0.087	0.186	0.114	0.057
T5-Small	0.501	0.415	0.602	0.446	0.545	0.532	0.526
MolT5-Small	0.519	0.436	0.620	0.469	0.563	0.551	0.540
T5-Base	0.511	0.423	0.607	0.451	0.550	0.539	0.523
MolT5-Base	0.540	0.457	0.634	0.485	0.578	0.569	0.547
T5-Large	0.558	0.467	0.630	0.478	0.569	0.586	0.563
MolT5-Large	**0.594**	**0.508**	**0.654**	**0.510**	**0.594**	**0.614**	**0.582**

Table 1: Molecule captioning results on the test split of CheBI-20. Rouge scores are F1 values.

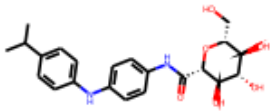
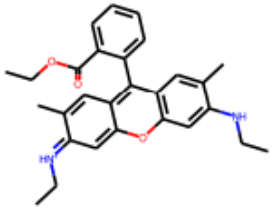
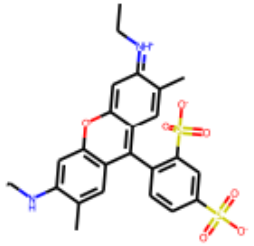
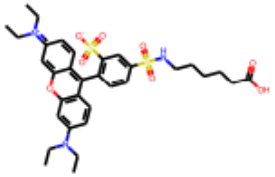
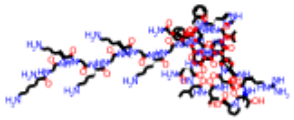
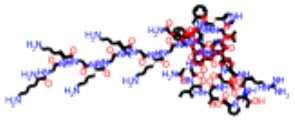
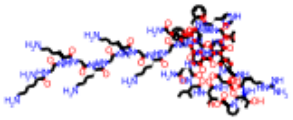
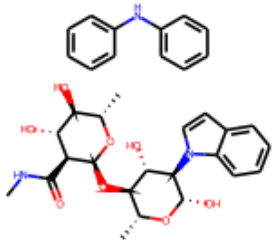
Model	BLEU↑	Exact↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	FCD↓	Text2Mol↑	Validity↑
Ground Truth	1.000	1.000	0.0	1.000	1.000	1.000	0.0	0.609	1.0
RNN	0.652	0.005	38.09	0.591	0.400	0.362	4.55	0.409	0.542
Transformer	0.499	0.000	57.66	0.480	0.320	0.217	11.32	0.277	**0.906**
T5-Small	0.741	0.064	27.703	0.704	0.578	0.525	2.89	0.479	0.608
MolT5-Small	0.755	0.079	25.988	0.703	0.568	0.517	2.49	0.482	0.721
T5-Base	0.762	0.069	24.950	0.731	0.605	0.545	2.48	0.499	0.660
MolT5-Base	0.769	0.081	24.458	0.721	0.588	0.529	2.18	0.496	0.772
T5-Large	0.854	0.279	16.721	0.823	0.731	0.670	1.22	0.552	0.902
MolT5-Large	**0.854**	**0.311**	**16.071**	**0.834**	**0.746**	**0.684**	**1.20**	**0.554**	0.905

Table 2: Molecule generation results on the test split of CheBI-20. Except for BLEU, Exact, Levenshtein, and Validity, other metrics are computed using only syntactically valid molecules, as in (Campos and Ji, 2021).

Case Study: Molecule Captioning

	Input	RNN	Transformer	T5	MolT5	Ground Truth
1		the molecule is a gdp - d - glucoside - a - [...]	the molecule is the stable isotope of helium with relative atomic mass 3. 016029. the least abundant (0. 000137 atom percent) isotope of naturally occurring helium.	The molecule is a GDP-D-glucose in which the anomeric centre of the pyranose fragment has alpha-configuration. It is a GDP-D-glucose and a ribonucleoside 5'-diphosphate-alpha-D-glucose. It is a conjugate acid of a GDP-alpha-D-glucose(2-).	The molecule is a GDP-L-galactose in which the anomeric oxygen is on the same side of the fucose ring as the methyl substituent. It has a role as a plant metabolite and a mouse metabolite. It is a conjugate acid of a GDP-beta-L-galactose(2-).	The molecule is a GDP-L-galactose having beta-configuration at the anomeric centre of the L-galactose fragment. It is a conjugate acid of a GDP-beta-L-galactose(2-).
2	^{39}Ar	the molecule is stable metallic [...]	the molecule is the stable isotope of thallium with relative atomic mass 202. 9723. the least abundant (29. 524 atom percent) isotope of naturally occurring thallium.	The molecule is the radioactive isotope of chromium with relative atomic mass 39.98286 and half-life of 138.376 days; the only naturally occurring isotope of chromium.	The molecule is the stable isotope of rubidium with relative atomic mass 44.955910, 100 atom percent natural abundance and nuclear spin 7/2.	The molecule is a trace radioisotope of argon with atomic mass of 38.964313 and a half-life of 269 years. It has a role as an isotopic tracer.
3		the molecule is a cationic fluorescent dye having 2, 3 - dimethyl - 1, 2, 3, 4, 6 - tetrahydro - 1h - 1, 2, 3, 4, 6 - tetrahydropyridin - 1 - yl] amino } amino group, respectively. it has a role as a fluorochrome.	the molecule is a deuterated compound that is is is is an isotopologue of chloroform in which the four hydrogen atoms have been replaced by deuterium. it is a deuterated compound and an alpha, omega - dicarboxylic acid.	The molecule is a quaternary ammonium ion and a member of phenanthridines. It has a role as an intercalator and a fluorochrome.	The molecule is an organic cation that is phenoxazin-5-ium substituted by amino and methylamino groups at positions 3 and 7 respectively. The chloride salt is the histological dye 'azure C'.	The molecule is an organic cation that is phenoxazin-5-ium substituted by methyl, amino and diethylamino groups at positions 2, 3 and 7 respectively. The tetrachlorozincate salt salt is the histological dye 'brilliant cresyl blue'.

