



THRUST 1

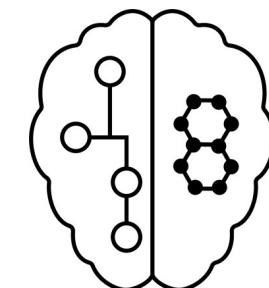
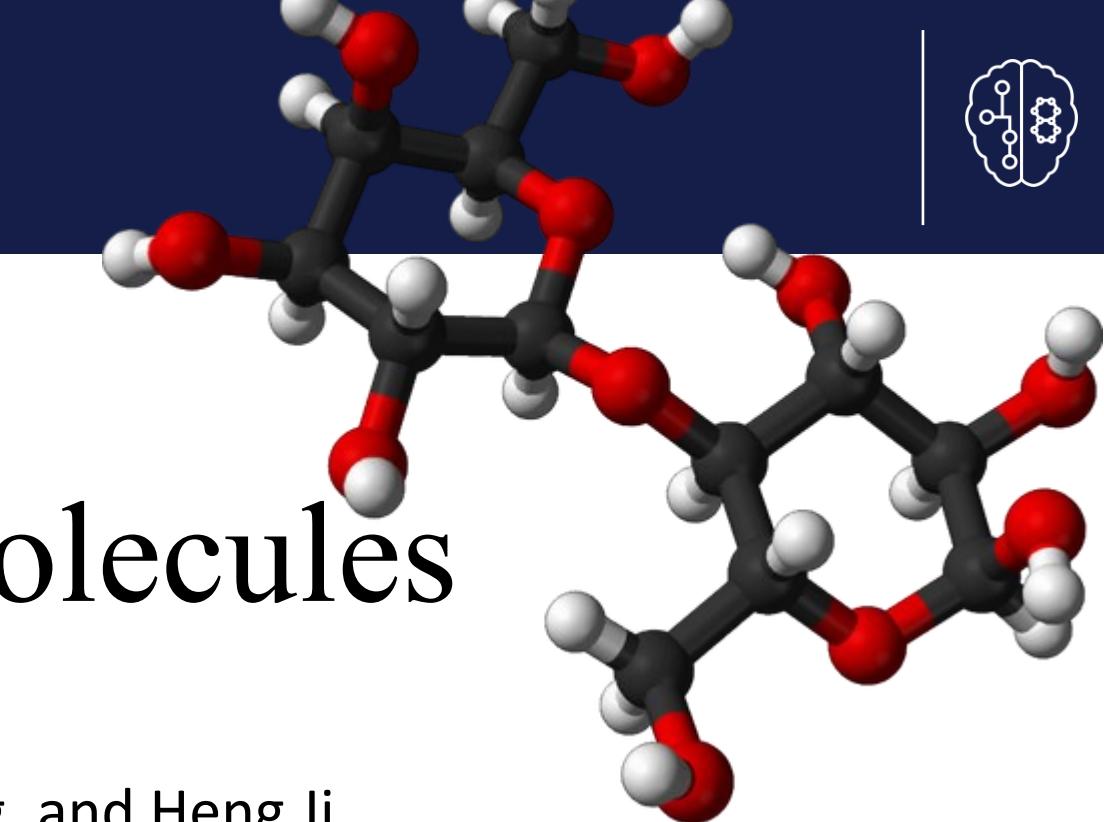
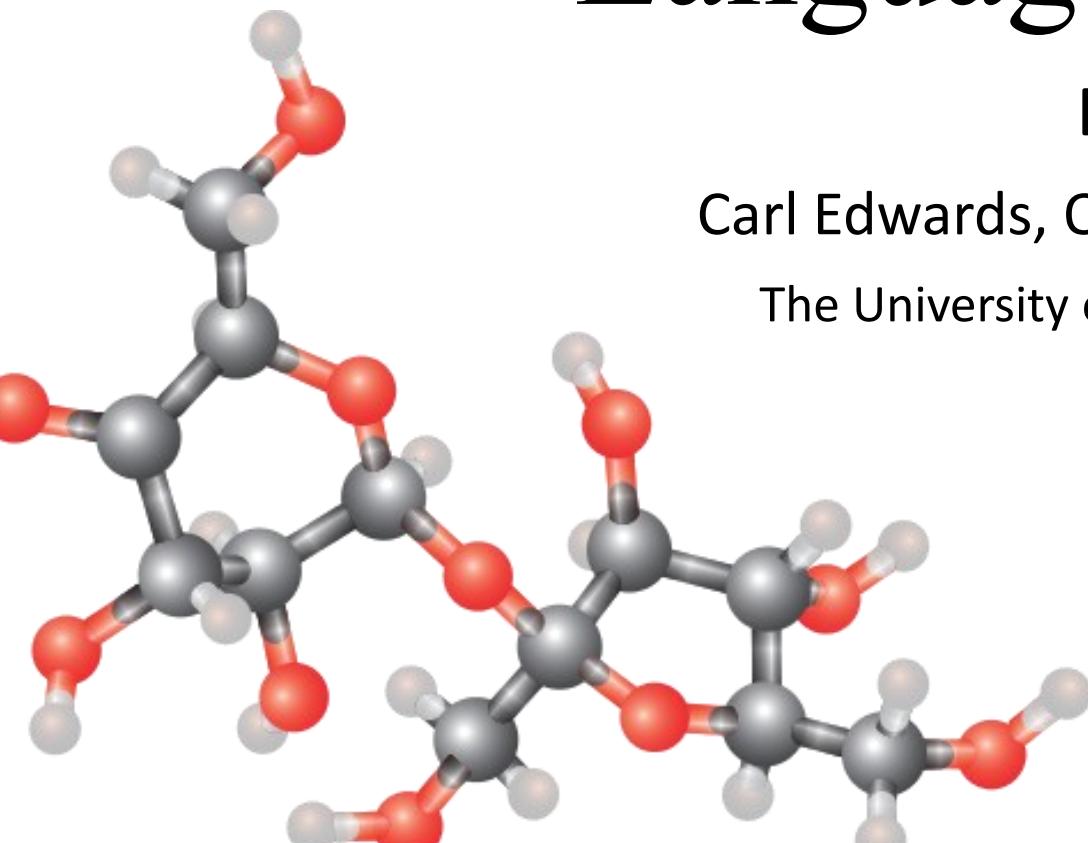


Language + Molecules

EACL 2024

Carl Edwards, Qingyun Wang, and Heng Ji

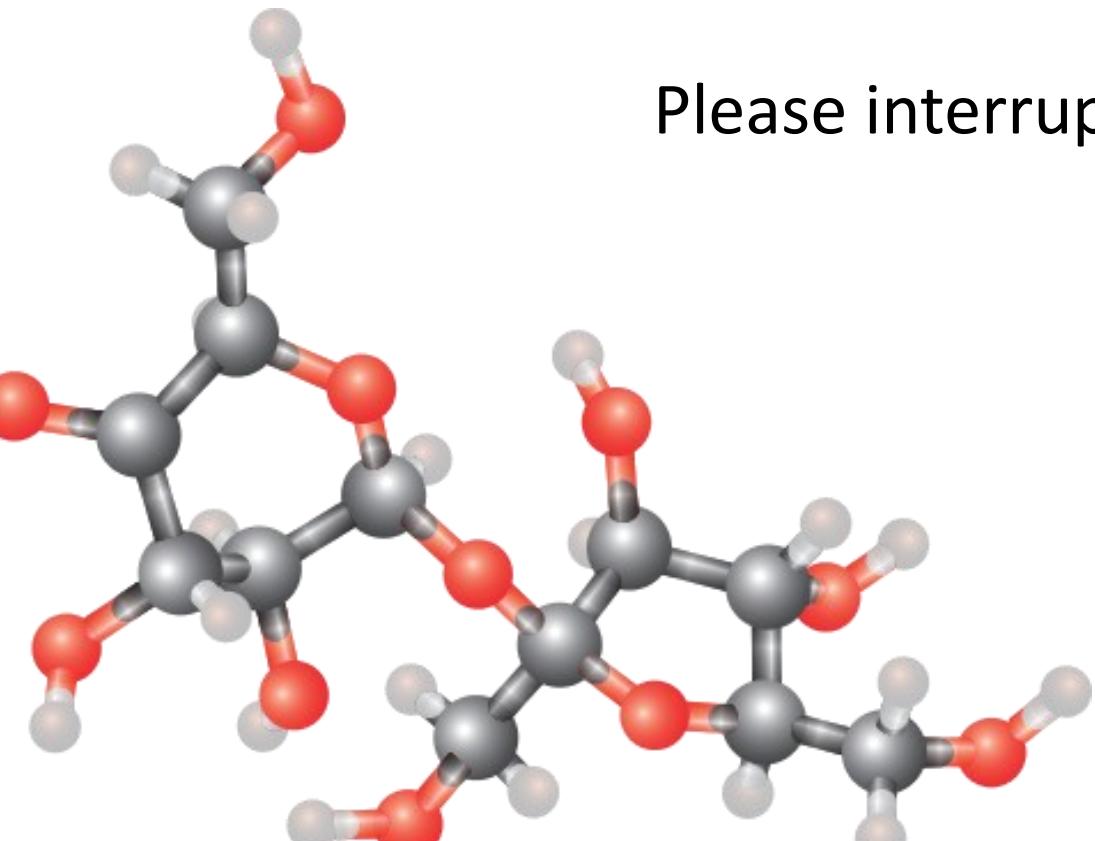
The University of Illinois Urbana-Champaign



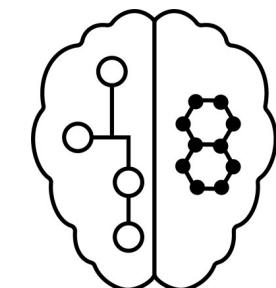
MOLECULE
MAKER LAB
INSTITUTE



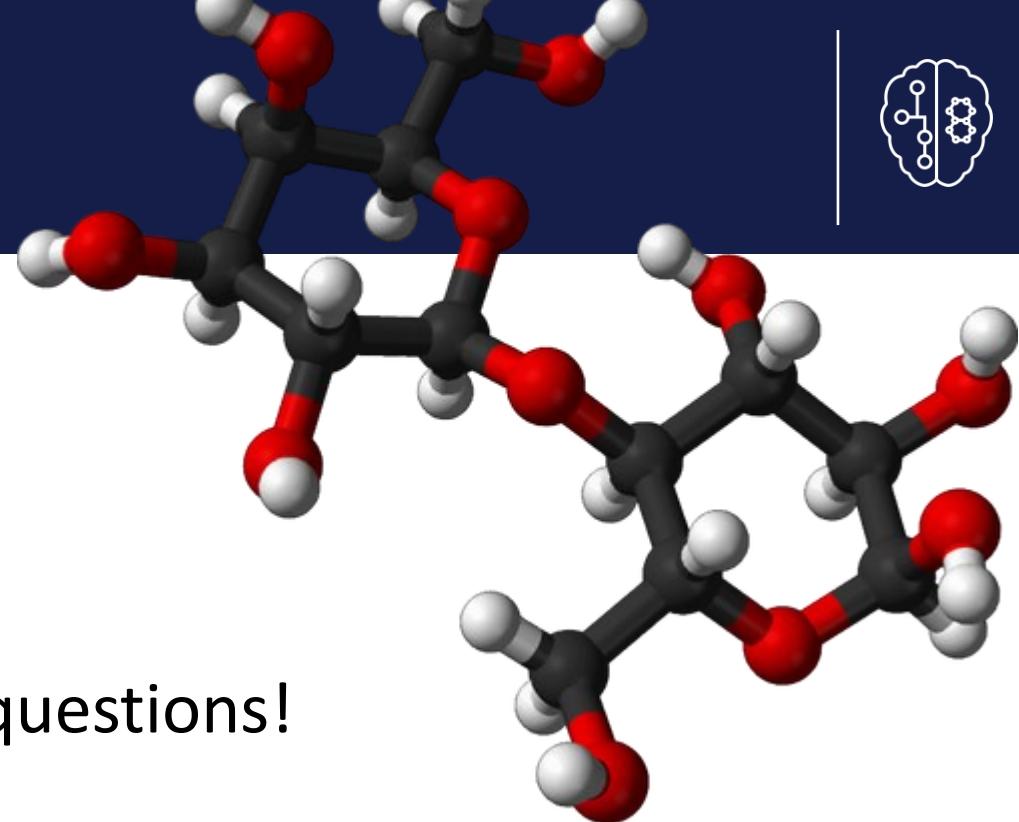
THRUST 1



Please interrupt if you have questions!



MOLECULE
MAKER LAB
INSTITUTE





THRUST 1

Outline



Background
and
Motivation

- How have we been using NLP to accelerate science?
- What are molecules and why do we care?
- How do we discover new drugs and medicines?

Integrating Language
with Molecules

- What does natural language have to offer?
- How do we actually integrate modalities?
- A real world example

Applications and Recent
Trends

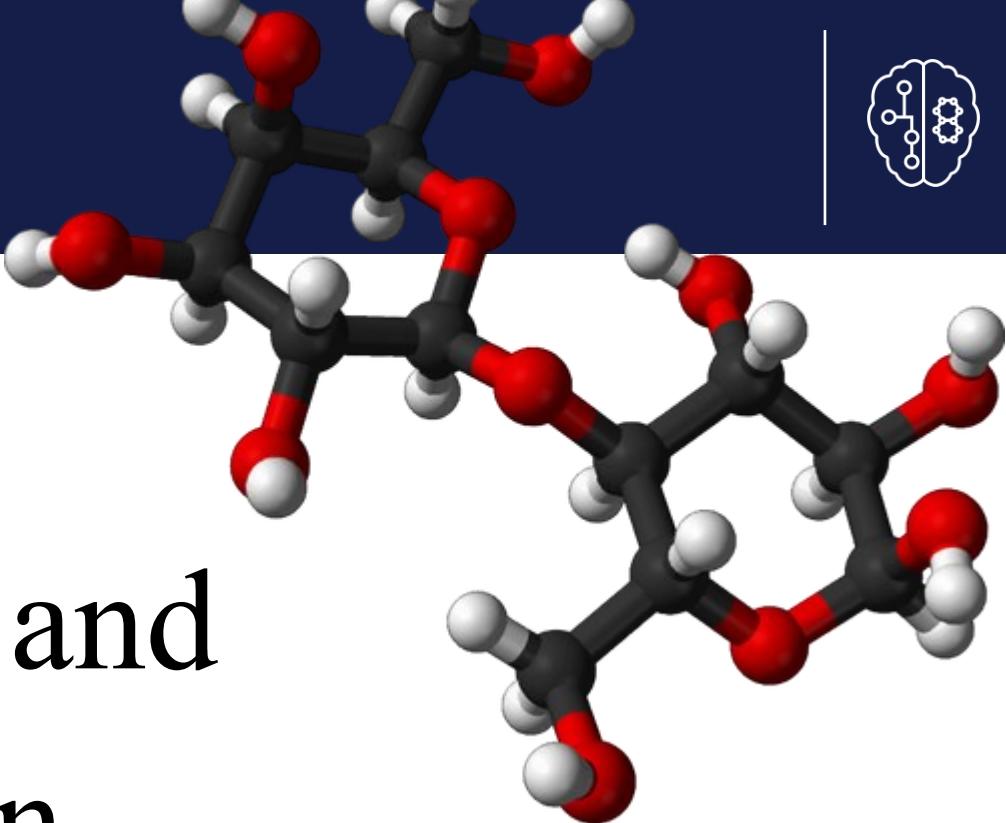
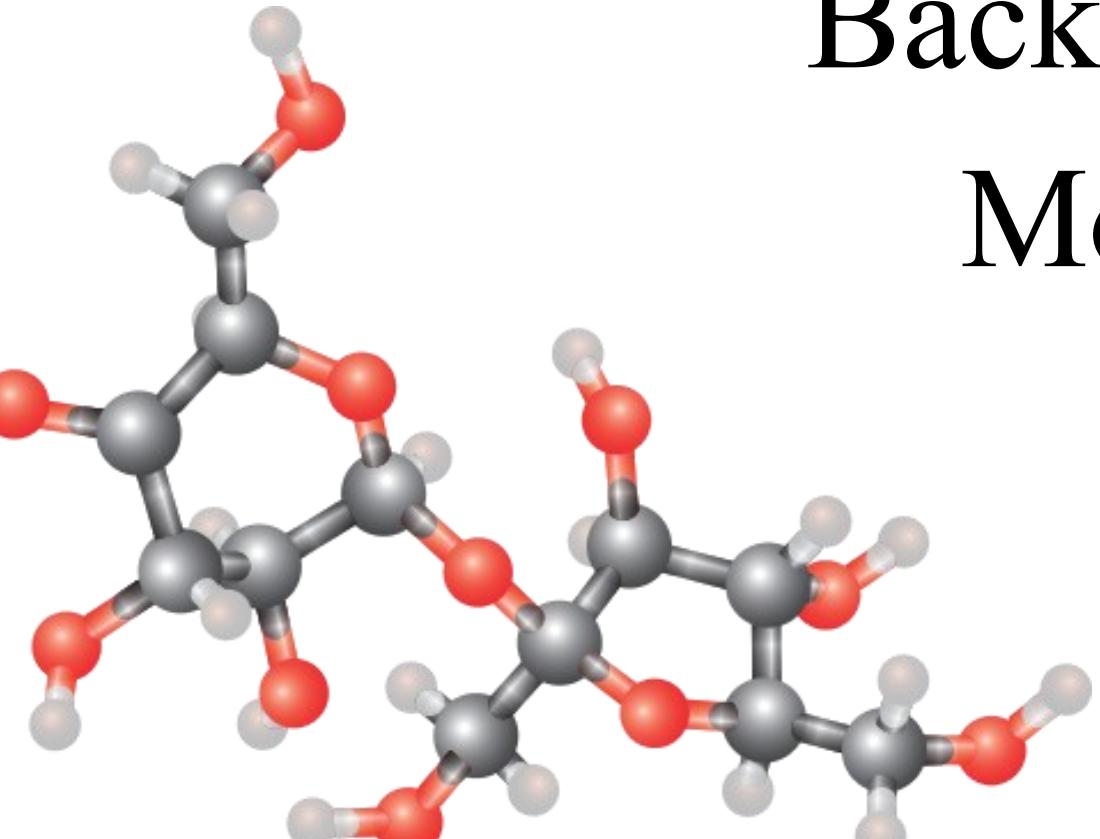
- What fields of chemistry and medicine can benefit from our systems?
- Where is the field going now?



THRUST 1



Background and Motivation



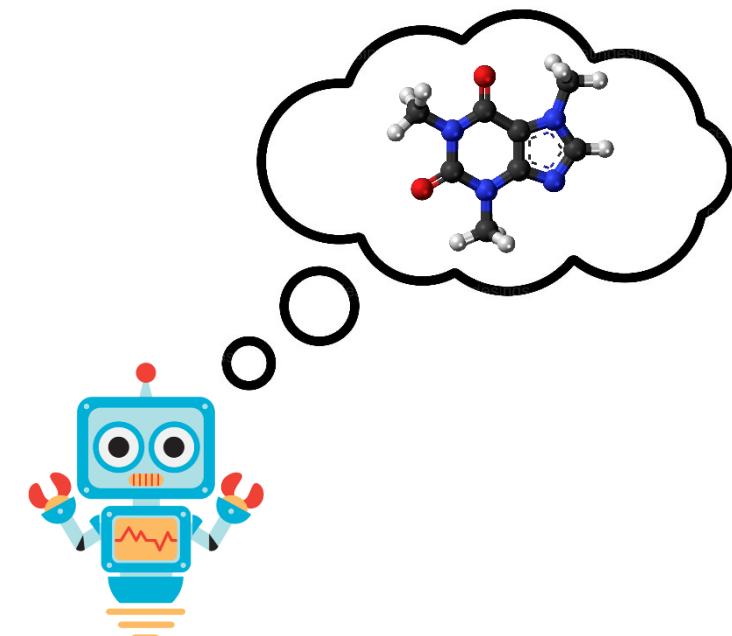


THRUST 1

How can we use natural language for chemistry?



1. Extracting and aggregating information across scientific documents
2. Enabling a new method of control for molecule design
 - a) Here, language model “hallucination” is actually a strength!



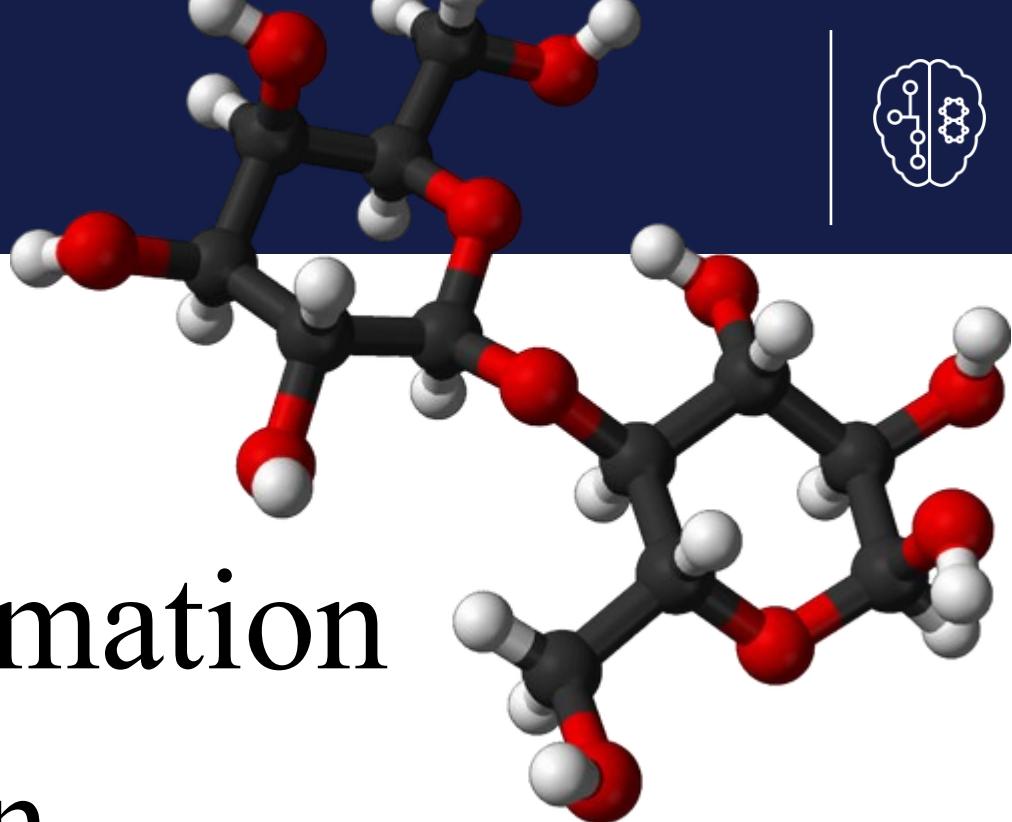
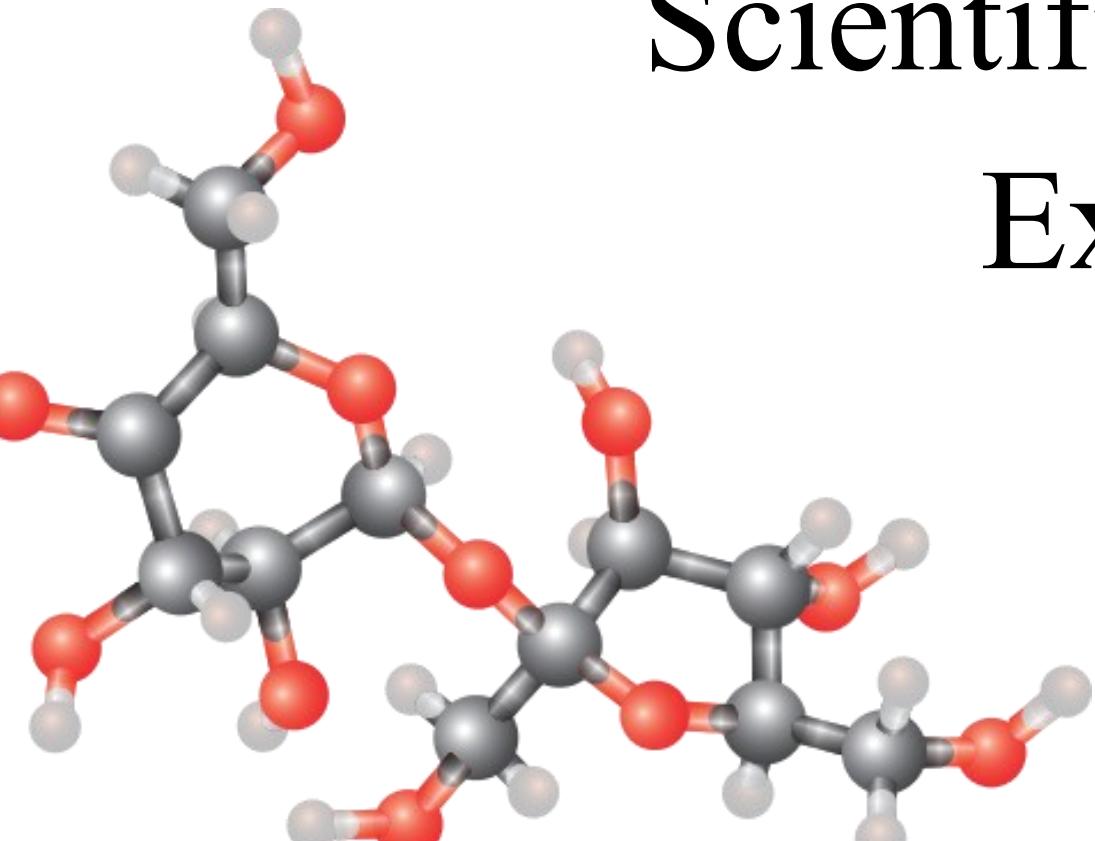


THRUST 1

Background and Motivation



Scientific Information Extraction



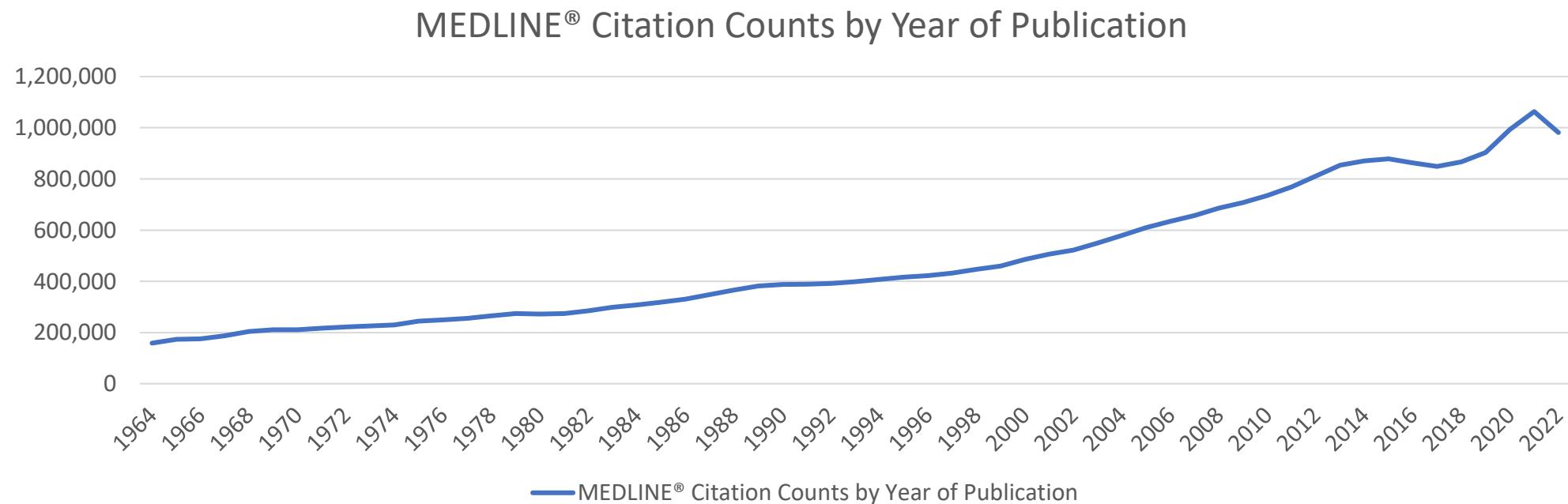


THRUST 1

INFORMATION OVERLOAD



- We are drowning in a deluge of messy, inconsistent, and badly formatted scientific information
- Human scientists cannot keep up without help



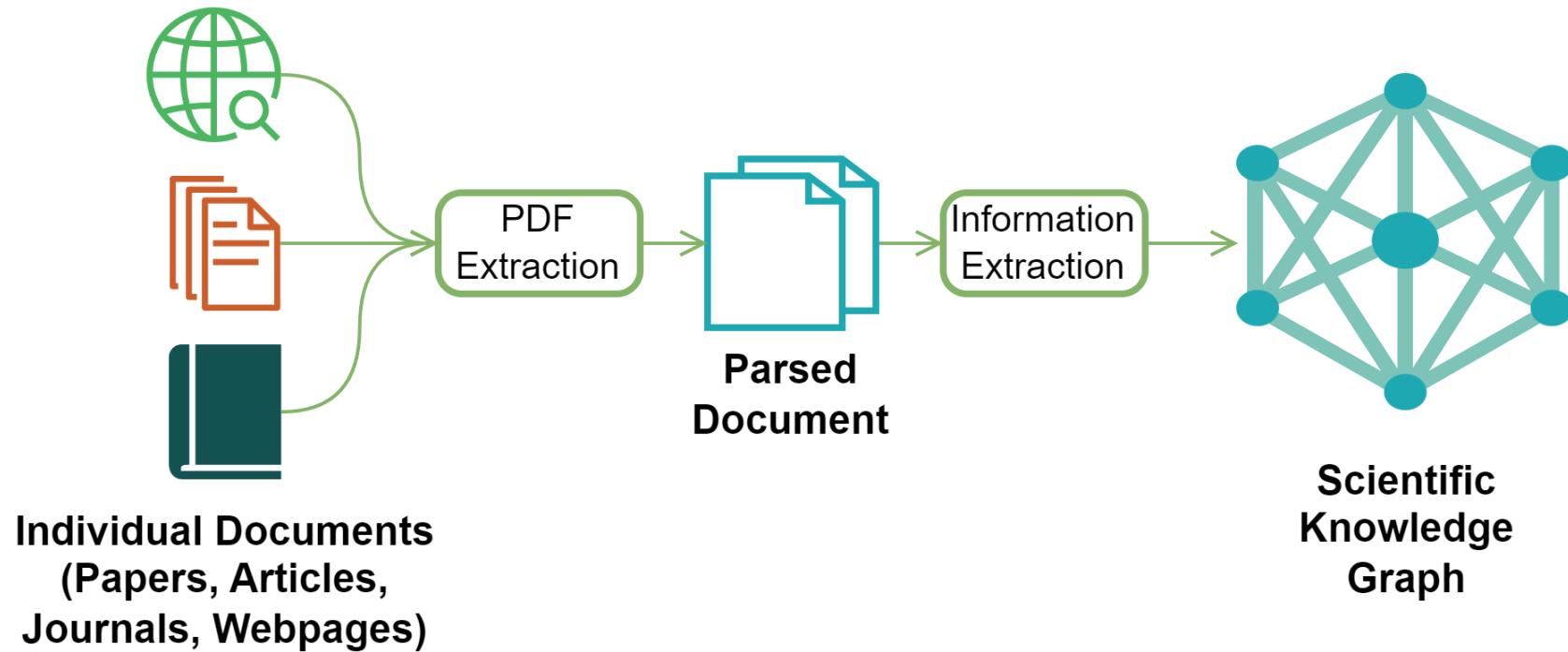
[1] https://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html

[2] Landhuis, E. (2016). Scientific literature: Information overload. *Nature*, 535(7612), 457-458.



THRUST 1

SCIENTIFIC INFORMATION EXTRACTION PIPELINE





THRUST 1

PDF EXTRACTION



- Extract structured data from a paper PDF

Full-Resolution Residual Networks for Semantic Segmentation in Street Scenes

Tobias Pohlen Alexander Hermans Markus Mathias Bastian Leibe
Visual Computing Institute
RWTH Aachen University, Germany
tobias.pohlen@rwth-aachen.de {hermans, mathias, leibe}@vision.rwth-aachen.de

Abstract

Semantic image segmentation is an essential component of modern autonomous driving systems, as an accurate understanding of the surrounding scene is crucial to navigation and action planning. Current state-of-the-art approaches in semantic image segmentation rely on pre-trained networks that were initially developed for classifying images as a whole. While these networks exhibit outstanding recognition performance (i.e., what is visible?), they lack localization accuracy (i.e., where precisely is something located?). Therefore, additional processing steps have to be performed in order to obtain pixel-accurate segmentation masks at the full image resolution. To alleviate this problem we propose a novel ResNet-like architecture that exhibits strong localization and recognition performance. We combine multi-scale context with pixel-level accuracy by using two processing streams within our network: One stream carries information at the full image resolution, enabling precise adherence to segment boundaries. The other stream undergoes a sequence of pooling operations to obtain robust features for recognition. The two streams are coupled at the full image resolution using residuals. Without additional processing steps and without pre-training, our approach achieves an intersection-over-union score of 71.8% on the Cityscapes dataset.

1. Introduction

Recent years have seen an increasing interest in self driving cars and in driver assistance systems. A crucial aspect of autonomous driving is to acquire a comprehensive understanding of the surroundings in which a car is moving. Semantic image segmentation [49, 38, 21, 53, 33], the task of assigning a set of predefined class labels to image pixels, is an important tool for modeling the complex relationships of the semantic entities usually found in street scenes, such as cars, pedestrians, road, or sidewalks. In automotive scenarios it is used in various ways, e.g. as a pre-processing step to discard image regions that are unlikely to contain objects of

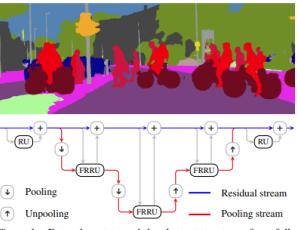


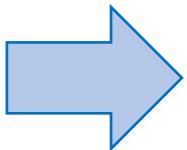
Figure 1. Example output and the abstract structure of the full-resolution residual network. The network has two processing streams. The residual stream (blue) stays at the full image resolution. The pooling stream (red) undergoes a sequence of pooling and unpooling operations. The two processing streams are coupled using full-resolution residual units (FRRUs).

interest [42, 15], to improve object detection [4, 23, 24, 58], or in combination with 3D scene geometry [32, 17, 35]. Many of those applications require precise region boundaries [20]. In this work, we therefore pursue the goal of achieving high-quality semantic segmentation with precise boundary adherence.

Current state-of-the-art approaches for image segmentation all employ some form of *fully convolutional network* (FCN) [38] that takes the image as input and outputs a probability map for each class. Many papers rely on network architectures that have already been proven successful for image classification such as variants of the ResNet [25] or the VGG architecture [30]. Starting from pre-trained nets, where a large number of weights for the target task can be pre-set by an auxiliary classification task, reduces training time and often yields superior performance compared to training a network from scratch using the (possibly limited amount of) data of the target application. However, a main limitation of using such pre-trained networks is that they

4151

Input / Paper PDF



{

“title”: “Full-Resolution Residual Networks for Semantic Segmentation in Street Scenes”,
“authors”: “Tobias Pohlen, Alexander Hermans, Markus Mathias, Bastian Leibe”,
“sections”: [...]

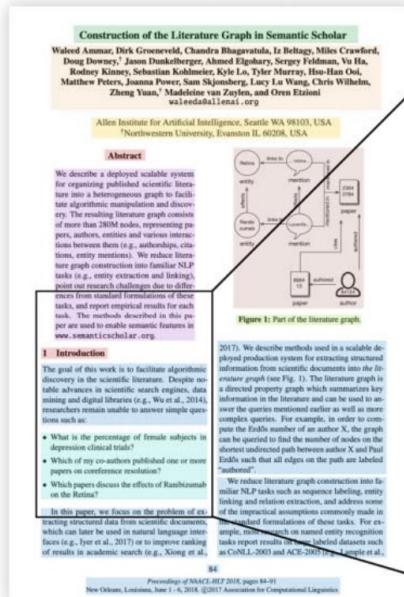
}

Output / Paper Metadata JSON



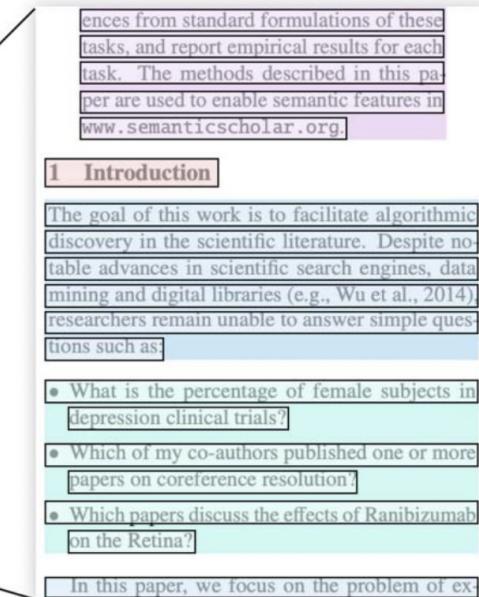
THRUST 1

PDF EXTRACTION



(a) Paper PDF Screenshot

The text blocks are highlighted in colored boxes.



(b) Examples of Visual Layout Groups

Tokens in the same text lines or blocks usually have the same semantic category.

- Previous PDF extractors ignore layout groups and require pretraining.
- A document page can be segmented into visual groups of tokens.
- Tokens within each group generally have the same semantic category.

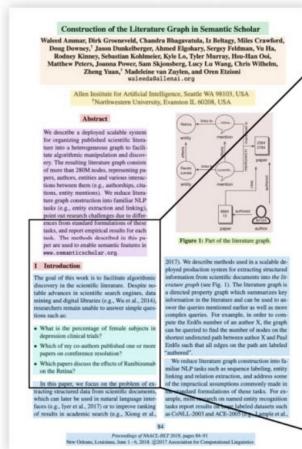


THRUST 1

PDF EXTRACTION

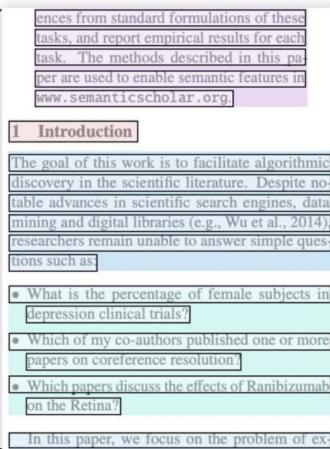


- Model visual layout groups to improve PDF extraction
 - Inject layout indicator tokens to input texts



(a) Paper PDF Screenshot

The text blocks are highlighted in colored boxes.



(b) Examples of Visual Layout Groups

Tokens in the same text lines or blocks usually have the same semantic category.

described in this paper are used to enable semantic features in www.semanticscholar.org. [BLK] 1

Introduction [BLK] The goal of this work is to facilitate algorithmic discovery in the scientific literature. Despite notable advances in scientific search engines, data mining and digital libraries (e.g., Wu et al., 2014), researchers remain unable to answer simple questions such as:

- What is the percentage of female subjects in depression clinical trials?
- Which of my co-authors published one or more papers on coreference resolution?
- Which papers discuss the effects of Ranibizumab on the Retina?

In this paper, we focus on the problem of ex-

methods described in this paper are used to enable semantic features in www.semanticscholar.org. [BLK] 1

Introduction [BLK] The goal of this work is to facilitate algorithmic discovery in the scientific literature. Despite notable advances in scientific search engines, data mining and digital libraries (e.g., Wu et al., 2014), researchers remain unable to answer simple questions such as: [BLK] What is the percentage of female subjects in depression clinical trials? Which of my

False Positive [BLK] Despite notable advances in scientific search engines, data mining and digital libraries (e.g., Wu et al., 2014), researchers remain unable to answer simple questions such as: What is the percentage of female subjects in depression clinical trials? [BLK]

False Negative [BLK] Despite notable advances in scientific search engines, data mining and digital libraries (e.g., Wu et al., 2014), researchers remain unable to answer simple questions such as: What is the percentage of female subjects in depression clinical trials? [BLK]



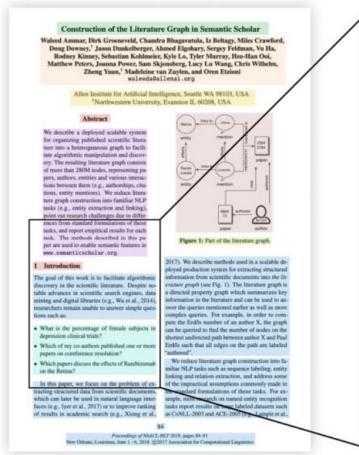
THRUST 1

PDF EXTRACTION



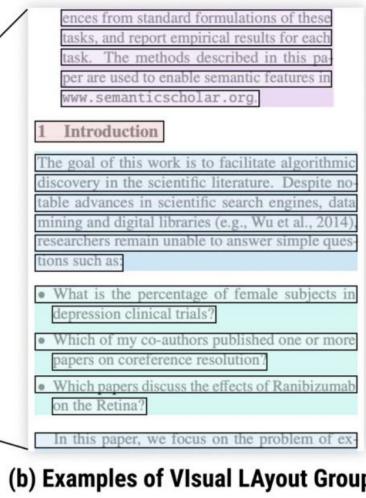
- Model visual layout groups to improve PDF extraction
 - Inject layout indicator tokens to input texts
 - Hierarchical encode layout groups

✓ Doesn't require pretraining



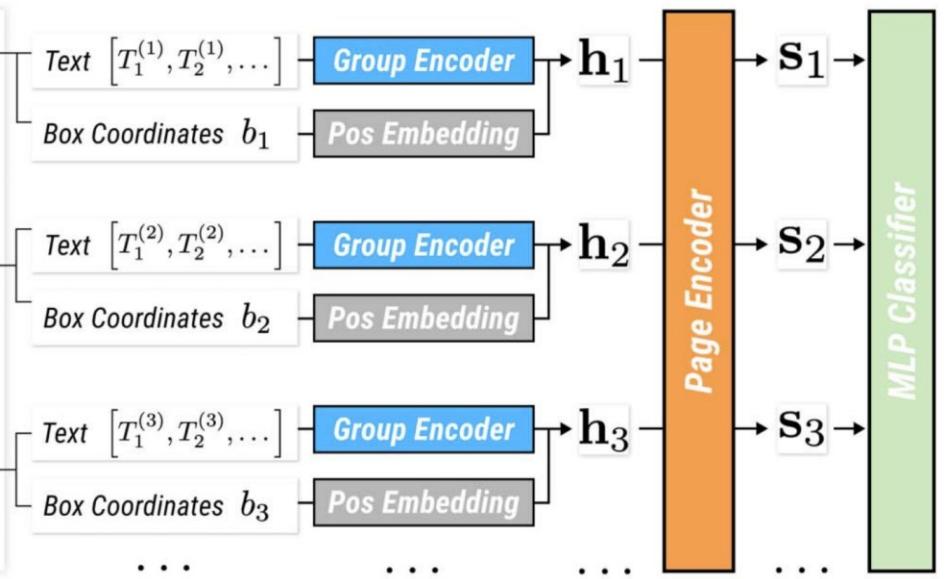
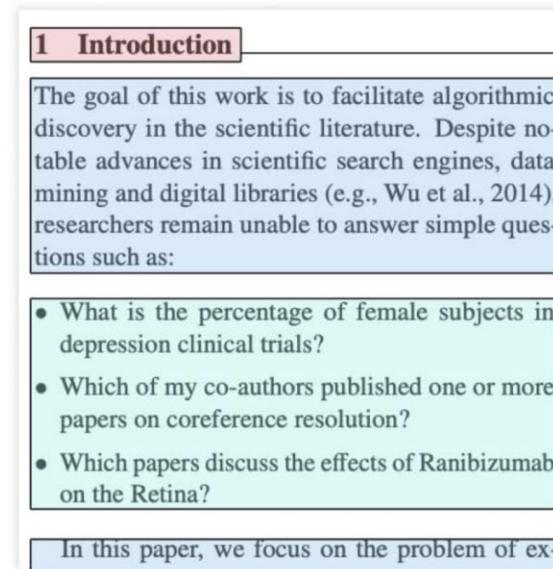
(a) Paper PDF Screenshot

The text blocks are highlighted in colored boxes.



(b) Examples of Visual Layout Groups

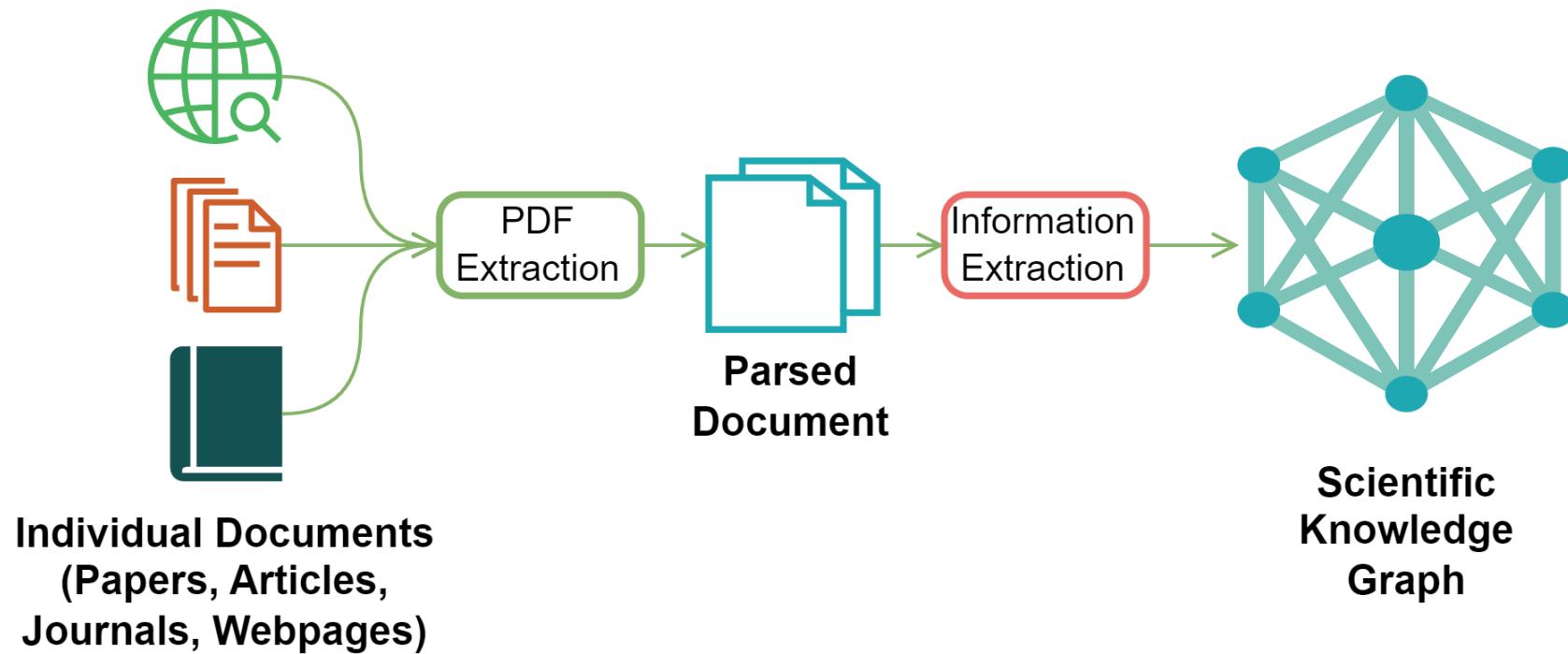
Tokens in the same text lines or blocks usually have the same semantic category.





THRUST 1

SCIENTIFIC INFORMATION EXTRACTION PIPELINE





- Transform unstructured or semi-structured scientific texts into a structured form that can be more easily understood
 - Identifying and structuring
 - Entities
 - Quantities (and their scope)
 - Events
 - Relations, etc.

BaCl₂ 2H₂O and FeCl₃ 6H₂O are used as initial materials. In a typical experiment, 0.487 g of FeCl₃ 6H₂O and 0.054 g of BaCl₂ 2H₂O (Fe/Ba mole ratio: N=8) were dissolved in 21 mL of distilled water. Then, 4.0 g of KOH was added to the solution with continuous stirring. Finally, the solution was put into a 30 mL of Teflon-lined stainless-steel autoclave and kept it in an oven at 220 °C for 24 h, followed by furnace cooling to room temperature.



- Why do we need to extract information?
 - Handling Information Overload
 - Facilitating Knowledge Discovery
 - Improving Accessibility of Scientific Knowledge
 - Updating Knowledge Base

BaCl₂ 2H₂O and FeCl₃ 6H₂O are used as initial materials. In a typical experiment, 0.487 g of FeCl₃ 6H₂O and 0.054 g of BaCl₂ 2H₂O (Fe/Ba mole ratio: N=8) were dissolved in 21 mL of distilled water. Then, 4.0 g of KOH was added to the solution with continuous stirring. Finally, the solution was put into a 30 mL of Teflon-lined stainless-steel autoclave and kept it in an oven at 220 °C for 24 h, followed by furnace cooling to room temperature.

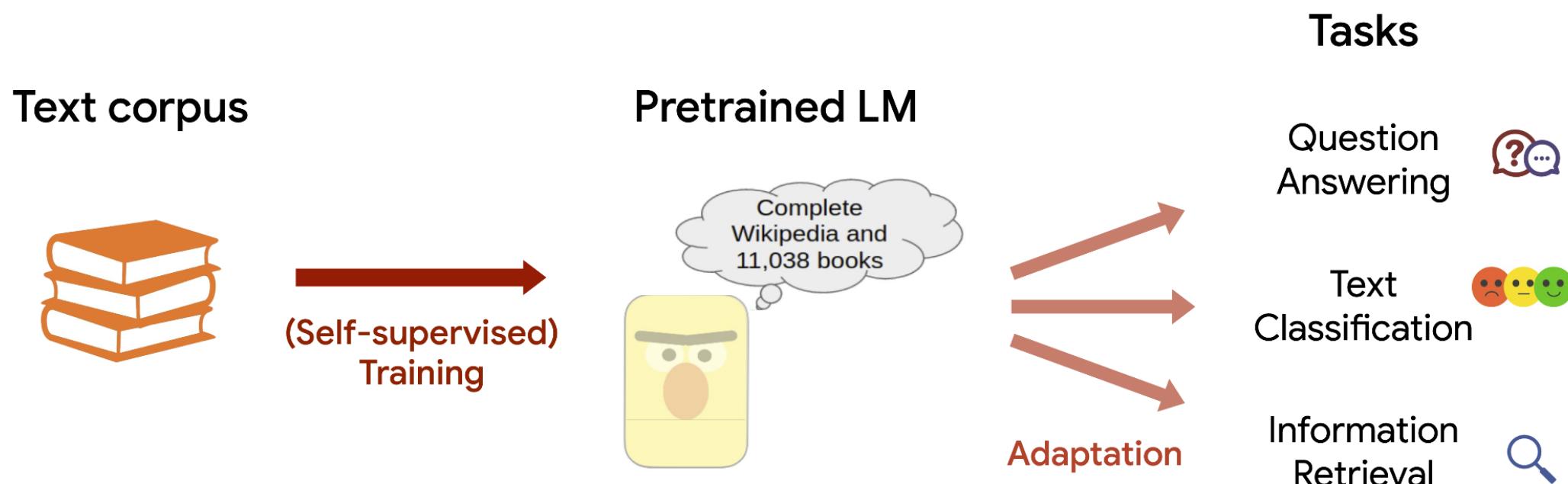


THRUST 1

SCIENTIFIC LARGE LANGUAGE MODEL



- Scientific information extraction benefits from pretrained large language models





THRUST 1

KNOWLEDGE-ENHANCED SCIENTIFIC LLM



- What knowledge is useful for scientific large language model?
 - Domain/Task knowledge from pretraining corpus
 - Cross document knowledge such as the links between different papers



THRUST 1

KNOWLEDGE-ENHANCED SCIENTIFIC LLM

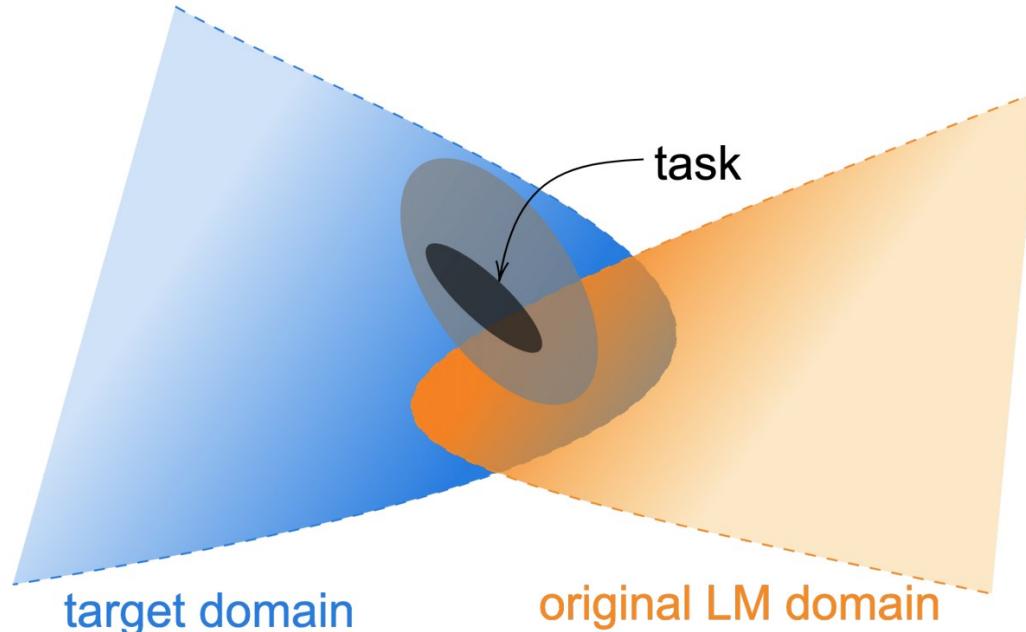


- What knowledge is useful for scientific large language model?
 - Domain/Task knowledge from pretraining corpus
 - Cross document knowledge such as the links between different papers



THRUST 1

WHY DO WE NEED DOMAIN/TASK PRETRAINING

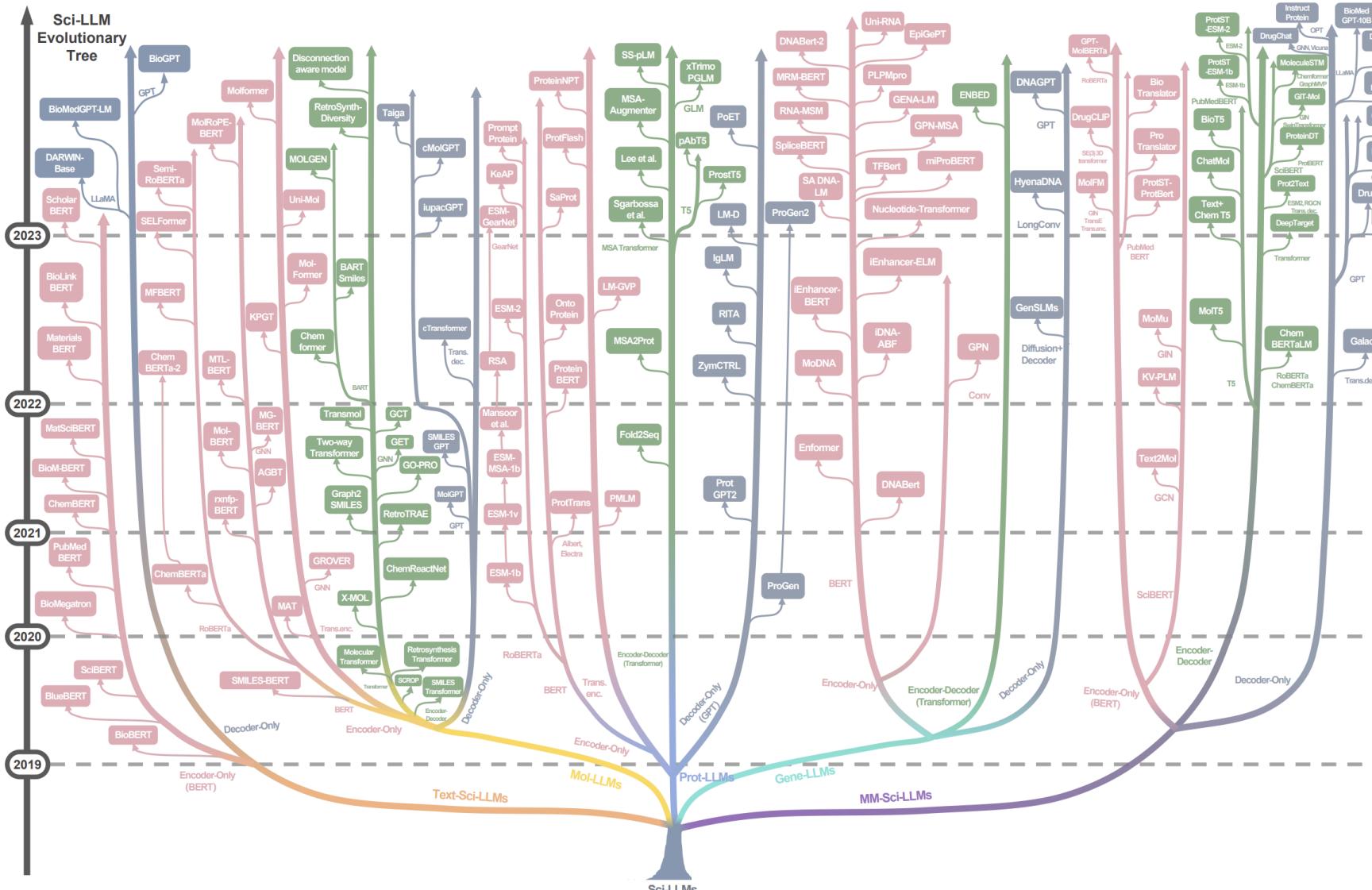


- Language models benefit from continued pretraining on the domain
- Task-adaptive pretraining on a smaller but directly task-relevant corpus can still boost performance
- Automatic or human-curated unlabeled data related to the task can further improve performance in low-resources setting



THRUST 1

HISTORY OF SCIENTIFIC LARGE LANGUAGE MODEL





THRUST 1

KNOWLEDGE-ENHANCED SCIENTIFIC LLM



- What knowledge is useful for scientific large language model?
 - Domain/Task knowledge from pretraining corpus
 - Cross document knowledge such as the links between different papers



THRUST 1

INCORPORATING DOCUMENT LINKS



Document



Linked document
(e.g. hyperlink, citation)

[Tidal Basin, Washington D.C.]

The **Tidal Basin** is a man-made reservoir located between It is part of West Potomac Park, is near the National Mall and is a focal point of **the National Cherry Blossom Festival** held each spring. The Jefferson Memorial,

[The National Cherry Blossom Festival] ... It is a spring celebration commemorating the March 27, 1912, gift of **Japanese cherry trees** from Mayor of Tokyo City Yukio Ozaki to the city of Washington, D.C. Mayor Ozaki gifted the trees to enhance ...

Multi-hop knowledge

(e.g. **Tidal Basin** has **Japanese cherry trees**)

- Previous large language model pretraining methods only consider **a single document**
- Document links can provide **salient multi-hop knowledge**
 - Useful for downstream tasks such as QA, knowledge discovery
- Types of links
 - Hyperlinks/citations
 - Easy to access in scale
 - High precision of relevance
 - Can capture relevant documents which are not obvious through lexical similarity
 - Lexical similarity

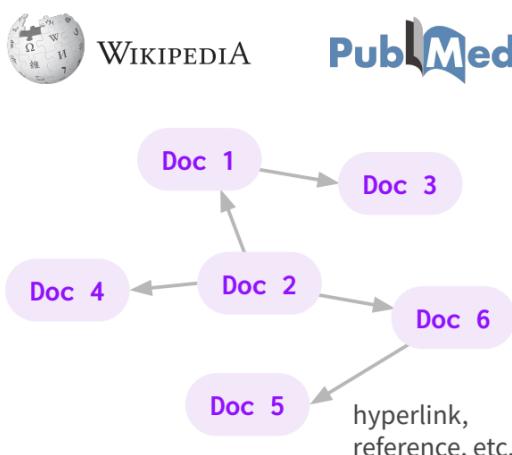


THRUST 1

INCORPORATING HYPERLINK/CITATION LINKS



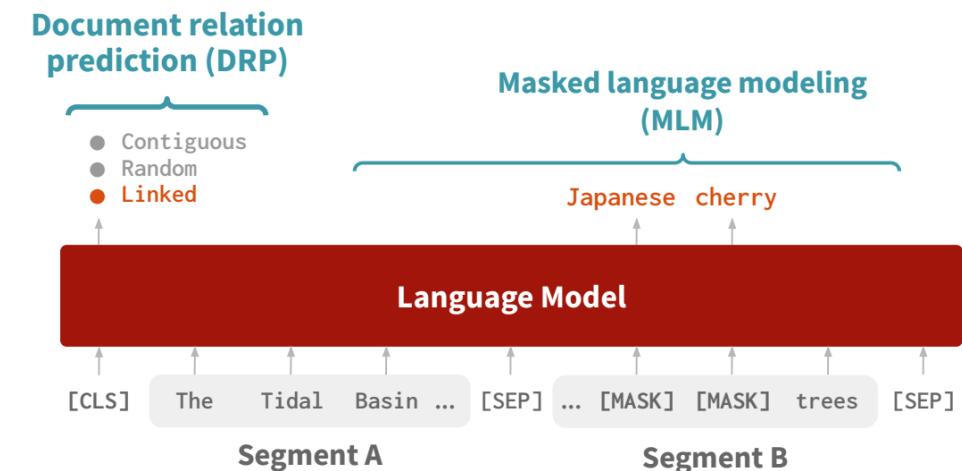
- ✓ Link related documents together to provide new knowledge
- ✗ Document links can be **broken** due to removed contents
- ✗ Hyperlinks/Citations might be **subjective**, influenced by authors' knowledge and preferences



Corpus of linked documents



Create LM inputs



Pretrain the LM



THRUST 1

INCORPORATING DOCUMENT LINKS



Document



Linked document
(e.g. hyperlink, citation)

[Tidal Basin, Washington D.C.]

The **Tidal Basin** is a man-made reservoir located between It is part of West Potomac Park, is near the National Mall and is a focal point of **the National Cherry Blossom Festival** held each spring. The Jefferson Memorial,

[The National Cherry Blossom Festival] ... It is a spring celebration commemorating the March 27, 1912, gift of **Japanese cherry trees** from Mayor of Tokyo City Yukio Ozaki to the city of Washington, D.C. Mayor Ozaki gifted the trees to enhance ...

- Previous large language model pretraining methods only consider **a single document**
- Document links can provide **salient multi-hop knowledge**
 - Useful for downstream tasks such as QA, knowledge discovery
- Types of links
 - Hyperlinks/citations
 - **Lexical similarity**

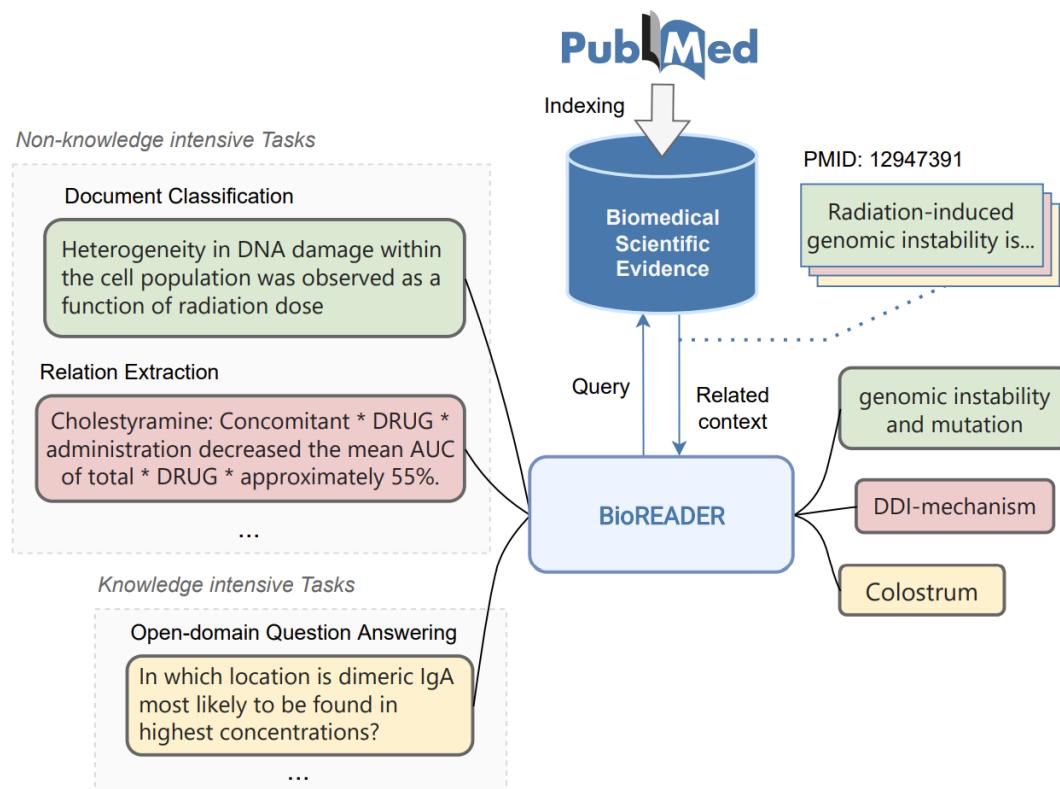
Multi-hop knowledge

(e.g. **Tidal Basin** has **Japanese cherry trees**)



THRUST 1

INCORPORATING SEMANTIC LINKS



- It's **impossible** to store all factual and domain-specific knowledge into parameters.
- Retrieval is a **complementary path** to architectural scaling.
 - Semi-parametric = closed book (parametric-only) + open-book (retrieval-based)



THRUST 1 CHALLENGES



- Few benchmarks are publicly available

Datasets

9,274 machine learning datasets

Share your dataset with the ML community!

5 dataset results for `chemical AND Named Entity Recognition (NER)`

Dataset	Description	Papers	Benchmarks
BC4CHEMD (BioCreative IV Chemical compound and drug name recognition)	Introduced by Krallinger et al. in The CHEMDNER corpus of chemicals and drugs and its annotation principles BC4CHEMD is a collection of 10,000 PubMed abstracts that contain a total of...	4 PAPERS	3 BENCHMARKS
BC5CDR (BioCreative V CDR corpus)	BC5CDR corpus consists of 1500 PubMed articles with 4409 annotated chemicals , 5818 diseases and 3116 chemical -disease interactions.	170 PAPERS	7 BENCHMARKS
BC7 NLM-Chem (BioCreative VII NLM-Chem)	Full-text chemical identification and indexing in PubMed articles. Identifying named entities is an important building block for many complex knowledge extraction tasks. Chemical entities...	2 PAPERS	3 BENCHMARKS
BioRED	BioRED is a first-of-its-kind biomedical relation extraction dataset with multiple entity types (e.g. gene/protein, disease, chemical) and relation pairs (e.g. gene-disease; chemical-chemi ...)	13 PAPERS	3 BENCHMARKS
JNLPBA	...From this search 2,000 abstracts were selected and hand annotated according to a small taxonomy of 48 classes based on a chemical classification. 36 terminal classes were used to anno...	17 PAPERS	3 BENCHMARKS



- Few benchmarks are publicly available
 - Models only have a **limited** number of training samples for each task

Entity Extraction	Relation Extraction	Event Extraction	Entity Linking
BC4CHEMD	Abroad-RE	BioNLP 09	XL-BEL
BC5CDR	BC5CDR	Genia 2013	Mantra GSC
BC7 NLM-Chem	BC6 ChemProt	Genia 2016	
BioRED	BC7 DrugProt	MatSci-NLP	
CHEMET	BioRED		
Chem-FINESE	BioRelEx		
JNLPBA	JNLPBA		
MatSci-NLP	MatSci-NLP		
PolymerAbstracts			



THRUST 1

CHALLENGES

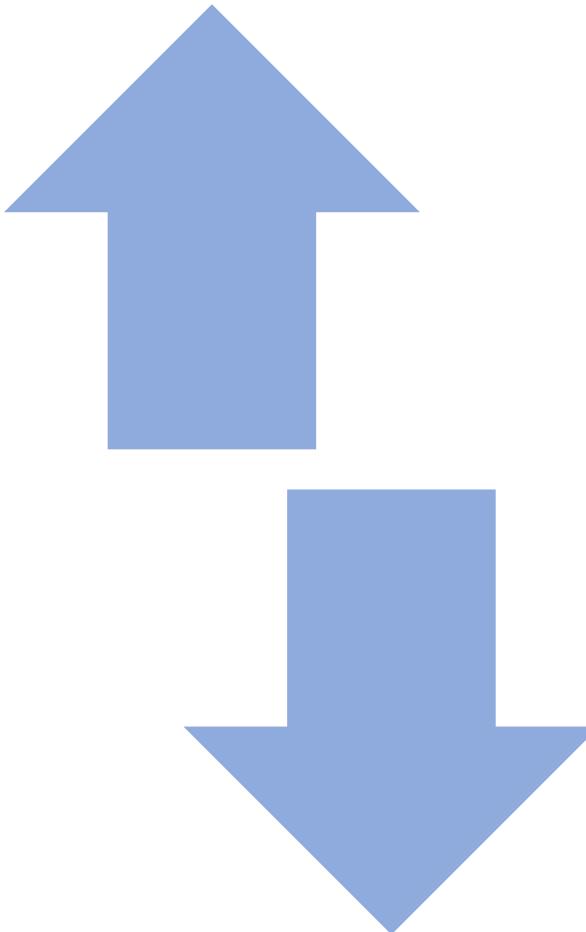


- Few benchmarks are publicly available
- Comprehending scientific papers requires **external knowledge**
- *Acetylsalicylic acid* has both anti-inflammatory and antipyretic effects.
- *Aspirin* is a Nonsteroidal Anti-inflammatory Drug and Platelet Aggregation Inhibitor.
- Both *Aspirin* and *Acetylsalicylic acid* refer to $\text{CH}_3\text{COOC}_6\text{H}_4\text{COOH}$.



THRUST 1

EXTERNAL KNOWLEDGE ENHANCED IE



How to obtain external knowledge?

How to leverage external knowledge in IE?



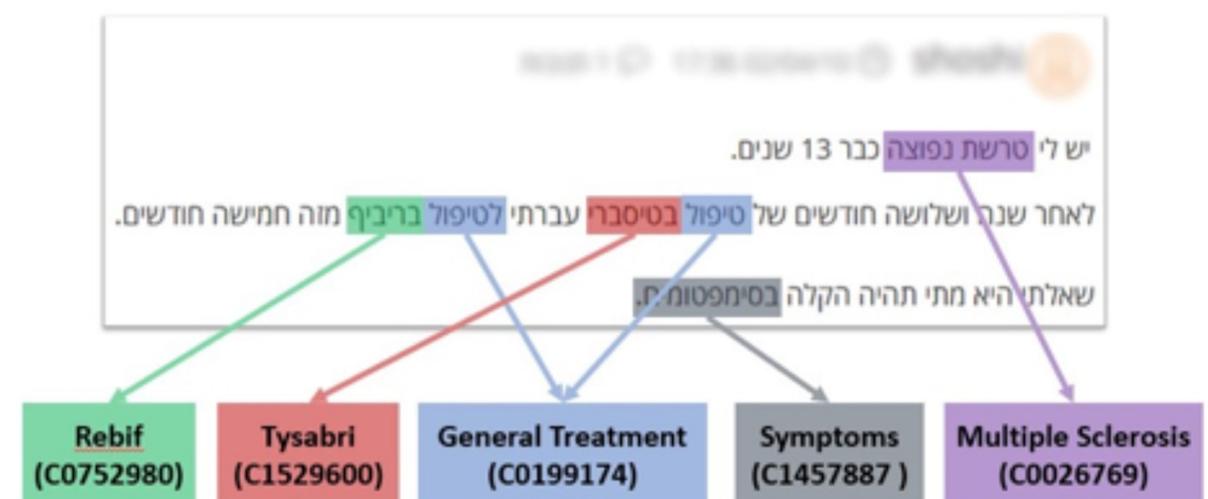


THRUST 1

SCIENTIFIC CROSS-LINGUAL ENTITY LINKING



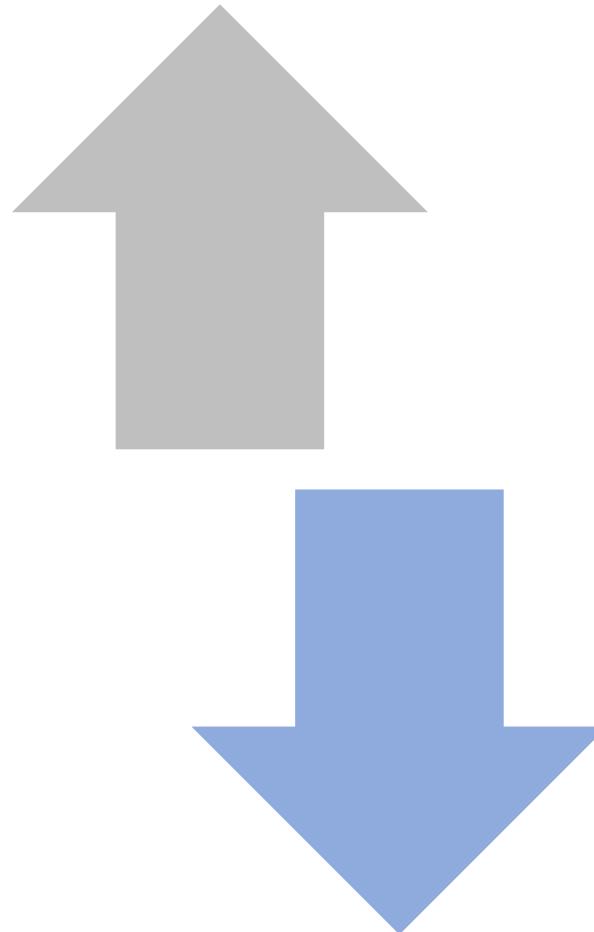
- Scientific Entity Linking [1][2]
- Goal:
 - Mapping entity mentions in text of a source language to entities in a knowledge base, (e.g., UMLS), in a target language





THRUST 1

EXTERNAL KNOWLEDGE ENHANCED IE



How to obtain external knowledge?

How to leverage external knowledge in IE?

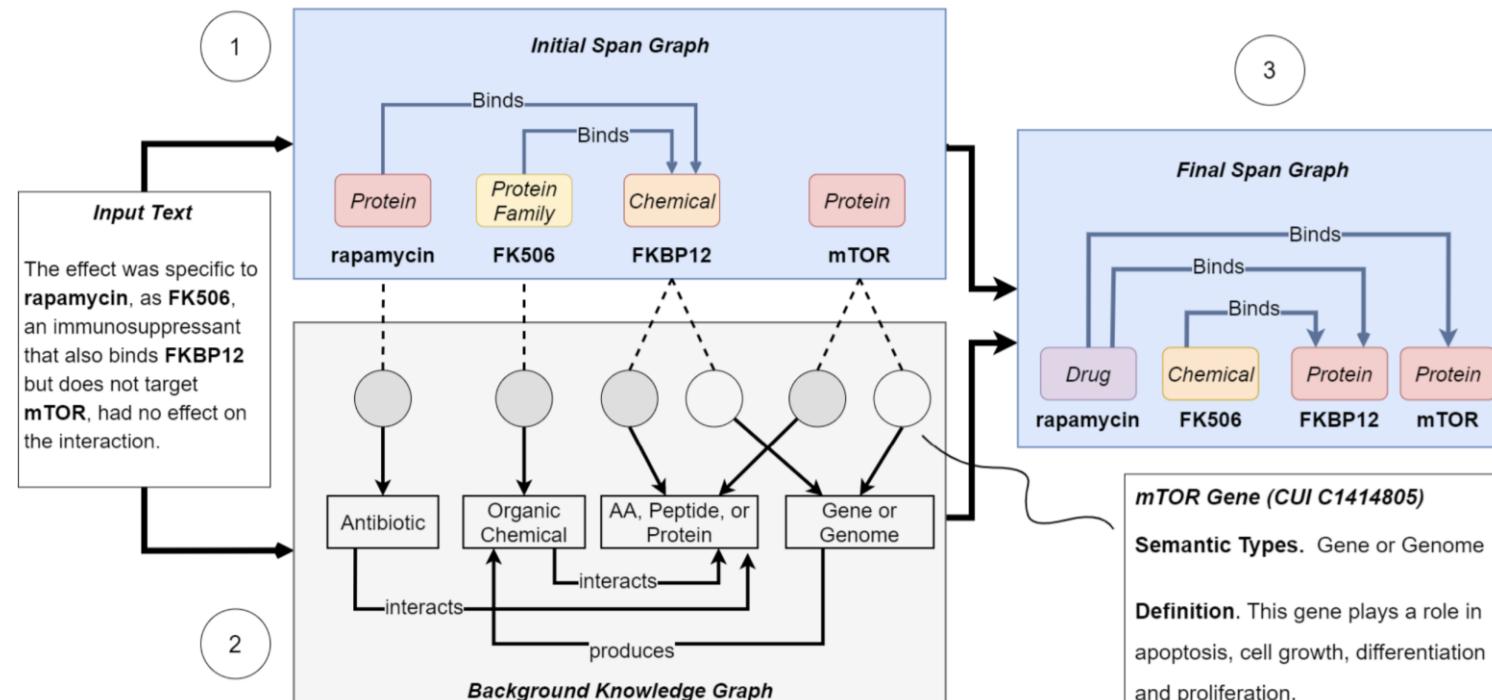


THRUST 1

EXTERNAL KNOWLEDGE ENHANCED IE



- ✓ Utilize an entity linker to transfer background knowledge from external KB to jointly extract entities and relations.
- ✗ Struggle to deal with long distance between knowledge elements.





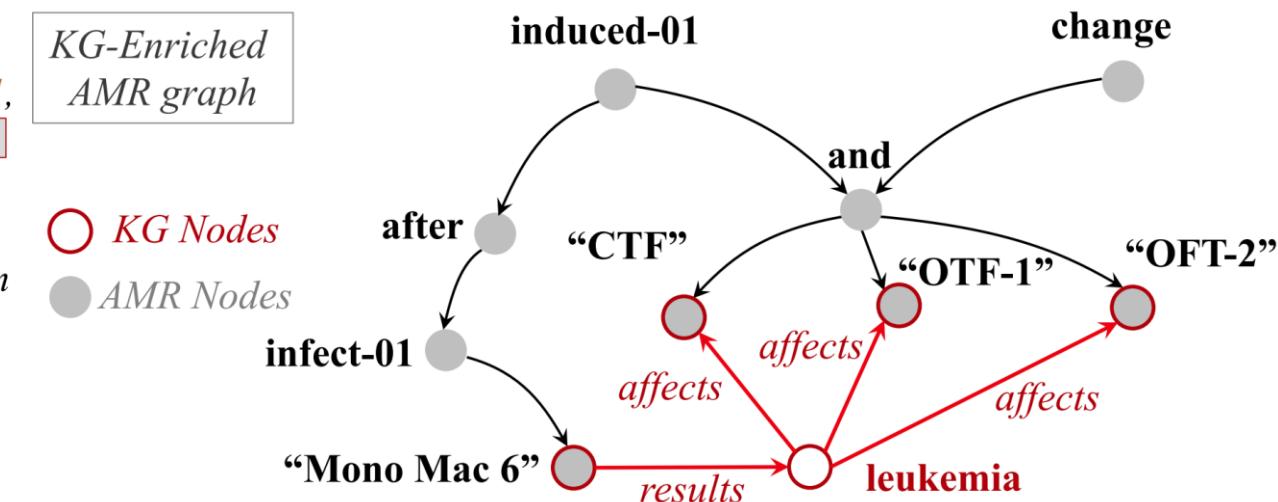
THRUST 1

AMR ENHANCED IE



- Long distance between knowledge elements. (ACE05-E 56 words vs GENIA-2011 77 words)
- Difficult to understand the scientific concepts and relations.
- Compress the context using Abstract Meaning representation (AMR) to create **rich semantic graphs**.
- Enrich the model with **external knowledge**.

We identified a cell-type-specific differential response: CREB, CTF, OTF-1, OFT-2, and NF-kappa B genes were strongly **induced** 1 to 4 hours after influenza A virus infection in the monocytic cell line Mono Mac 6, while in freshly prepared human monocytes no significant **changes** were detected.



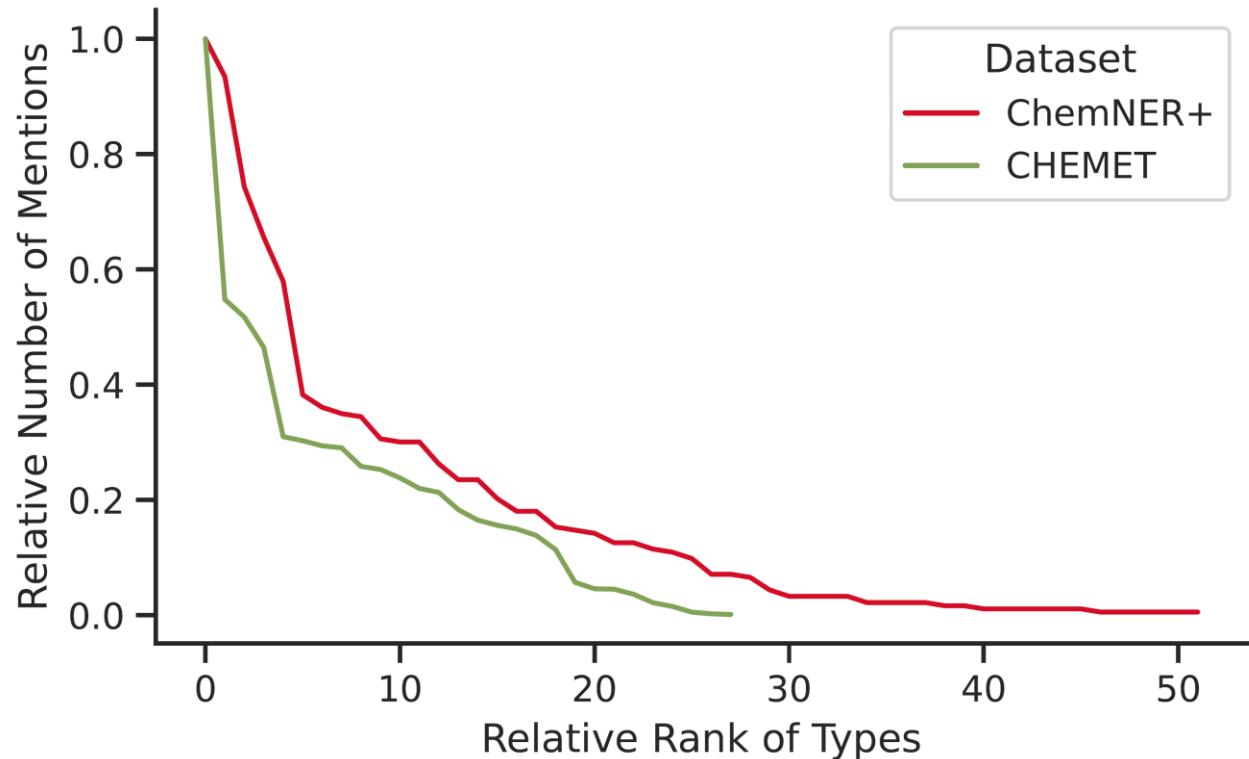


THRUST 1

CHALLENGES



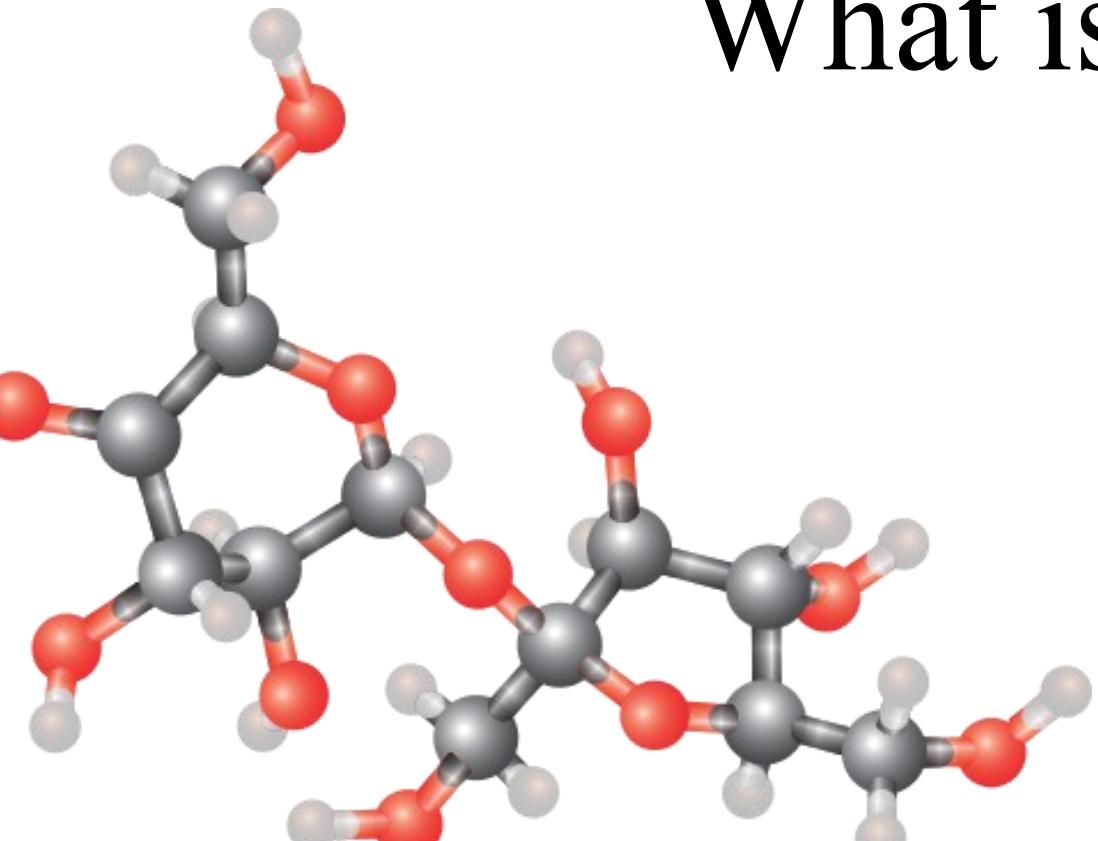
- Few benchmarks are publicly available
- Comprehending scientific papers requires **external knowledge**
- Scientific documents contain **more entities** per sentence compared the sentence in general domain
- Long-tail problems are **more prevalent** in scientific domain compared to general domain



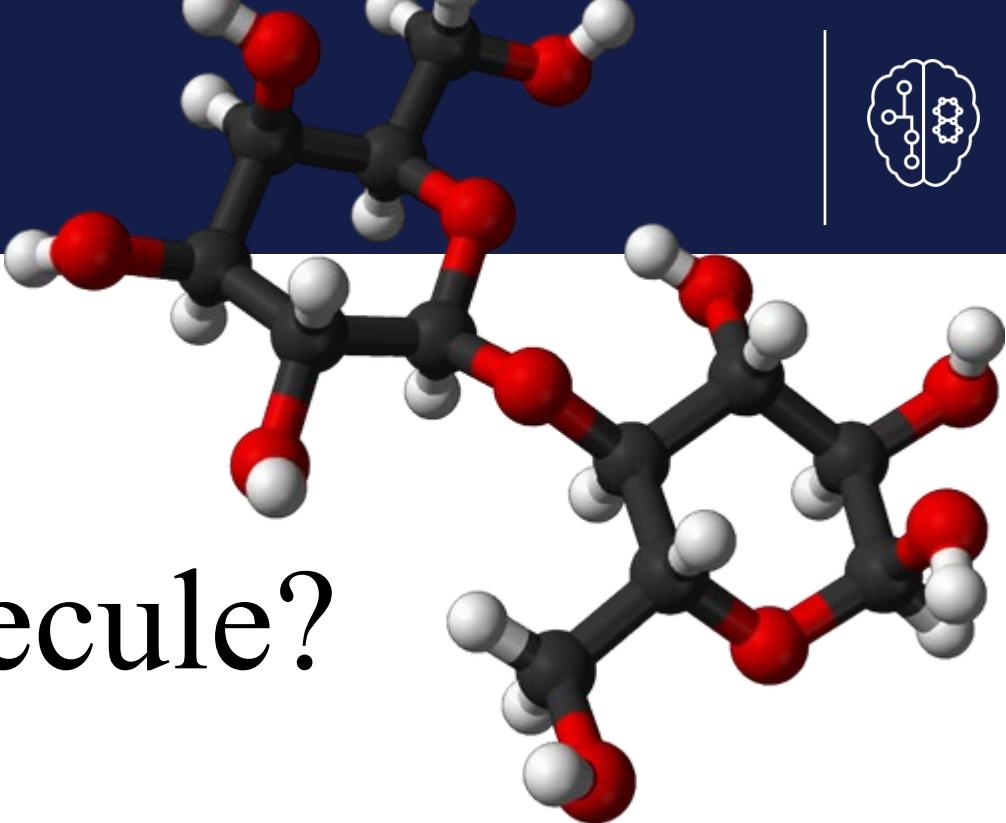


THRUST 1

Background and Motivation



What is a Molecule?

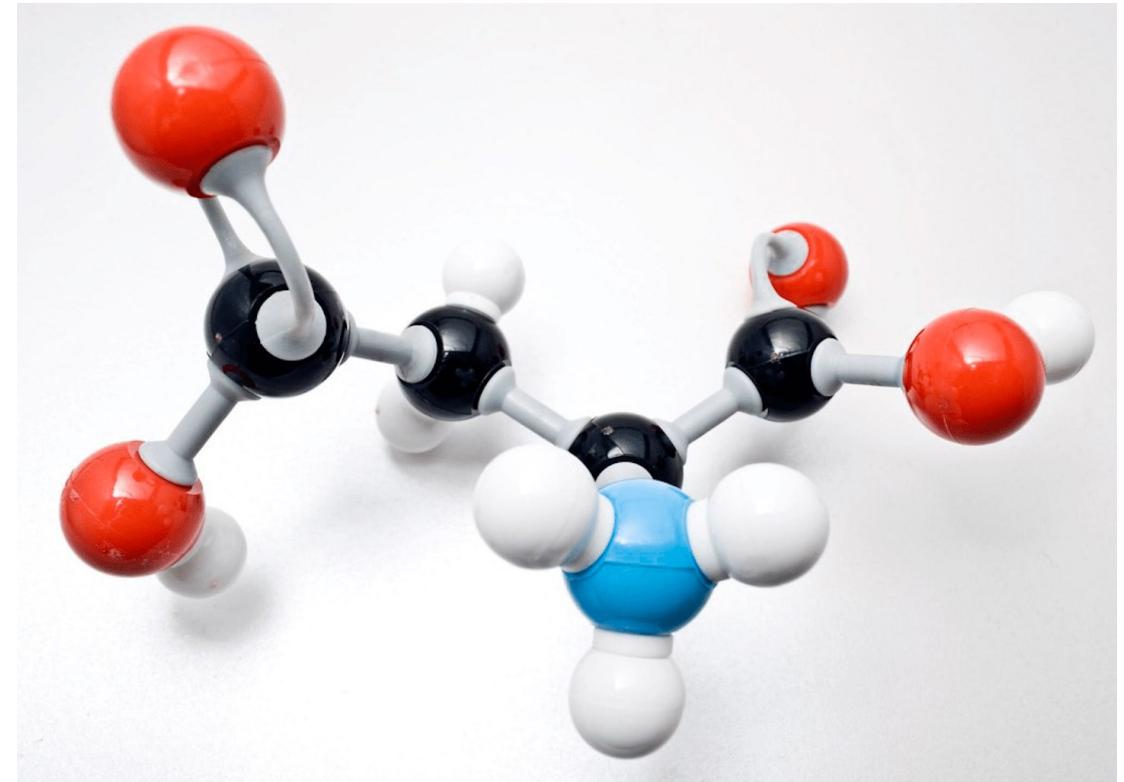




A graph– but more than a graph



- A molecule is a group of two or more atoms held together by attractive forces.
 - Molecules are the smallest unit of a substance which still retain its chemical and compositional properties.
 - These forces are called chemical bonds.
- Molecules form the basis of everything we see around us!