Case Study: Molecule Generation

	Input	RNN	Transformer	T5	MolT5	Ground Truth
1	The molecule is a sulfonated xanthene dye of absorption wavelength 573 nm and emission wavelength 591 nm. It has a role as a fluorochrome.	Invalid				
2	The molecule is a linear 27-membered polypeptide comprising the sequence Lys-Gly-Lys-Gly-Lys-Gly-Lys-Gly-Lys-Gly-Glu-Asn-Pro-Val-Val-His-Phe-Phe-Tyr-Asn-Ile-Val-Thr-Pro-Arg-Thr-Pro. Corresponds to the sequence of the myelin basic protein 83-99 (MBP83-99) immunodominant epitope with the lysyl residue at position 91 replaced by tyrosyl [MBP83-99(Y(91))] and with an (L-lysylglycyl)5 [(KG5)] linker attached to the glutamine(83) (E(83)) residue.	Invalid	Invalid			
3	The molecule is a hydrate that is the dihydrate form of manganese(II) chloride. It has a role as a MRI contrast agent and a nutraceutical. It is a hydrate, an inorganic chloride and a manganese coordination entity.	$^{\cdot}O^{-}$ H_2O H_2O $H_2OH_2OH_2OH_2O$ Na^+		$Mn^{2+\cdot}$ Cl^- H_2OH_2O Cl^-		

Take-Away Messages

- Unlike CLIP, for sequence-to-sequence models (e.g., BART and T5), even if you do not have paired data, you can still pre-train the model within each modality using self supervision and then fine-tune the model using a small amount of paired data.
  - Can we apply the same idea to DNA/RNA language models?

- Limitations:
  - The model still relies on paired (molecule, text) data
    - Can we just pre-train the model on mixed molecule and text sequences using next token prediction?
    - Can we expect text-to-molecule and molecule-to-text translation to be an emergent ability?
    - *CM3: A Causal Masked Multimodal Model of the Internet*. arXiv 2022.

Agenda

- Using a Graph Encoder
  - Text2Mol: CLIP
- Using SMILES/SELFIES to Describe Molecules
  - MolT5: Encoder-Decoder
  - LlaSMol: Decoder-Only + Instruction Tuning
- Using Natural Language to Describe Molecules
  - CrystalLLM: Decoder-Only + Instruction Tuning

How do off-the-shelf LLMs perform in chemistry tasks?

Molecule Generation (MG)

Query: Give me a molecule that satisfies the conditions outlined in the description: The molecule is a member of the class of tripyrroles that is a red-coloured pigment with antibiotic properties produced by Serratia marcescens. It has a role as an antimicrobial agent, a biological pigment, a bacterial metabolite, an apoptosis inducer and an antineoplastic agent. It is a tripyrrole, an aromatic ether and a ring assembly.

Response: Here is a potential molecule:

<SMILES> CCCCCC1=C(C)NC(/C=C2\N=C(C3=CC=CN3)C=C2OC)=C1 </SMILES>

Method	BLEU (↑)	Exact (↑)	Levenshtein (↓)	Validity (↑)	MACCS FTS (↑)	RDK FTS (↑)	Morgan FTS (↑)	FCD (↓)
MolT5-Large [17]	0.601	0.290	41.600	0.940	0.879	0.797	0.752	5.394
GPT-4 (zero-shot)	0.490±0.017	0.046±0.009	47.418±1.668	0.758±0.015	0.733±0.020	0.514±0.021	0.432±0.014	11.913±0.972
GPT-4 (Scaffold, k=10)	**0.816±0.004**	**0.174±0.029**	**21.160±0.600**	0.888±0.023	**0.867±0.005**	0.738±0.010	**0.672±0.013**	6.224±0.449
GPT-4 (Scaffold, k=5)	0.815±0.011	0.164±0.018	21.862±1.768	0.874±0.030	0.865±0.015	**0.741±0.023**	0.670±0.028	**5.843±0.515**
GPT-4 (Random, k=10)	0.602±0.016	0.060±0.007	42.390±1.008	0.770±0.030	0.762±0.013	0.548±0.017	0.475±0.015	10.594±0.414
GPT-3.5 (Scaffold, k=10)	0.479±0.156	0.094±0.011	82.008±40.354	0.854±0.059	0.833±0.006	0.686±0.016	0.585±0.013	8.341±0.607
Davinci-003 (Scaffold, k=10)	0.741±0.011	0.100±0.010	25.648±2.186	0.936±0.009	0.783±0.014	0.648±0.004	0.560±0.010	8.335±0.310
Llama2-13B-chat (Scaffold, k=10)	0.626±0.013	0.020±0.000	33.956±2.648	0.782±0.008	0.679±0.015	0.568±0.014	0.454±0.009	12.387±0.437
GAL-30B (zero-shot)	0.004±0.000	0.000±0.000	2738.136±166.093	**0.956±0.011**	0.233±0.011	0.109±0.006	0.053±0.002	35.091±0.774

What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks. NeurIPS 2023.

How do off-the-shelf LLMs perform in chemistry tasks?

Molecule Captioning (MC)

Query: Describe this molecule:
<SMILES> CCOC(=O)C1=CN=CN1[C@H](C)C1=CC=CC=C1 </SMILES>

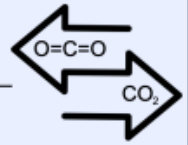
Response: The molecule is an imidazole derivative with short-acting sedative, hypnotic, and general anesthetic properties. Etomidate appears to have gamma-aminobutyric acid (GABA) like effects, mediated through GABA-A receptor. The action enhances the inhibitory effect of GABA on the central nervous system by causing chloride channel opening events which leads to membrane hyperpolarization.