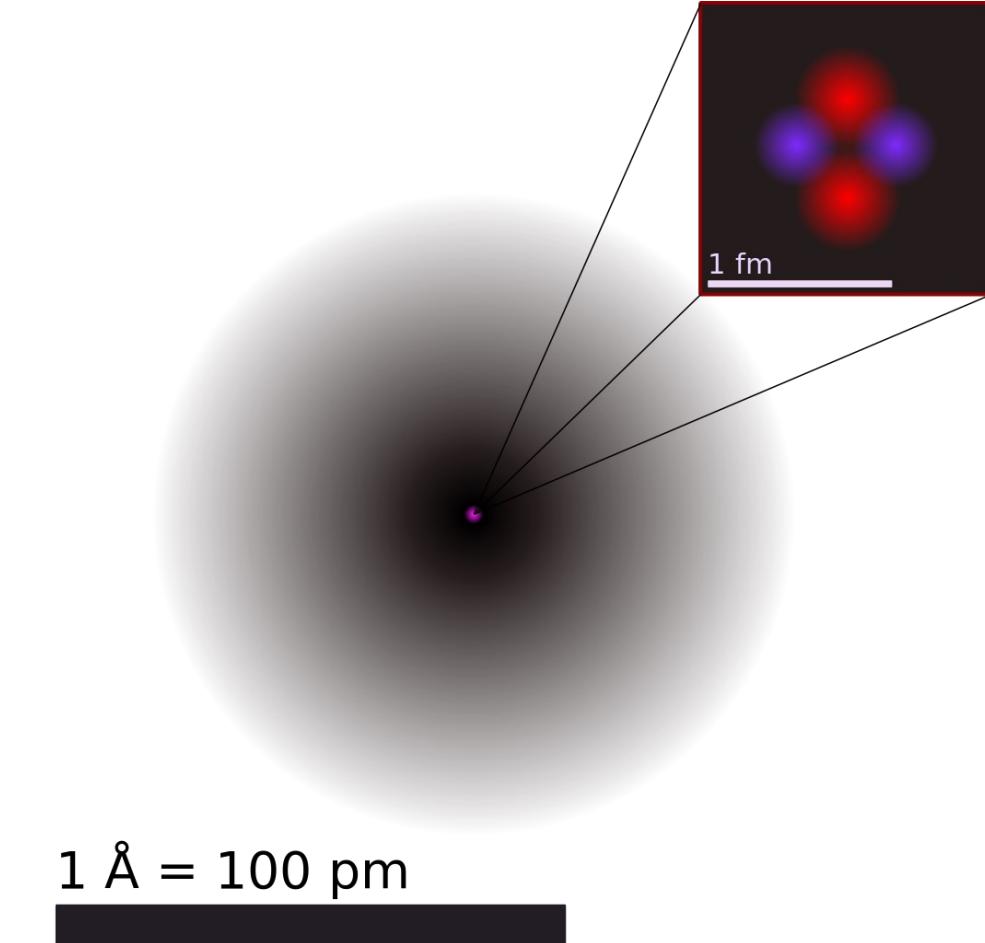


THRUST 1

What is an atom?



- Atoms are the basic particles of the chemical elements.
- They consist of protons, neutrons, and electrons.
 - Atoms have an elemental type determined by their number of protons.
- They are the nodes in the molecule graph

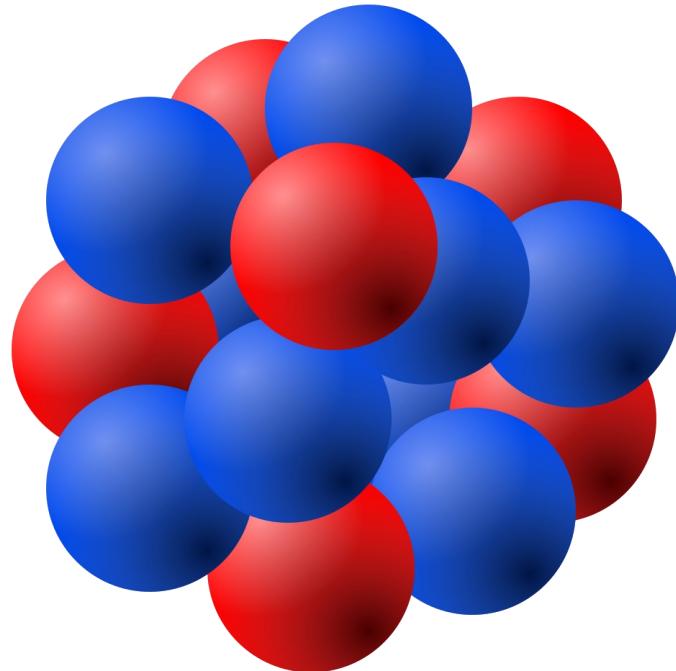




Electrons, protons, and neutrons



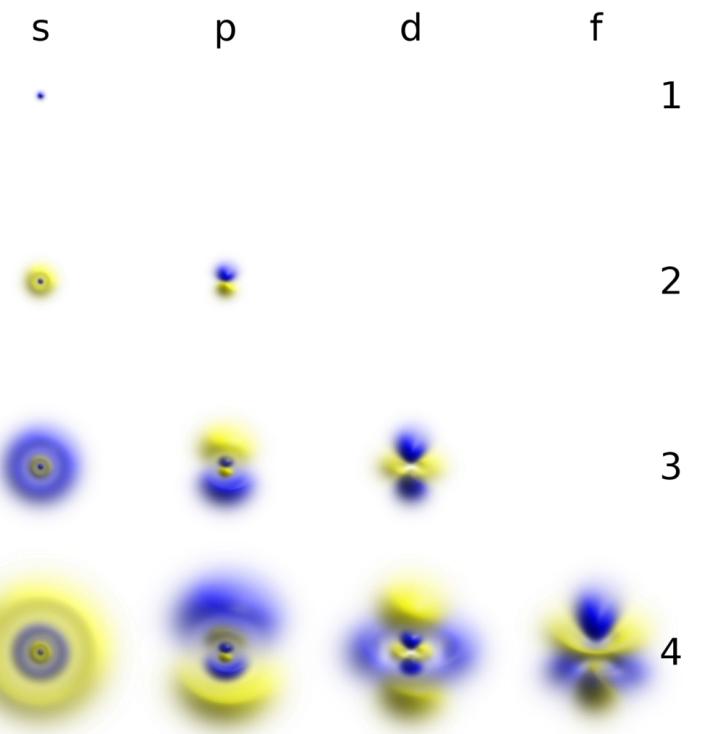
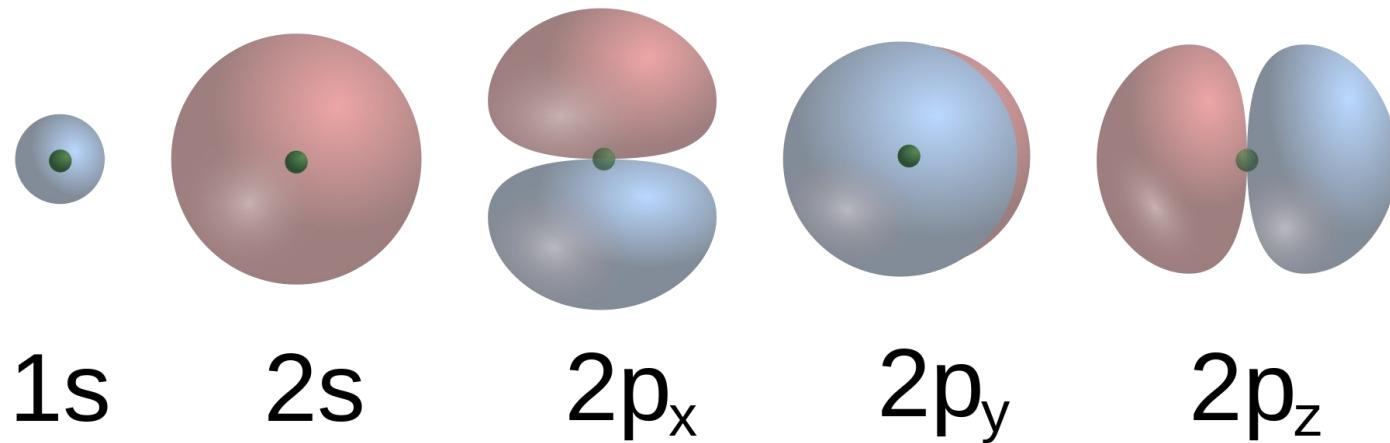
- Protons are positively charged particles which sit in the nucleus (the center of the atom)
- Neutrons are neutrally charged particles which also exist inside the nucleus.
 - They help separate the protons.
- Electrons are negative particles that exist as densities around the nucleus.





THRUST 1

What do atoms look like though?





What are the different elements?

- What element an atom is tells us about its behavior and interaction with other atoms

Group ►	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18		
Period ▼																				
Nonmetals	1 H	Some elements near the dashed staircase are sometimes called <i>metalloids</i>																		
Metals	2 Li Be	2 He																		
3	11 Na	12 Mg	Transition metals (sometimes excl. group 12)																	
4	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr		
5	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe		
6	55 Cs	56 Ba	La to Yb		71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	
7	87 Fr	88 Ra	Ac to No		103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Nh	114 Fl	115 Mc	116 Lv	117 Ts	118 Og
	s-block (incl. He)		f-block		d-block								p-block (excl. He)							
Lanthanides	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb						
Actinides	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No						

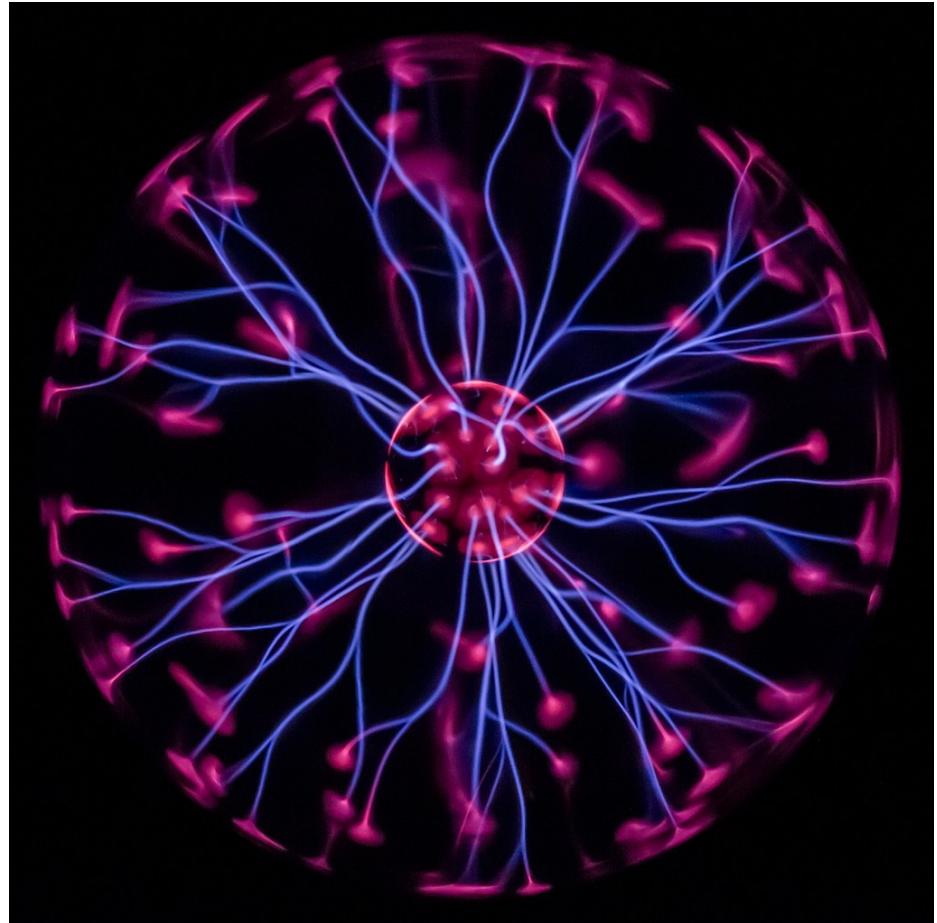
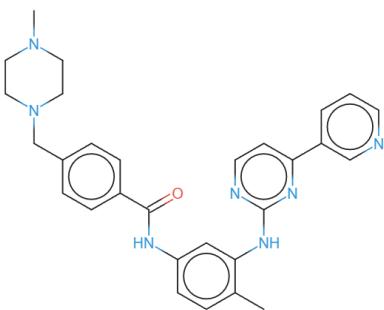


THRUST 1

What are chemical forces?



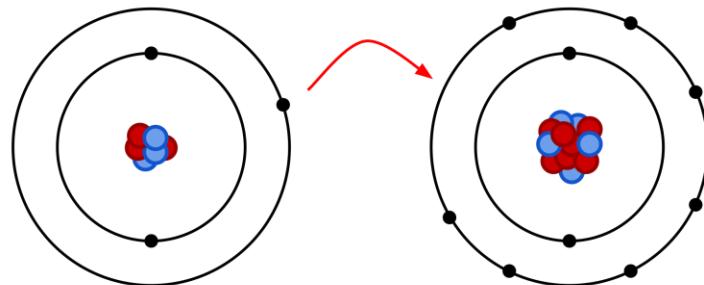
- Chemical forces are essentially the interactions between these atoms
- There are intramolecular forces (inside the molecules) and intermolecular forces (between different molecules)
 - Intramolecular forces are the edges in the molecular graph



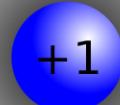


Intramolecular forces – bonds between the atoms

- Covalent bonds – two atoms “team up” and share electrons
 - There can be single, double, and triple bonds.



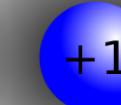
1e



•H

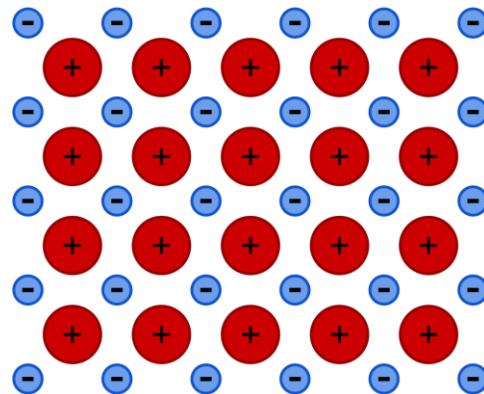


2e⁻



H:H
H - H

- Ionic bonds – two oppositely charged atoms come together to have a neutral charge

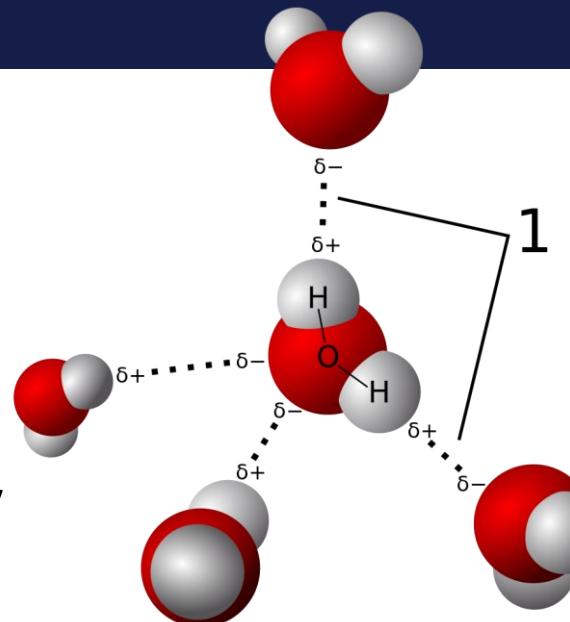


- Metallic Bonds – Pure metals will have electrons delocalize and form an “electron sea”



Intermolecular forces – attractions between molecules

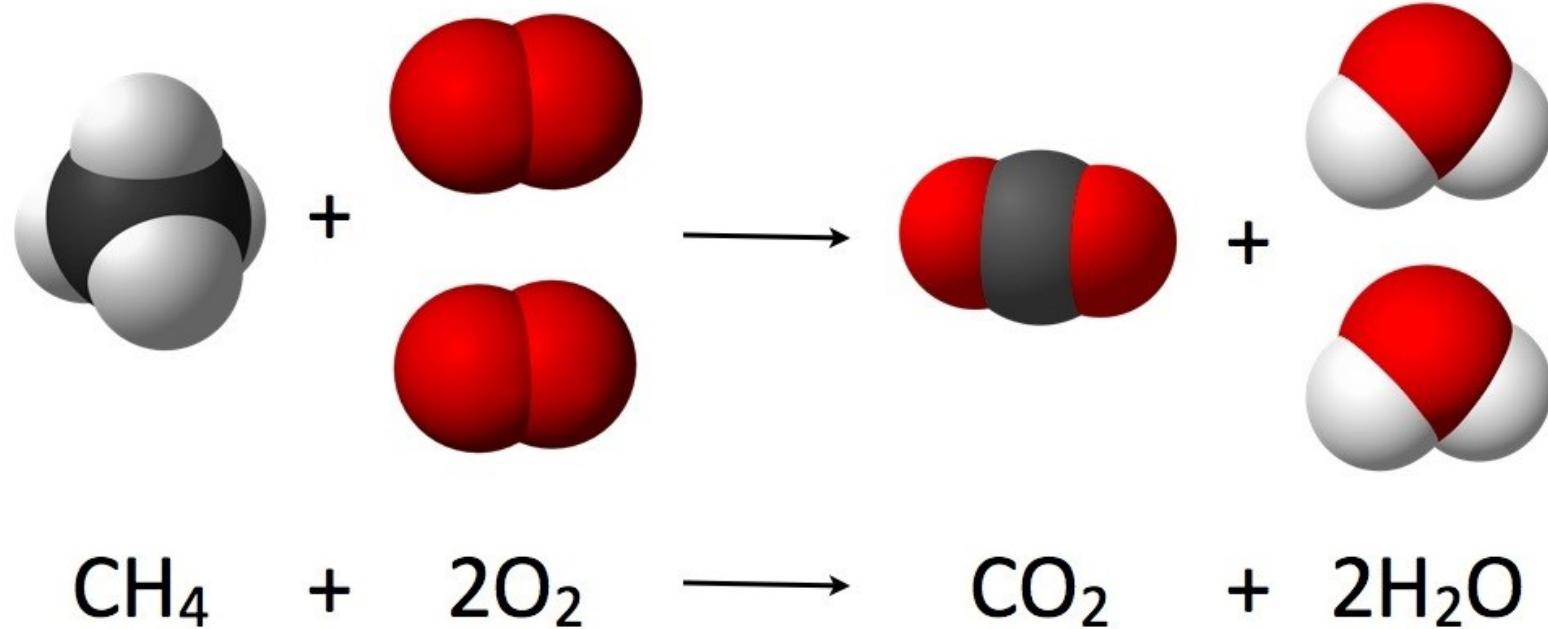
- These are generally much weaker than intramolecular forces
 - Hydrogen bonding – Covalently bonded hydrogens become charged and attracted to other atoms
 - Van de Waals forces – weak attractive forces between neutral atoms due to random shifts in electron density





Chemical Reactions– how we get new molecules

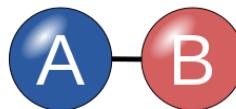
- A chemical reaction transforms one set of molecules into another
- There are countless chemical reactions, but generally, they either require energy (e.g., cooking) or produce energy (e.g., a car combustion engine)
- Our modern world is dependent on these processes



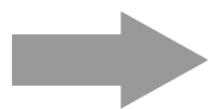


THRUST 1

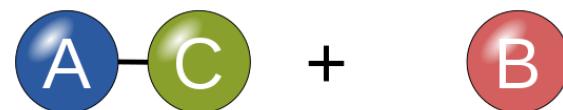
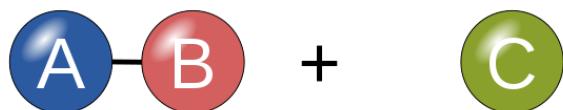
Chemical Reactions– The four basic types



- Synthesis



- Decomposition



- Single Replacement



- Double Replacement



THRUST 1

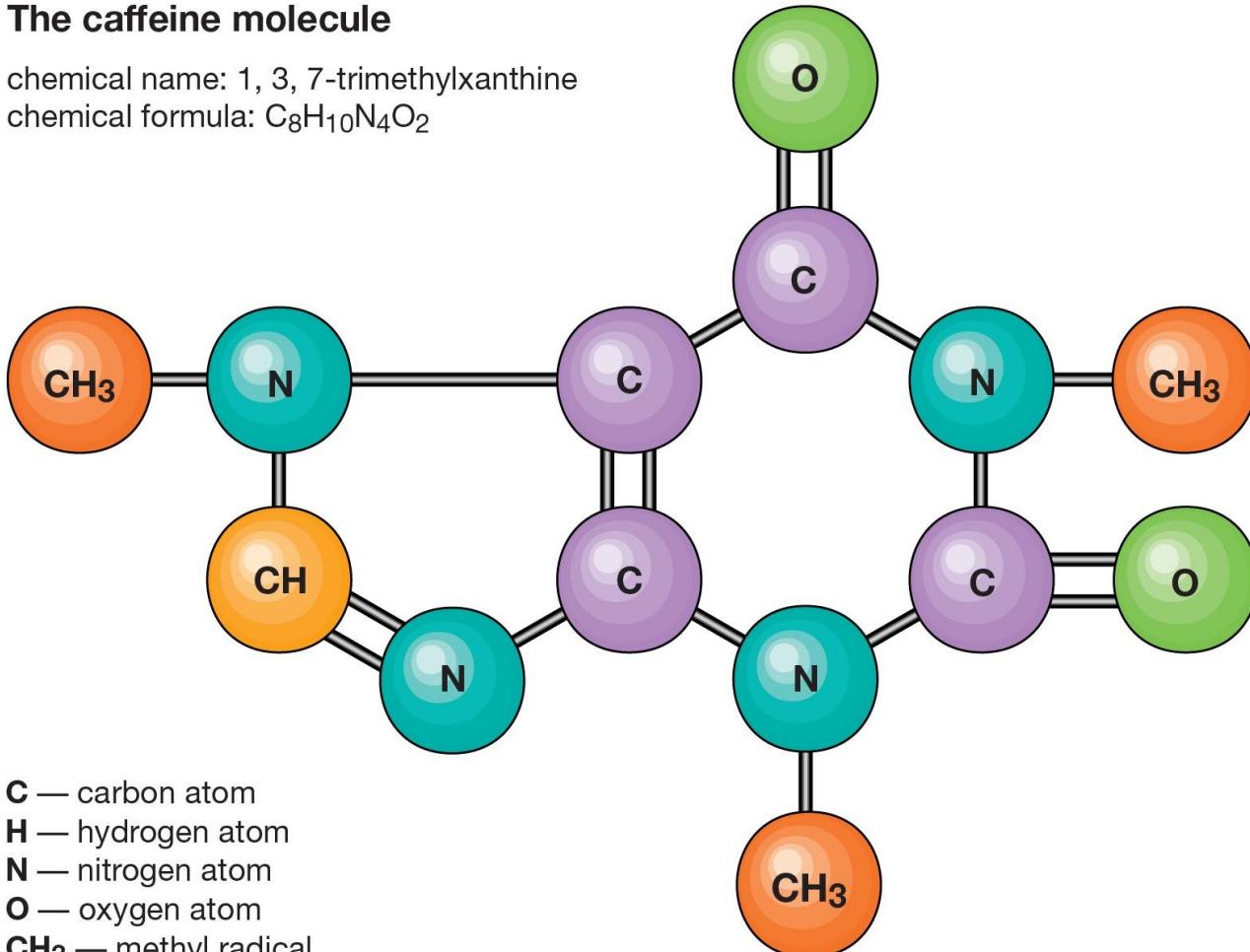
Let's see an example!



The caffeine molecule

chemical name: 1, 3, 7-trimethylxanthine

chemical formula: C₈H₁₀N₄O₂

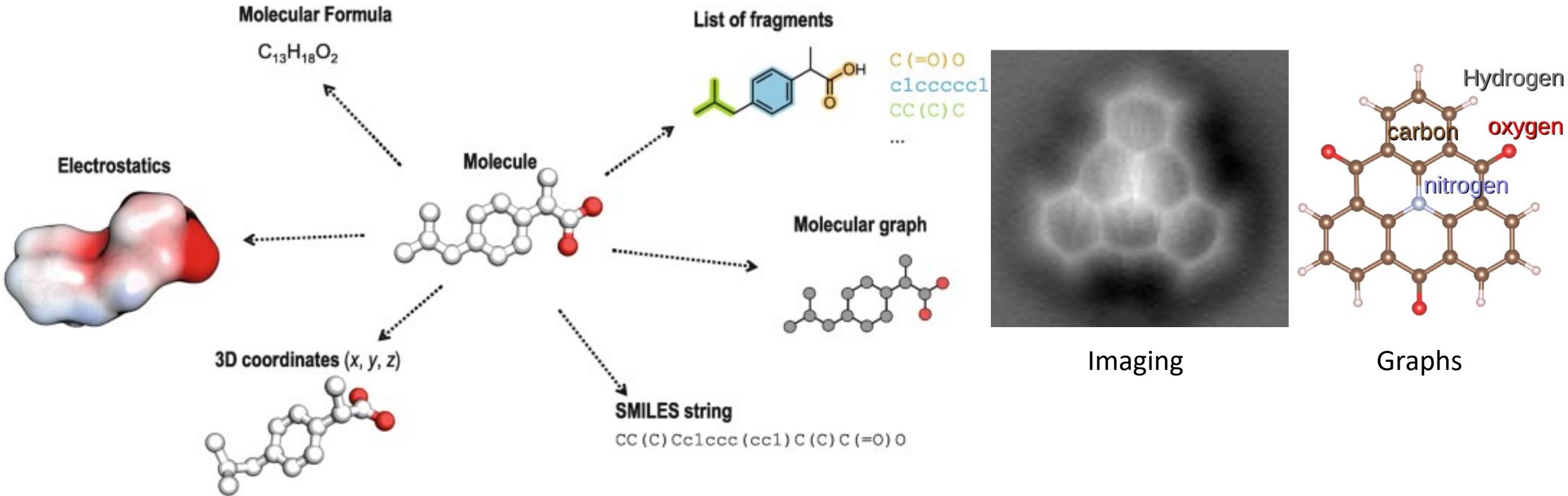


© 2010 Encyclopædia Britannica, Inc.



THRUST 1

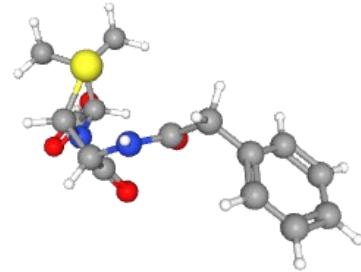
How do scientists represent these?



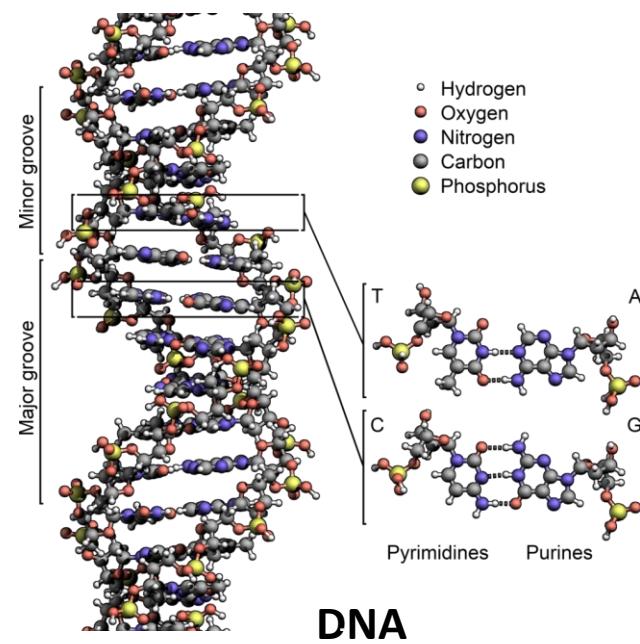


THRUST 1

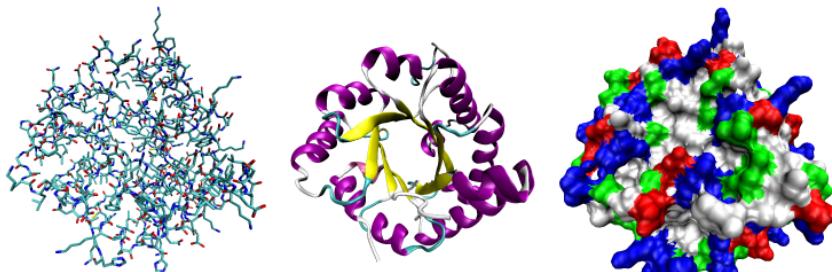
Types of molecules



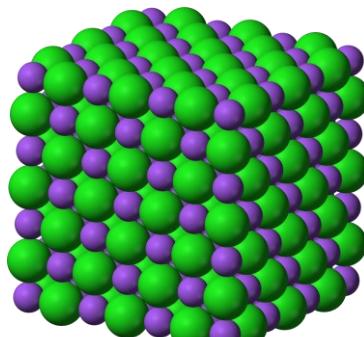
Organic Molecules



DNA



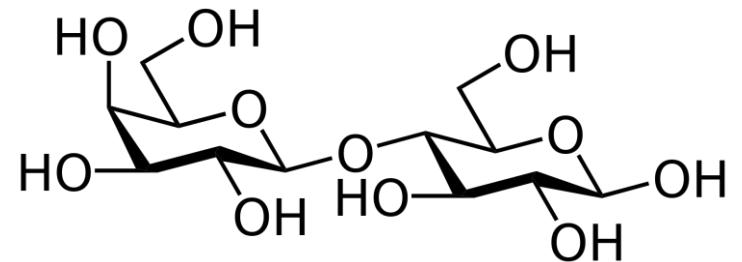
Proteins



Ionic Compounds



Metals

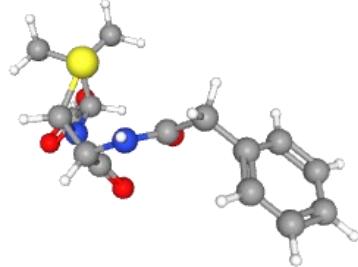


Carbohydrates

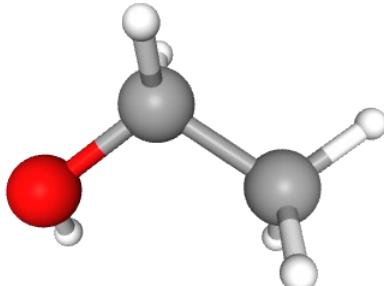


THRUST 1

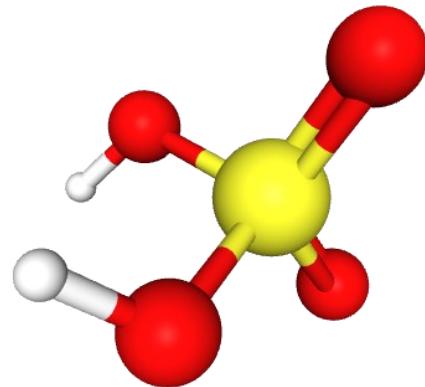
World-changing Applications!



Penicillin – antibiotics have saved an estimated 200 million lives



Ethanol – What's used to make alcoholic drinks used by billions

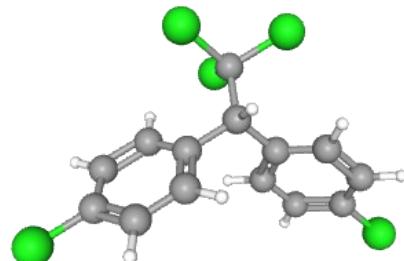


Sulfuric Acid – Used in mining, fertilizers, detergents, and plastic.

Source: Cosmos



Silicon – enabled the computer revolution and solar panels



DDT – Used to eliminate malaria from Europe and North America, but with health consequences.



THRUST 1

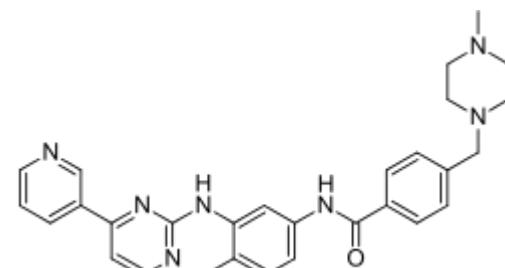
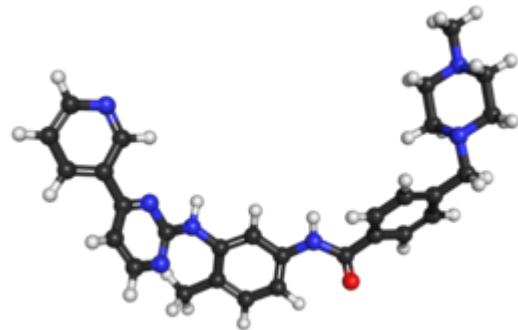
Chemistry is a Language!



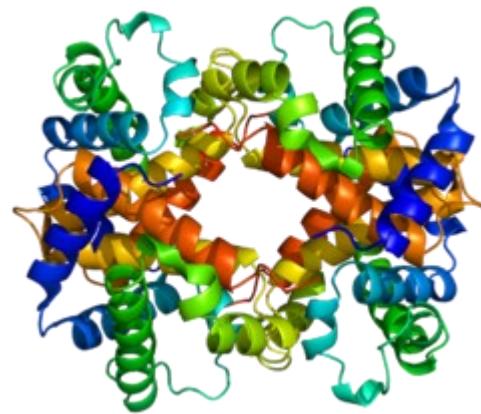
Chemistry is a language

Machine
Translation

Chemie ist eine Sprache



Cc1ccc(cc1Nc2nccc(n2)c3cccnc3)NC(=O)c4ccc(cc4)CN5CCN(CC5)C



VLSPADKTNVKAAW
GKVGAGHAGEYGAE
ALERMFLSFPTTKTY
FPHFDLSHGSAQV...

actttctggtccccacagactcag
agagaaccaccatggtgctgtct
cctgccgacaa gaccaacgtca
aggccgcctgggtaaggt ...



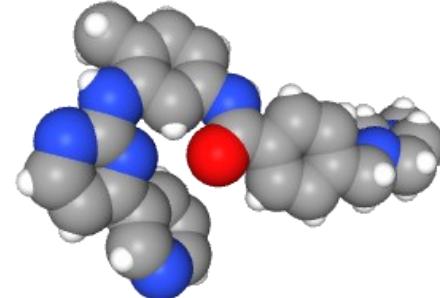
THRUST 1

Different representation formats for ML models

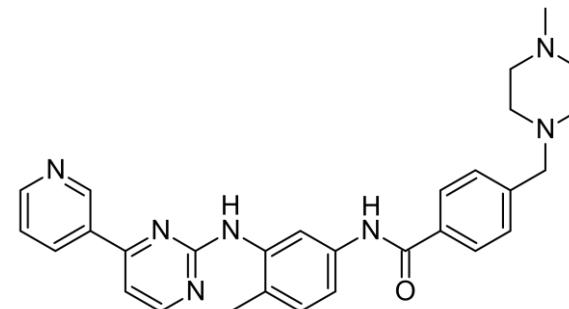


Cc1ccc(cc1Nc2nccc(n2)c3cccnc3)NC(=O)c4cc(c(cc4)CN5CCN(CC5)C

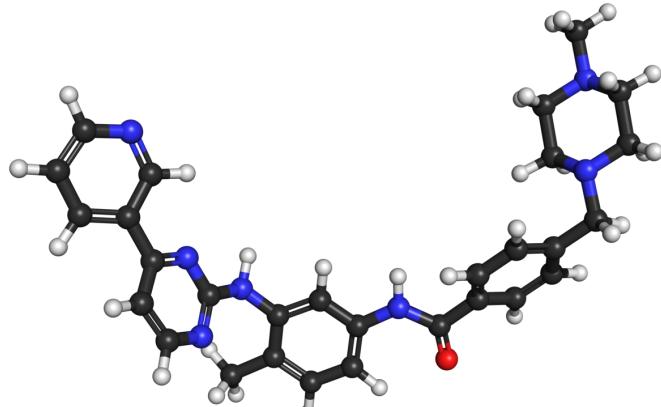
Sequence



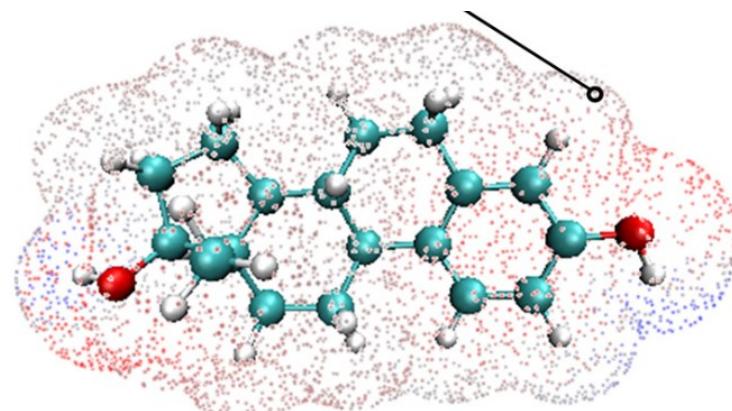
Space-Filling



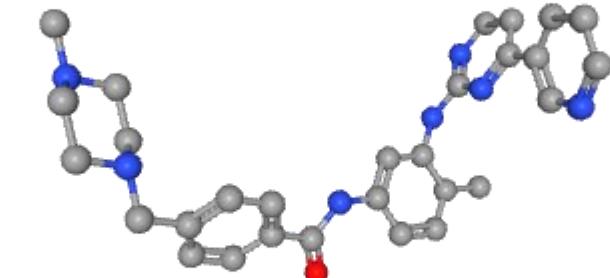
Graph



3D Graph



Point Cloud



Without hydrogens

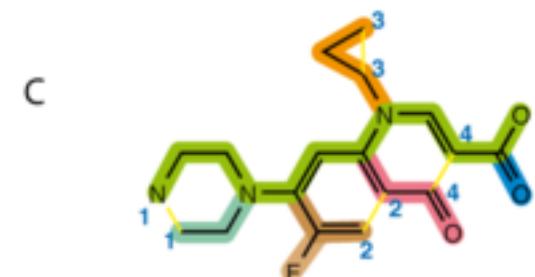
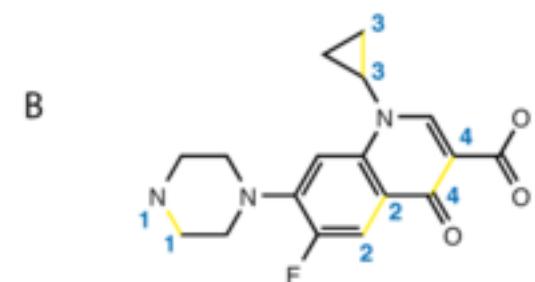
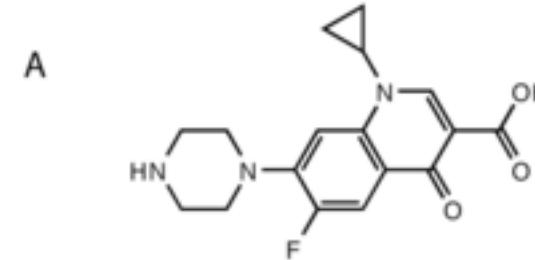


THRUST 1

SMILES (Weininger 1988)



- The best known string representation for molecules is *Simplified molecular-input line-entry system*, commonly known as SMILES.
- SMILES strings have been used to store molecular structures for decades.
 - GPT models have been exposed to SMILES in their training data.
- How does SMILES work?
 - Atoms are represented as 1-2 characters, such as 'C' for carbon, 'Br' for bromine, and 'F' for fluorine.
 - Rings are created through the use of numbers.
 - Branches are created with parenthesis.
 - Hydrogens are usually implicit.



D

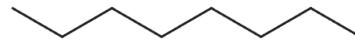
N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O



THRUST 1

Let's play with SMILES

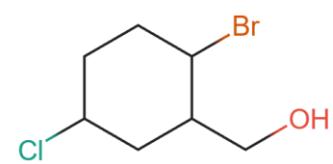
1. Let's start with a chain of carbons:



2. Now, let's add some other elements:

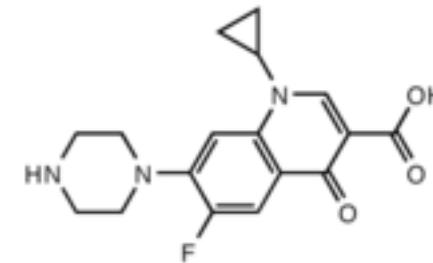


3. Let's create some ring structures:

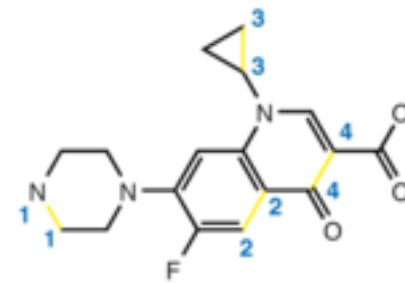


- Try for yourself at:
https://doc.gdb.tools/smilesDrawer/sd/example/index_light.html

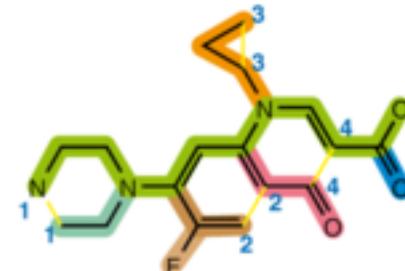
A



B



C



D



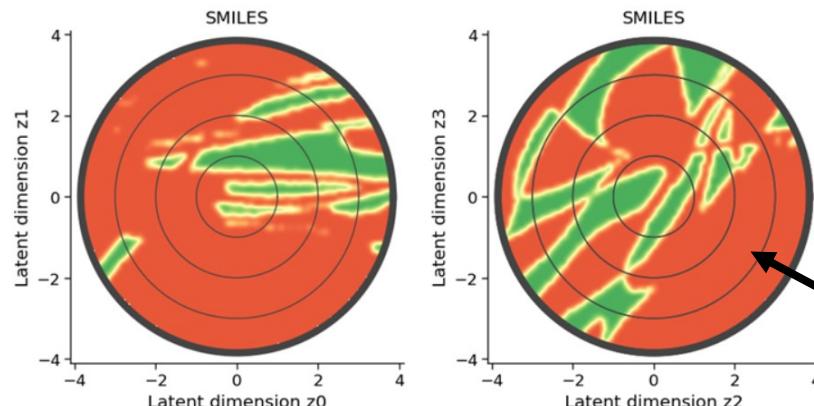


THRUST 1

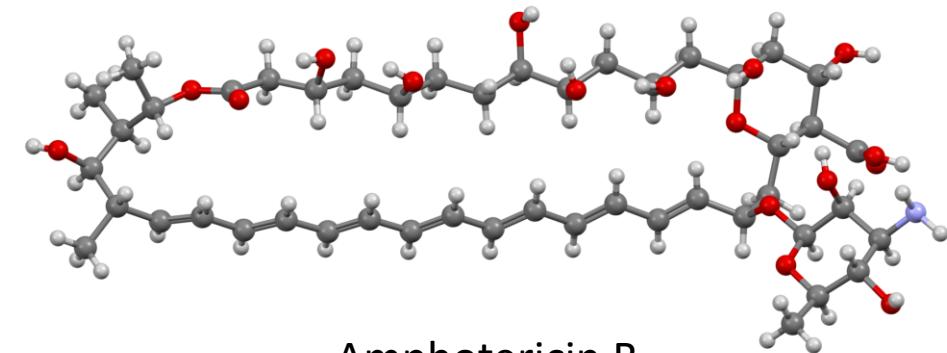
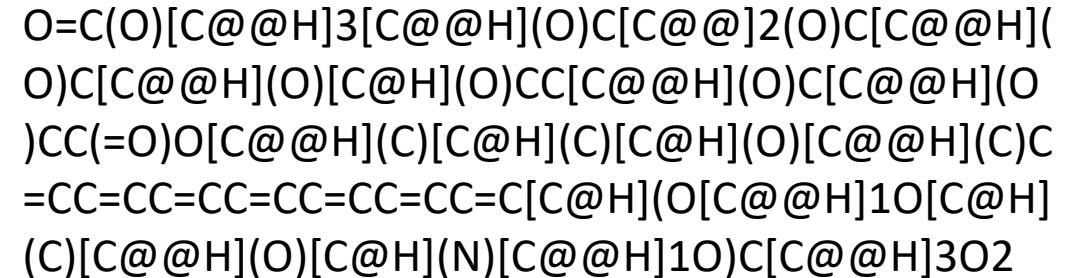
SMILES – What's wrong with it?



- Minimal 3D information
 - SMILES do not contain 3D information— just limited stereochemistry (left vs. right handedness)
- There are many invalid sequences
 - SMILES is not made for generative modeling!



These regions of a VAE latent space generate invalid sequences!



Amphotericin B

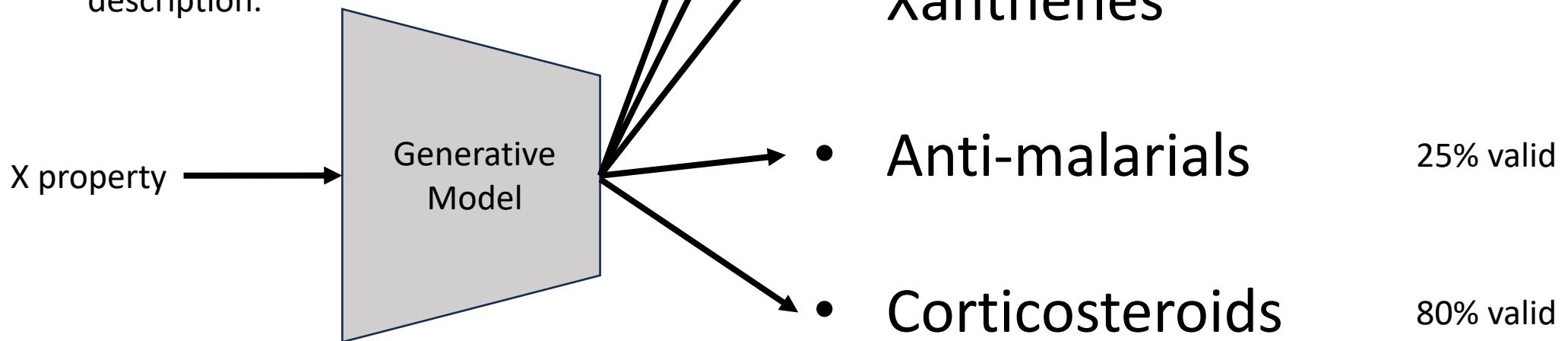


THRUST 1

SMILES – A secret benefit?



- Do invalid SMILES tell us when a generative model doesn't understand a region of chemical space?
 - This is especially the case in conditional generative modeling, such as text-guided molecule description.