Method	BLEU-2 (↑)	BLEU-4 (↑)	ROUGE-1 (↑)	ROUGE-2 (↑)	ROUGE-L (↑)	METEOR (↑)
MolT5-Large [17]	0.482	0.383	0.574	0.410	0.516	0.530
GPT-4 (zero-shot)	0.062±0.001	0.013±0.001	0.192±0.002	0.040±0.002	0.125±0.002	0.209±0.002
GPT-4 (Scaffold, k=10)	0.464±0.008	0.365±0.008	**0.545±0.003**	**0.362±0.003**	0.459±0.007	**0.519±0.005**
GPT-4 (Scaffold, k=5)	0.456±0.003	0.357±0.004	0.540±0.005	0.355±0.007	0.455±0.005	0.505±0.005
GPT-4 (Random, k=10)	0.260±0.007	0.140±0.007	0.393±0.004	0.180±0.006	0.309±0.004	0.320±0.007
GPT-3.5 (Scaffold, k=10)	0.468±0.010	0.368±0.010	0.534±0.005	0.355±0.007	0.457±0.006	0.497±0.005
Davinci-003 (Scaffold, k=10)	**0.488±0.011**	**0.391±0.012**	0.532±0.008	0.359±0.010	**0.465±0.008**	0.478±0.011
Llama2-13B-chat (Scaffold, k=10)	0.197±0.005	0.140±0.004	0.331±0.005	0.193±0.005	0.265±0.005	0.372±0.006
GAL-30B (zero-shot)	0.008±0.000	0.002 ± 0.000	0.019±0.002	0.004±0.000	0.015±0.002	0.043±0.002

What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks. NeurIPS 2023.

How do off-the-shelf LLMs perform in chemistry tasks?

Name Conversion



IUPAC to Molecular Formula (NC-I2F)

Query: What is the molecular formula of the compound with this IUPAC name
<IUPAC> 2,5-diphenyl-1,3-oxazole </IUPAC> ?
Response: <MOLFORMULA> C15H11NO </MOLFORMULA>

IUPAC to SMILES (NC-I2S)

Query: Could you provide the SMILES for
<IUPAC> 4-ethyl-4-methyloxolan-2-one </IUPAC> ?
Response: Of course. It's <SMILES> CCC1(C)COC(=O)C1 </SMILES>

SMILES to Molecular Formula (NC-S2F)

Query: Given the SMILES representation
<SMILES> S=P1(N(CCCl)CCCl)NCCCO1 </SMILES>, what would be its molecular formula?
Response: It is <MOLFORMULA> C7H15Cl2N2OPS </MOLFORMULA> .

SMILES to IUPAC (NC-S2I)

Query: Translate the given SMILES formula of a molecule
<SMILES> CCC(C)C1CNCCCNC1 </SMILES> into its IUPAC name.
Response: <IUPAC> 3-butan-2-yl-1,5-diazocane </IUPAC>

Method	smiles2iupac	iupac2smiles	smiles2formula	iupac2formula
STOUT [47]	0.55	0.7	-	-
GPT-4 (zero-shot)	0	0.008±0.008	0.048± 0.022	0.092±0.018
GPT-4 (Scaffold, k=5)	0	**0.014±0.009**	0.058±0.015	**0.118±0.022**
GPT-4 (Scaffold, k=20)	0	0.012±0.004	**0.086±0.036**	0.084±0.005
GPT-4 (Random, k=20)	0	0.010±0.007	0.070±0.032	0.076±0.011
GPT-3.5 (Scaffold, k=20)	0	0.010±0.000	0.052±0.004	0.044±0.009
Davinci-003 (Scaffold, k=20)	0	0	0.006±0.005	0.018±0.004
Llama2-13B-chat (Scaffold, k=20)	0	0	0.010±0.007	0
GAL-30B (Scaffold, k=10)	0	0	0	0

What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks. NeurIPS 2023.

How do off-the-shelf LLMs perform in chemistry tasks?

Property Prediction

ESOL (PP-ESOL)

Query: How soluble is <SMILES> CC(C)Cl </SMILES> ?
Response: Its log solubility is <NUMBER> -1.41 </NUMBER> mol/L.

LIPO (PP-LIPO)

Query: Predict the octanol/water distribution coefficient logD under the circumstance of pH 7.4 for <SMILES> NC(=O)C1=CC=CC=C1O </SMILES> .
Response: <NUMBER> 1.090 </NUMBER>

BBBP (PP-BBBP)

Query: Is blood-brain barrier permeability (BBBP) a property of <SMILES> CCNC(=O)/C=C/C1=CC=CC(Br)=C1 </SMILES>?
Response: <BOOLEAN> Yes </BOOLEAN>

ClinTox (PP-ClinTox)

Query: Is <SMILES> COC[C@@H](NC(C)=O)C(=O)NCC1=CC=CC=C1 </SMILES> toxic?
Response: <BOOLEAN> No </BOOLEAN>

HIV (PP-HIV)

Query: Can <SMILES> CC1=CN(C2C=CCCC2O)C(=O)NC1=O </SMILES> serve as an inhibitor of HIV replication?
Response: <BOOLEAN> No </BOOLEAN>

SIDER (PP-SIDER)

Query: Are there any known side effects of <SMILES> CC1=CC(C)=C(NC(=O)CN(CC(=O)O)CC(=O)O)C(C)=C1Br </SMILES> affecting the heart?
Response: <BOOLEAN> No </BOOLEAN>