THRUST 1

DeepSMILES (O'Boyle et al. 2018)



- Can we remove the problem with matching parenthesis or numbers?
- In DeepSMILES:
 - specify ring numbers by specifying how many atoms prior should be connected
 - A number of parenthesis based on how many characters should surround them.
 - Some invalid strings still exist, such as “C))C”

SMILES	DeepSMILES
C1CCCC1	CCCC5
C1CCCCCCCC1	CCCCCC%10
C(O)C	CO)C
C(OF)C	COF))C
C(F)(F)C	CF)F)C
C(=O)Cl	C=O)Cl
C(OC(=O)Cl)I	COC=O)Cl)))I
C1CC(OC)CC1	CCCOC))CC5
C1=C/CCCCC/1	C=C/CCCCC/8
C\1=C/CCCCC1	C=C/CCCCC/8
B(c1ccccc1)(O)O	Bcccccc6))))))O)O
Cn1cccc-2nccc12	Cnccccnccc9-5
C1N[C@@]12CO2	CN[C@@]3CO3
[C@@]12(NC1)CO2	[C@@]NC3))CO3
CC1CCCO[C@]21CCCCO2	CCCCO[C@@]6CCCCO6
CC1CCCO[C@@]12CCCCO2	CCCCO[C@@]6CCCCO6
NC[C@]12CCCC1C3CC2CC3	NC[C@]CCCC5CCC8CC5
NC[C@]12CCCC2C3CC1CC3	NC[C@@]CCCC5CCC8CC5

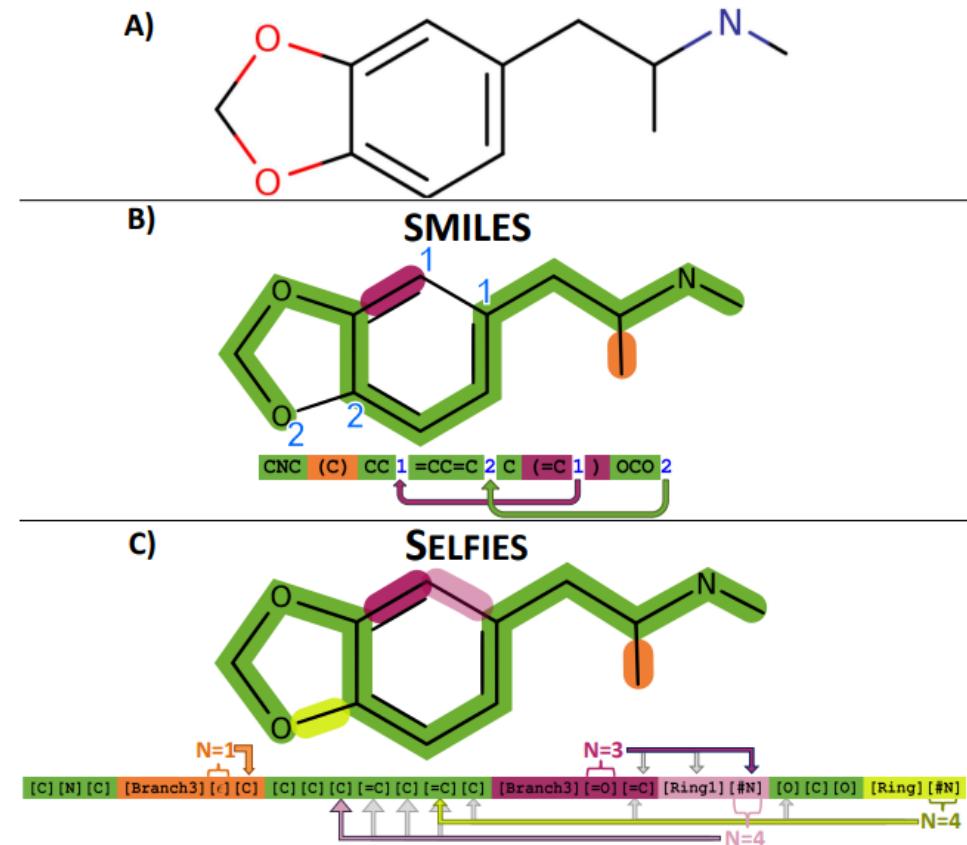
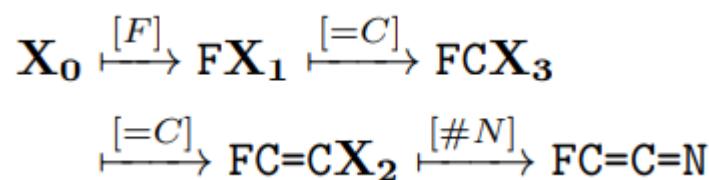


THRUST 1

SELFIES (Krenn et al. 2019)



- Can we integrate chemical graph constraints using a grammar-based approach?
 - A Chomsky type-2, context-free grammar for molecules





THRUST 1

SELFIES (Krenn et al. 2019)

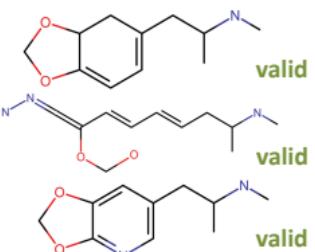


- How does this affect generative modeling and validity?

SMILES

A) Single Mutation

CNC (C) CC1=CC=CNC (=C1) OCO2
syntactically invalid

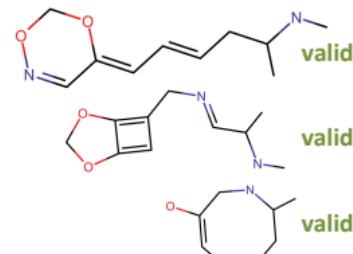


CNC (C) CC1=CC=C2C (=C1FOCO2
syntactically invalid

CFC (C) CC1=CC=C2C (=C1) OCO2
semantically invalid

B) Double Mutation

CNC (C) OC1=CC=C2C (=C1COCO2
syntactically invalid

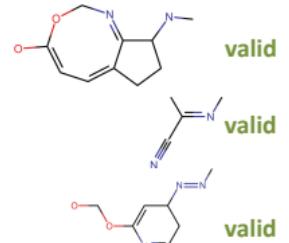


CNC (C) CC1=CCOCCC (=C1) OCO2
syntactically invalid

CNC (C) #C1=CC=C2C (=C1) OCON
syntactically & semantically invalid

C) Triple Mutation

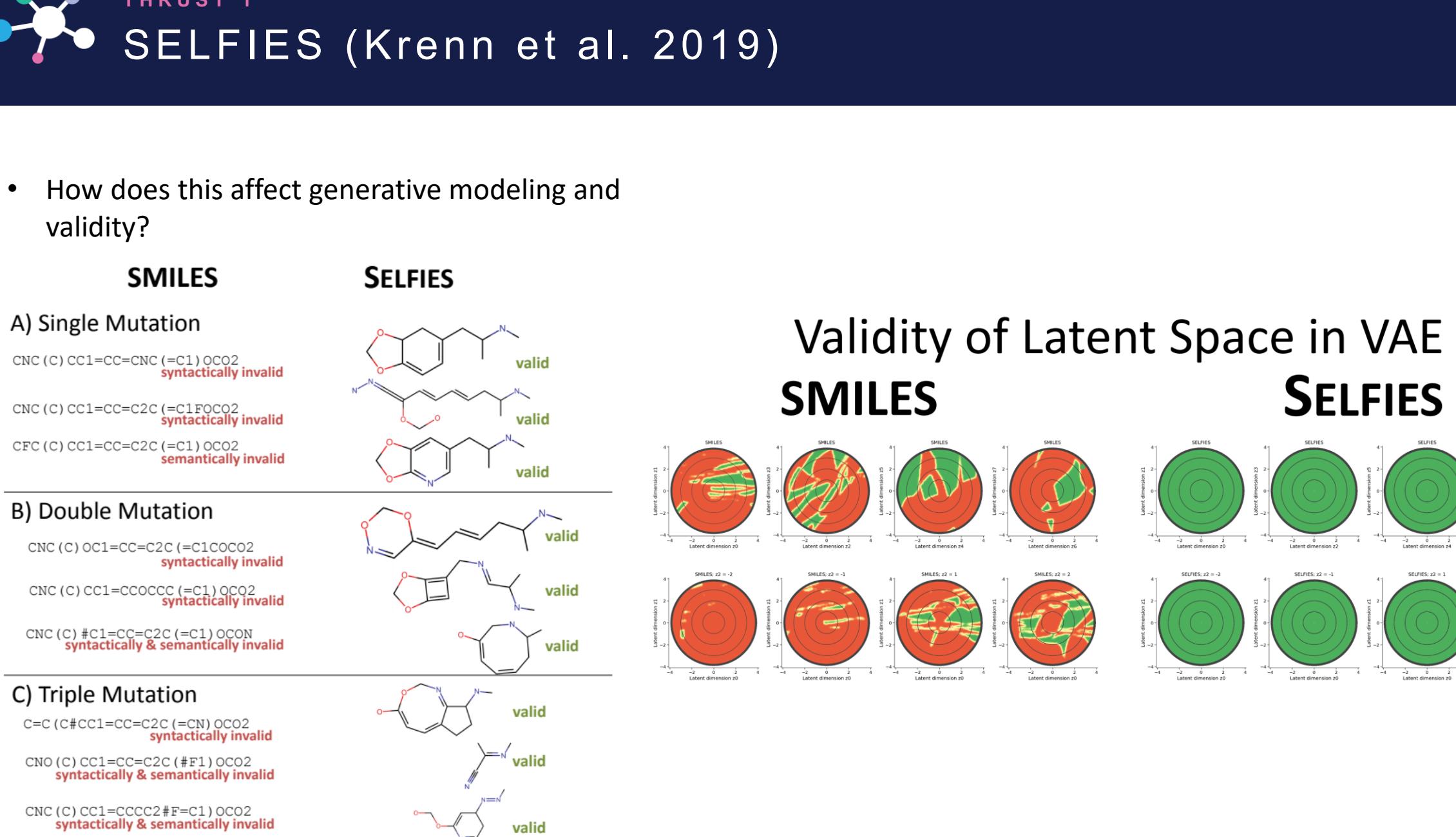
C=C (C#CC1=CC=C2C (=CN) OCO2
syntactically invalid



CNO (C) CC1=CC=C2C (#F1) OCO2
syntactically & semantically invalid

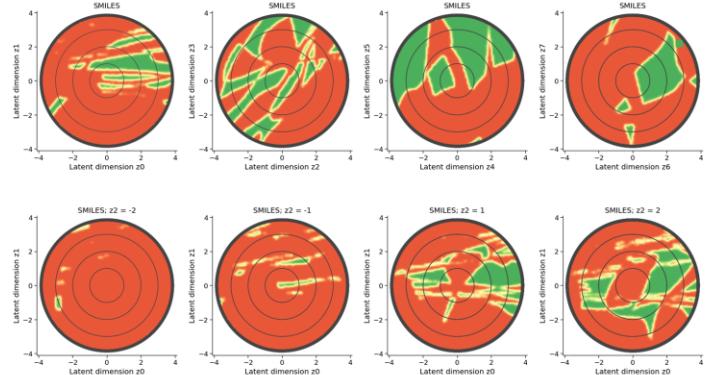
CNC (C) CC1=CCCC2#F=C1) OCO2
syntactically & semantically invalid

SELFIES

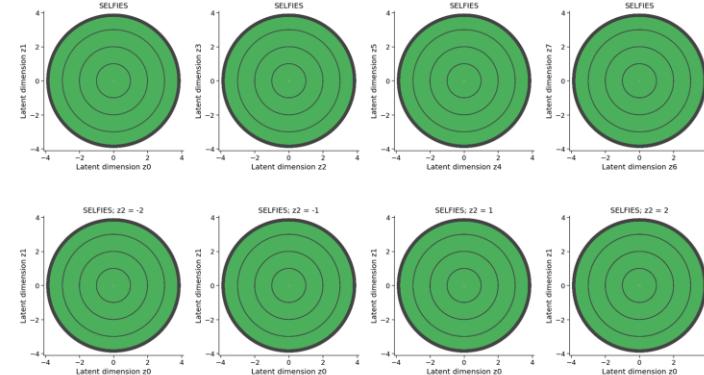


Validity of Latent Space in VAE

SMILES



SELFIES



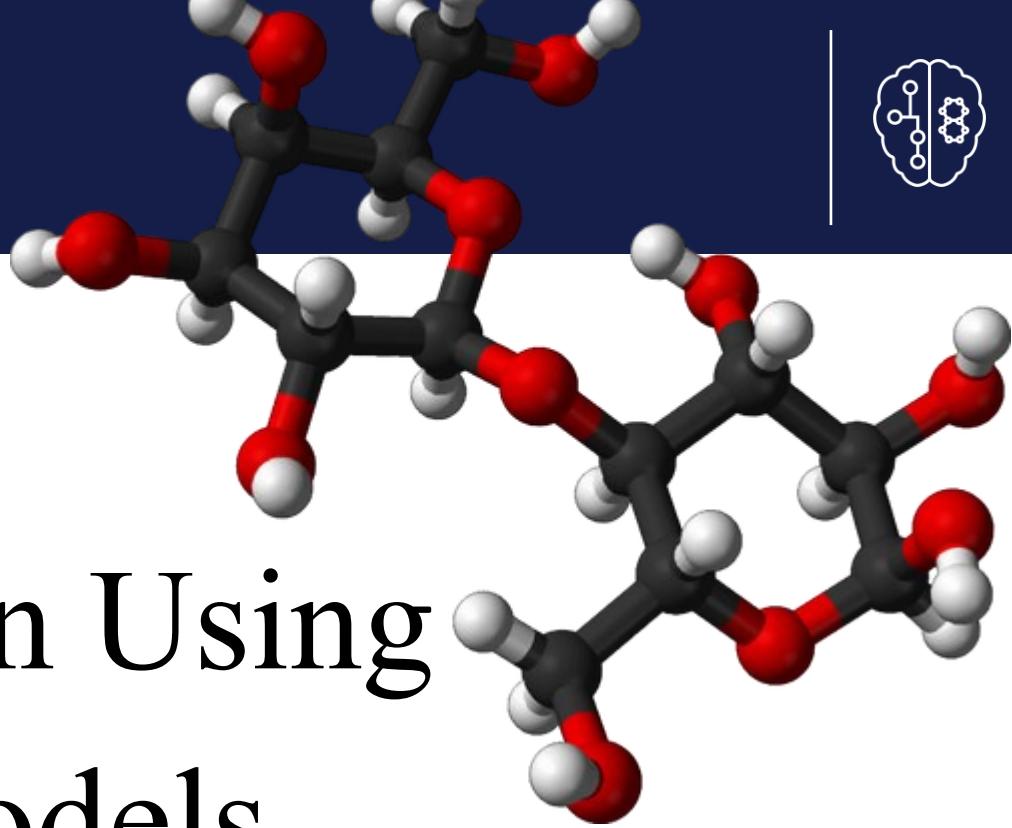
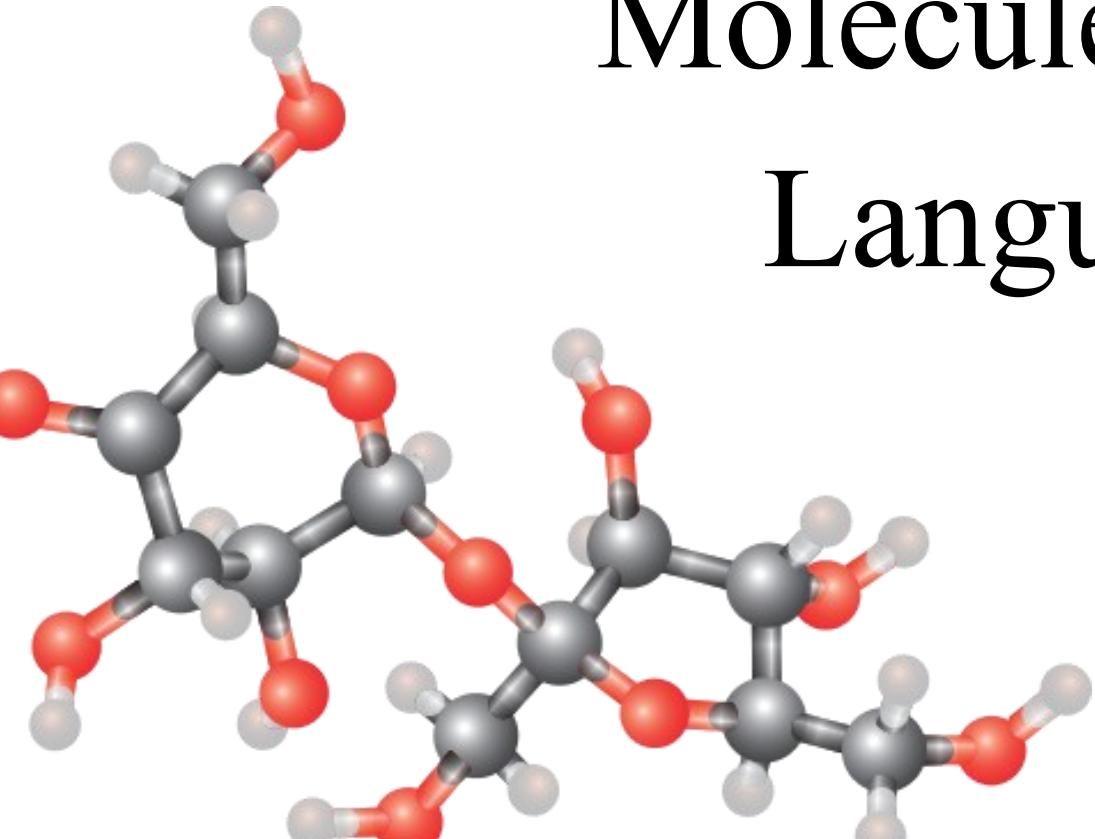


THRUST 1

Background and Motivation



Molecule Design Using Language Models



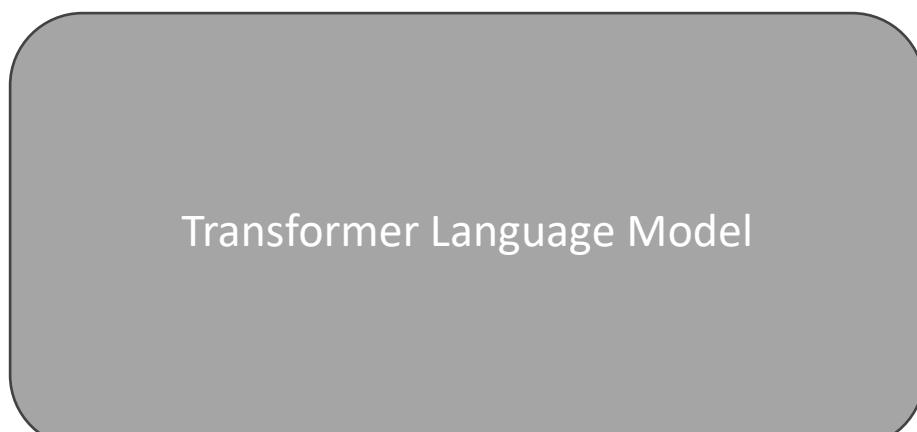


THRUST 1

Applying language models to molecule strings



- Breakthroughs in language modeling have been applied to the language of molecules.
- Models are pretrained using typical masked language modeling objectives or property prediction.



CCc(c1)ccc2[n+]1ccc3c2[nH]c4c3cccc4



MLM: Masked Language Model

C1CC1(C(=O)NC2=CC=C(C=C2)OC3=C4C=C(NC4=NC=C3)C(=O)NCCN5CC0CC5)C(=O)NC6=CC=C(C=C6)F
↓
C1CC1(C(=O)NC2=CC=C(C=C2)OC3=C4C=C(NC4=NC=C3)C(=O)NCCN5CC0CC5)C(=O)NC6=CC=C(C=C6)F

MTR: Multi-Task Regression

C1CC1(C(=O)NC2=CC=C(C=C2)OC3=C4C=C(NC4=NC=C3)C(=O)NCCN5CC0CC5)C(=O)NC6=CC=C(C=C6)F
↓
MW, TPSE, FSP3C, #ROTBONDS, etc ...

ChemBERTa-2. Ahmad et al. 2022



THRUST 1

ChemBERTa 1 and 2



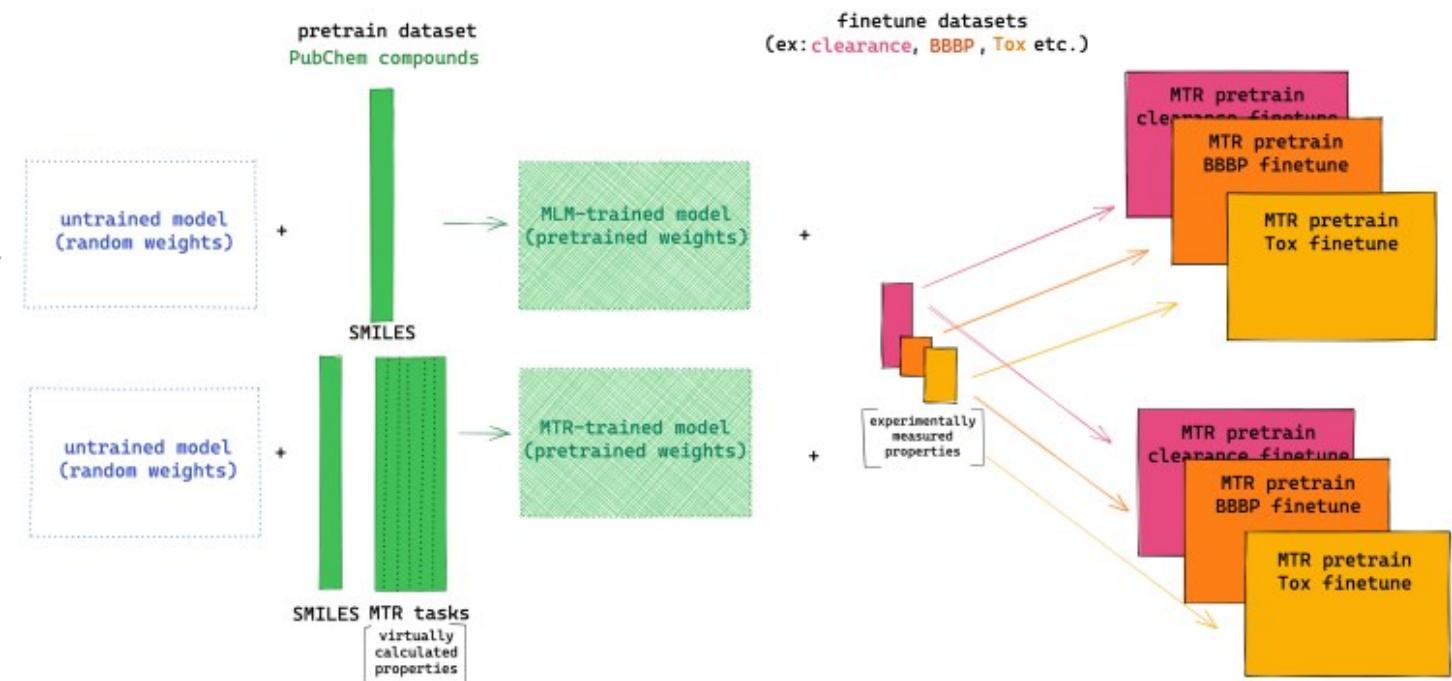
MLM: Masked Language Model

C1CC1(C(=O)NC2=CC=C(C=C2)OC3=C4C=C(NC(=N)C3)C(=O)NCCN5CCOCC5)C(=O)NC6=CC=C(C=C6)F
↓
C1CC1(C(=O)NC2=CC=C(C=C2)OC3=C4C=C(NC(=N)C3)C(=O)NCCN5CCOCC5)C(=O)NC6=CC=C(C=C6)F

MTR: Multi-Task Regression

C1CC1(C(=O)NC2=CC=C(C=C2)OC3=C4C=C(NC(=N)C3)C(=O)NCCN5CCOCC5)C(=O)NC6=CC=C(C=C6)F
↓
MW, TPSE, FSP3C, #ROTBONDS, etc ...

(a) MLM vs. MTR

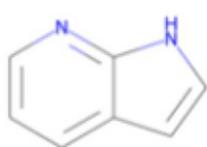
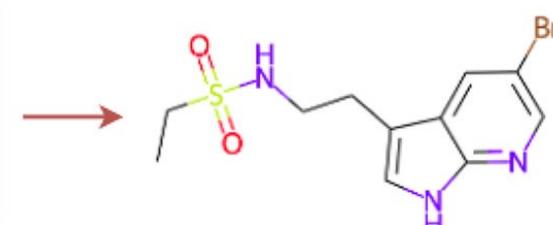
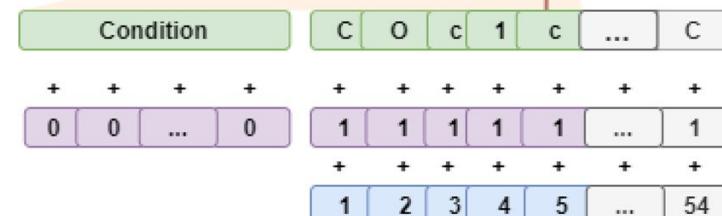
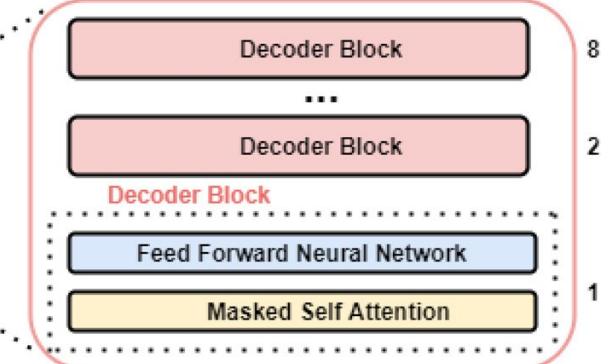
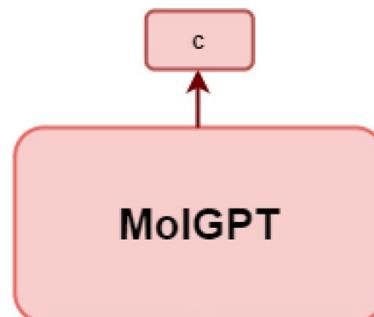


(b) Training pipeline



THRUST 1

MoIGPT

**Scaffold**LogP
TPSA
SAS
QED**Properties****Generated Molecule**

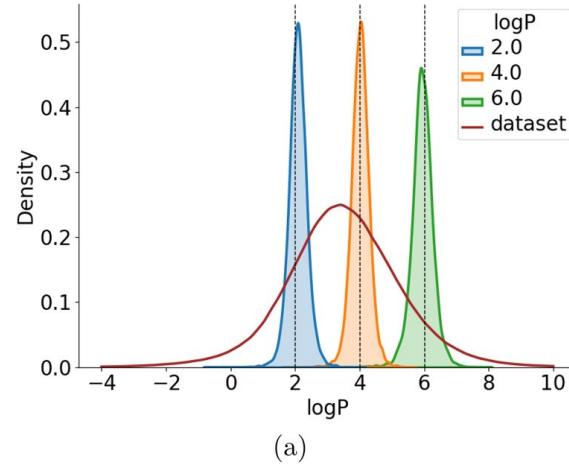
→ Segment Token

→ Position Token

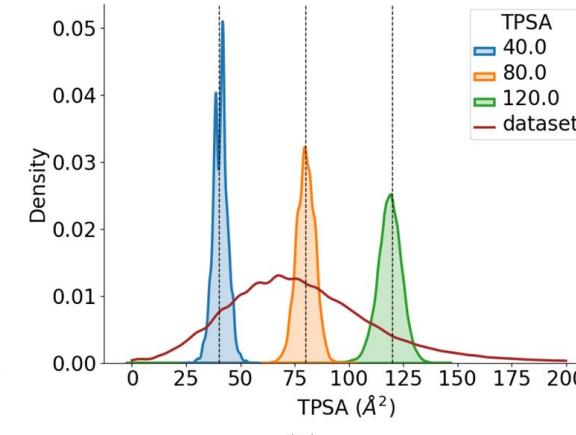


THRUST 1

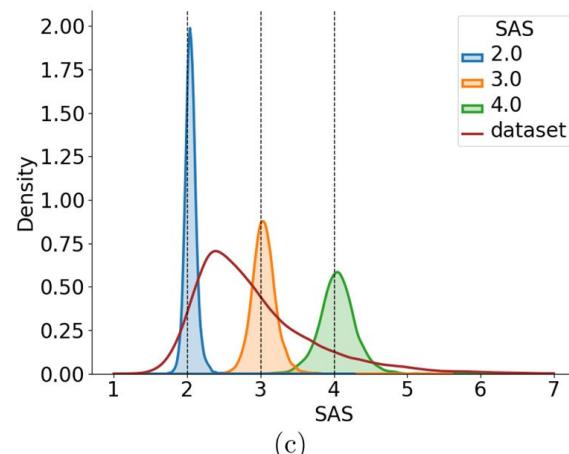
MoIGPT



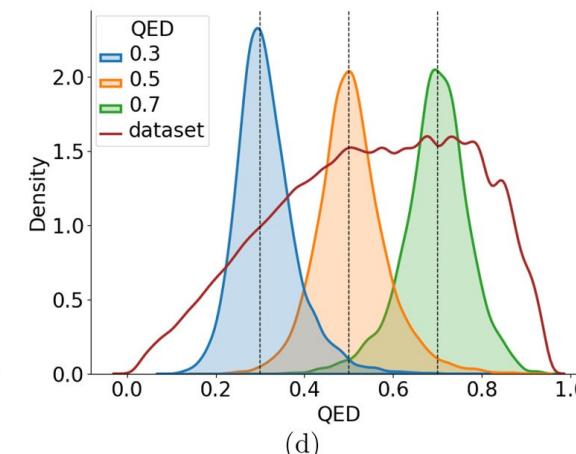
(a)



(b)



(c)



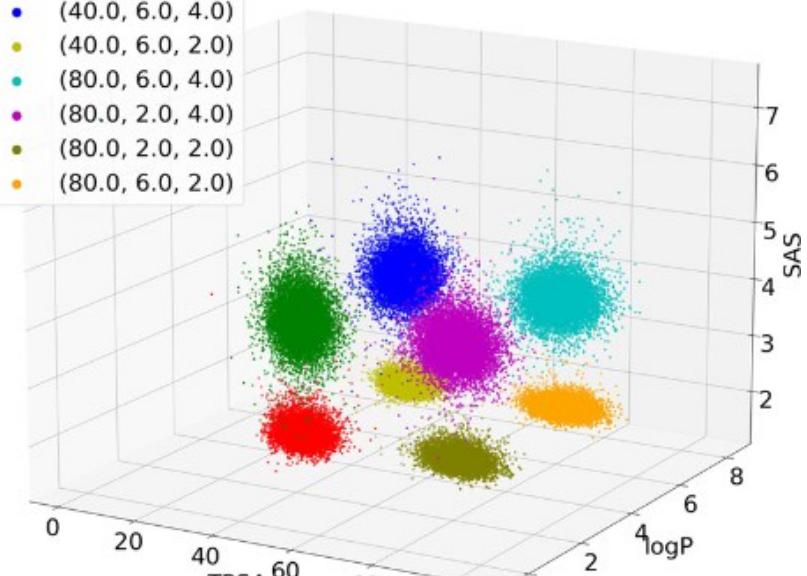
(d)

Conditioning on properties:

(a) $\log P$, (b) TPSA, (c) SAS, and (d) QED.

Bagal et al. 2022

- (TPSA, $\log P$, SAS)
- (40.0, 2.0, 2.0)
 - (40.0, 2.0, 4.0)
 - (40.0, 6.0, 4.0)
 - (40.0, 6.0, 2.0)
 - (80.0, 6.0, 4.0)
 - (80.0, 2.0, 4.0)
 - (80.0, 2.0, 2.0)
 - (80.0, 6.0, 2.0)



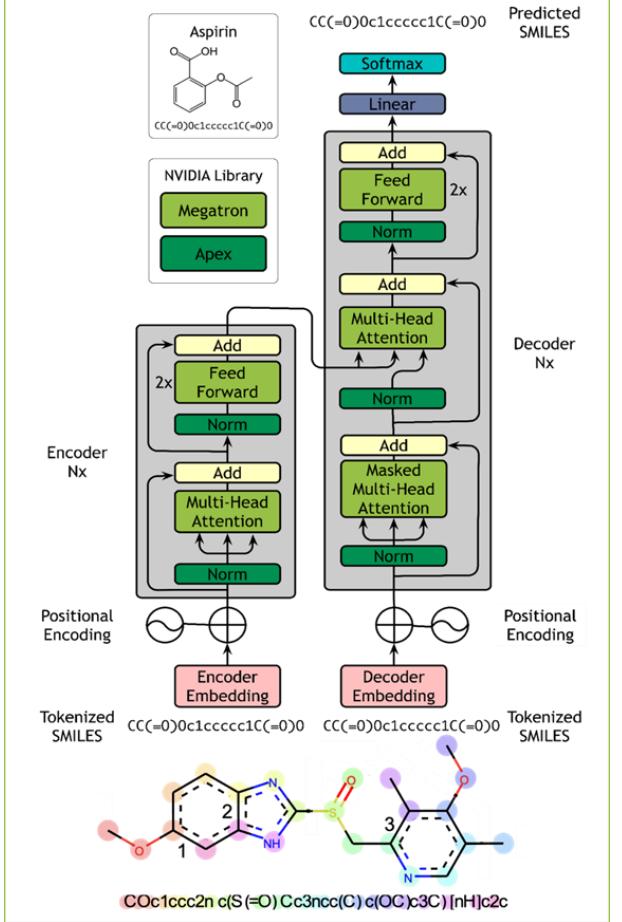
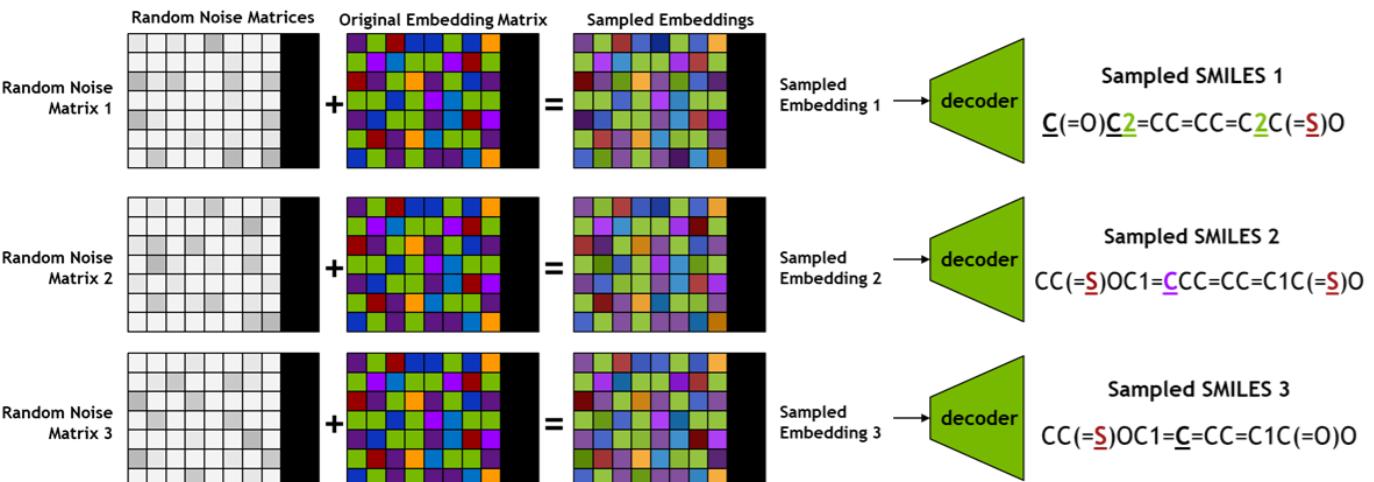
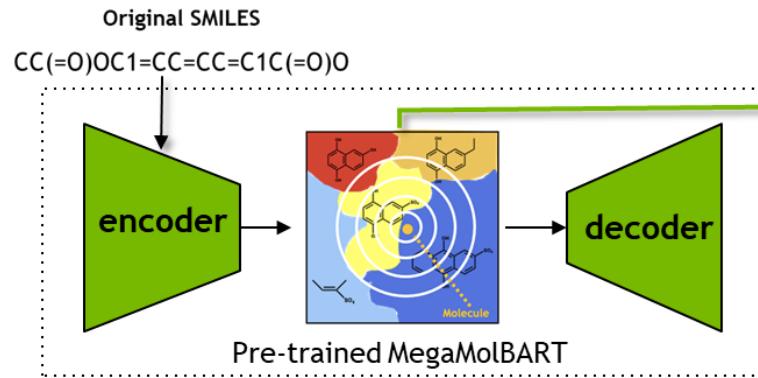
Conditioning on multiple properties at once



THRUST 1

MegaMolBART v1 and v2

- Trained by NVIDIA using their BioNeMo framework on 1.45 billion SMILES strings



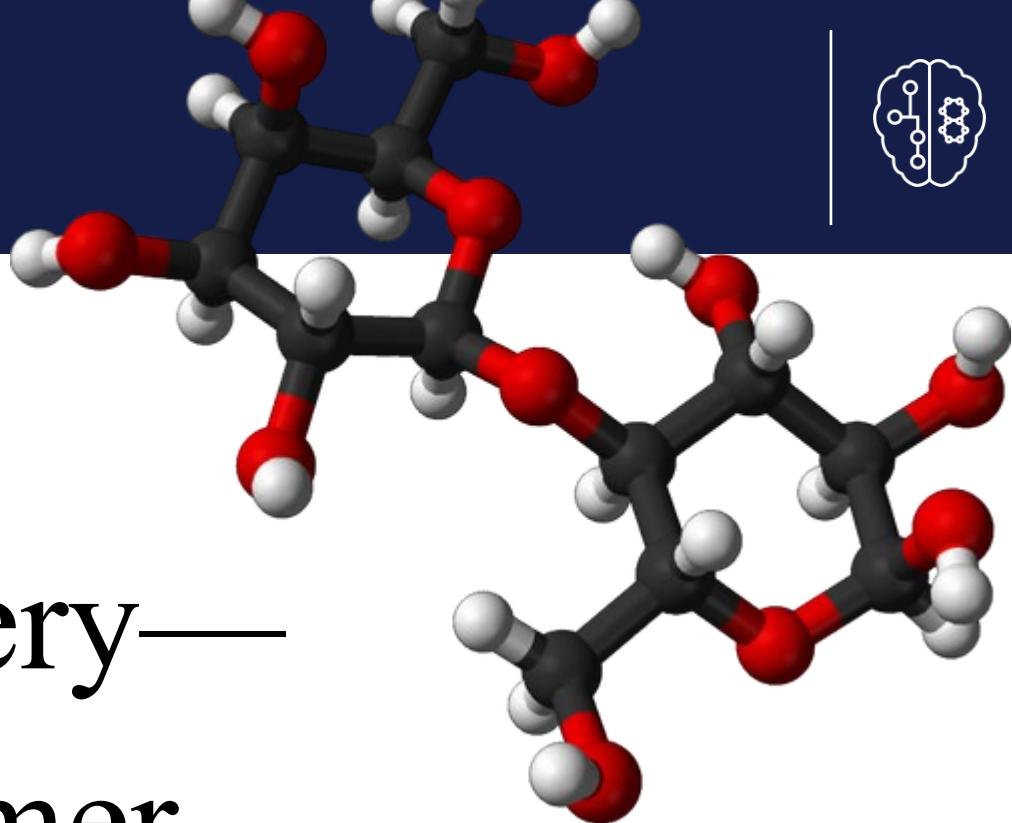
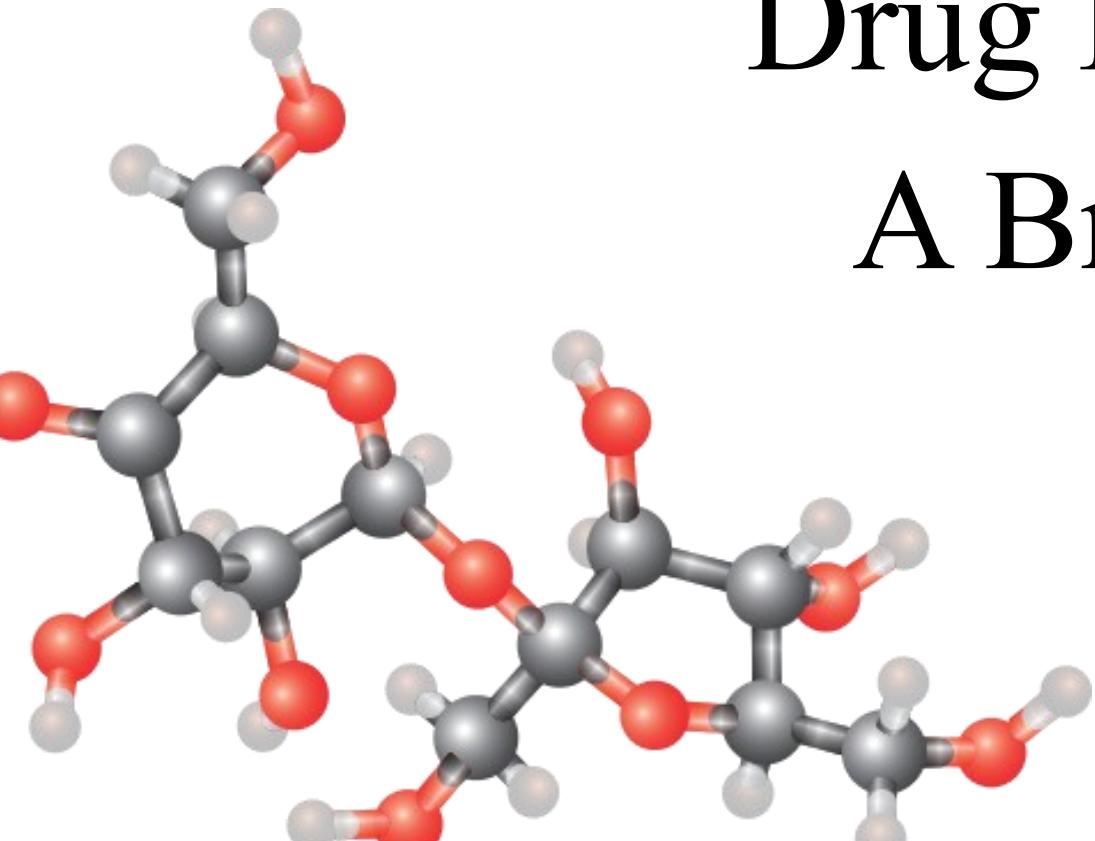


THRUST 1

Background and Motivation



Drug Discovery— A Brief Primer



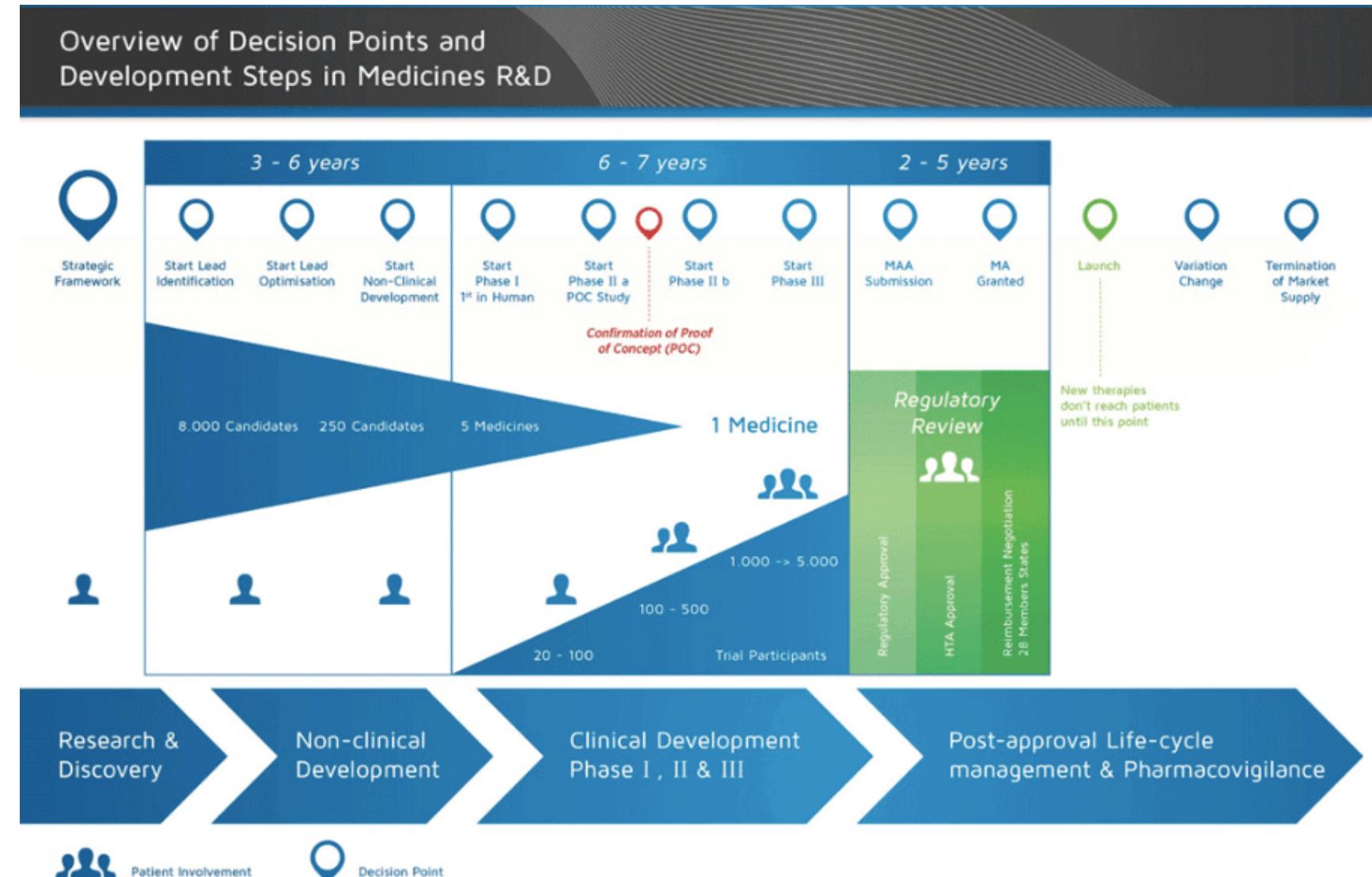


THRUST 1

Drug discovery: the big, long, expensive funnel



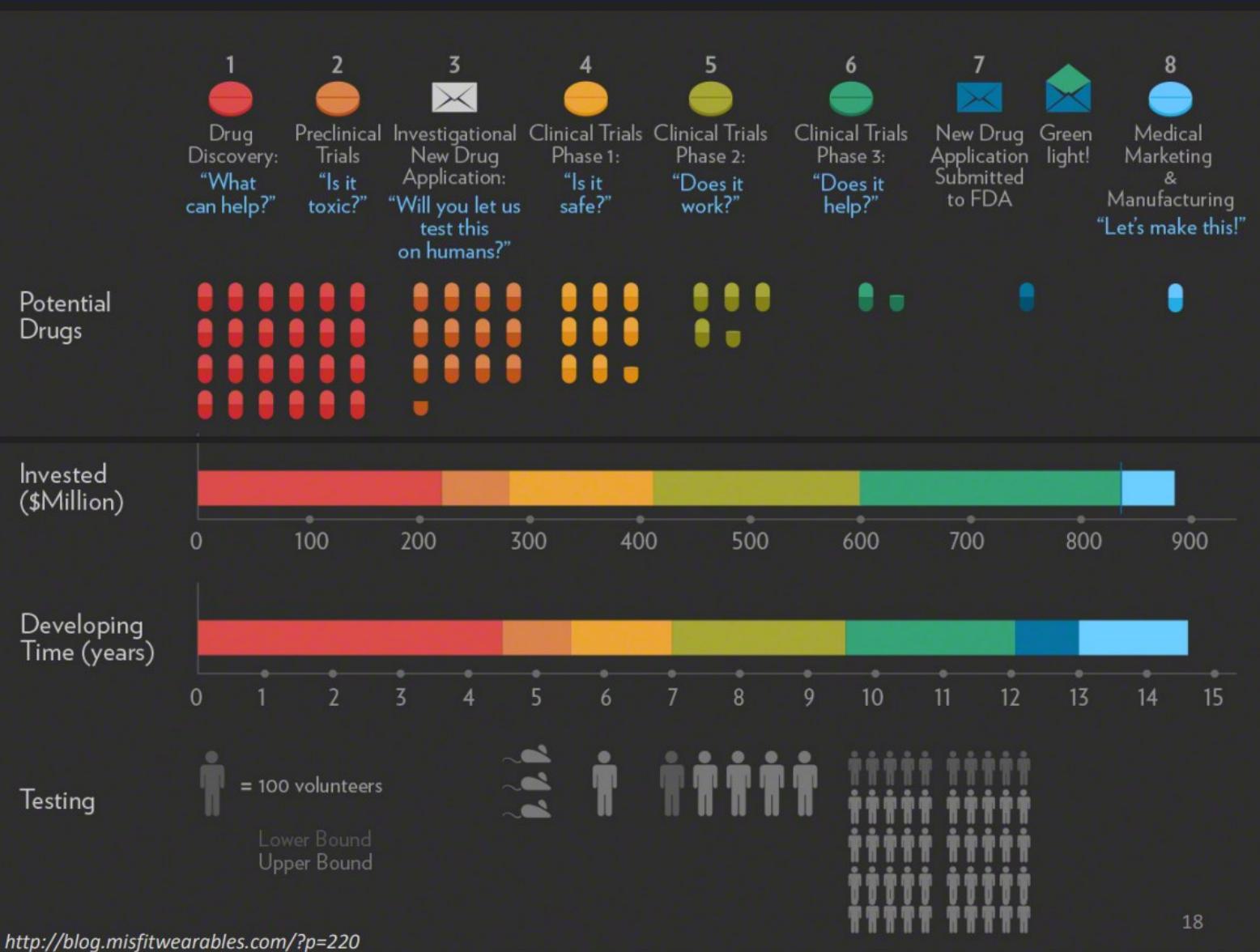
- Drug discovery is a complex process of finding a molecule which can change the way the body works—usually to treat a disease, disorder, or malignancy.
- Drug discovery is essentially a funnel—a drug must pass increasingly challenging tests to become approved.