	BBBP	BACE	HIV	Tox21	ClinTox
RF	0.881	0.758	0.518	0.260	0.461
XGBoost	0.897	0.765	0.551	0.333	0.620
GPT-4 (zero-shot)	0.560 ±0.034	0.322±0.018	0.977±0.013	0.489±0.018	0.555±0.043
GPT-4 (Scaffold, k= 4)	0.498 ± 0.028	0.516 ± 0.024	0.818 ± 0.015	0.444 ± 0.004	0.731 ± 0.035
GPT-4 (Scaffold, k= 8)	**0.587±0.018**	**0.666±0.023**	0.797 ± 0.021	**0.563±0.008**	0.736±0.033
GPT-4 (random, k= 8)	0.469 ± 0.025	0.504 ± 0.020	**0.994 ± 0.006**	0.528±0.003	**0.924±0.000**
GPT-3.5 (Scaffold, k= 8)	0.463 ± 0.008	0.406 ± 0.011	0.807 ± 0.021	0.529 ± 0.021	0.369 ± 0.029
Davinci-003 (Scaffold, k= 8)	0.378 ± 0.024	0.649 ± 0.021	0.832 ± 0.020	0.518±0.009	0.850 ± 0.020
Llama2-13B-chat (Scaffold, k= 8)	0.002 ± 0.001	0.045 ± 0.015	0.069 ± 0.033	0.047 ± 0.013	0.001 ± 0.003
GAL-30B (Scaffold, k= 8)	0.074 ± 0.019	0.025 ± 0.013	0.014 ± 0.016	0.077 ± 0.046	0.081 ± 0.015

What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks. NeurIPS 2023.

How do off-the-shelf LLMs perform in chemistry tasks?

Property Prediction

ESOL (PP-ESOL)

Query: How soluble is <SMILES> CC(C)Cl </SMILES> ?
Response: Its log solubility is <NUMBER> -1.41 </NUMBER> mol/L.

LIPO (PP-LIPO)

Query: Predict the octanol/water distribution coefficient logD under the circumstance of pH 7.4 for <SMILES> NC(=O)C1=CC=CC=C1O </SMILES> .
Response: <NUMBER> 1.090 </NUMBER>

BBBP (PP-BBBP)

Query: Is blood-brain barrier permeability (BBBP) a property of <SMILES> CCNC(=O)/C=C/C1=CC=CC(Br)=C1 </SMILES>?
Response: <BOOLEAN> Yes </BOOLEAN>

ClinTox (PP-ClinTox)

Query: Is <SMILES> COC[C@@H](NC(C)=O)C(=O)NCC1=CC=CC=C1 </SMILES> toxic?
Response: <BOOLEAN> No </BOOLEAN>

HIV (PP-HIV)

Query: Can <SMILES> CC1=CN(C2C=CCCC2O)C(=O)NC1=O </SMILES> serve as an inhibitor of HIV replication?
Response: <BOOLEAN> No </BOOLEAN>

SIDER (PP-SIDER)

Query: Are there any known side effects of <SMILES> CC1=CC(C)=C(NC(=O)CN(CC(=O)O)CC(=O)O)C(C)=C1Br </SMILES> affecting the heart?
Response: <BOOLEAN> No </BOOLEAN>

	BBBP	BACE	HIV	Tox21	ClinTox
RF	0.881	0.758	0.518	0.260	0.461
XGBoost	0.897	0.765	0.551	0.333	0.620
GPT-4 (zero-shot)	0.560 ±0.034	0.322±0.018	0.977±0.013	0.489±0.018	0.555±0.043
GPT-4 (Scaffold, k= 4)	0.498 ± 0.028	0.516 ± 0.024	0.818 ± 0.015	0.444 ± 0.004	0.731 ± 0.035
GPT-4 (Scaffold, k= 8)	**0.587±0.018**	**0.666±0.023**	0.797 ± 0.021	**0.563±0.008**	0.736±0.033
GPT-4 (random, k= 8)	0.469 ± 0.025	0.504 ± 0.020	**0.994 ± 0.006**	0.528±0.003	**0.924±0.000**
GPT-3.5 (Scaffold, k= 8)	0.463 ± 0.008	0.406 ± 0.011	0.807 ± 0.021	0.529 ± 0.021	0.369 ± 0.029
Davinci-003 (Scaffold, k= 8)	0.378 ± 0.024	0.649 ± 0.021	0.832 ± 0.020	0.518±0.009	0.850 ± 0.020
Llama2-13B-chat (Scaffold, k= 8)	0.002 ± 0.001	0.045 ± 0.015	0.069 ± 0.033	0.047 ± 0.013	0.001 ± 0.003
GAL-30B (Scaffold, k= 8)	0.074 ± 0.019	0.025 ± 0.013	0.014 ± 0.016	0.077 ± 0.046	0.081 ± 0.015

What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks. NeurIPS 2023.

How do off-the-shelf LLMs perform in chemistry tasks?

Forward Synthesis (FS)

Query: <SMILES> NC1=CC=C2OCOC2=C1.O=CO</SMILES>

Based on the reactants and reagents given above, suggest a possible product.

Response: A possible product can be <SMILES> O=CNC1=CC=C2OCOC2=C1 </SMILES>

Method	Top-1 Accuracy (↑)	Invalid SMILES (↓)
Chemformer [26]	0.938	0%
GPT-4 (zero-shot)	0.004 ± 0.005	17.4% ± 3.9%
GPT-4 (Scaffold, k=20)	**0.230 ± 0.022**	7.0% ± 1.6%
GPT-4 (Random, k=20)	0.012 ± 0.008	8.4% ± 1.5%
GPT-4 (Scaffold, k=5)	0.182 ± 0.015	6.6% ± 1.5%
GPT-3.5 (Scaffold, k=20)	0.184 ± 0.005	15.6% ± 2.3%
Davinci-003 (Scaffold, k=20)	0.218 ± 0.008	11.4% ± 2.7%
Llama2-13B-chat (Scaffold, k=20)	0.032 ± 0.013	27.8% ± 5.5%
GAL-30B (Scaffold, k=5)	0.036 ± 0.011	**5.2% ± 1.5%**

What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks. NeurIPS 2023.