THRUST 1

Why do we have all these stages?



Key observation: This is a field where LLM hallucination isn't the worst thing ever!

- Guarantees do not need to come from the model. They come from testing, which is sufficiently generalized and established.



THRUST 1

Balancing the right properties

- Modern drug discovery depends on multi-objective optimization— there are so many properties we want our drugs to have (or to not have).
- Can we maximize the effect of the drug while minimizing its toxicity and side effects?

Permeability

Solubility

Selectivity

Potency

Stability



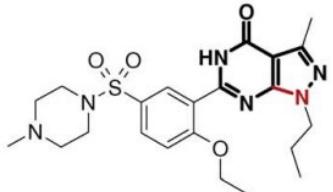


THRUST 1

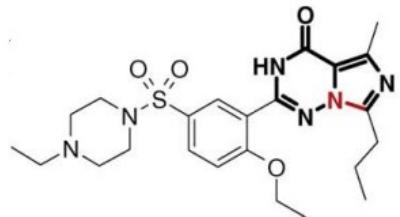
Let's focus on small-molecule drug discovery



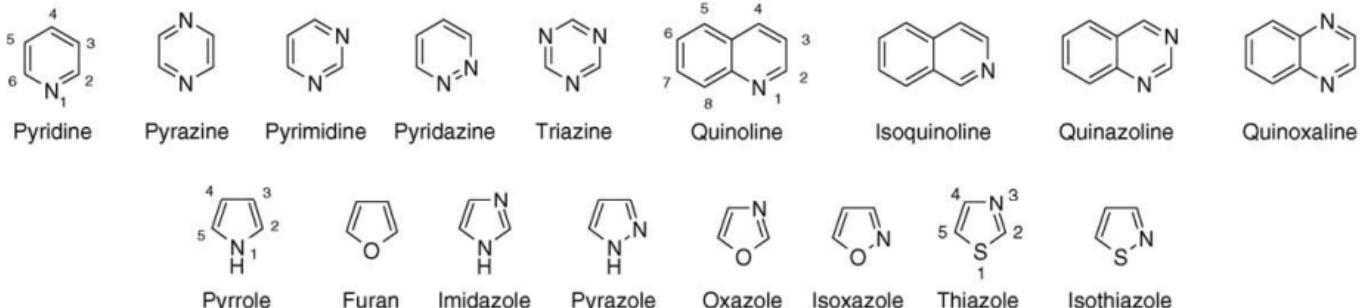
- Most drugs can be described as a “core scaffold” decorated with “functional groups”
 - “Functional groups” are atoms or groups of atoms which give a molecule its characteristic properties
 - The “core scaffold” is a usually flat, rigid structure that these groups are arranged around.



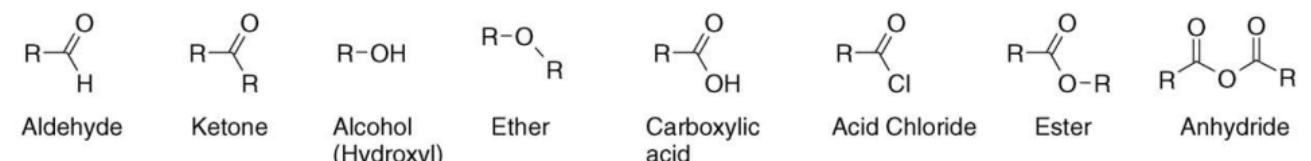
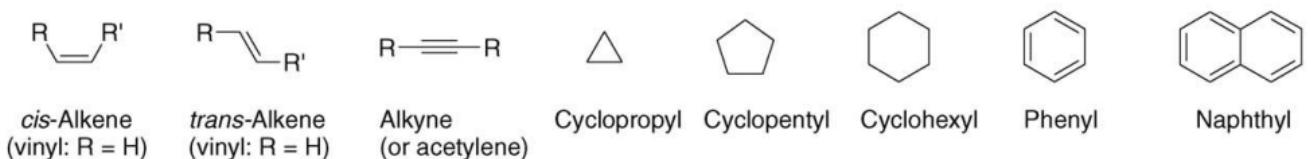
Sildenafil
Viagra®
Pfizer
March 1998



Vardenafil
Levitra®
Bayer
August 2003



Example core scaffolds.



Example functional groups.

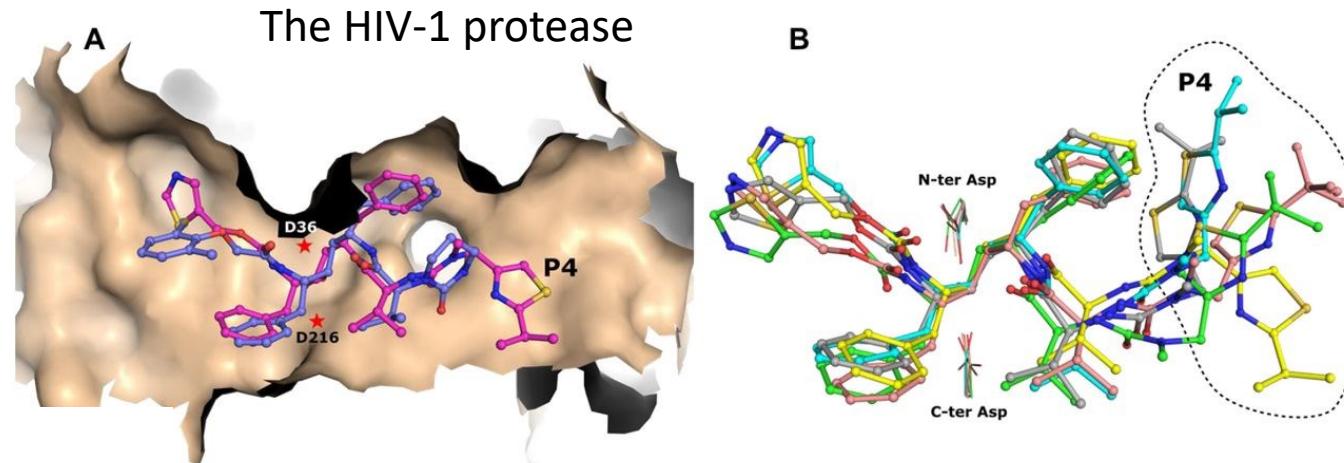


THRUST 1

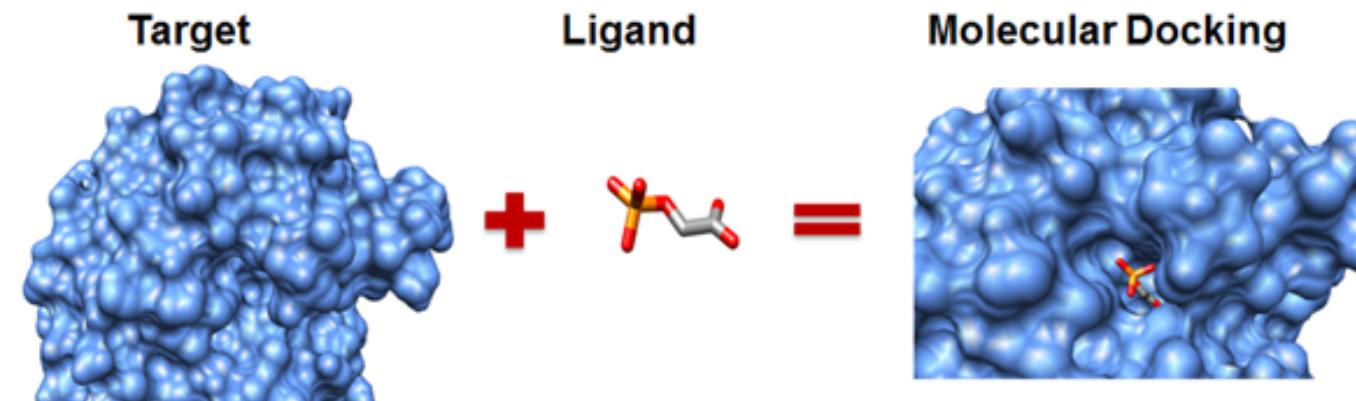
Interacting with a target



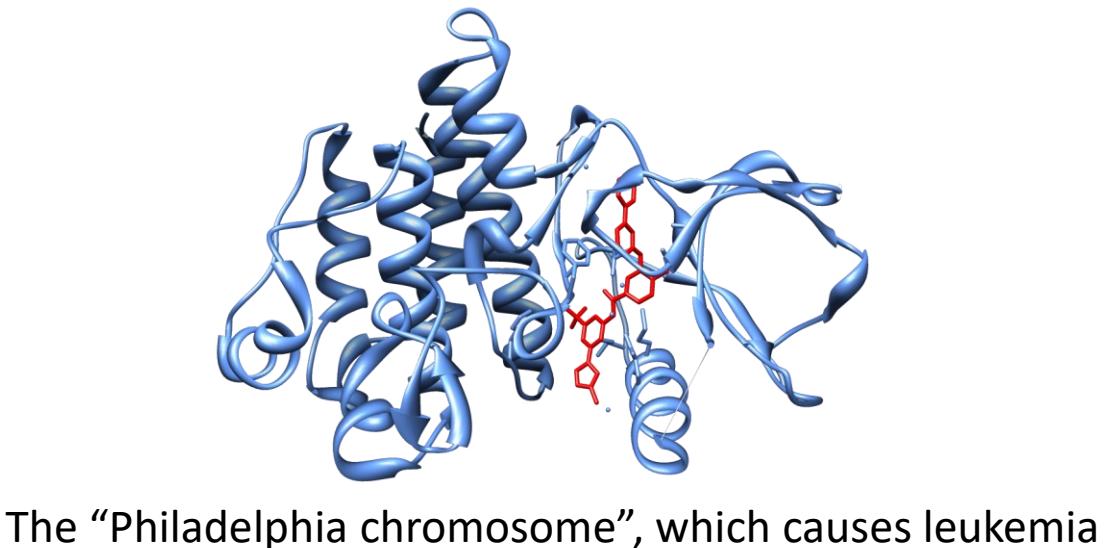
- We want to find a drug (ligand) which will interact with a target protein to modify its behavior.
 - This can include disrupting, increasing, or decreasing its behavior



Ref: bioRxiv 10.1101/2021.09.27.461917



Ref: en-lifesci.tau.ac.il



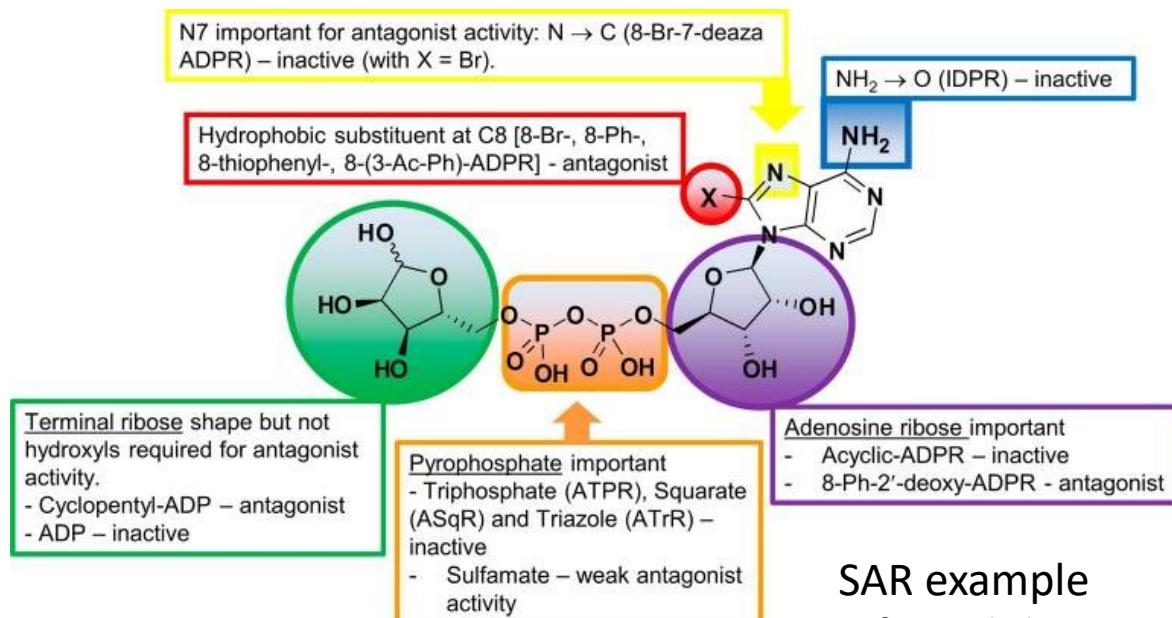


THRUST 1

Pharmacophores and structure-activity relationships

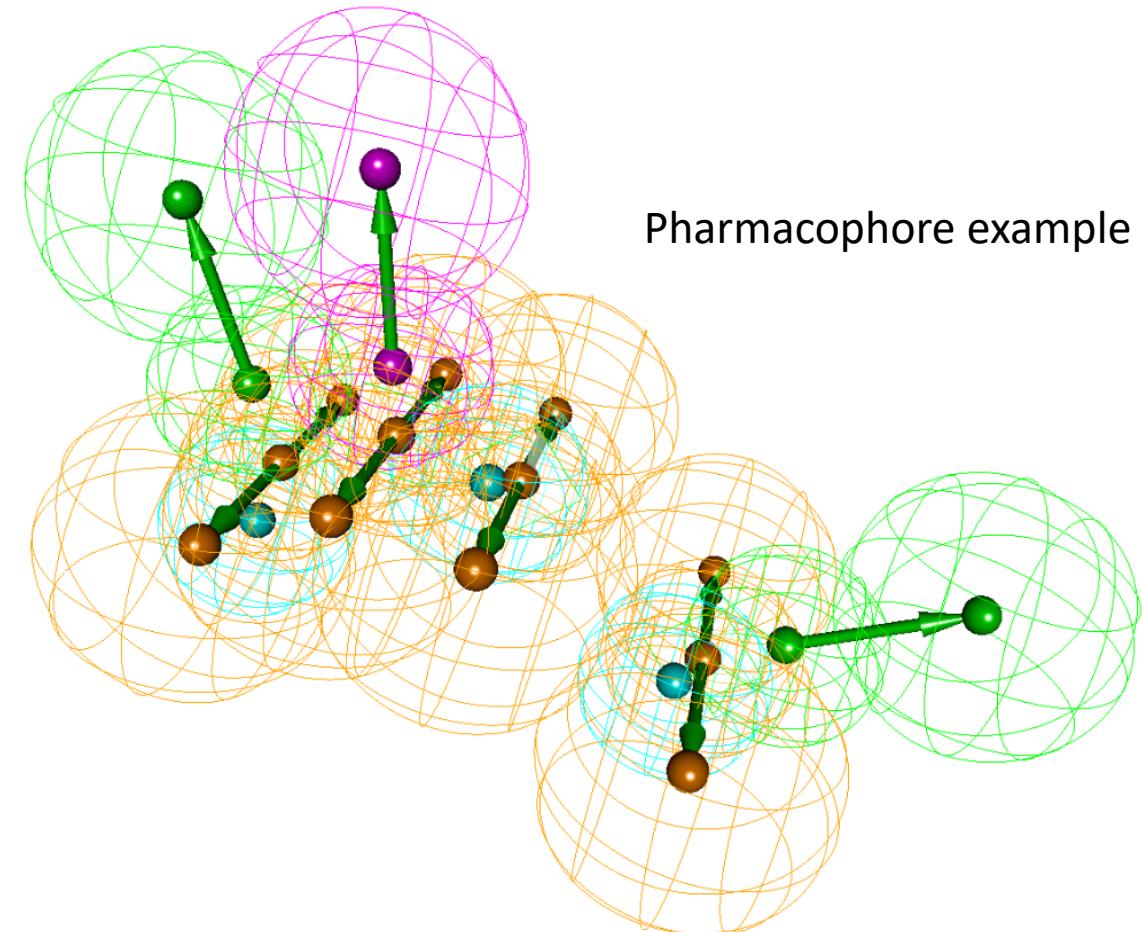


- A “pharmacophore” is the spatial orientation of functional groups necessary for activity at a biomolecular target
- “Structure-activity relationship” (SAR) is the correlation between certain structures or groups with a certain biological activity.



SAR example

Ref: J Med Chem. 2013 Dec 27; 56(24): 10079–10102





THRUST 1

Toxicophores: an example



- Toxicophores are substructures which are related to toxic properties in molecules, such as mutagenicity.
- Given a candidate drug, we might want to replace one of these substructures with something else.

Toxicophore name	Substructure representation	Example compound
aromatic nitro	$\text{O}_2\text{N}^+ \text{O}^-$ aro	
aromatic amine	NH_2 aro	
three-membered heterocycle	$\text{NH}_2\text{O}, \text{S}$	
nitroso	$\text{O}=\text{N}$	
unsubstituted heteroatom-bonded heteroatom	NH_2OH N, O	
azo-type	$\text{N}=\text{N}$	
aliphatic halide	$\text{Cl}, \text{Br}, \text{I}$	
polycyclic aromatic system	aro aro arom. rings arom. rings	



THRUST 1

(Some) Important properties to consider in drug discovery



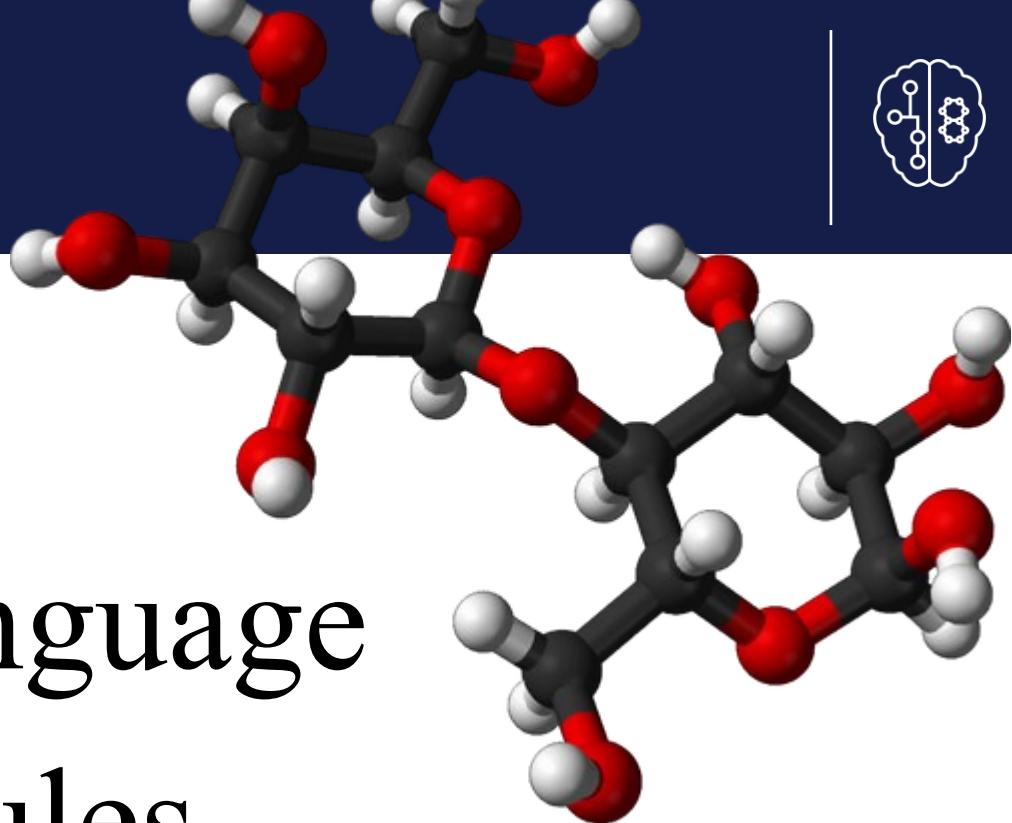
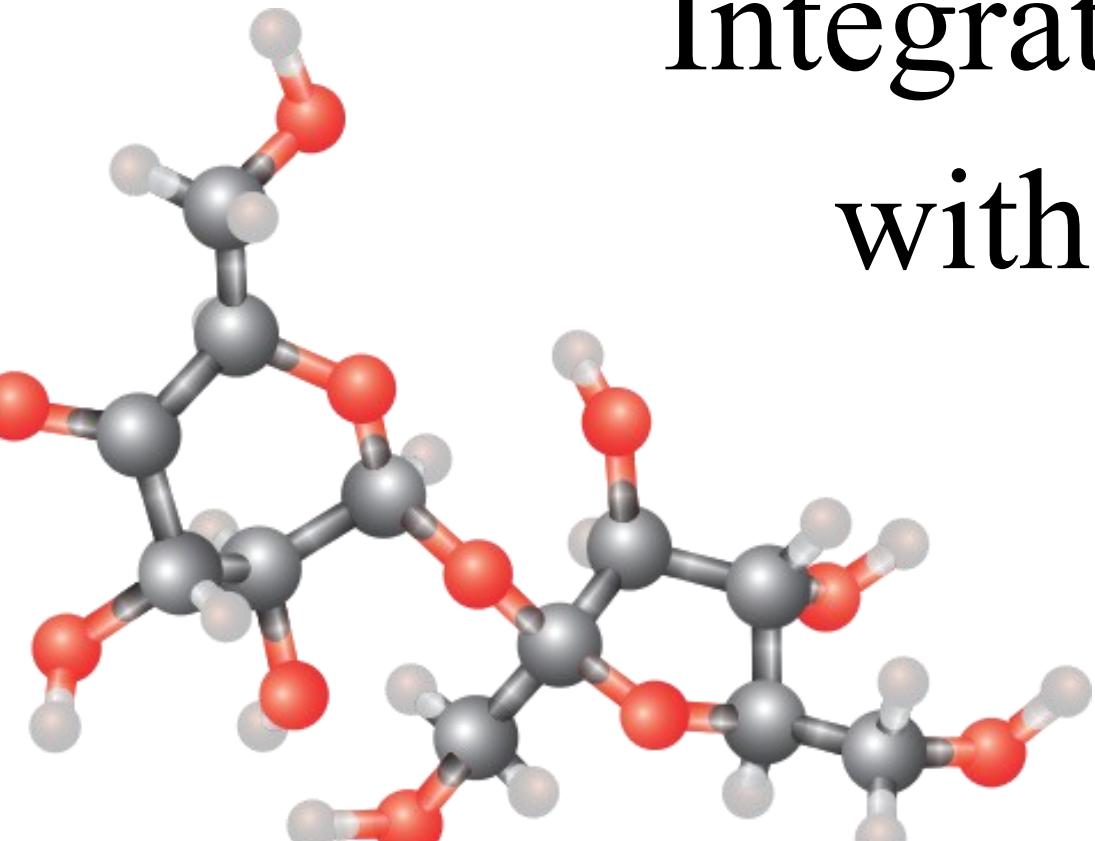
- Number of hydrogen bond donors
- Number of hydrogen bond acceptors
- Molecular weight
- Lipophilicity
- Solubility
- Molar refractivity
- Number of atoms
- Bioavailability
- Blood brain barrier penetration
- Toxicity
- Target selectivity
- ADME
 - Adsorption
 - Distribution
 - Metabolism
 - Excretion
- ...



THRUST 1



Integrating Language with Molecules



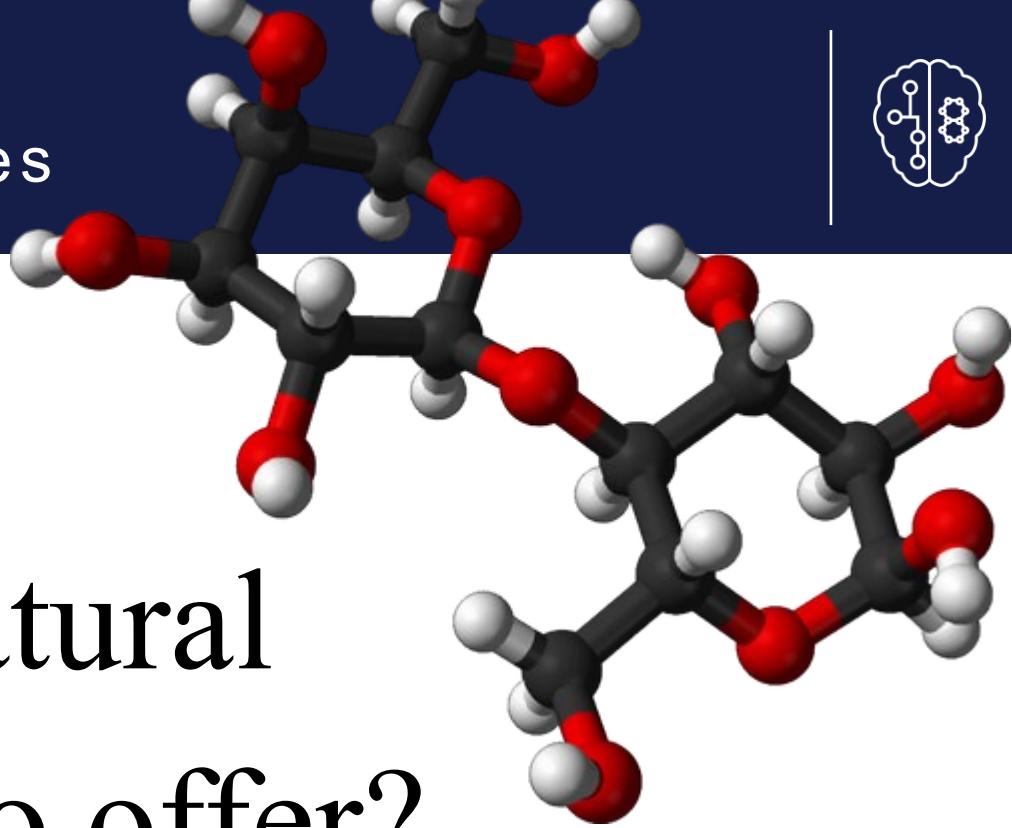
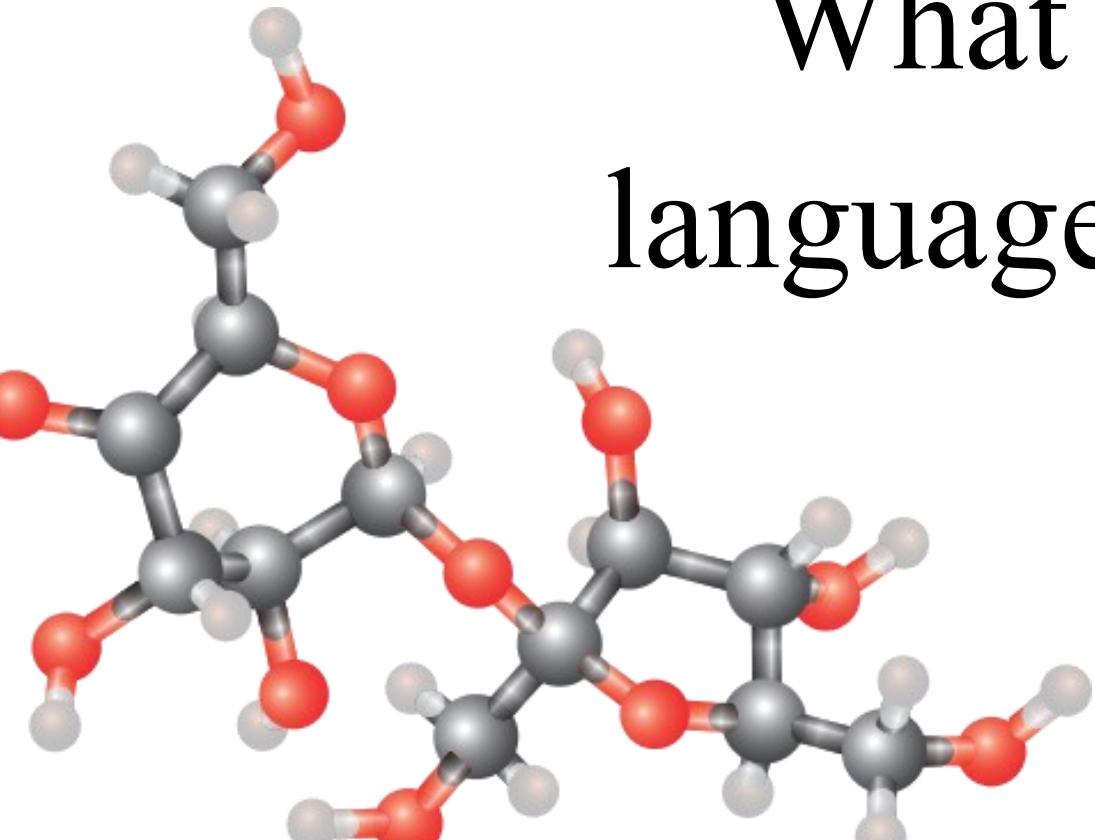


THRUST 1

Integrating Language with Molecules



What does natural
language have to offer?





THRUST 1

Generative Modeling



- One of the largest problems in current LLMs—hallucination---becomes a strength for discovering molecules with high-level functions, abstract properties, and composition of many properties.
- Can we achieve the power and flexibility of language-guided image generation on other modalities which can have large real-world impact?

“Generate a drug that binds with the MYC protein and describe ...”

“Does [chemical structure] satisfy the criteria of ...?”



“ [Chemical structure] might be an effective treatment for malaria..”



DALLE-3’s take on “A cat holding a flaming sword riding a unicorn on the moon, photorealistic.”

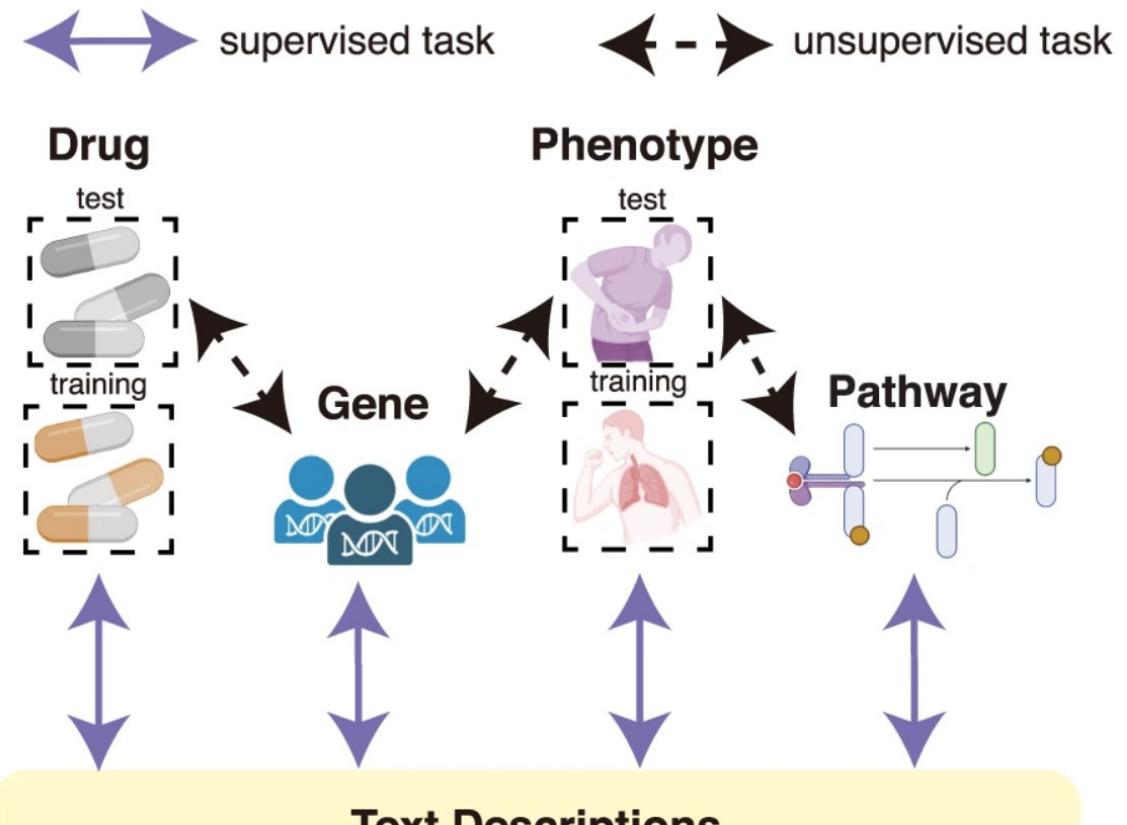


THRUST 1

Bridging Modalities



- One of the biggest problems in NLP research is getting the enough data to train your model to do what you want.
 - This is especially true in the scientific domain where low-data scenarios are common
 - e.g., enzymes for a reaction or drugs which cure a disease
- Can we use language to connect these different types of data— and what they do— together?
 - For example, connecting cellular pathways and the drugs the change how they work.





THRUST 1

Domain Understanding

- How can we expect language models to understand the real-world without being exposed to it?
- Grounding language models into external real-world knowledge can improve understanding of unseen molecules and help justify proposed molecules across many tasks.





THRUST 1

Automation



- Instruction-following, dialogue-capable, and tool-equipped models can guide automated discovery in silico and in robotic labs.
- These models can also be used to reason about and discover new chemical knowledge.

Q: What are the top three catalysts for the reverse water gas reaction? Let's think step-by-step...

A: Pt, Pd, Ru. Because...

Q: What are the top three catalysts like Pt, Pd, Ru for the RWGS reaction? Include catalysts with [low cost].

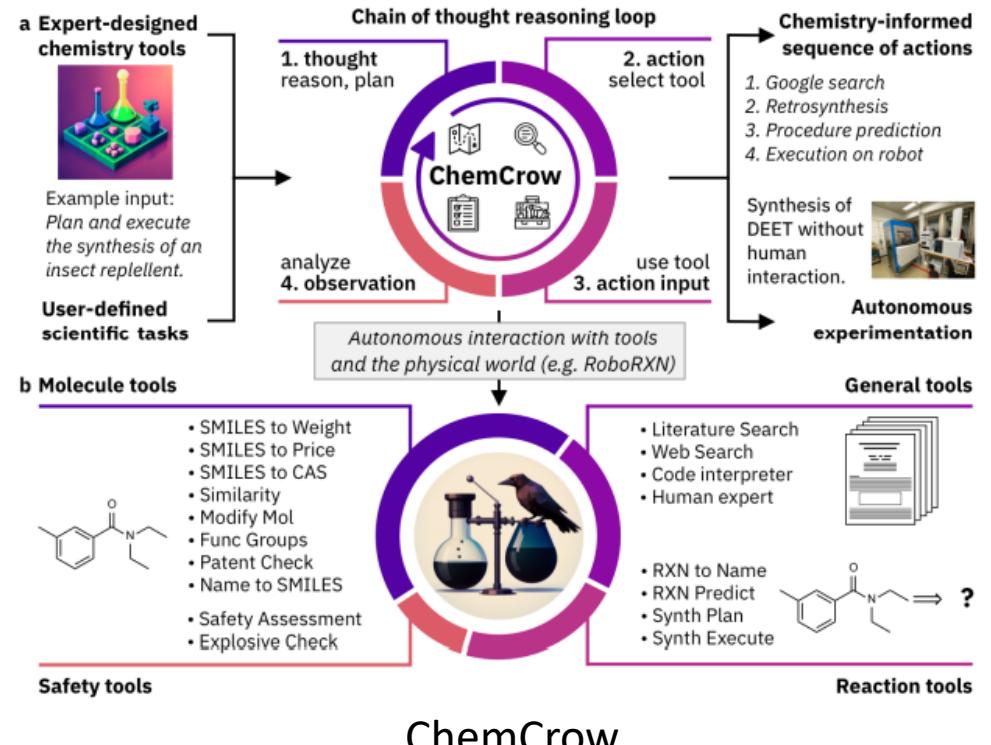
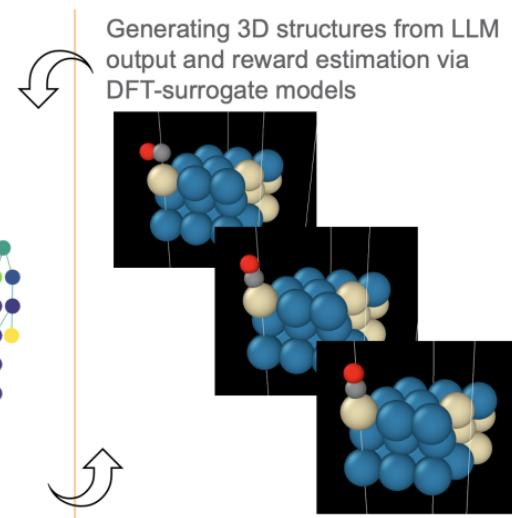
A: Cu, Ni, Co. Because...

Q: Provide three metal oxide catalysts that include Cu, Ni, Co for the reverse water gas reaction. Consider catalysts with [low cost, high activity].

A: NiO, NiCuO, Co₃O₄. Because...

Explore the chemical space via LLM-based Heuristic Search methods

ChemReasoner

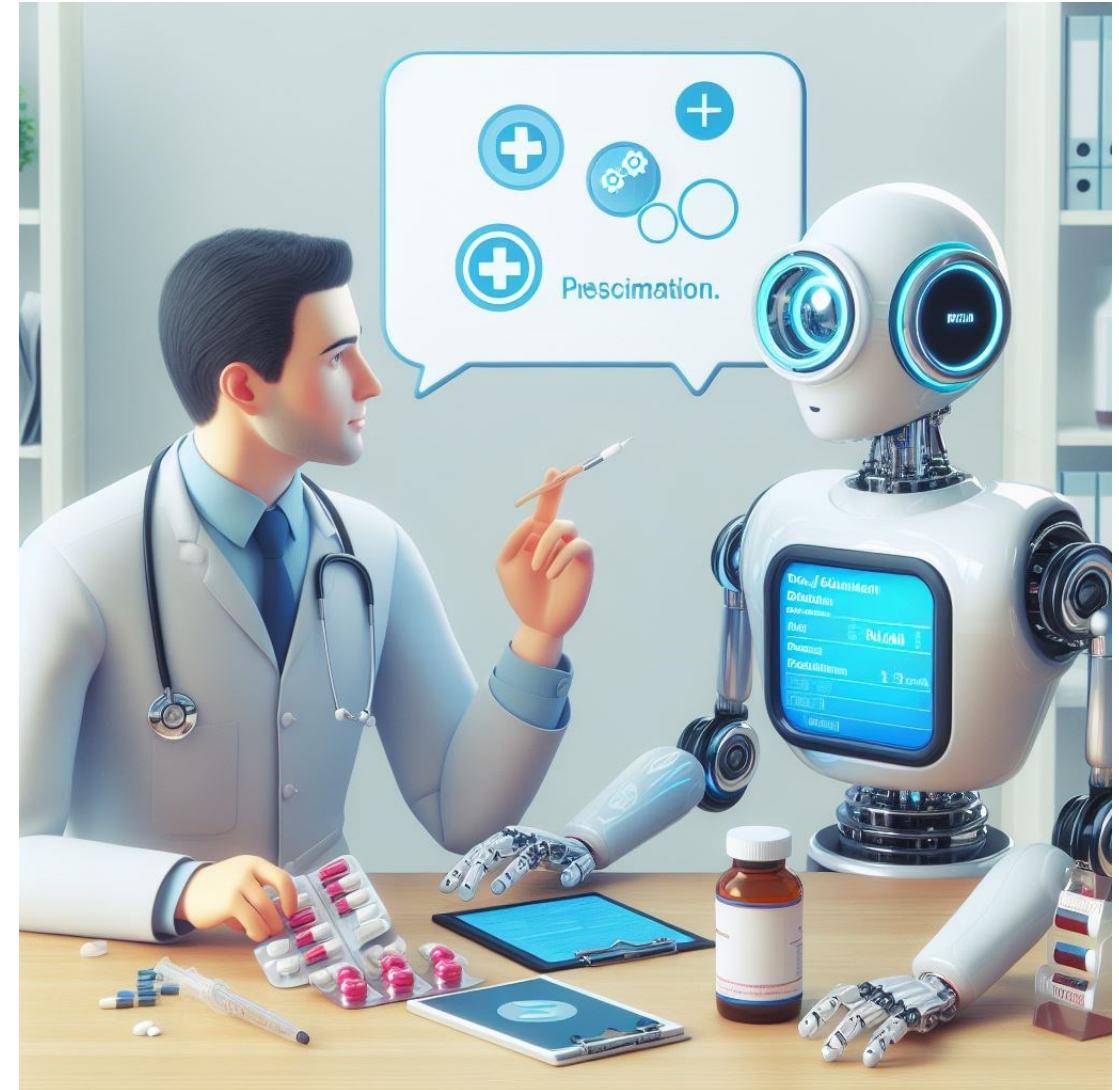




THRUST 1

Democratization

- Language enables scientists without computational expertise to leverage advances in scientific AI.
 - Imagine a future where a doctor can write a few sentences describing a patient's symptoms and then receive exact structure of the drugs necessary to treat that patient's ailment (taking into account the patient's genotype, phenotype, and medical history)





THRUST 1

Language for controlling and interfacing with chemistry (the goal)



1. Language can enable abstract, functional, and compositional control over complex properties when designing novel molecules .
2. Language can serve as a “bridge” between modalities
 - (e.g., cellular pathways and drugs).
3. Tool-enabled language models hold promise for chemical reasoning and even directing laboratory experiments.
4. Language makes chemistry AI more accessible.

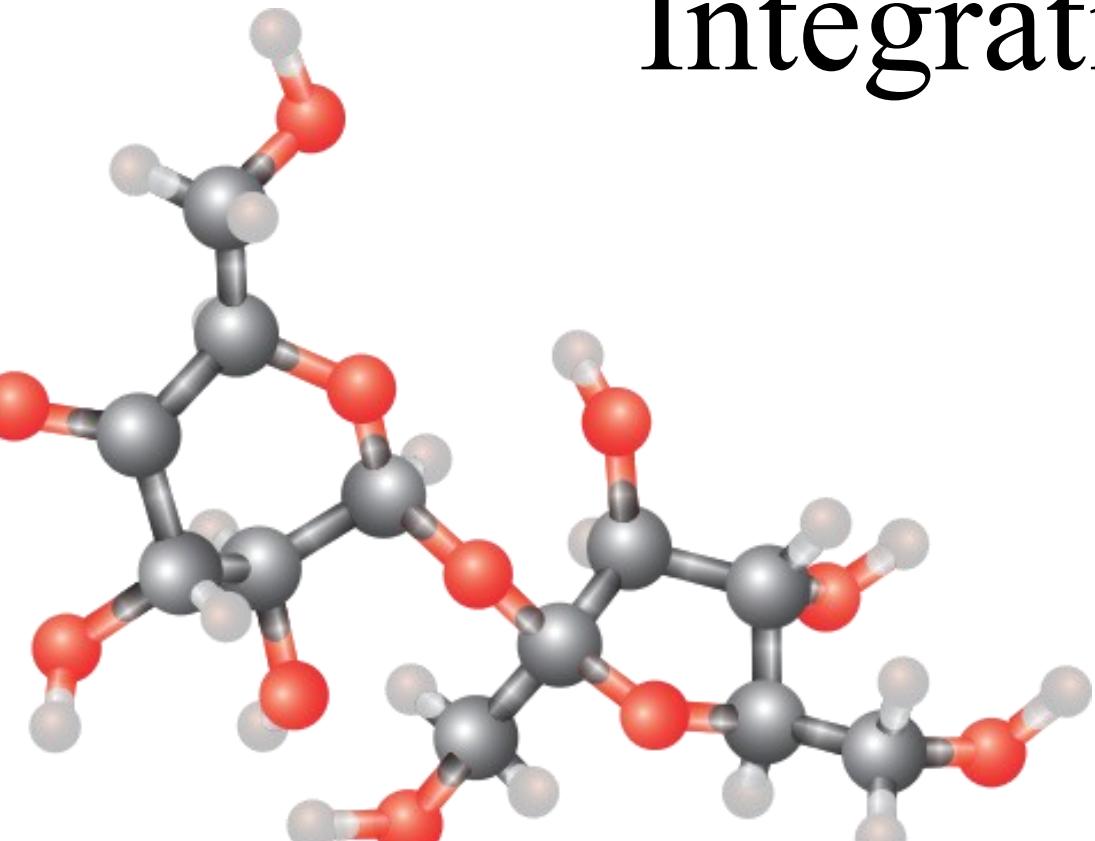
Language is a glue—between data types, robots, and people.

Language has been developed as the method by and for humans to abstractly reason about the world. In much the same way that science often relies on natural phenomenon (e.g., penicillin) for innovation, we can rely on natural linguistic phenomenon for abstraction and connection.