How do off-the-shelf LLMs perform in chemistry tasks?

Retrosynthesis (RS)

Query: Identify possible reactants that could have been used to create the specified product.
<SMILES> CC1=CC=C(N)N=C1N </SMILES>
Response: <SMILES> CC(C#N)CCC#N.N </SMILES>

Method	Top-1 Accuracy (↑)	Invalid SMILES (↓)
Chemformer [26]	0.536	0%
GPT-4 (zero-shot)	0.006 ± 0.005	20.6% ± 4.7%
GPT-4 (Scaffold, k=20)	0.096 ± 0.013	10.4% ± 3.4%
GPT-4 (Scaffold, k=5)	0.114 ± 0.013	11.0% ± 1.2%
GPT-4 (Random, k=20)	0.012 ± 0.011	18.2% ± 4.2%
GPT-3.5 (Scaffold, k=20)	0.022 ± 0.004	6.4% ± 1.3%
Davinci-003 (Scaffold, k=20)	**0.122 ± 0.013**	6.0% ± 1.2%
Llama2-13B-chat (Scaffold, k=20)	0	27.2% ± 1.5%
GAL-30B (Scaffold, k=5)	0.016 ± 0.005	**5.2% ± 1.8%**

What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks. NeurIPS 2023.

Observation

- Off-the-shelf LLMs (+ few-shot in-context learning) cannot outperform task-specific supervised models with much fewer parameters in most cases.
- Can we instruction-tune an LLM on a wide range of chemistry tasks?

Task	Task abbr.	#Train	#Valid	#Test	#All	Qry.	Resp.
Name Conversion. Data Source: PubChem							
IUPAC to Molecular Formula	NC-I2F	300,000	1,497	2,993	304,490	84	25
IUPAC to SMILES	NC-I2S	299,890	1,496	2,993	304,379	82	59
SMILES to Molecular Formula	NC-S2F	299,890	1,496	2,993	304,379	68	26
SMILES to IUPAC	NC-S2I	299,890	1,496	2,993	304,379	72	68
Property Prediction. Data Source: MoleculeNet							
ESOL	PP-ESOL	888	111	112	1,111	43	22
Lipo	PP-Lipo	3,360	420	420	4,200	80	11
BBBP	PP-BBBP	1,569	196	197	1,962	68	11
ClinTox	PP-ClinTox	1,144	143	144	1,431	69	11
HIV	PP-HIV	32,864	4,104	4,107	41,075	63	11
SIDER	PP-SIDER	22,820	2,860	2,860	28,540	82	11
Molecule Description. Data Source: Mol-Instructions, ChEBI-20							
Molecule Captioning	MC	56,498	1,269	2,538	60,305	83	102
Molecule Generation	MG	56,498	1,269	2,493	60,260	117	75
Chemical Reaction. Data Source: USPTO-full							
Forward Synthesis	FS	971,809	2,049	4,062	977,920	98	52
Retrosynthesis	RS	941,735	2,092	4,156	947,983	77	70
Overall		3,288,855	20,498	33,061	3,342,414	83	55

LlaSMol: Advancing Large Language Models for Chemistry with a Large-Scale, Comprehensive, High-Quality Instruction Tuning Dataset. COLM 2024.

Performance of LlaSMol

Table 1: Results for name conversion (NC) and property prediction (PP) tasks. Metrics EM, Valid, and Acc are in percentage.

Model	NC					PP					
	I2F	I2S		S2F	S2I	ESOL	Lipo	BBBP	Clintox	HIV	SIDER
	EM	EM	Valid	EM	EM	RMSE↓	RMSE↓	Acc	Acc	Acc	Acc
Task-Specific, Non-LLM Based Models											
SoTA	97.9	73.5	99.4	100.0	56.5	0.819	0.612	85.3	92.4	97.0	70.0
Existing LLMs without fine-tuning on SMolInstruct											
GPT-4	8.7	3.3	84.2	4.8	0.0	2.570	1.545	62.9	50.0	59.6	57.6
Claude 3 Opus	34.6	17.7	90.2	9.2	0.0	**1.036**	1.194	**75.1**	41.7	76.4	67.0
Galactica	9.1	9.7	95.6	0.0	0.0	4.184	2.979	69.0	92.4	**96.7**	68.1
Llama 2	0.0	0.0	18.3	0.0	0.0	3.287	1.634	58.9	45.1	93.3	61.9
Code Llama	0.0	0.0	81.0	0.0	0.0	3.483	1.733	58.9	85.4	91.8	60.2
Mistral	0.0	0.0	40.3	0.0	0.0	3.079	1.730	40.6	15.3	7.1	38.1
Molinst (chemistry LLM)	0.0	0.0	96.2	0.0	0.0	2.271	1.691	60.9	6.3	4.5	52.4
ChemLLM (chemistry LLM)	0.8	0.3	3.9	0.0	0.0	1.946	1.797	22.3	75.7	72.9	32.6
Our LlaSMol Series											
$\text{LlaSMol}_{\text{Galactica}}$	83.2	58.7	99.4	91.2	18.3	1.959	1.213	69.0	**93.1**	**96.7**	70.1
$\text{LlaSMol}_{\text{Llama 2}}$	73.8	46.6	99.0	87.0	12.9	2.791	1.338	69.0	92.4	**96.7**	68.7
$\text{LlaSMol}_{\text{Code Llama}}$	75.4	49.9	99.3	88.6	15.5	2.959	1.203	69.0	**93.1**	**96.7**	69.9
$\text{LlaSMol}_{\text{Mistral}}$	**87.9**	**70.1**	**99.6**	**93.2**	**29.0**	1.150	**1.010**	74.6	**93.1**	**96.7**	**70.7**

Performance of LlaSMol

Table 2: Results for molecule captioning (MC), molecule generation (MG), forward synthesis (FS), and retrosynthesis (RS). Metrics EM, FTS, and Valid are in percentage.