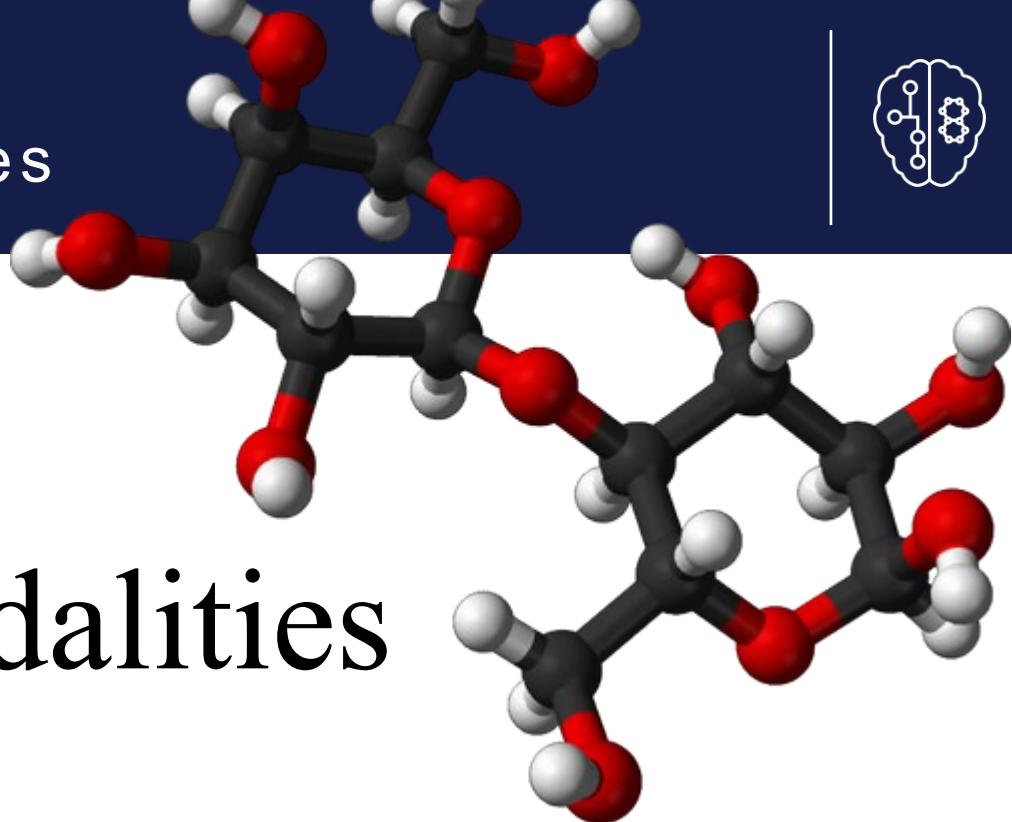


THRUST 1

Integrating Language with Molecules



Integrating Modalities





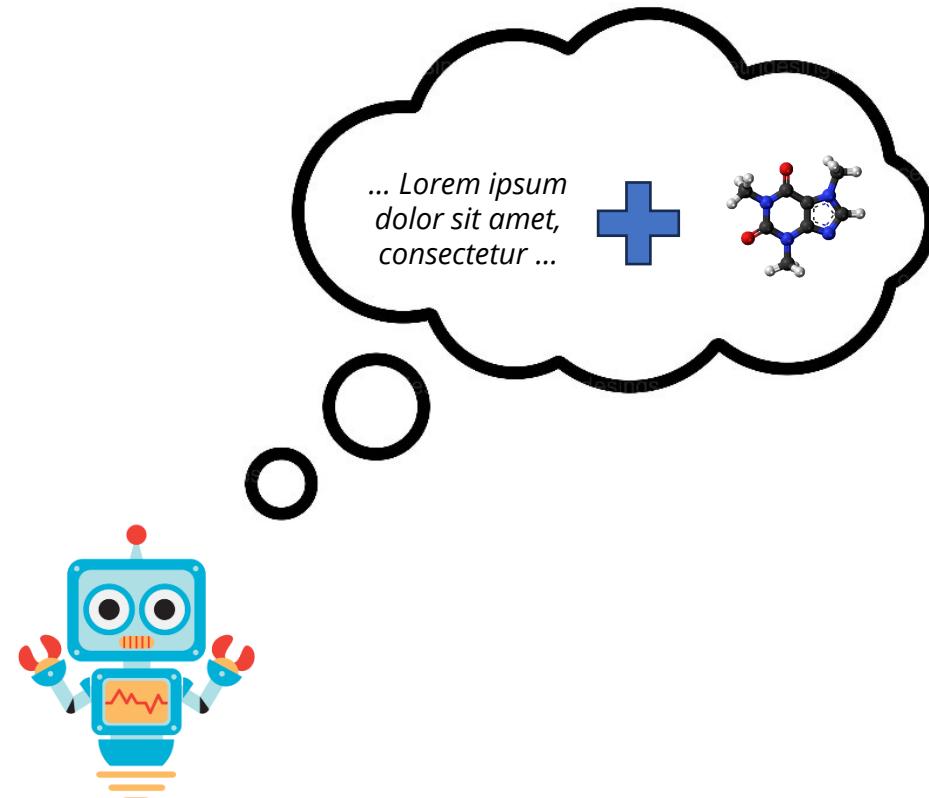
Do I want multimodality?



- Before we dive into molecule-language multimodality, do we even need it?
- Generally, given sufficient data for training, adding modalities to language will likely be helpful simply by grounding the model's understanding into the real world.
 - However, obtaining multimodal data can be challenging in practice.

Here are some general rules of thumb we can follow:

1. Is multimodality a core part of my task, such as molecule captioning?
2. Should I add language to typically molecule-only tasks? In other words, do I need the level of control and abstraction offered by natural language; or is there complementary information available as text?
3. Will I meaningfully benefit from anything beyond language, or is all the information I need expressed as text?





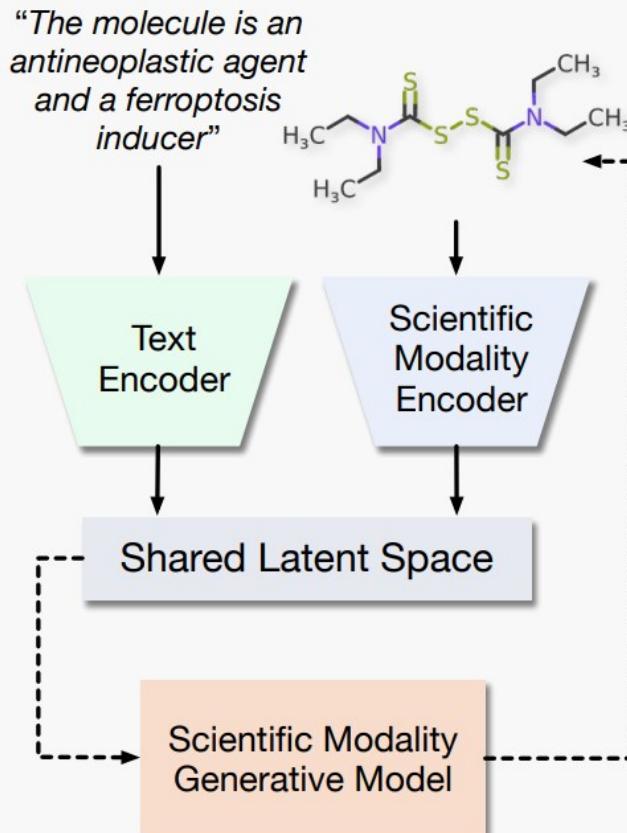
THRUST 1

Types of Models

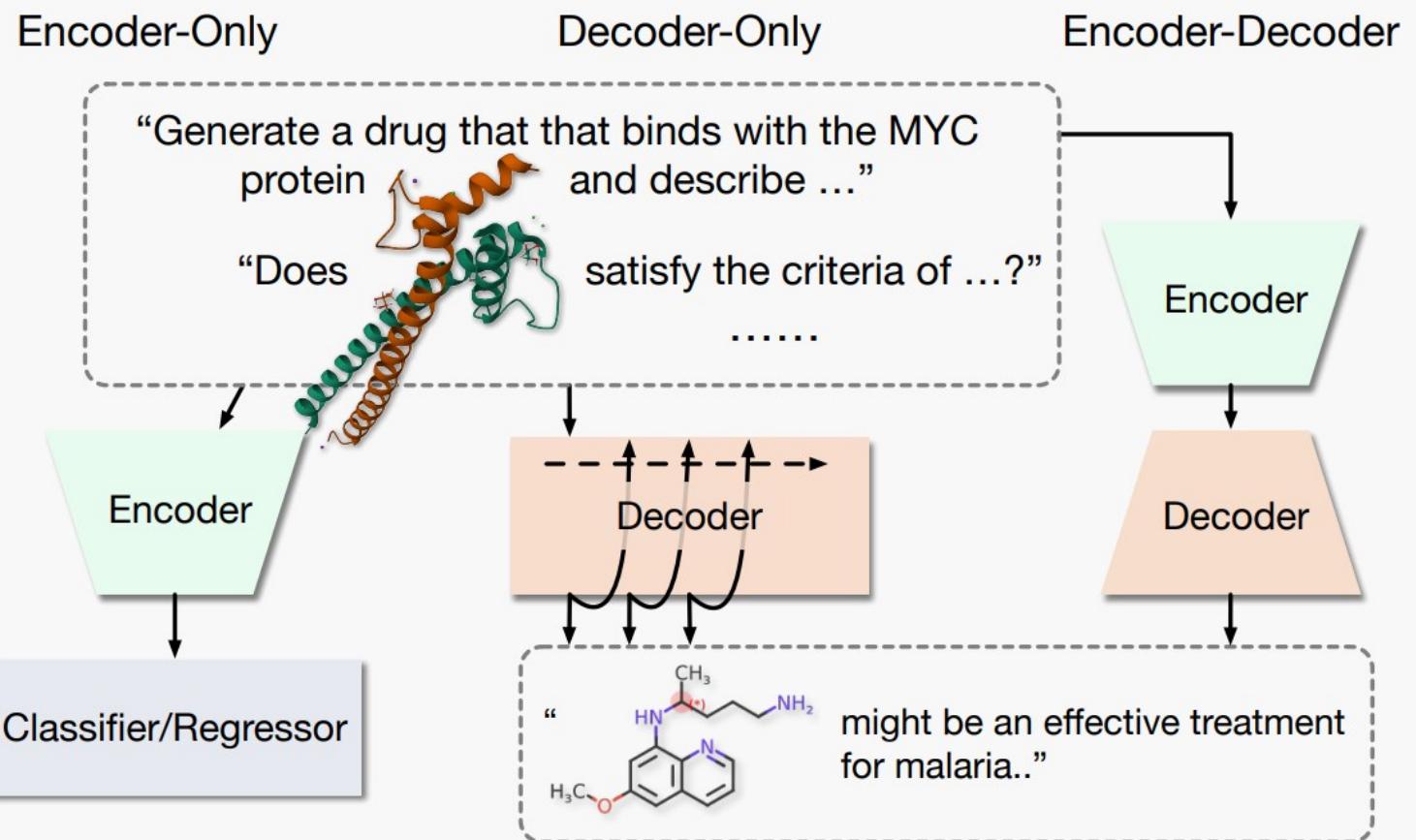


Multimodal LLMs for Science

Bi-Encoder Models



Joint-Representation Models



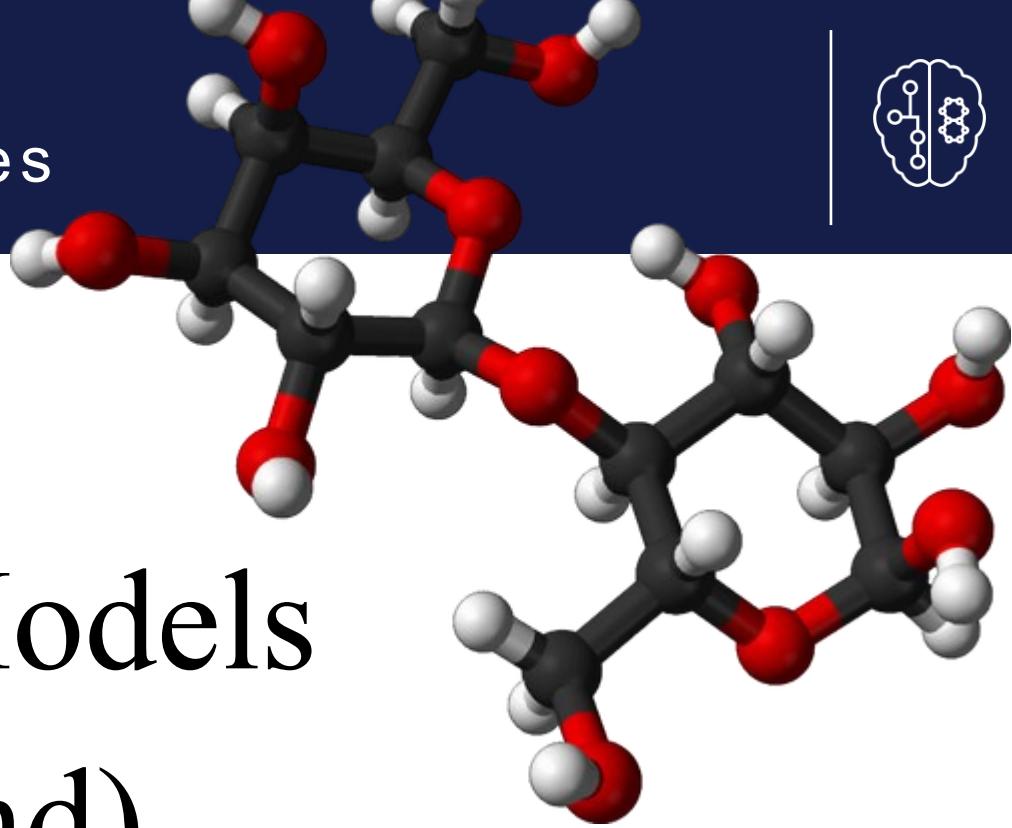
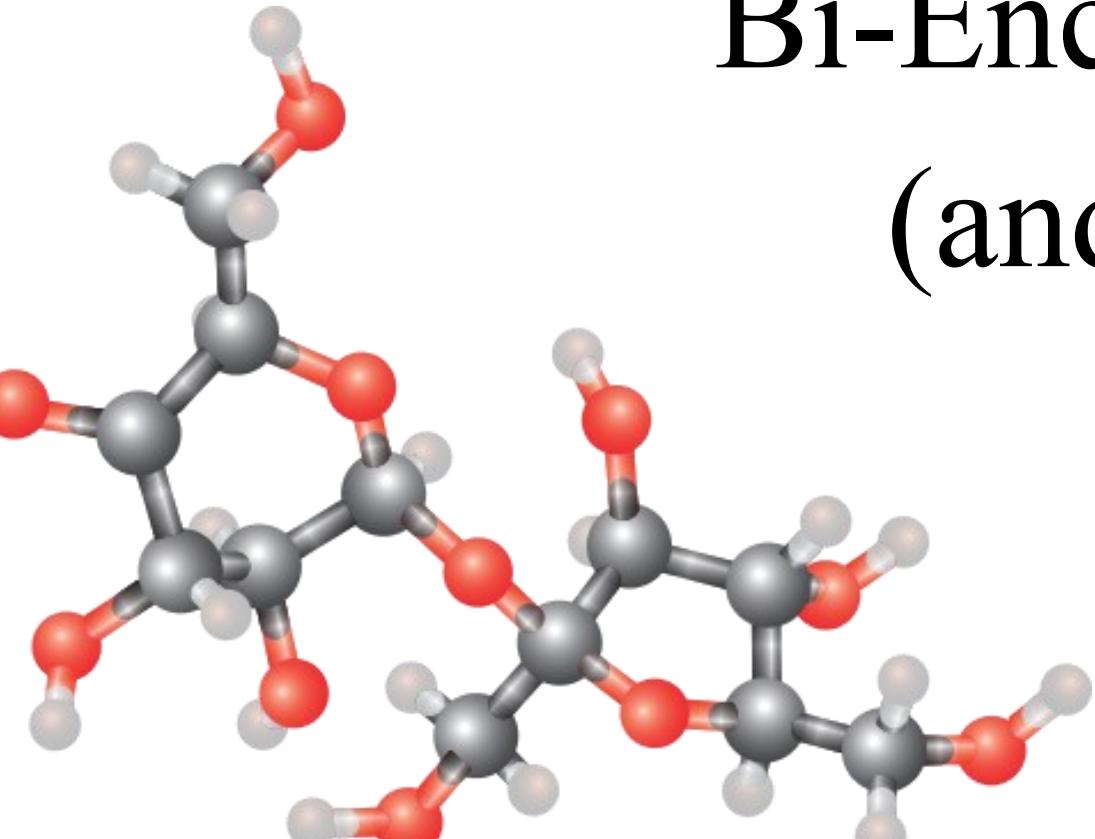


THRUST 1

Integrating Language with Molecules



Bi-Encoder Models (and Beyond)



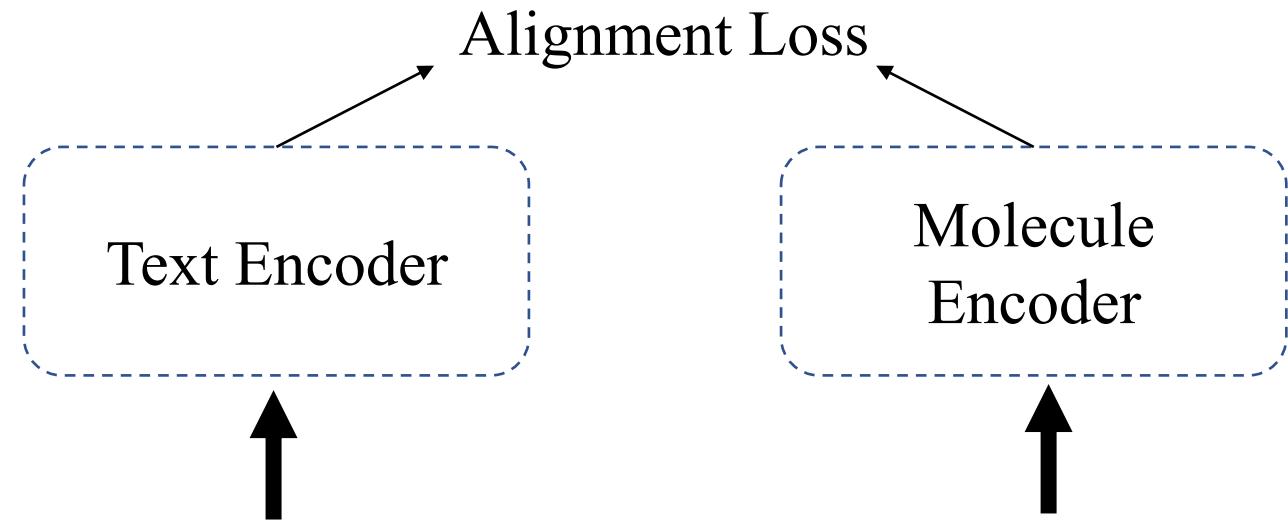
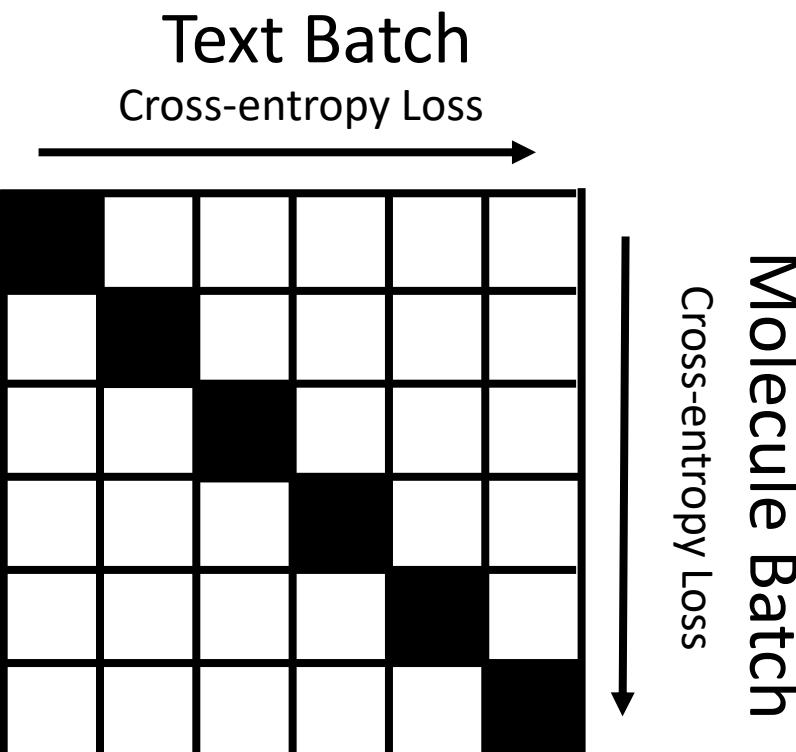


THRUST 1

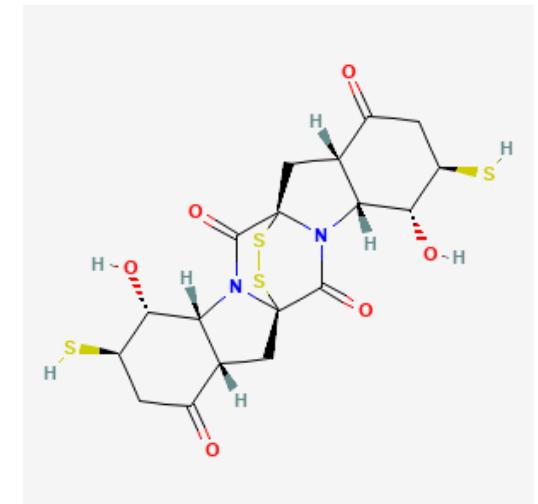
CONNECTING LANGUAGE AND MOLECULE: BI-ENCODER MODELS



- Need to connect information in two very different modalities.
 - Bi-encoder solution: create aligned embedding space.



Rostratin D is an organic disulfide isolated from the whole broth of the marine-derived fungus *Exserohilum rostratum* and has been shown to exhibit antineoplastic activity. [...] It is a bridged compound, a cyclic ketone, a lactam, an organic disulfide, an organic heterohexacyclic compound, a secondary alcohol, a dithiol and a diol.





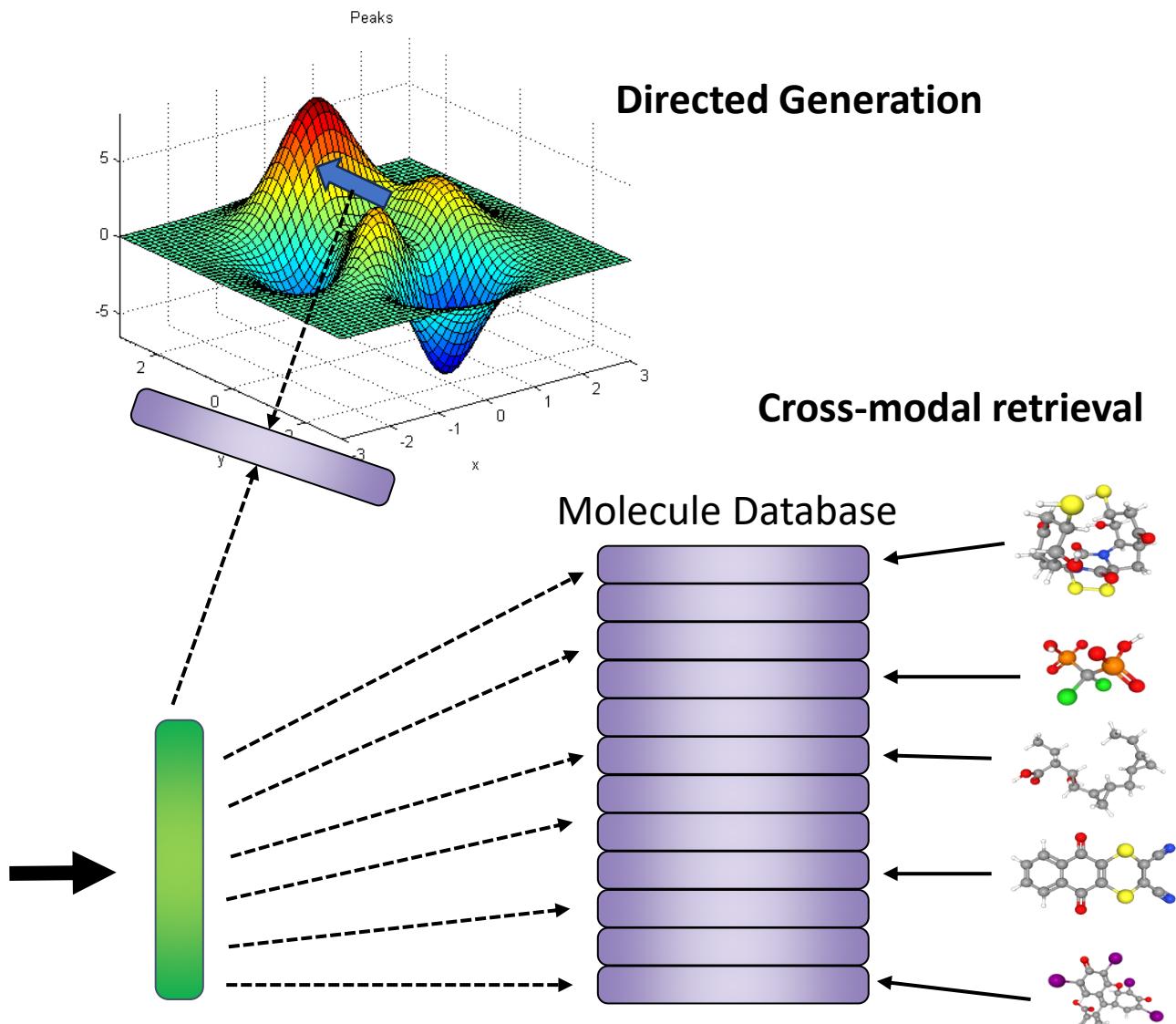
THRUST 1

BI-ENCODER MODELS



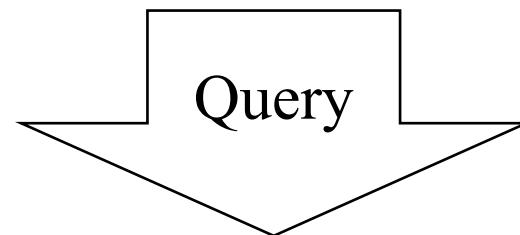
- Cross-modal retrieval (initial task)
 - (Text2Mol, Edwards et al. 2021)
- Integration into molecule and protein generation frameworks
 - (MoMu, Su et al. 2022)
 - (ProteinCLAP, Liu et al. 2023)
- Drug editing
 - (MoleculeSTM, Liu et al. 2022)
- Assay activity prediction
 - (CLAMP, Seidl et al. 2023)
- Many-modal representations between text, drugs, proteins, phenotypes, cellular pathways, and gene expressions.
 - (BioTranslator, Xu et al. 2023)
- New applications all the time!

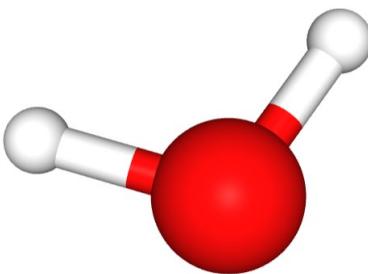
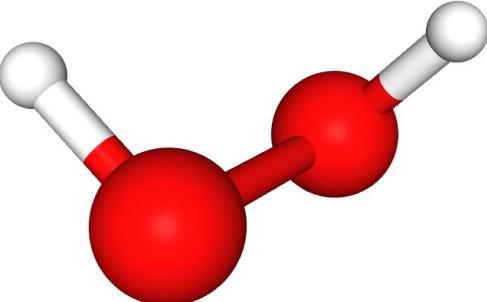
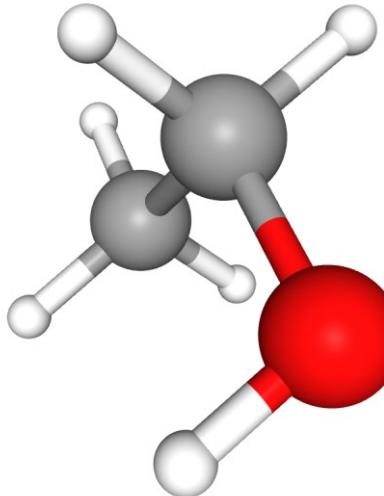
The molecule is an organic disulfide shown to exhibit antineoplastic activity. It has a role as [...]





Water is an oxygen hydride consisting of an oxygen atom that is covalently bonded to two hydrogen atoms.

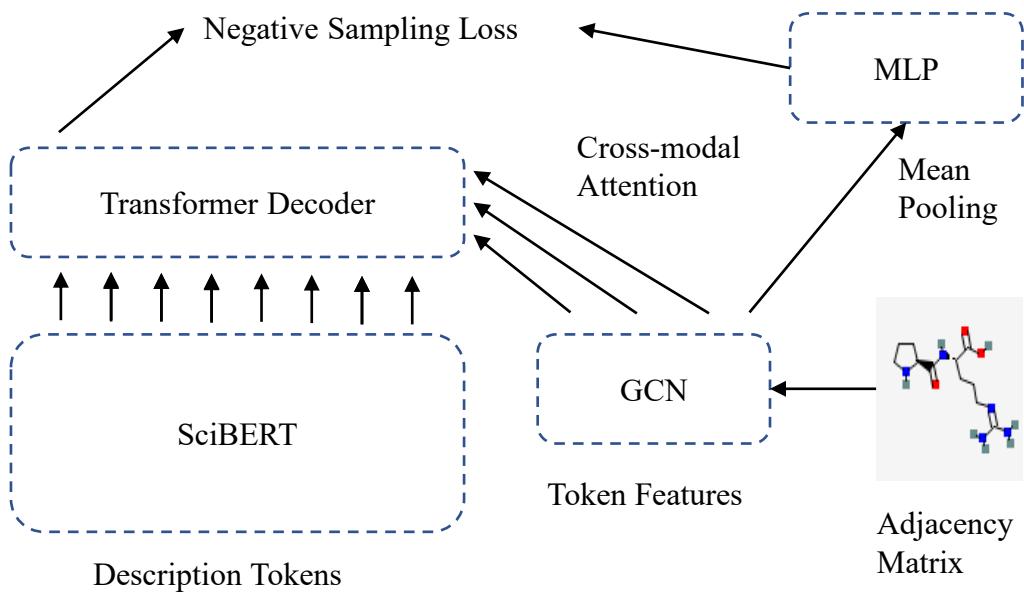


1. 
 H_2O
2. 
 H_2O_2
3. 
 $\text{C}_2\text{H}_6\text{O}$



THRUST 1

Text2mol: Learning associations from text-molecule data

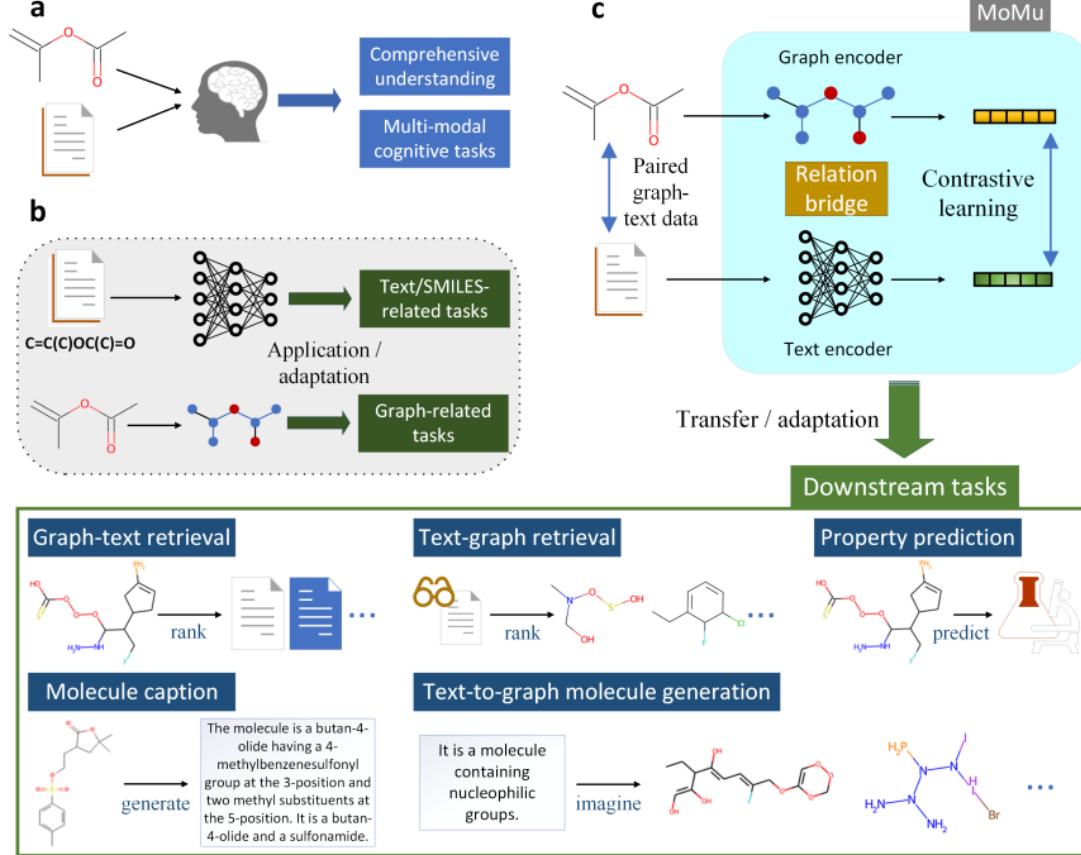


Token	Substructure	Supp	Conf
Titanium	Ti=O	1.29	0.65
Aluminium	Al ³⁺	4.31	0.23
Manganese	Mn ²⁺	10.08	0.30
Toluene	C – C=C	12.93	0.231
Toluene	C ₇ H ₈	23.79	0.425
##chloro	Cl – C	18.81	0.207
pollutant	F – C	3.097	0.208
chromatography	C – Si	2.976	0.271
acid	C – O – H	2398.7	0.078
crown	C – C – O	4.18	0.325

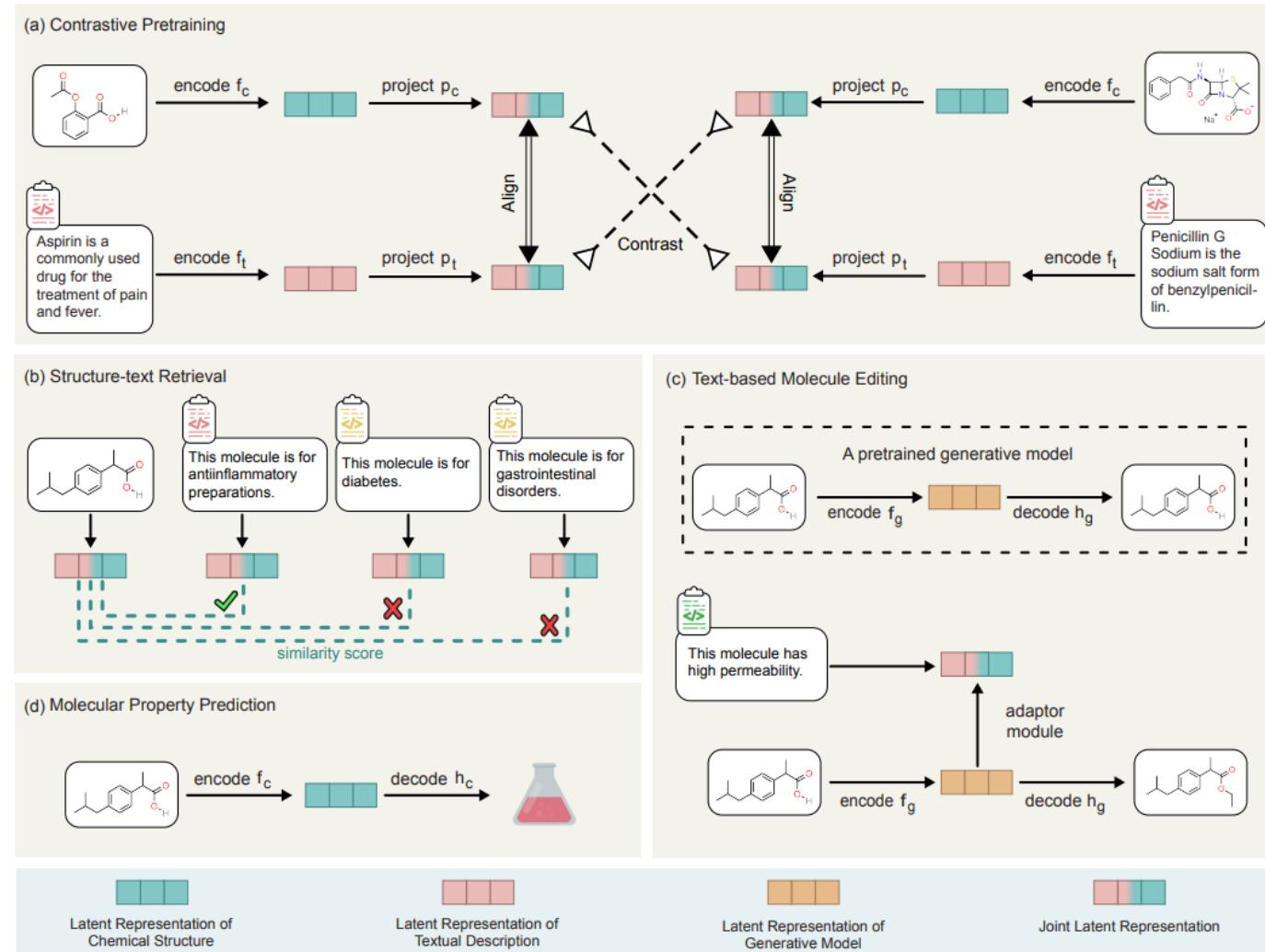


THRUST 1

MoMu and MoleculeSTM



Su et al. 22



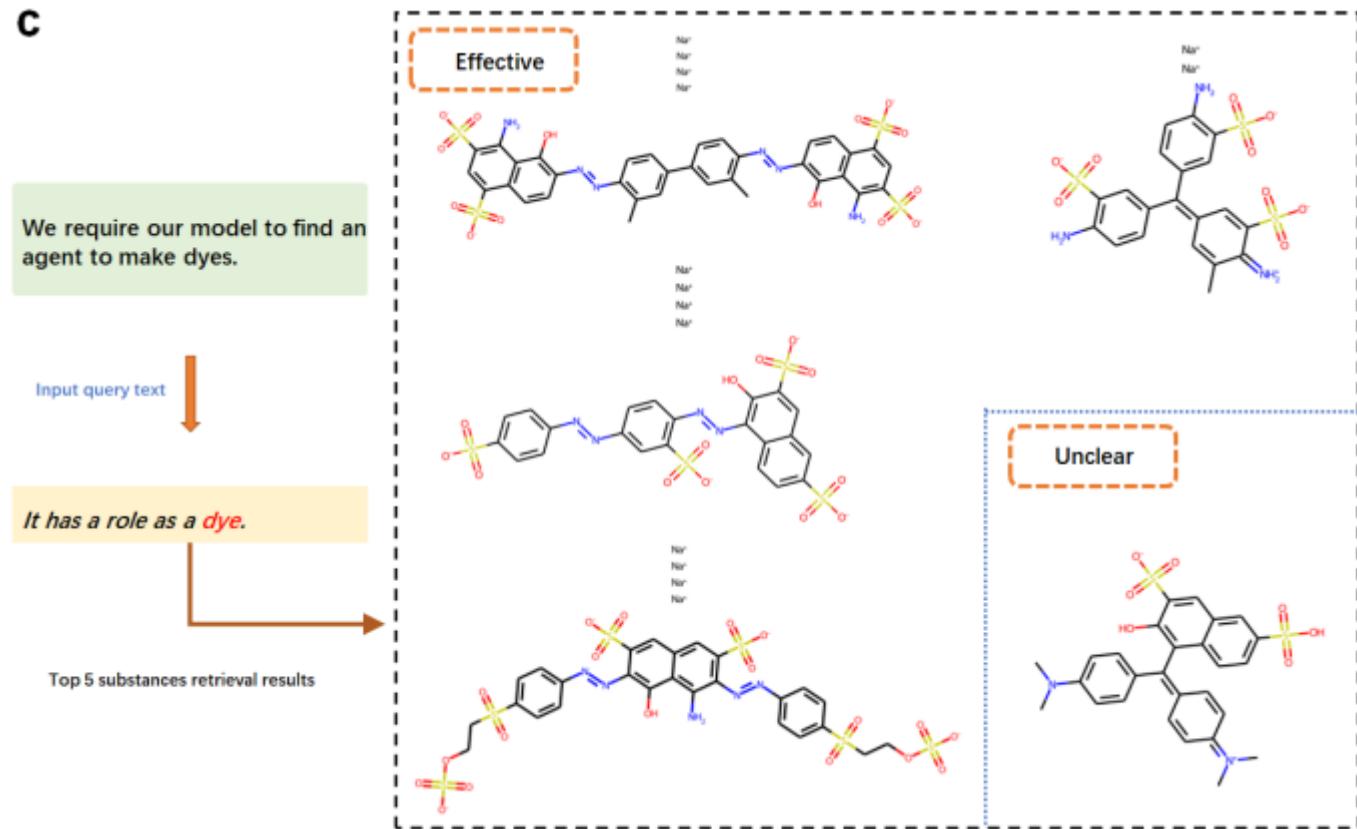
Liu et al. 22



Looking for Function (Retrieval, MoMu)

Key Problem: It's hard to evaluate how good the model is on unknown molecules

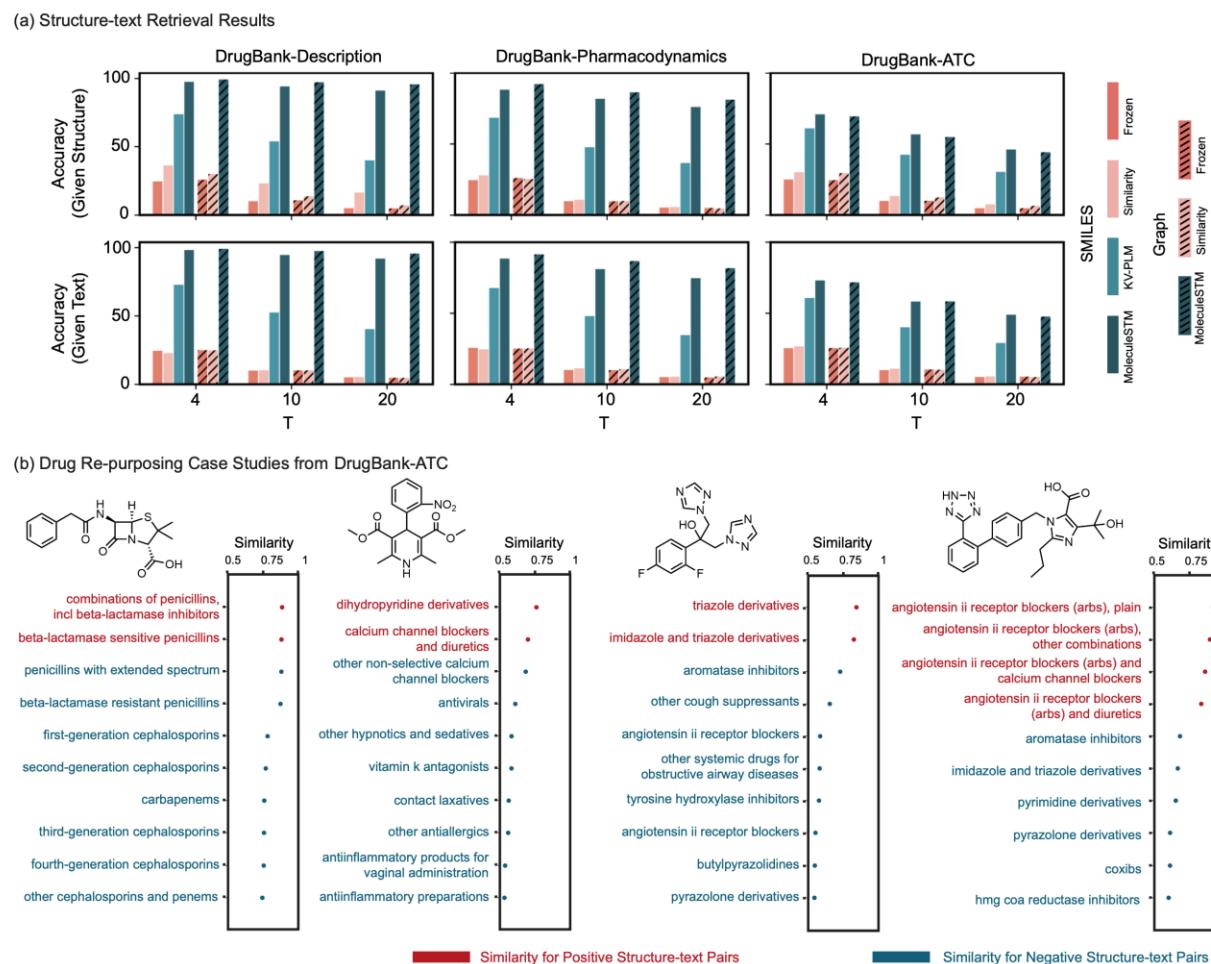
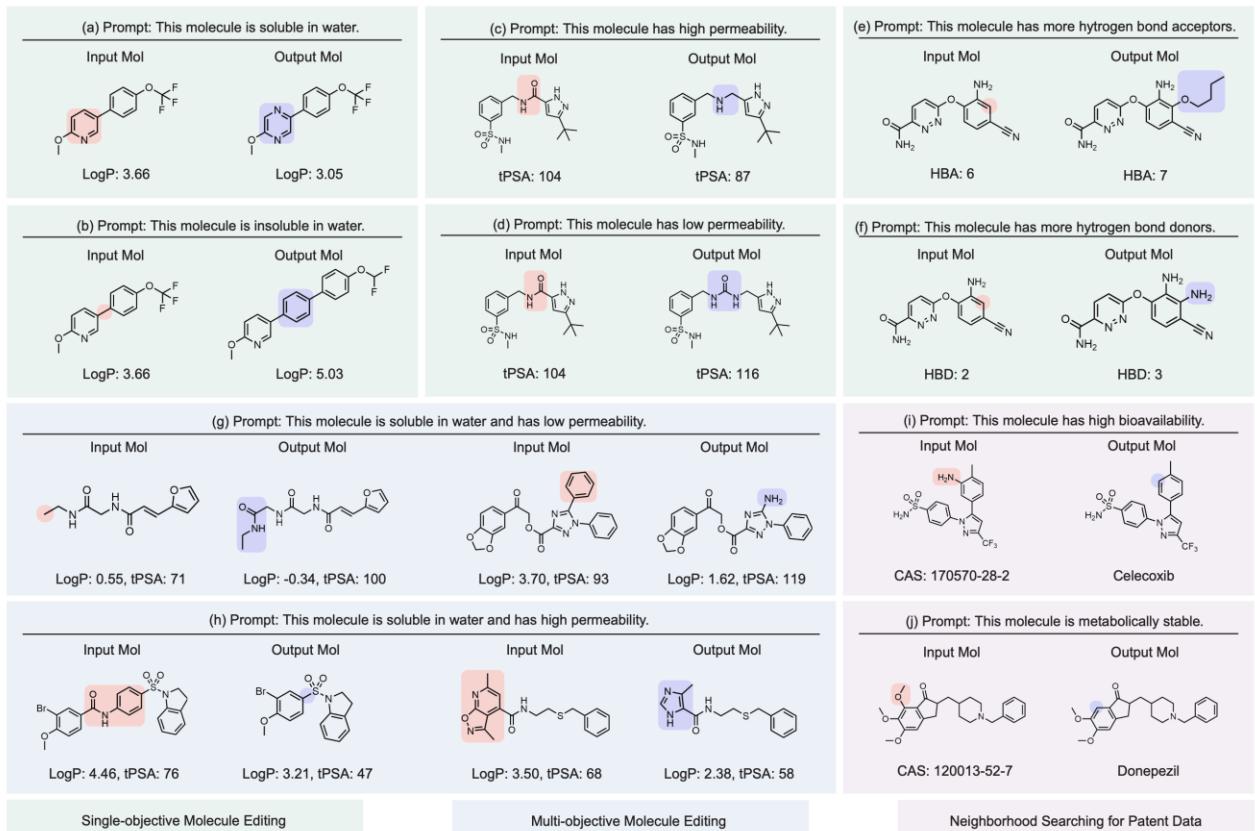
- This is both good and bad, but unlike images, it isn't easy for a human to check the results.





THRUST 1

Molecule Editing (MoleculeSTM)



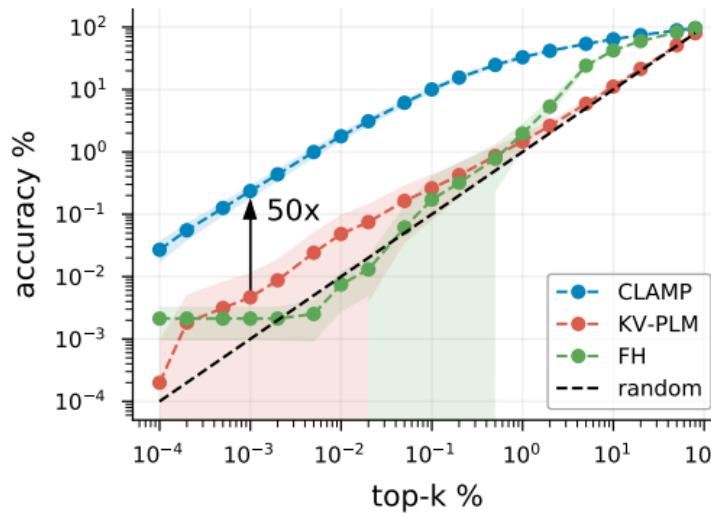


THRUST 1

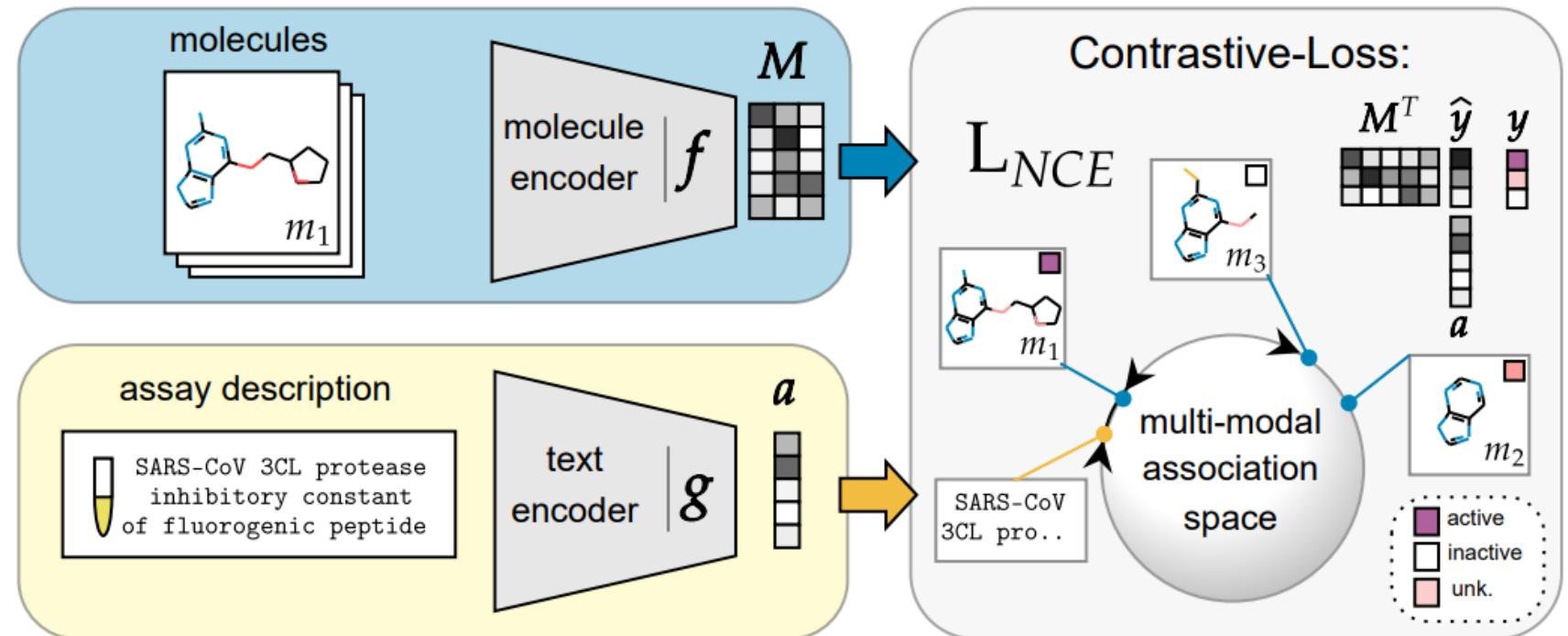
CLAMP (Seidl et al. 2023)



- Can we use the textual description of laboratory experiments to predict results for certain molecules?



Zero-shot retrieval results



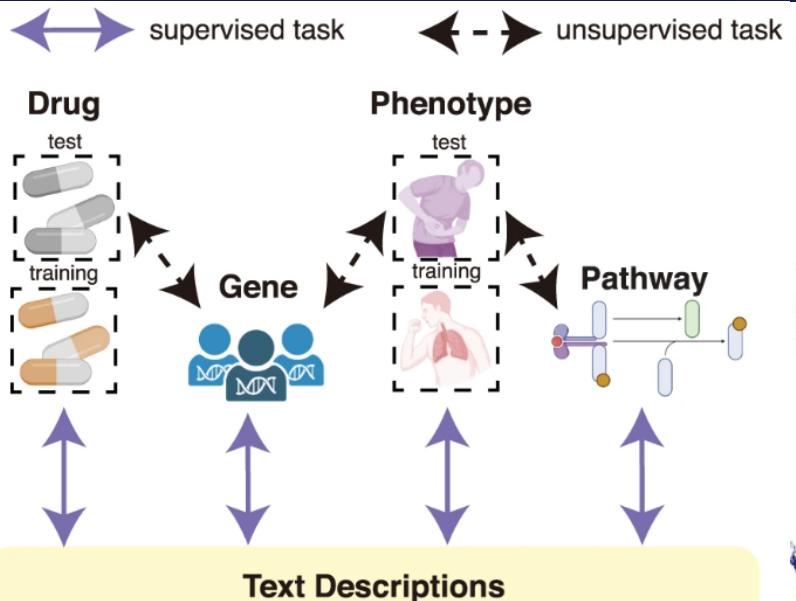
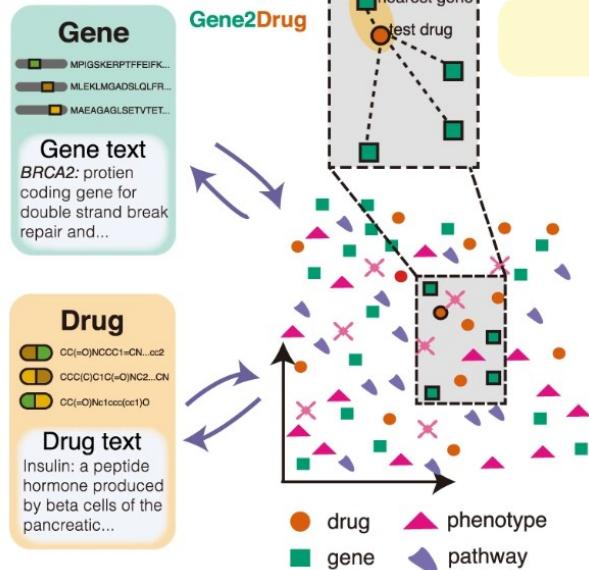


THRUST 1

BioTranslator



Learning a many-modal latent space with a contrastive loss.



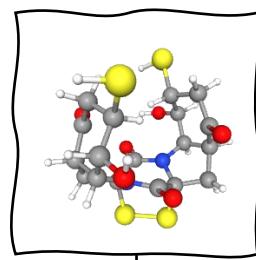
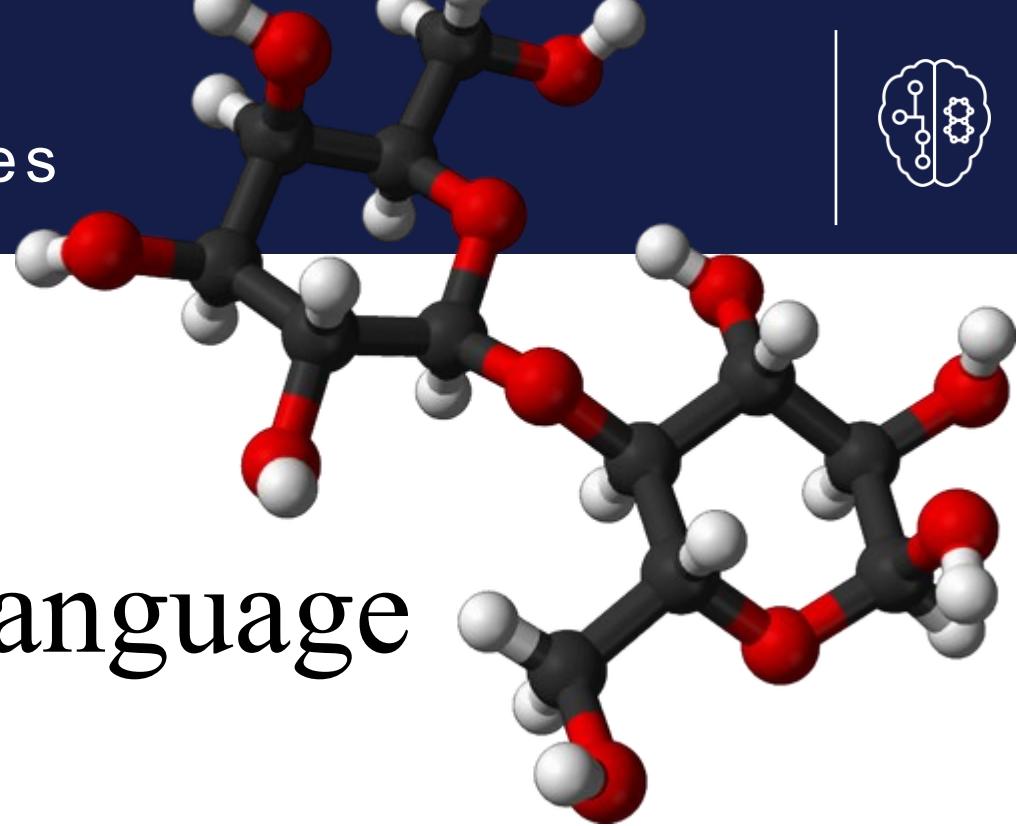


THRUST 1

Integrating Language with Molecules



Joint Molecule-Language Models



is

an

organic

disulfide

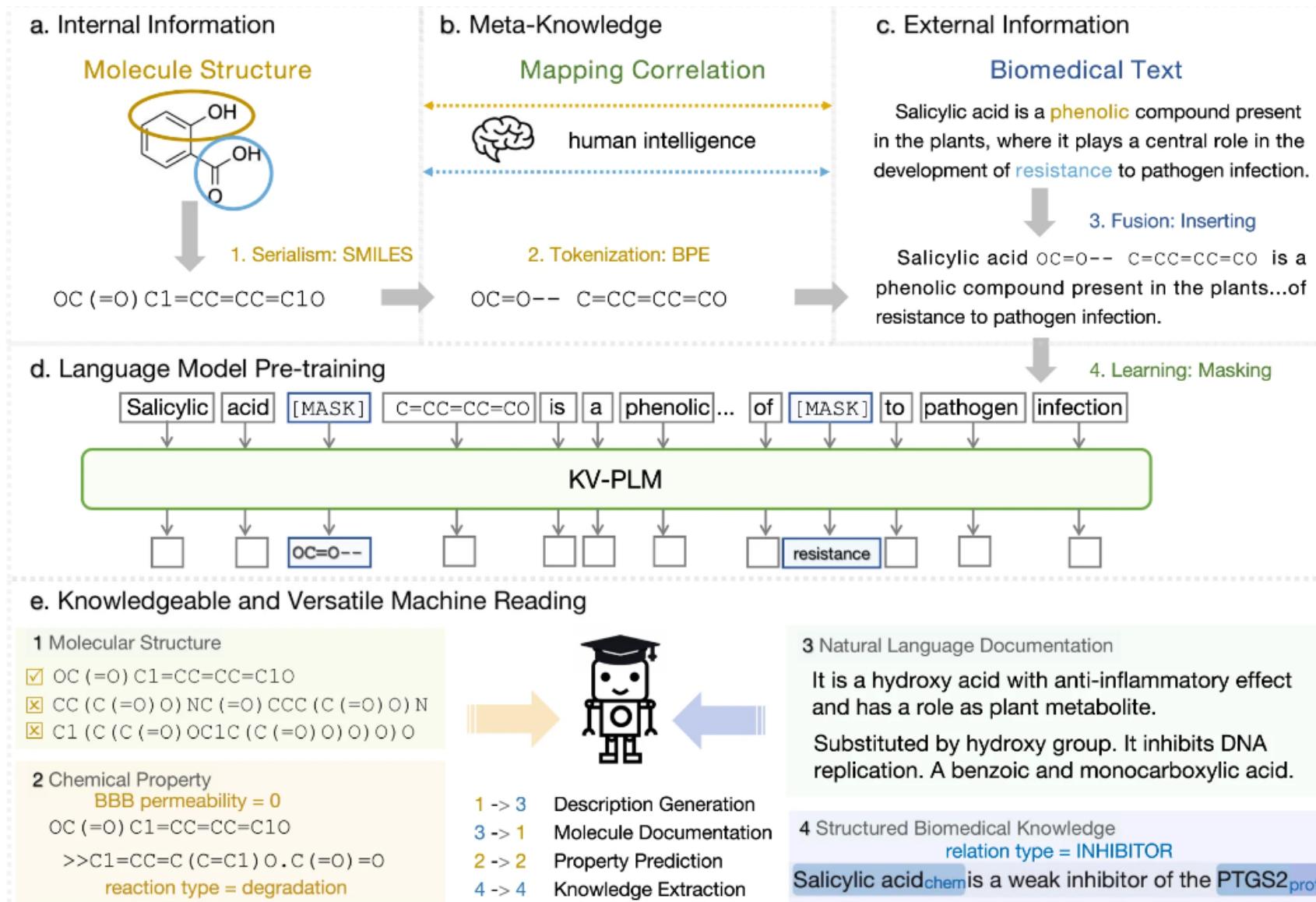
...

Large Language Model



THRUST 1

KV-PLM: An Encoder-Only Joint Model





THRUST 1

KV-PLM: Tasks



Molecular Structure	SMILES sub-groups	Property Prediction
	CO- C=CC=CC=C- C=C- N- C=O CO- C=CC=CC=C- C=C- N- C=O CO- C=CC=CC=C- C=C- N- C=O CO- C=CC=CC=C- C=C- N- C=O	colorless gas; toxic by ingestion; decomposed when heated; substituted by methyl group; an alcohol ; inorganic weakly acidic; pleasant sweet odor; found in fruits and herbs; aromatic ether ; contains methoxy groups; member of benzenes inhibits DNA replication ; plant metabolite ; overexpressed in cancer cells; aromatic ether; aromatic compound; member of benzenes inhibitor; yellow crystalline ; antifungal agrochemical ; aromatic compound; organic compound; aromatic ether
COC1=CC=C(C=C1)C=CNC=O	Tuberin: It has a role of inhibitor. It is a yellow crystalline . It has a role as antioxidant . It is an aromatic compound. It is an organic compound. It is an aromatic ether .	
	C=CC=CC=CC=O- C=C- C=CC=CC=CO C=CC=CC=CC=O- C=C- C=CC=CC=CO	toxic by ingestion; irritate skin; decomposed when heated; bacterial metabolite.; aromatic compound ; benzoic acid ; organic compound . relatively neutral ; a bitter taste.; irritate skin; a fruity taste ; aromatic compound ; benzenes ; organic compound .
C1=CC=C(C=C1)C(=O)C=CC2=CC=C(C=C2)O	4-hydroxychalcone: It has a role as plant metabolite . It is a protein kinase inhibitor . It has a role as prevent oxidation and free radical formation. It is an aromatic compound, a mer	
	CC- #N- Cl CC- #N- Cl C(C#N)Cl	colorless gas; toxic by ingestion; decompo colorless liquid; substituted by methyl gro . colorless gas; toxic by ingestion ; decompo colorless liquid; substituted by methyl gro. Chloroacetonitrile: It is a colorless gas. Very toxic by ingestion. It has a role as a pollutant . Substituted by methyl group. It is a chloride compound. It belongs to c

- Molecule Property Prediction

- NLP Tasks (Relation extraction)

- Cross-modal retrieval

Property query	Substances retrieval result
Anti-inflammatory	Effective: Elocalcitol ^a 53 , Fluocinolone, Fluocinonide, and Marinobufagenin ^a 54
	Unclear: Eribulin mesylate, U46619, Cholesteryl linoleate, Hallactone B, Leukotriene A4, and Npvvhffknivprtppps
Antineoplastic	Effective: Rebeccamycin ^a 55 , Idarubicin, Eribulin mesylate, Piroxantrone, and XC-302 free base
	Unclear: Trimethoprim, Cyclomontanin C, Hexamidin, Fosinopril, and Dabigatran
Antioxidant	Effective: Purpurin ^a 56 , Aromadendrin, Amburoside A, Dioxinodehydroeckol ^a 57 , and hematein ^a 58
	Unclear: Capensinidin, 2'-Hydroxygenistein, Hydramacrophyllol A, 23566-96-3, and Olivomycin

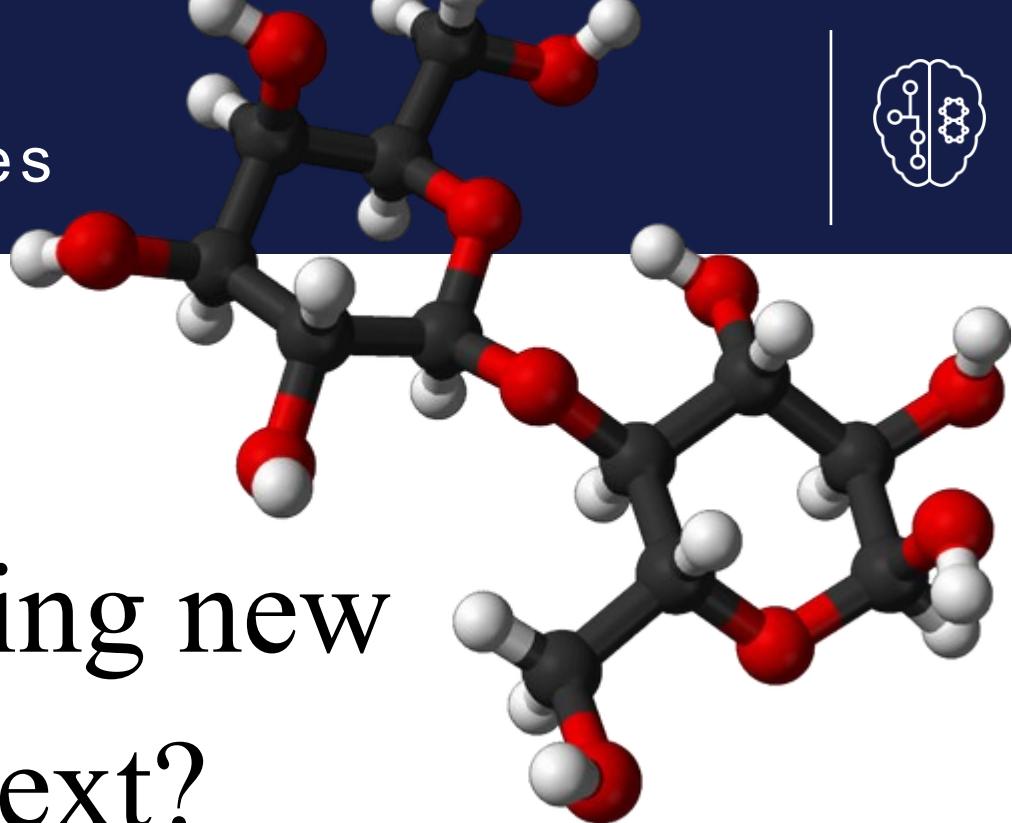
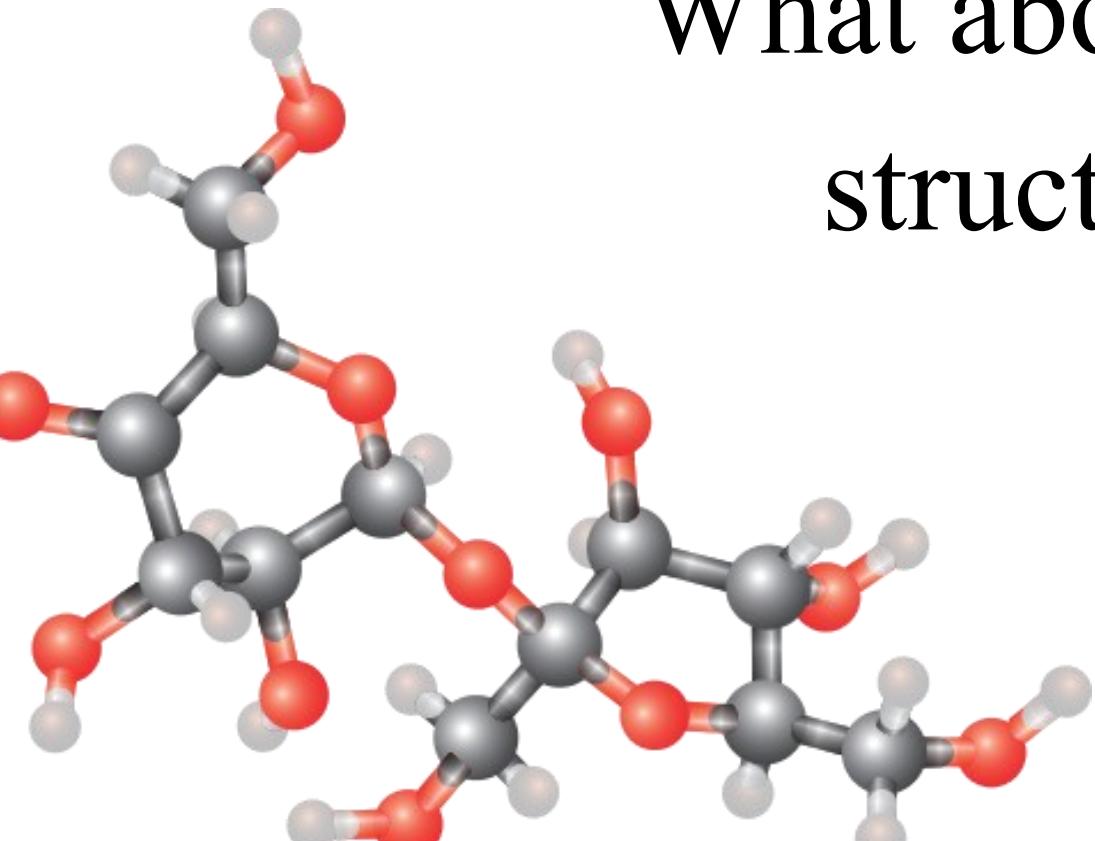


THRUST 1

Integrating Language with Molecules



What about creating new
structures or text?





THRUST 1

Describing a Molecule: Molecule Captioning



- There are an enormous number of possible molecules.
- There are an enormous number of properties that they can have.
- Some properties might not be easy to measure quantitatively.
- These can't all be tested in a lab.
- Can we describe molecules—at a high level— using natural language?

Image Captioning

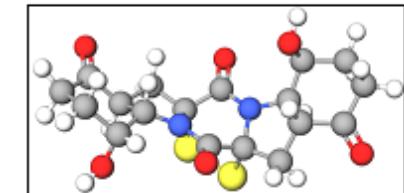


1. a cat sitting on top of an open laptop computer.
2. a cat that is sitting on top of a lap top.
3. a cat is sitting on the keyboard of a laptop.
4. a cat is sitting on an open laptop.
5. a striped cat sitting on top of a laptop

Captions from COCO

Molecule Captioning

C1CC(=O)C2CC34C(=O)N5C6C(CCC(=O)C6CC5(C(=O)N3C2C1O)SS4)O



SMILES representation

3D View

The molecule is an organic disulfide isolated from the whole broth of the marine-derived fungus *Exserohilum rostratum* and has been shown to exhibit antineoplastic activity. It has a role as a metabolite and an antineoplastic agent. It is a bridged compound, a lactam, an organic disulfide, an organic heterohexacyclic compound, a secondary alcohol, a cyclic ketone and a diol.

Caption



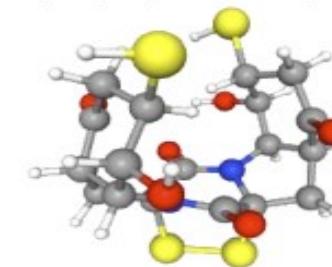
THRUST 1

Molecule Captioning

- Molecule captioning is hard!
- We can describe a molecule with:
 1. A chemical formula
 2. As one of many different synthetic routes from known precursor molecules
 3. In terms of properties (e.g. carcinogenic or lipophilic, absorbs wavelengths of 570 nm)
 4. In terms of applications (e.g. a dye, an antipneumonic, or an antifungal)
 5. In terms of its functional groups (e.g. “substituted by hydroxy groups at positions 5 and 7 and a methyl group at position 8”)
 6. Many other methods!



1. grey cat sits on laptop computer on the floor
2. a cat that is on top of a computer.
3. a gray and white cat is sitting on a laptop
4. a cat is sitting peacefully across a laptop.
5. a cat sleeping on top of an open laptop computer.



The molecule is an organic disulfide isolated from the whole broth of the marine-derived fungus *Exserohilum rostratum* and has been shown to exhibit antineoplastic activity. It has a role as a metabolite and an antineoplastic agent. It is a bridged compound, a cyclic ketone, a lactam, an organic disulfide, an organic heterohexacyclic compound, a secondary alcohol, a dithiol and a diol.





THRUST 1

Creating the right molecule: text-guided molecule generation



- Text-to-image models can generate high-level ideas or compose multiple functions and properties
 - Can we do this with molecules too?

TEXT PROMPT

an armchair in the shape of an avocado....

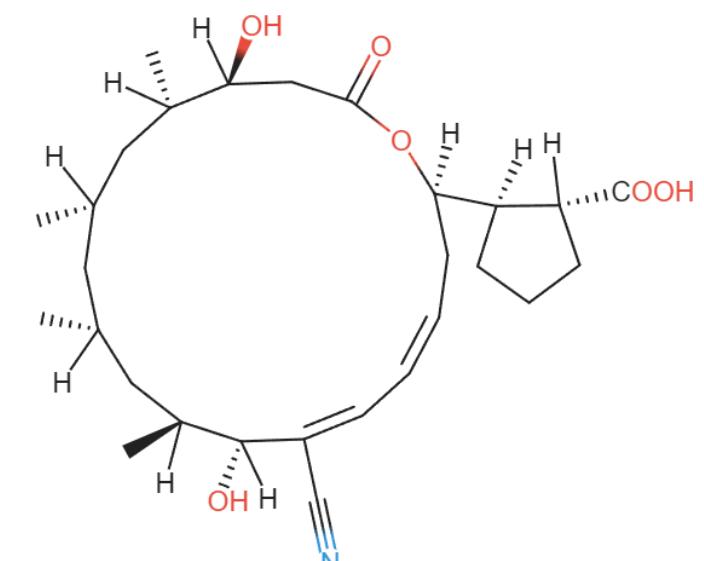
AI-GENERATED IMAGES



[Ramesh et al. 21, 22]

Generate a molecule which satisfies a given description:

The molecule is a macrolide that is isolated from several Streptomyces species and displays antibiotic, antineoplastic and antimarial properties. It has a role as a bacterial metabolite, an antimicrobial agent, an antifungal agent, an antineoplastic agent, an apoptosis inducer and an antimarial. It is a macrolide, a monocarboxylic acid, a secondary alcohol, a diol and an aliphatic nitrile.





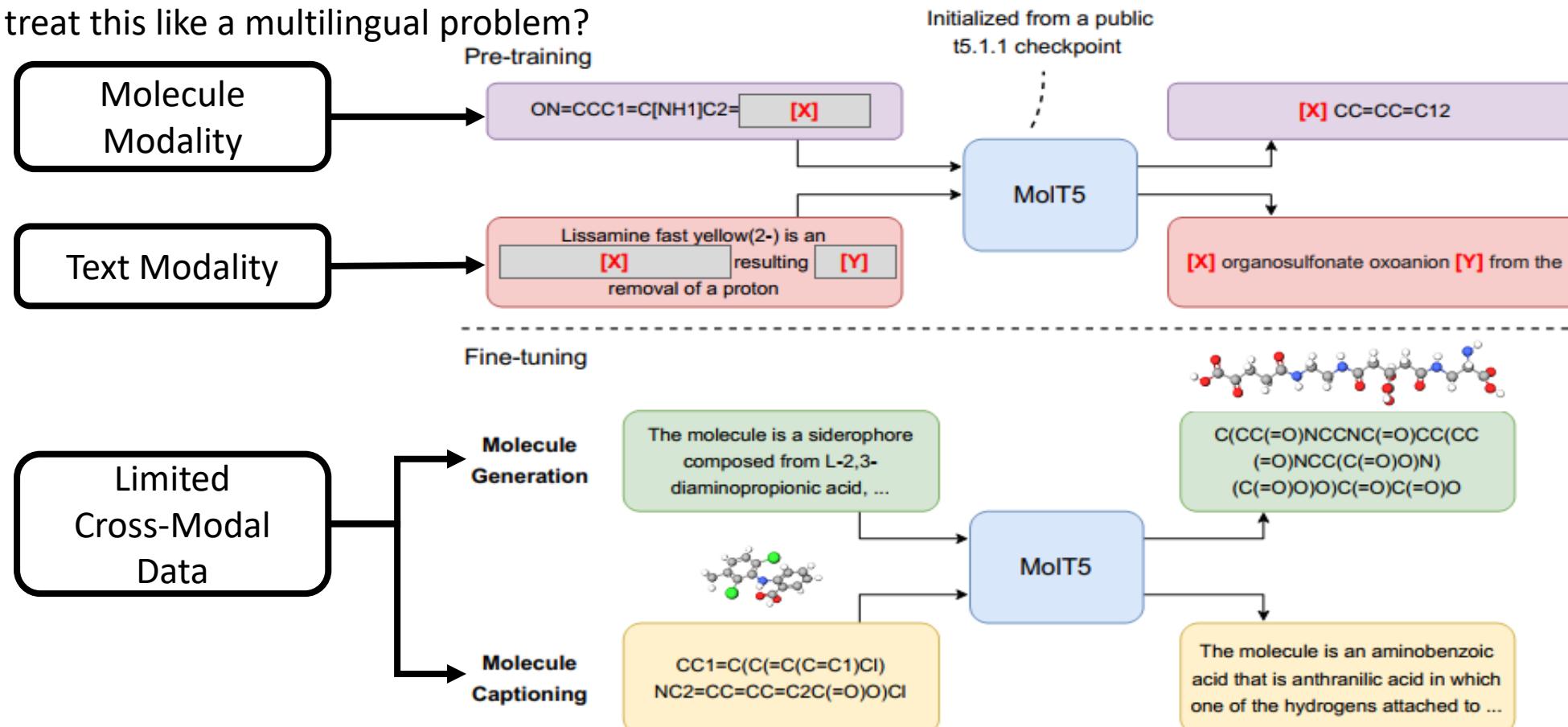
THRUST 1

MolT5: Training a model to translate between language and molecules



Data scarcity is a big potential issue

Can we treat this like a multilingual problem?





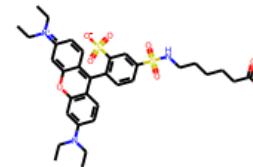
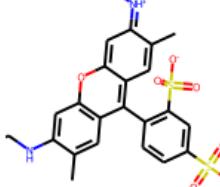
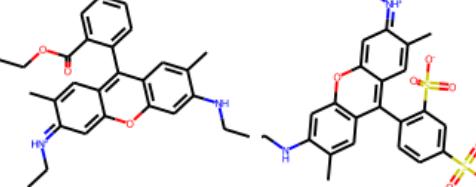
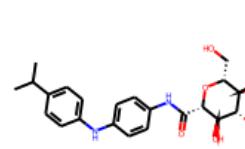
THRUST 1

MolT5: Generated Molecules

**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.

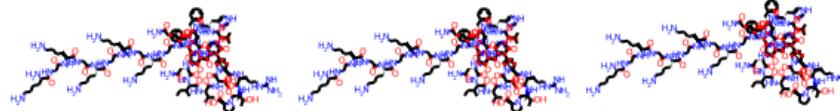
Invalid



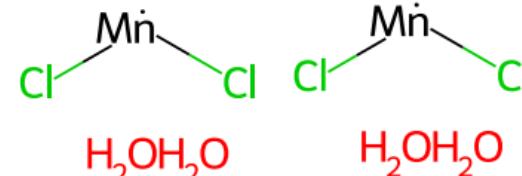
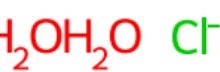
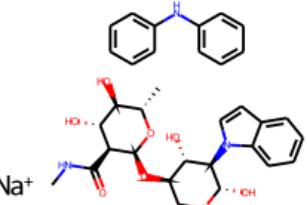
2 The molecule is a linear 27-membered polypeptide comprising the sequence Lys-Gly-Lys-Gly-Lys-Gly-Lys-Gly-Glu-Asn-Pro-Val-Val-His-Phe-Phe-Tyr-Asn-Ile-Val-Thr-Pro-Arg-Thr-Pro. Corresponds to the sequence of the myelin basic protein 83-99 (MBP83-99) immunodominant epitope with the lysyl residue at position 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



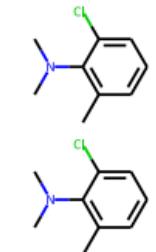
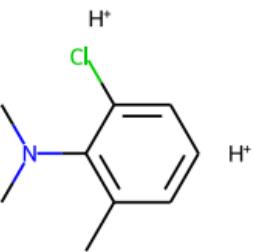
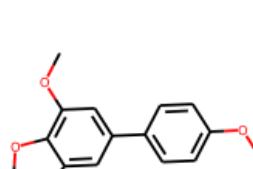
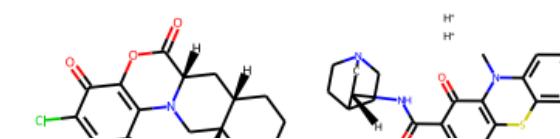
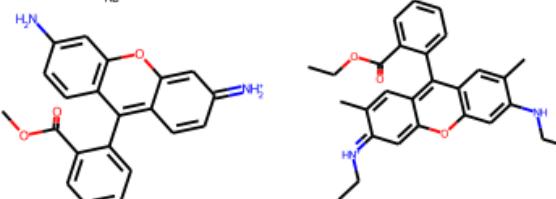
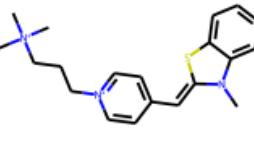
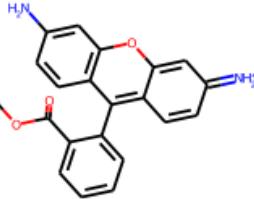
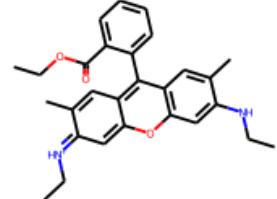
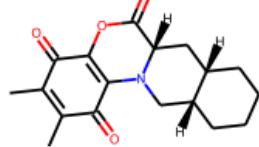
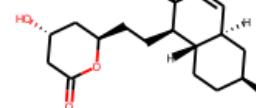
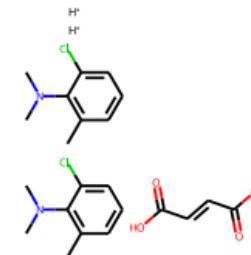
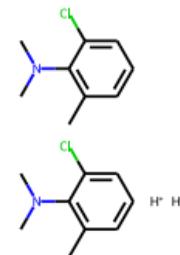
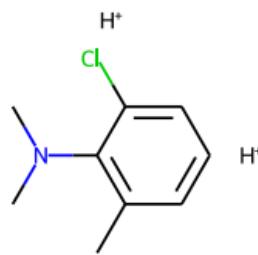
3 The molecule is a hydrate that is the dihydrate form of manganese(II) chloride. It has a role as a MRI contrast agent and a nutraceutical. It is a hydrate, an inorganic chloride and a manganese coordination entity.



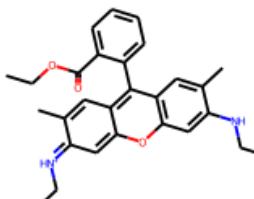
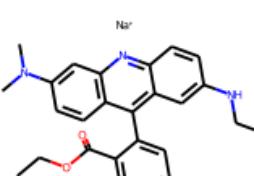
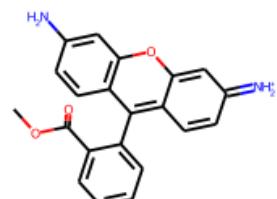


THRUST 1

MolT5: Searching for Function



Input: The molecule displays antimarial properties.

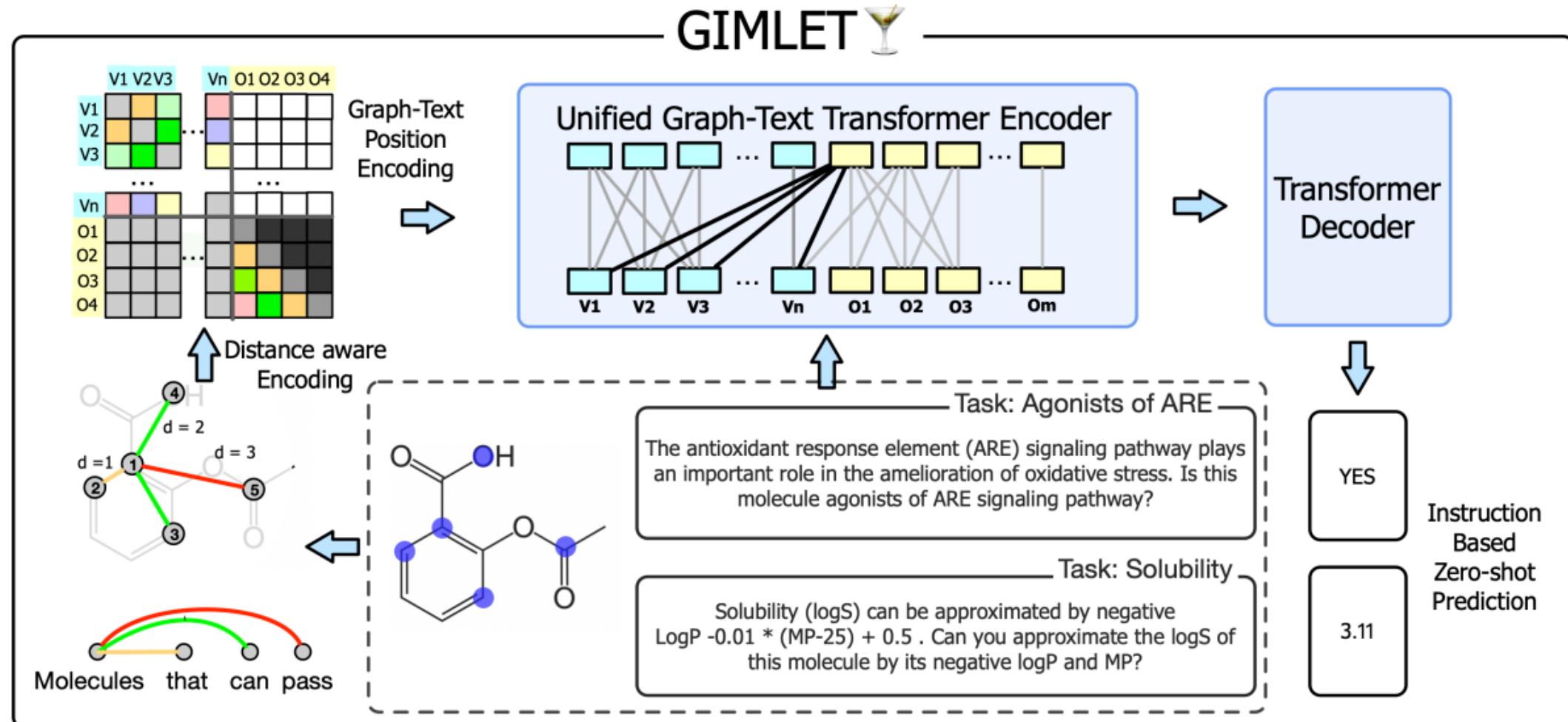


Input: The molecule is a blue dye.



THRUST 1

GIMLET

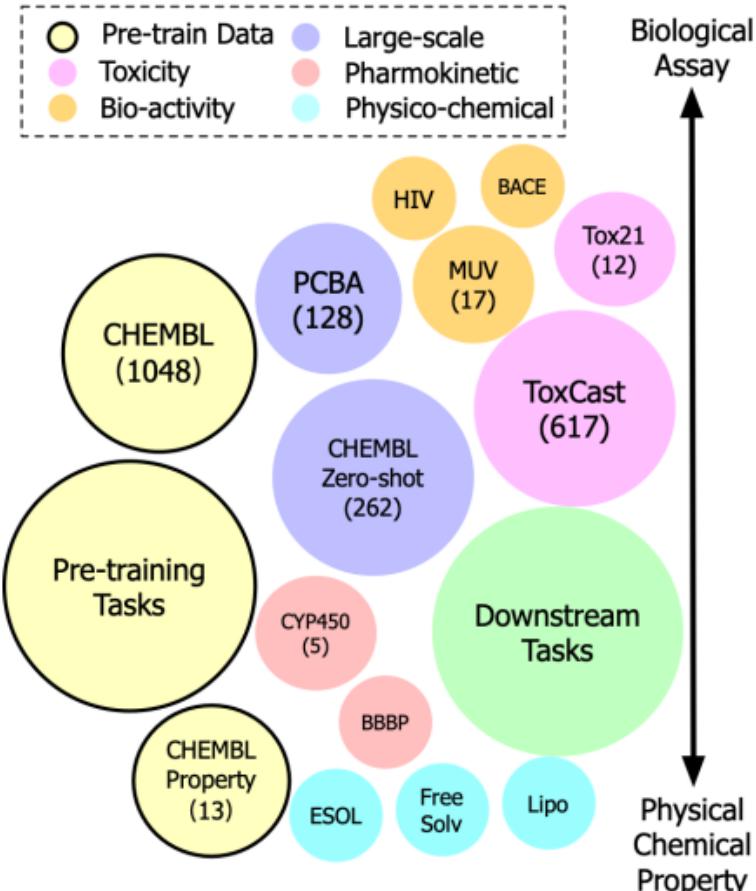


GIMLET: A Unified Graph-Text Model for Instruction-Based Molecule Zero-Shot Learning



THRUST 1

GIMLET



Heavy atoms counting (From ChEMBL property)

"Heavy atom refers to any atom that is not hydrogen. How many heavy atoms do the molecule have?"

Inhibitors of Schistosoma Mansoni Peroxiredoxins (From ChEMBL)

"The functional assay is named qHTS Assay for the Inhibitors of Schistosoma Mansoni Peroxiredoxins. It is related to two other pubchem assays, namely Confirmation Concentration-Response Assay for Inhibitors of the Schistosoma mansoni Redox Cascade and Schistosoma Mansoni Peroxiredoxins (Prx2) and thioredoxin glutathione reductase (TGR) coupled assay. The assay category is also confirmatory and it pertains to the Schistosoma mansoni organism. Is this molecule effective to the assay?"

Inhibition of receptor SF-1 (From MUV)

"The nuclear receptor SF-1 (steroidogenic factor-1) is expressed in the pituitary, testes, ovaries, and adrenal gland and regulates steroid hormone production at many levels, including direct regulation of expression of major P450 enzymes involved in steroid hormone synthesis. Is this molecule inhibitor of SF-1?"