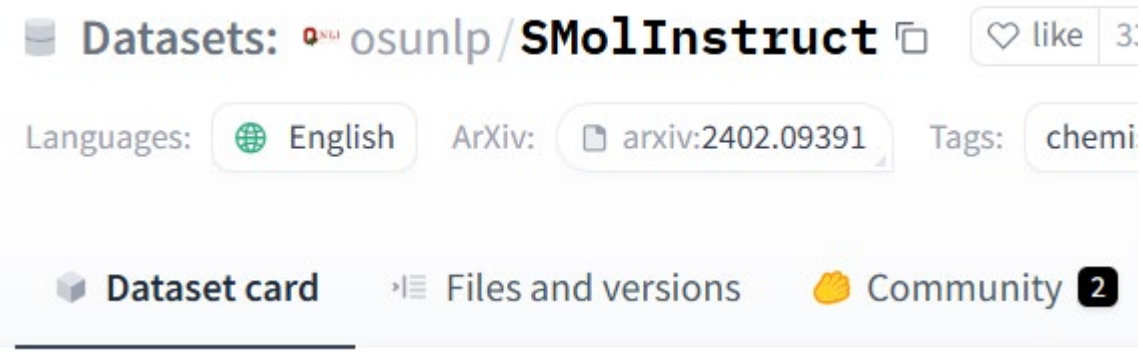
Model	MC	MG			FS			RS		
	METEOR	EM	FTS	Valid	EM	FTS	Valid	EM	FTS	Valid
Task-Specific, Non-LLM Based Models										
SoTA	0.515	31.7	73.2	95.3	78.7	92.2	100.0	47.0	77.5	99.7
Existing LLMs Without Fine-Tuning on SMolInstruct										
GPT-4	0.188	6.4	42.6	81.4	1.6	40.5	87.0	0.0	33.4	42.6
Claude 3 Opus	0.219	12.3	57.6	92.6	3.7	45.7	97.0	1.1	46.2	94.8
Galactica	0.050	0.0	11.6	94.7	0.0	25.9	83.7	0.0	34.6	93.0
Llama 2	0.150	0.0	4.8	93.5	0.0	13.7	97.7	0.0	27.5	87.7
Code Llama	0.143	0.0	8.5	95.2	0.0	15.8	99.6	0.0	25.3	97.1
Mistral	0.193	0.0	9.0	35.9	0.0	19.9	95.8	0.0	24.2	98.0
Molinst (chemistry LLM)	0.124	6.0	43.6	84.8	2.1	31.7	**99.8**	5.7	48.0	97.8
ChemLLM (chemistry LLM)	0.050	0.9	14.3	4.3	0.0	1.6	38.5	0.0	2.9	10.9
Our LlaSMol Series										
$\text{LlaSMol}_{\text{Galactica}}$	0.394	7.7	52.2	99.6	53.1	79.9	99.7	25.7	67.0	99.9
$\text{LlaSMol}_{\text{Llama 2}}$	0.377	6.4	47.1	99.6	47.1	76.9	**99.8**	22.5	65.2	99.9
$\text{LlaSMol}_{\text{Code Llama}}$	0.366	6.5	46.6	**99.7**	52.0	79.2	**99.8**	25.7	66.7	**100.0**
$\text{LlaSMol}_{\text{Mistral}}$	**0.452**	**19.2**	**61.7**	**99.7**	**63.3**	**84.9**	**99.8**	**32.9**	**70.4**	**100.0**

Take-Away Messages

- Instruction-tuning LLMs on a wide range of chemistry tasks significantly improves off-the-shelf LLMs (e.g., GPT-4).
- Despite the performance improvement, LLMs (+ few-shot in-context learning) still underperform task-specific supervised SOTA in most cases.
  - Building a generalist chemistry LLM is still a challenging task!

- Drawbacks:
  - No experiments on how the model can be generalized to unseen chemistry tasks.

https://huggingface.co/datasets/osunlp/SMolInstruct

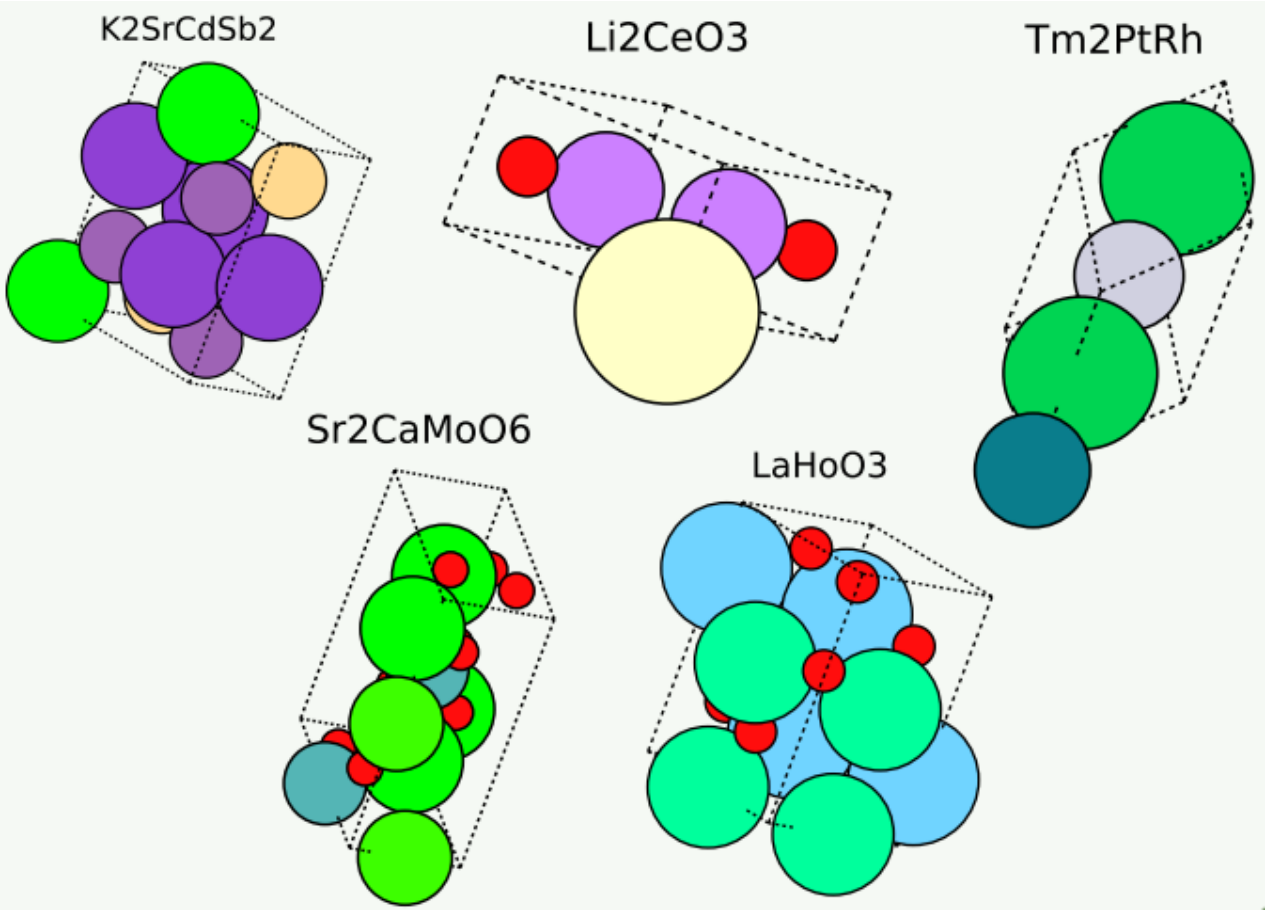


Agenda

- Using a Graph Encoder
  - Text2Mol: CLIP
- Using SMILES/SELFIES to Describe Molecules
  - MolT5: Encoder-Decoder
  - LlaSMol: Decoder-Only + Instruction Tuning
- Using Natural Language to Describe Molecules
  - CrystalLLM: Decoder-Only + Instruction Tuning

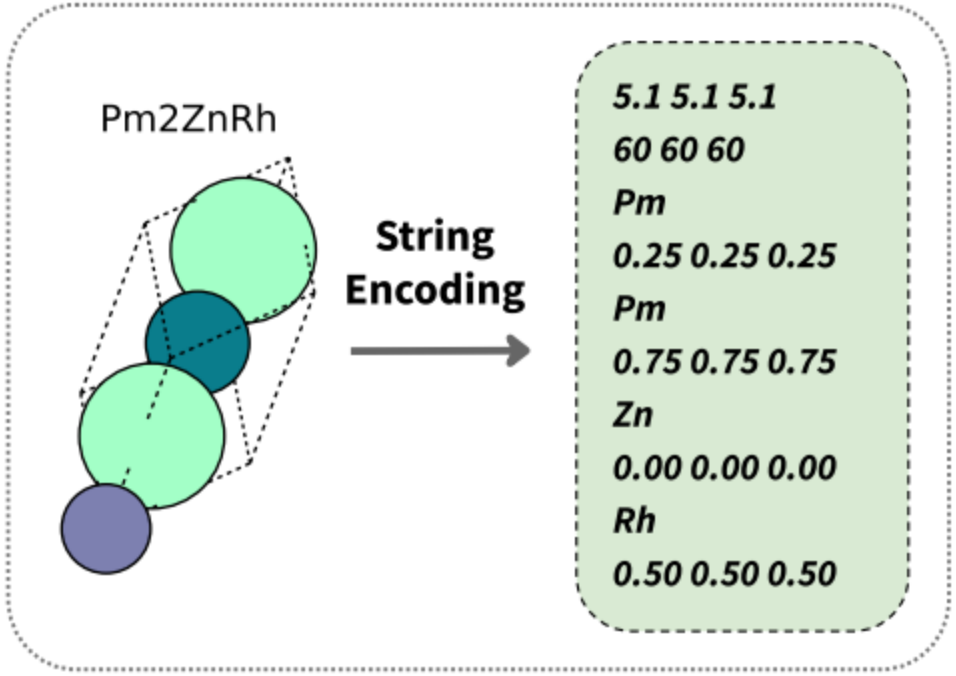
Representing Crystals

- Chemical formula (e.g., K2SrCdSb2)?
  - Too succinct!
- Crystals are periodic, so we only need to describe one cell.
- Key Idea: Use natural language to describe the coordinates of each atom in a cell.



Using Natural Language to Describe Crystals

$$C = (l_1, l_2, l_3, \theta_1, \theta_2, \theta_3, e_1, x_1, y_1, z_1, ..., e_N, x_N, y_N, z_N)\,.$$



Fine-Tuned Language Models Generate Stable Inorganic Materials as Text. ICLR 2024.

Fine-tuning Tasks: Generation and Infilling

Generation Prompt	Infill Prompt
<s>Below is a description of a bulk material. [The chemical formula is Pm2ZnRh]. Generate a description of the lengths and angles of the lattice vectors and then the element type and coordinates for each atom within the lattice: [Crystal string]</s>	<s>Below is a partial description of a bulk material where one element has been replaced with the string "[MASK]": [Crystal string with [MASK]s] Generate an element that could replace [MASK] in the bulk material: [Masked element]</s>

Blue text is optional and included to enable conditional generation. Purple text stands in for string encodings of atoms.

Fine-Tuned Language Models Generate Stable Inorganic Materials as Text. ICLR 2024.

Performance of CrystalLLM

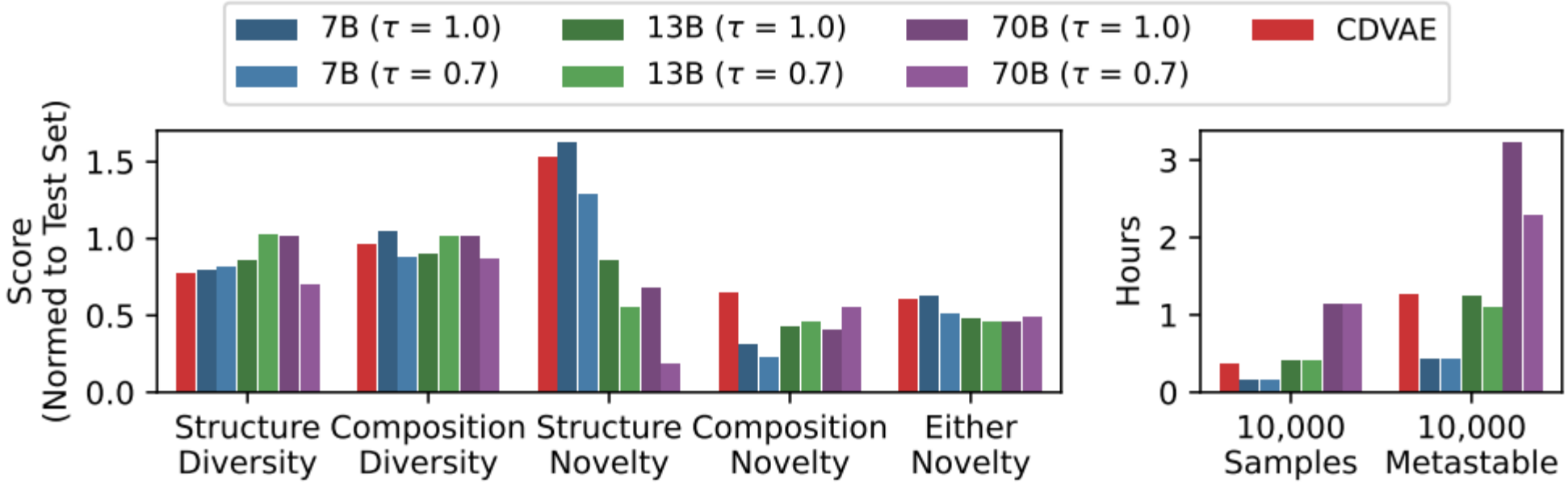
Table 1: Following prior work (Xie et al., 2021), we evaluate fine-tuned LLaMA-2 models using validity, which captures physical constraints, as well as coverage and property metrics, which capture alignment between the ground truth and sampling distribution. We add stability checks, which count the percentage of samples estimated to be stable by M3GNet (Chen & Ong, 2022) and DFT (Hafner, 2008) (details in Appendix B.2). LLaMA models generate a high percentage of both valid and stable materials.

Method	**Validity Check**		**Coverage**		**Property Distribution**		**Metastable**	**Stable**
	Structural↑	Composition↑	Recall↑	Precision↑	wdist (ρ)↓	wdist (N_{el})↓	M3GNet ↑	DFT† ↑
CDVAE	**1.00**	0.867	0.991	0.995	0.688	1.43	28.8%	5.4%
LM-CH	0.848	0.835	0.9925	0.9789	0.864	0.13	n/a	n/a
LM-AC	0.958	0.889	0.996	0.9855	0.696	0.09	n/a	n/a
LLaMA-2								
7B ($\tau = 1.0$)	0.918	0.879	0.969	0.960	3.85	0.96	35.1%	6.7%
7B ($\tau = 0.7$)	0.964	0.933	0.911	0.949	3.61	1.06	35.0%	6.2%
13B ($\tau = 1.0$)	0.933	0.900	0.946	0.988	2.20	0.05	33.4%	8.7%
13B ($\tau = 0.7$)	0.955	0.924	0.889	0.979	2.13	0.10	**38.0%**	14.4%
70B ($\tau = 1.0$)	0.965	0.863	0.968	0.983	1.72	0.55	**35.4%**	10.0%
70B ($\tau = 0.7$)	**0.996**	**0.954**	0.858	0.989	0.81	0.44	**49.8%**	10.6%

† Fraction of structures that are first predicted by M3GNet to have $E_{\text{hull}}^{\text{M3GNet}} < 0.1$ eV/atom, and then verified with DFT to have $E_{\text{hull}}^{\text{DFT}} < 0.0$ eV/atom.

Performance of CrystalLLM

- Fine-tuned LLaMA-2 outperforms CDVAE in generating novel and diverse samples as well as their overall speed.



Take-Away Messages

- LLMs can understand natural language descriptions of a crystal cell (i.e., side length, angle, and 3D coordinates) to for generating stable inorganic materials.
- By using a pre-trained LLM and simple fine-tuning, the approach avoids the need for crystal-specific tokenization or massive auxiliary datasets.
  - Can this idea be extended to other types of atomic structures like proteins or small molecules?
  - Can this idea be extended to non-periodic structures?

https://github.com/facebookresearch/crystal-text-llm

README | Code of conduct | License | Security

Fine-Tuned Language Models Generate Stable Inorganic Materials as Text

This repository contains the code for the paper *Fine-Tuned Language Models Generate Stable Inorganic Materials as Text* by Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C. Lawrence Zitnick, and Zachary Ward Ulissi (ICLR 2024).

Thank You!

Course Website: https://yuzhang-teaching.github.io/CSCE689-S25.html