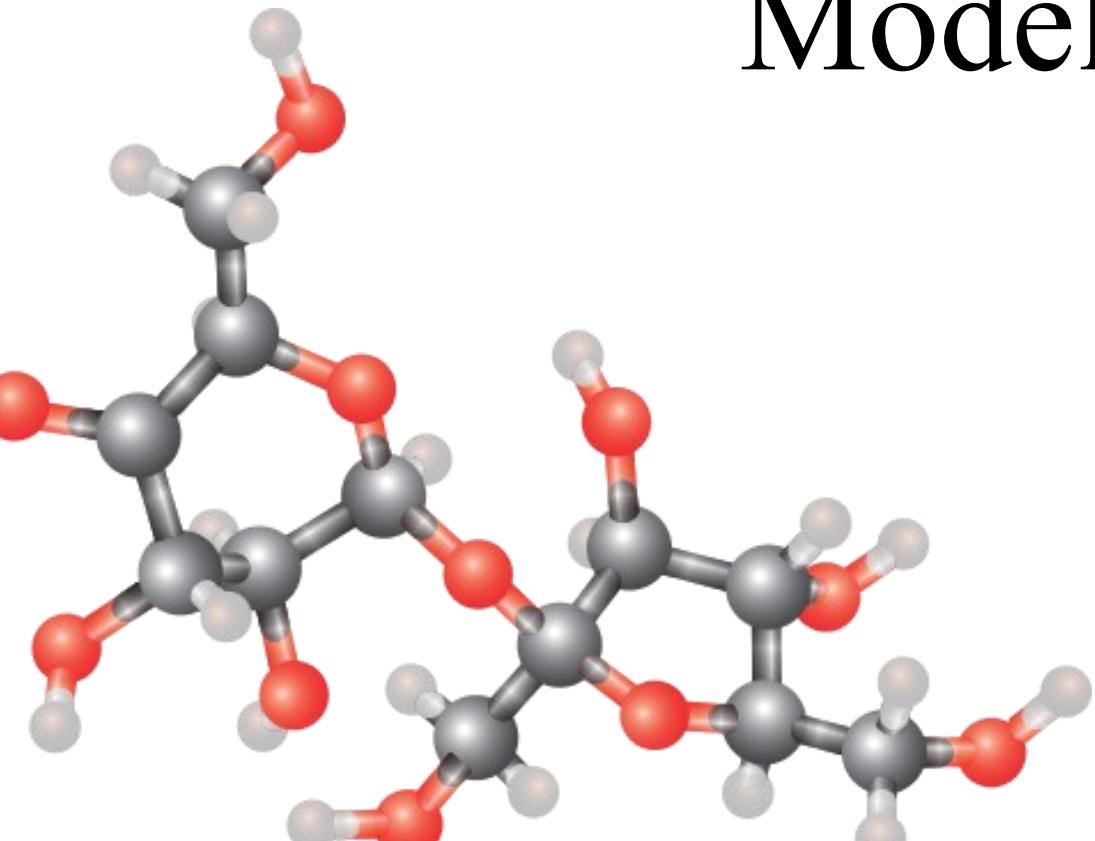
Toxicity to ARE signaling pathway (From Tox21)

"Oxidative stress has been implicated in the pathogenesis of a variety of diseases ranging from cancer to neurodegeneration. The antioxidant response element (ARE) signaling pathway is important in the amelioration of oxidative stress. Is this molecule agonists of antioxidant response element (ARE) signaling pathway?"

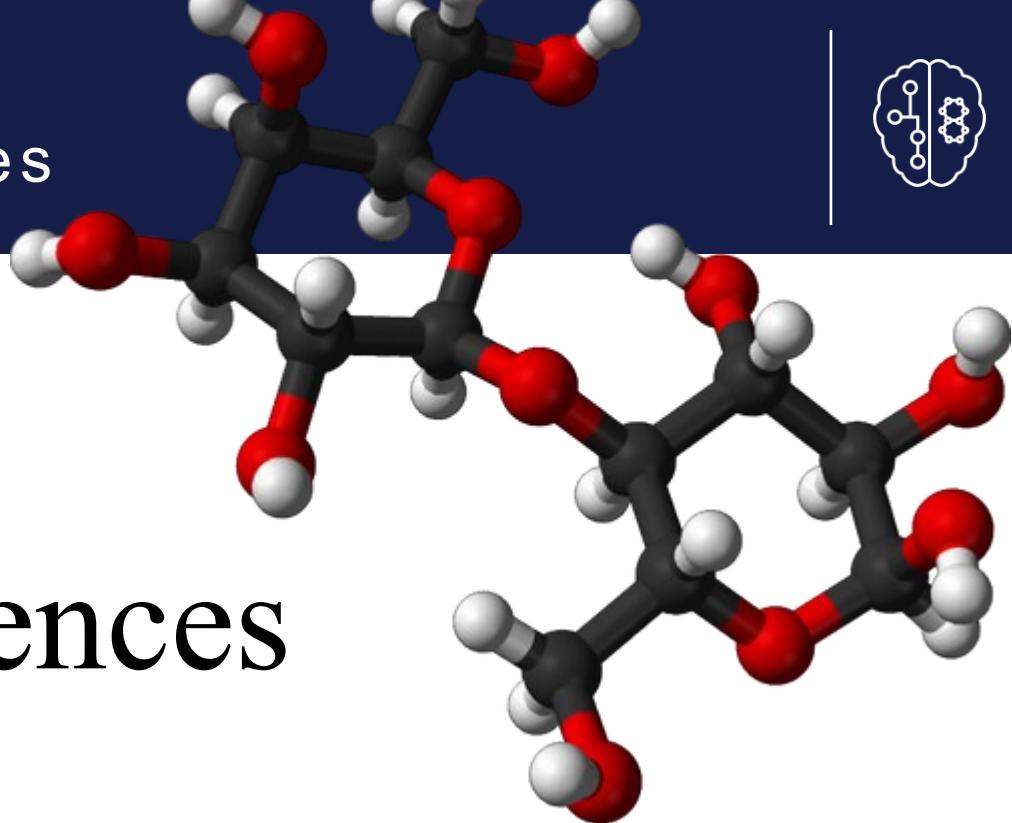


THRUST 1

Integrating Language with Molecules



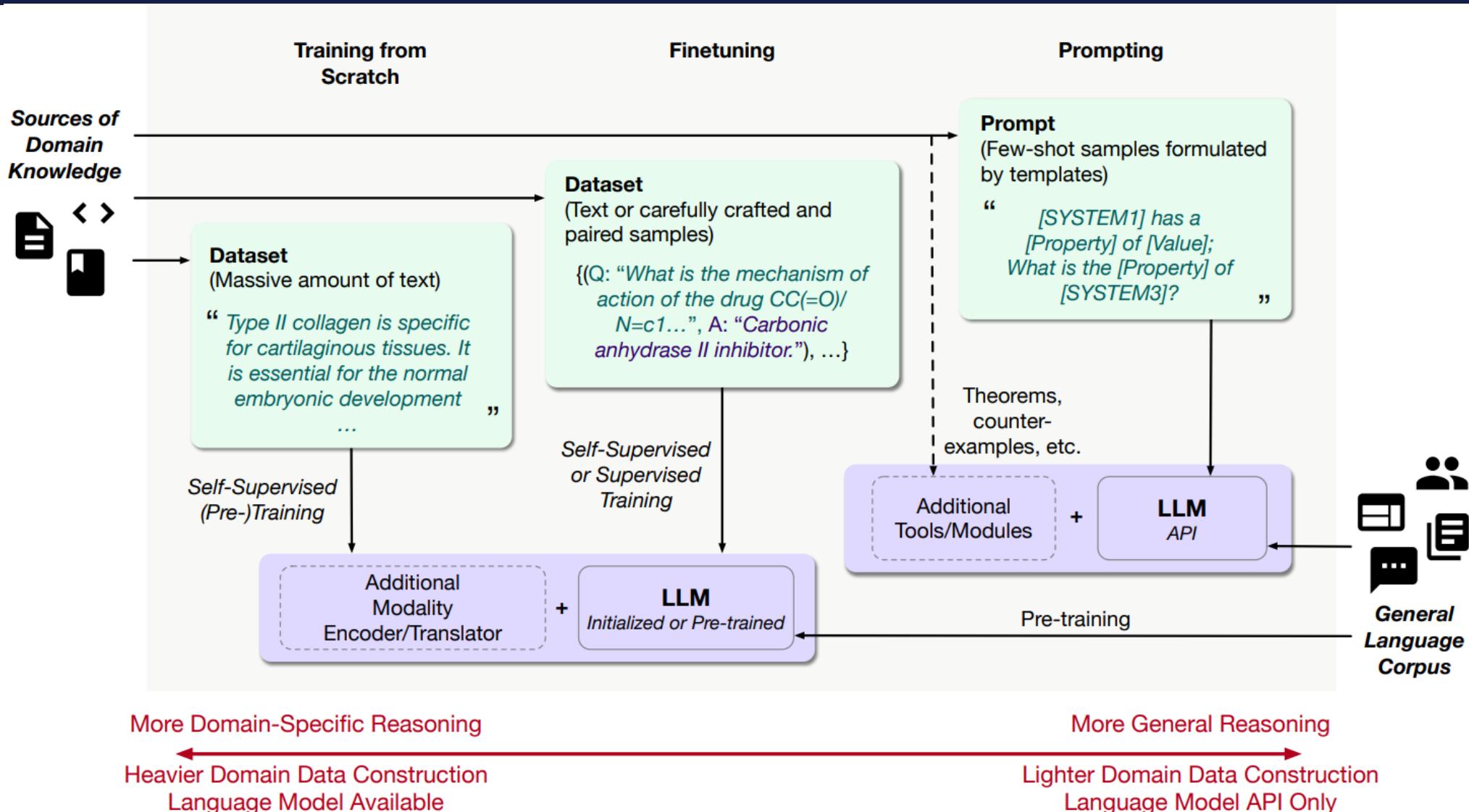
Model Differences





THRUST 1

Training Paradigms

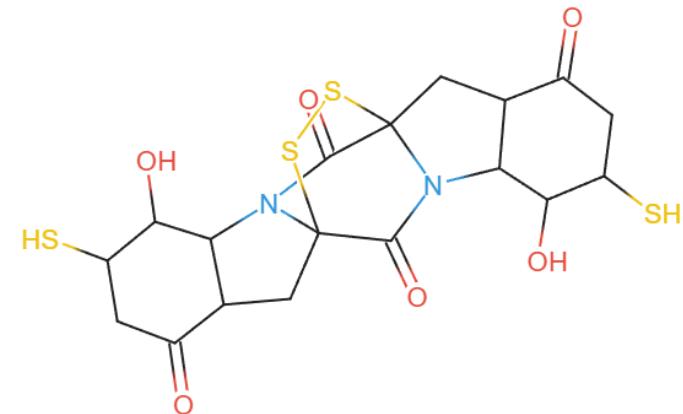




Data Sources

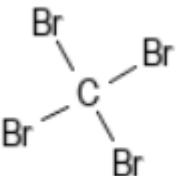


- a) High-quality scraped datasets
- b) Low-quality large-scale data via entity linking on the literature
 - a) Do we insert the molecule structure into the sentence?
- c) Templated-based conversion to language

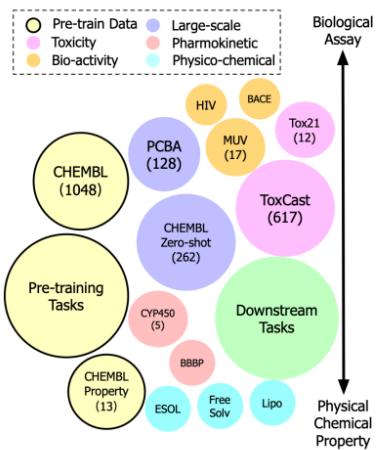


Rostratin D is an organic disulfide isolated from the whole broth of the marine-derived fungus *Exserohilum rostratum* and has been shown to exhibit antineoplastic activity. It has a role as a metabolite and an antineoplastic agent. It is a bridged compound, a cyclic ketone, a lactam, an organic disulfide, an organic heterohexacyclic compound, a secondary alcohol, a dithiol and a diol.

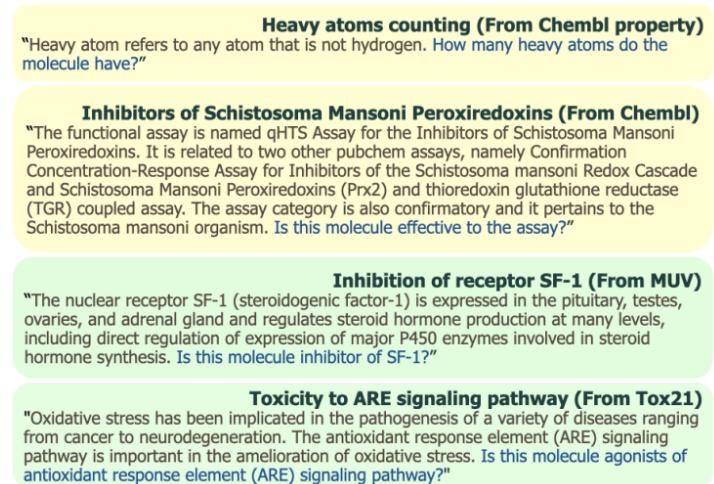
a) From Text2Mol



Heavily carbon-doped $In_{0.53}Ga_{0.47}As$ with hole densities between 5.6×10^{18} and $2.1 \times 10^{20} \text{ cm}^{-3}$ has been grown by solid source molecular beam epitaxy on InP. The dependence of [MASK] (CBr_4)-induced lattice mismatch upon the atomic carbon concentration has been determined from x-ray rocking curve measurements. It has been found by secondary ion mass spectroscopy that the alloy composition is altered by the preferential etching effect of CBr_4 . After taking into account this compositional change, the “intrinsic” lattice contraction solely due to carbon incorporation has been obtained, which obeys Vegard’s law.



b) From MoMu

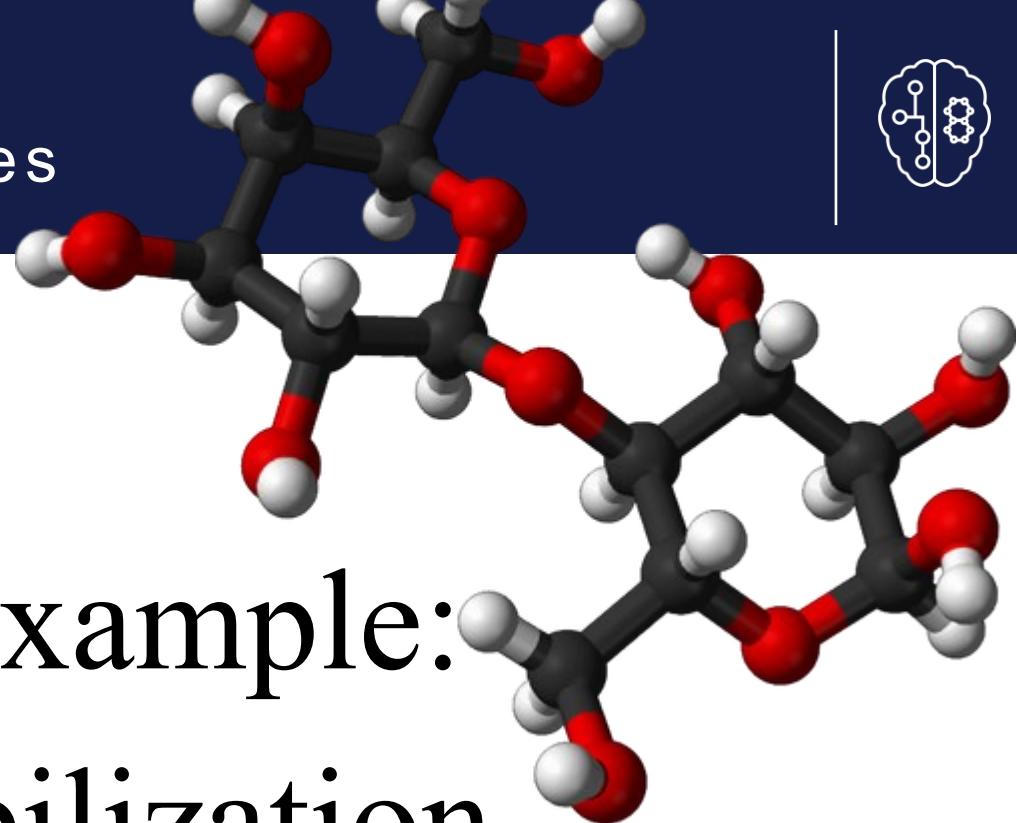


c) From GIMLET

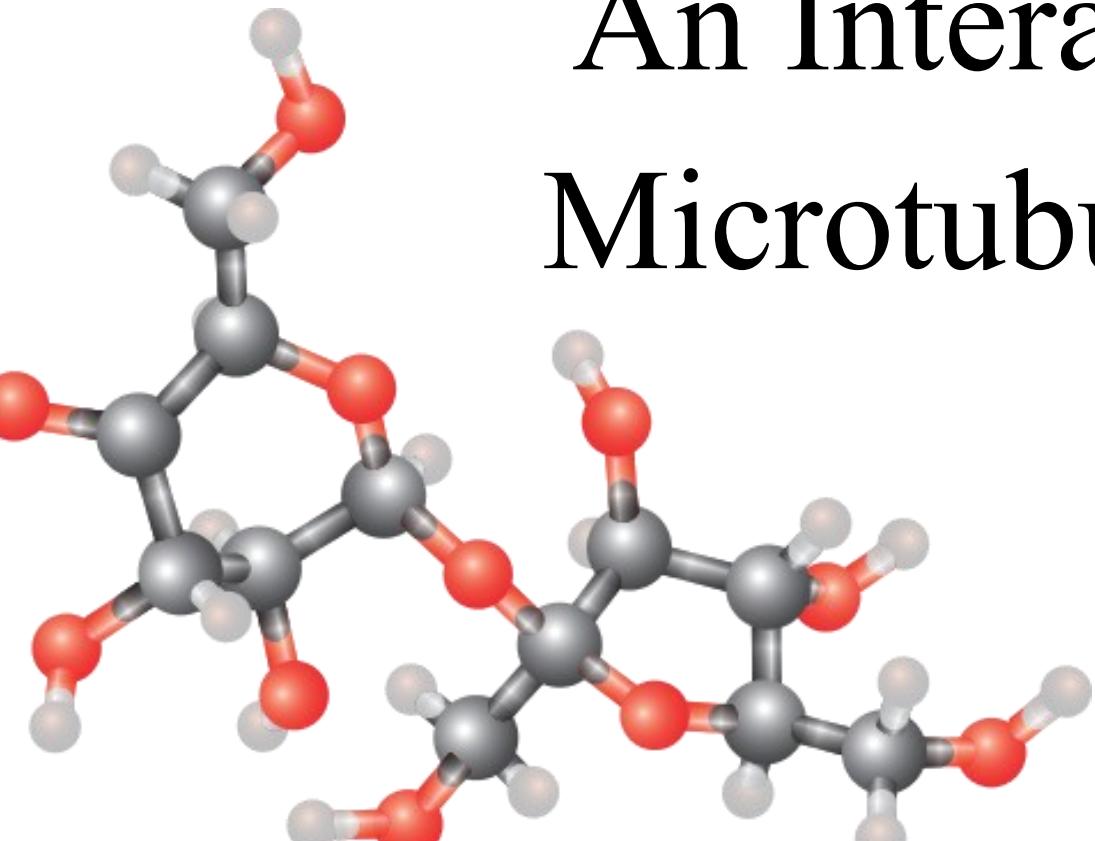


THRUST 1

Integrating Language with Molecules



An Interactive Example: Microtubule Stabilization





THRUST 1

Example Overview



We'll consolidate these ideas by exploring an interactive example of language-enabled molecule design.

Our hands-on example will consist of four components:

1. Language-enabled Molecule Design
2. Language-enabled Drug Design
3. Language-Guided Assay Testing
4. Interaction Prediction



Find Google Colab notebooks at

<https://github.com/language-plus-molecules/LPM-24-Tutorial>



THRUST 1

Open in Colab from Github



Welcome To Colaboratory

File Edit View Insert Runtime Tools Help

Table of contents X + Code + Text Copy to Drive

Getting started Data science Machine learning More Resources Featured examples + Section

Open notebook

Examples > Recent > Google Drive > GitHub > Upload >

Enter a GitHub URL or search by organization or user language-plus-molecules Search Include private repos

Repository: [language-plus-molecules/LPM-24-Tutorial](#) Branch: [main](#)

Path

L+M_EACL_tutorial.ipynb L+M_EACL_tutorial_local_ver.ipynb

+ New notebook Cancel

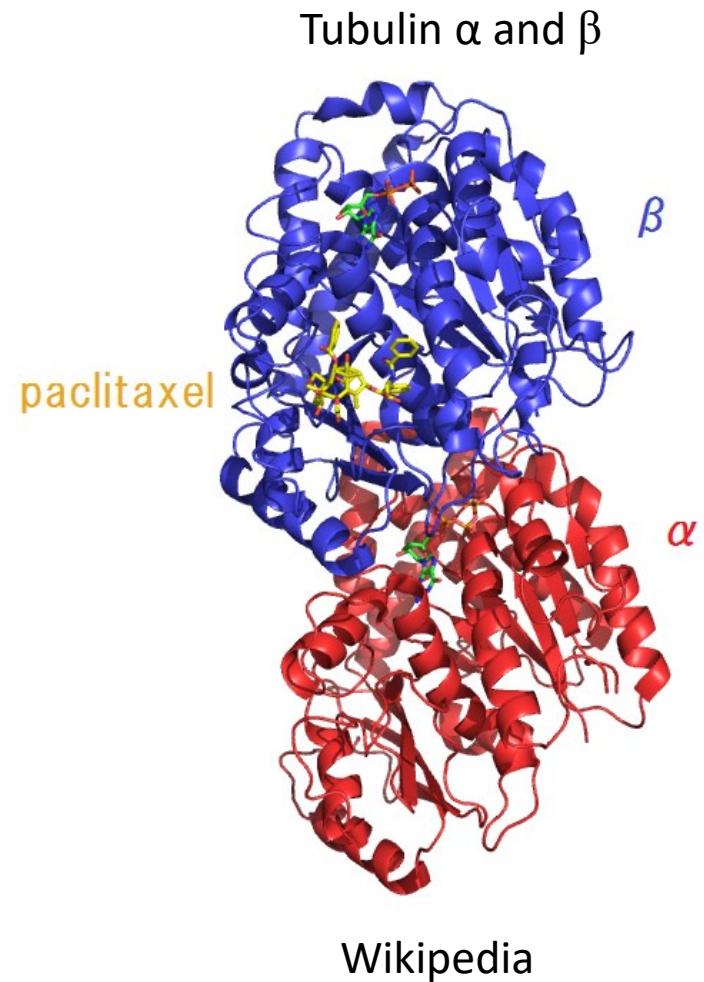
The screenshot shows the Google Colab interface with a dark theme. On the left, there's a sidebar with links like 'Getting started', 'Data science', 'Machine learning', 'More Resources', and 'Featured examples'. Below this is a button labeled '+ Section'. The main area is titled 'Open notebook' and contains a sidebar with 'Examples', 'Recent', 'Google Drive', 'GitHub', and 'Upload' options. The 'GitHub' section has a search bar with 'language-plus-molecules' entered, a 'Repository' dropdown set to 'language-plus-molecules/LPM-24-Tutorial', and a 'Branch' dropdown set to 'main'. A red box highlights the first item in the list below, which is 'L+M_EACL_tutorial.ipynb'. At the bottom of the dialog, there are '+ New notebook' and 'Cancel' buttons.



What are microtubules?



- These cellular structures play an important role in many processes such cell growth and division, and mutations can be oncogenic.
- In modern medicine, tumors such as pancreatic cancer are commonly treated by microtubule-targeting drugs such as paclitaxel (also known as Taxol).
- In our example, we will explore creating new drugs with this function using natural language instructions, which may be useful in cases of paclitaxel resistance.





THRUST 1

What is an assay?

- Generally, this is a test used to assess the presence, amount, or functional activity of a target.
- In particular, bioassays are important for drug discovery, e.g. :
 - Determining if “ligands” bind to a protein
 - Determining toxicity in human or animal cell lines (petri dishes of cells outside the body)
 - Mouse testing





THRUST 1

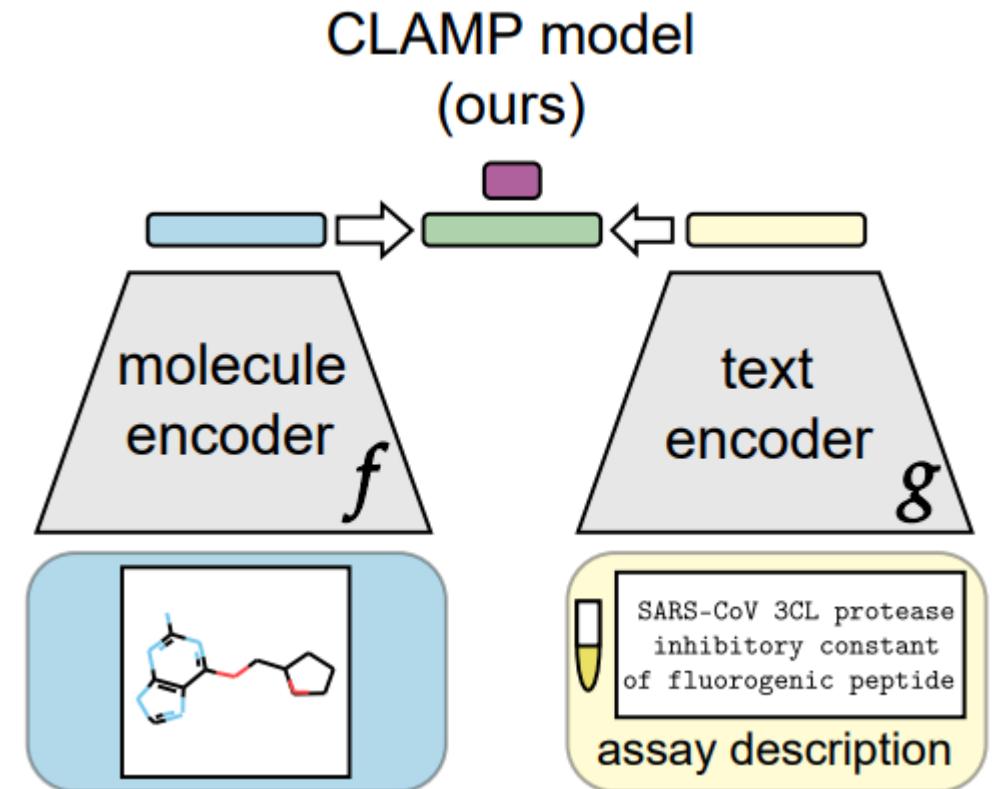
Language-Guided Assay Testing



- Let's use the CLAMP model to predict the function of our discovered drugs

A possible example we can use for our testing:

"The molecule is a microtubule stabilizer similar to Taxol. This causes it to display antineoplastic activity."





THRUST 1

Interaction Prediction



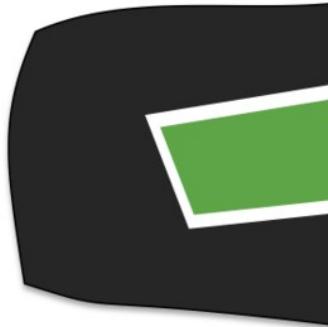
Target



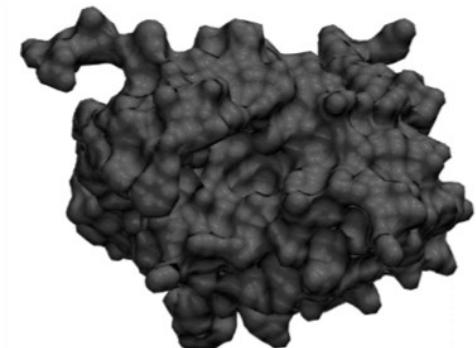
Ligand



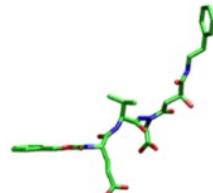
Complex



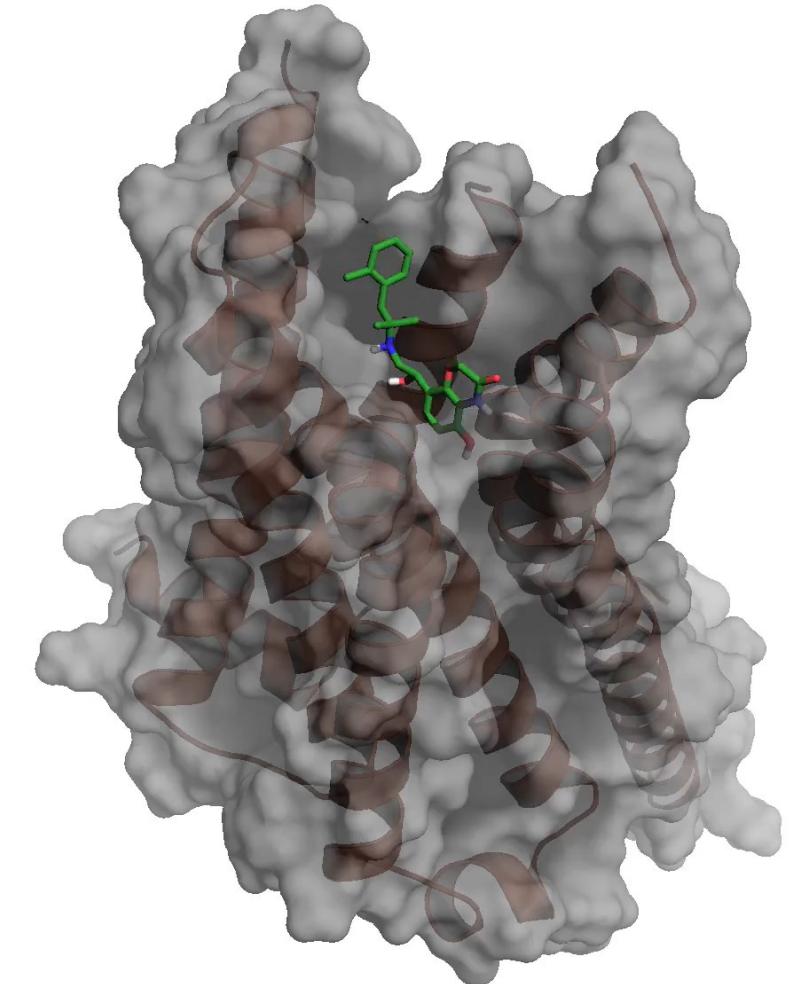
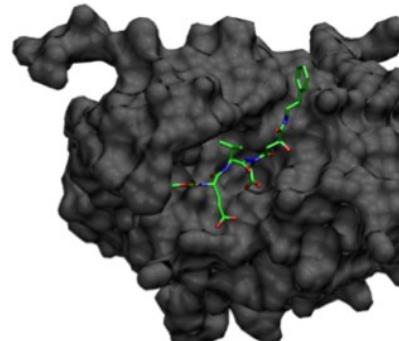
+
 docking →



+



docking →

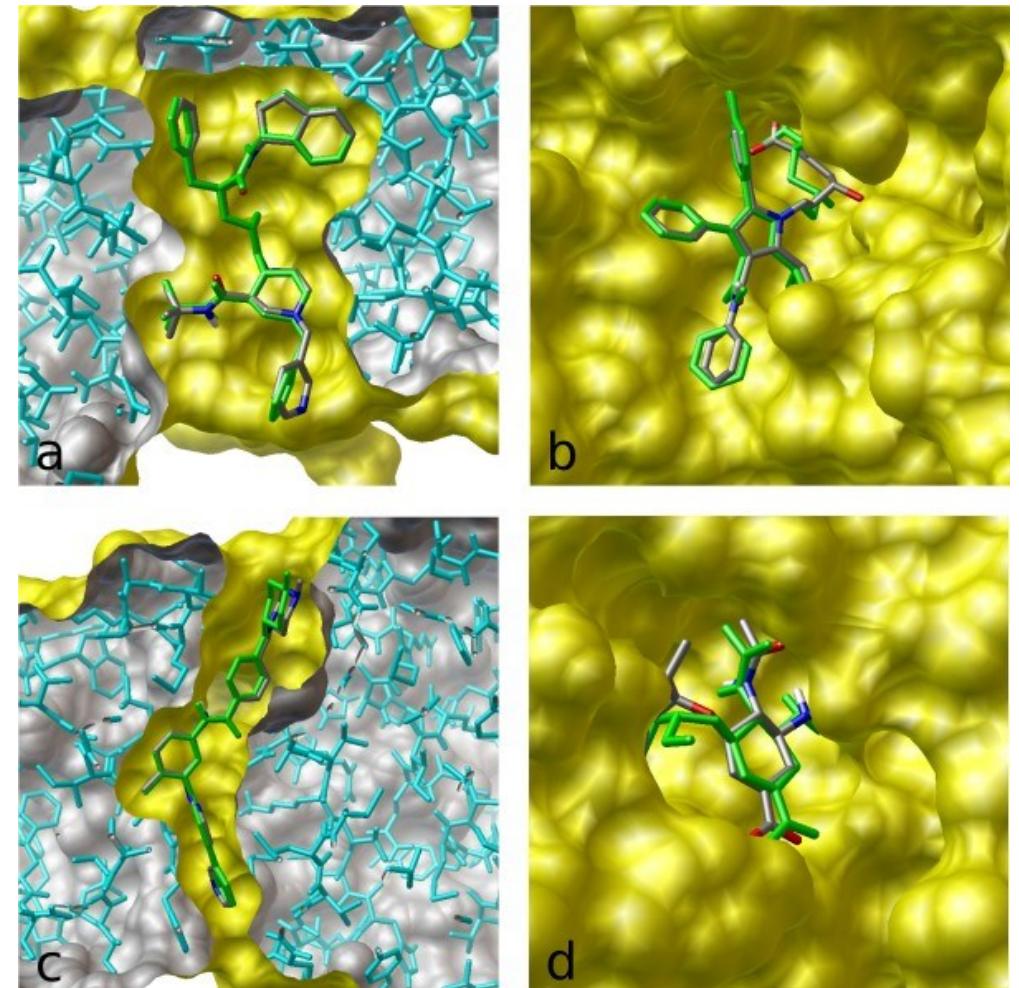
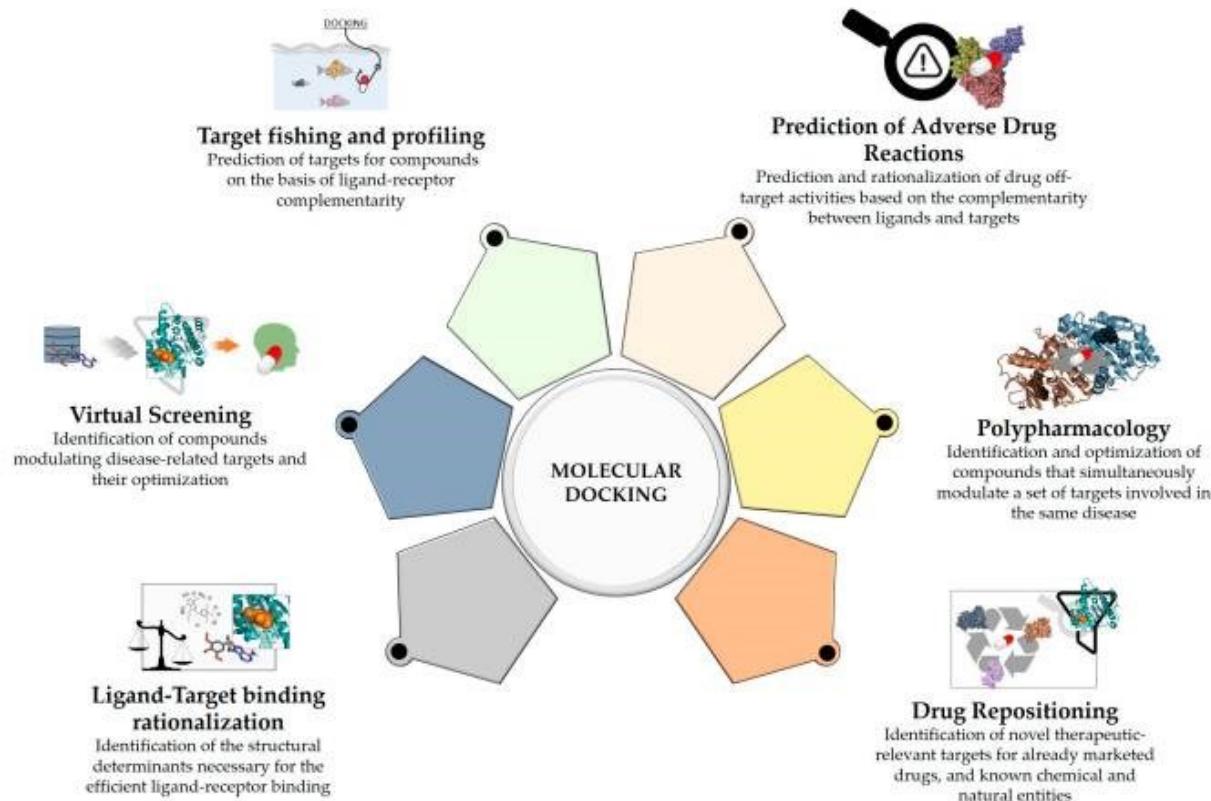


Wikipedia



What is drug docking and why does it matter?

- Docking scores: In drug discovery, docking scores are program-specific scores calculated to estimate how well a drug will bind with a target protein. This allows ranking drug candidates.

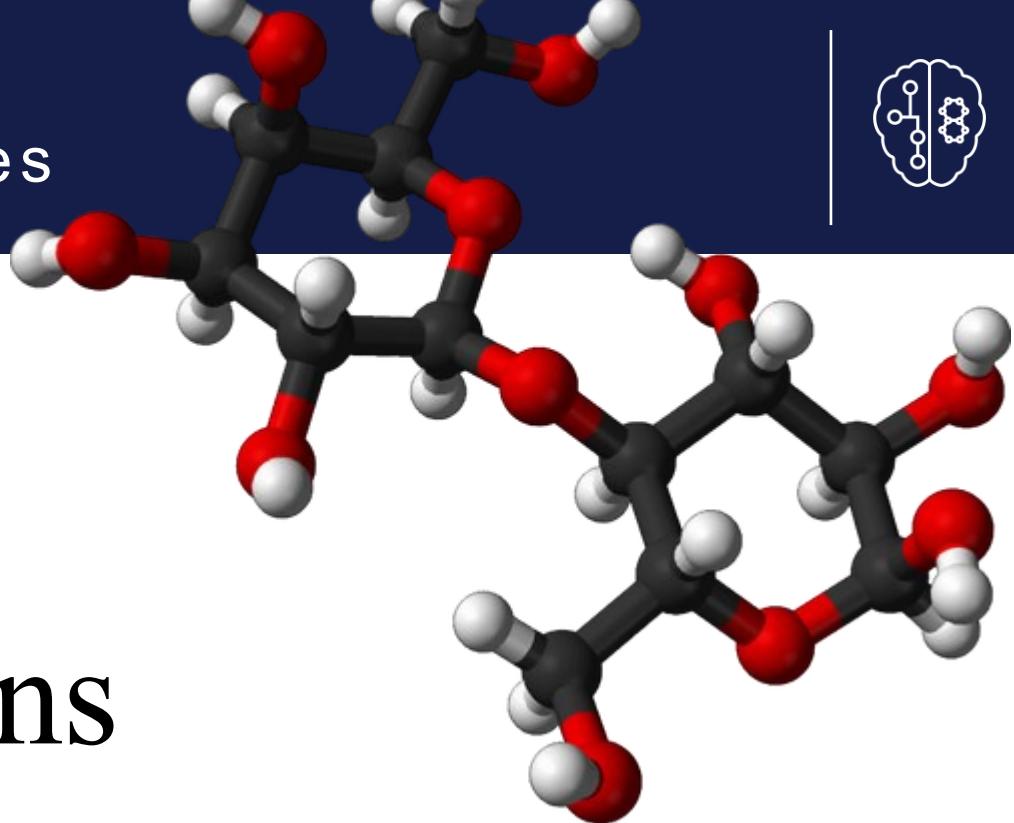


AutoDock Vina is a popular program for docking scores.

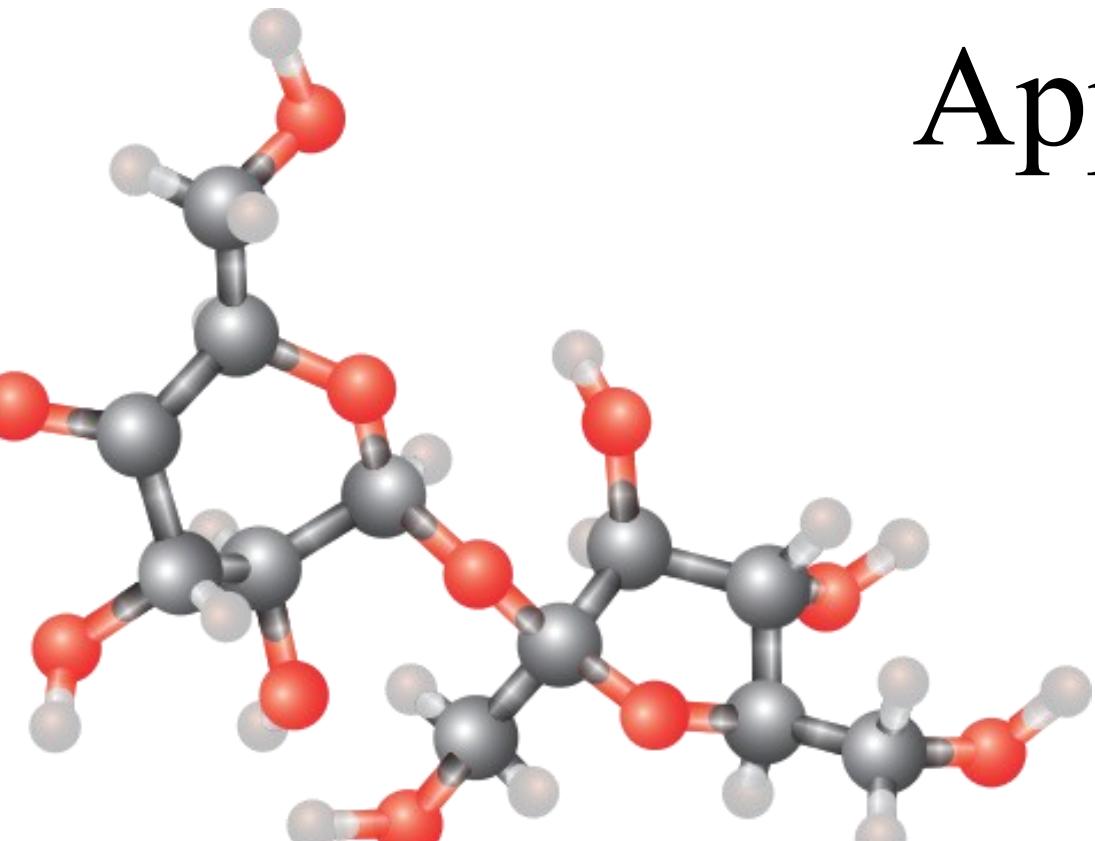


THRUST 1

Integrating Language with Molecules



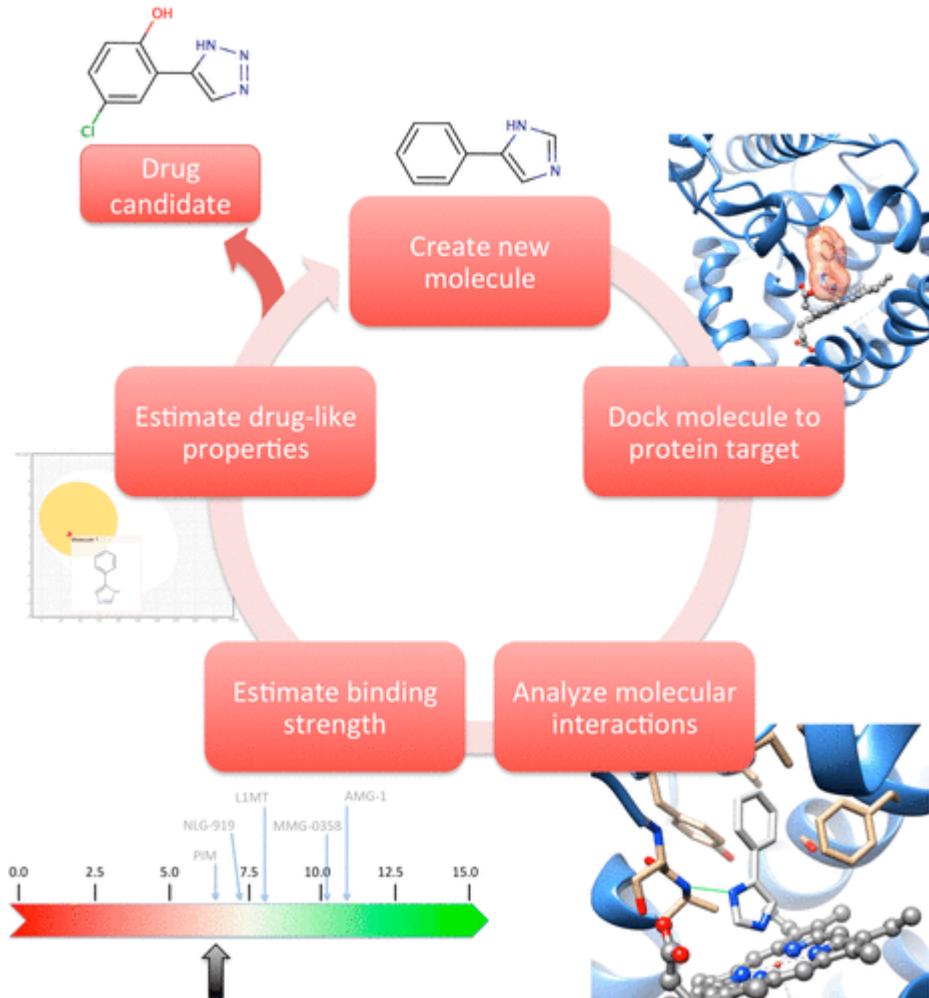
Applications





THRUST 1

Biomedicine



Daina et al. 2017 *J. Chem. Educ.*

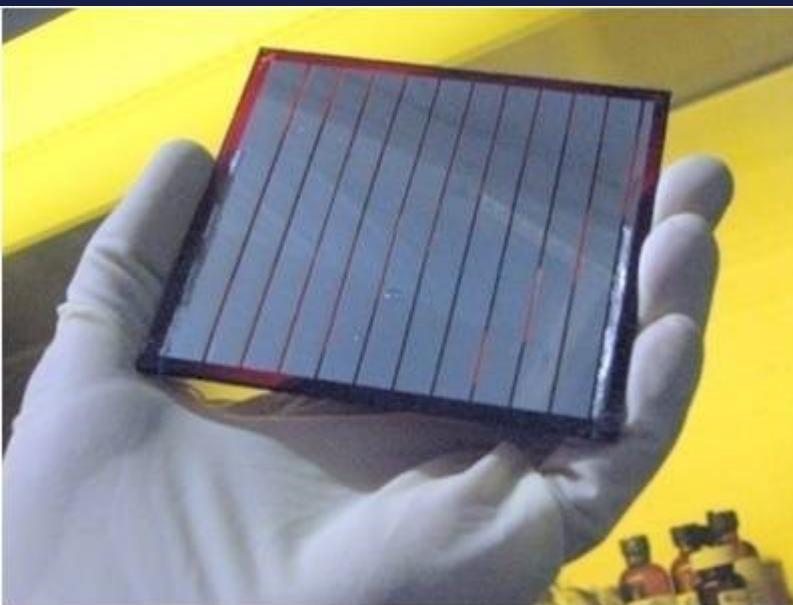
Salient Properties (a sample):

- Anti neoplastic – used in the treatment of cancer
- Glaucoma Treatment
- Asthma preventive
- Pyrogenic – induces feaver
- Circulatory stimulant – increases blood flow
- Glucocorticoid receptor modulator – an experimental drug class with anti-inflammatory properties
- Tyrosine kinase inhibitor – inhibits tyrosine kinases, with important applications in leukemia
- Serotonin agonist – binds and activates serotonin receptors, increasing the amount of the neurotransmitter in the body

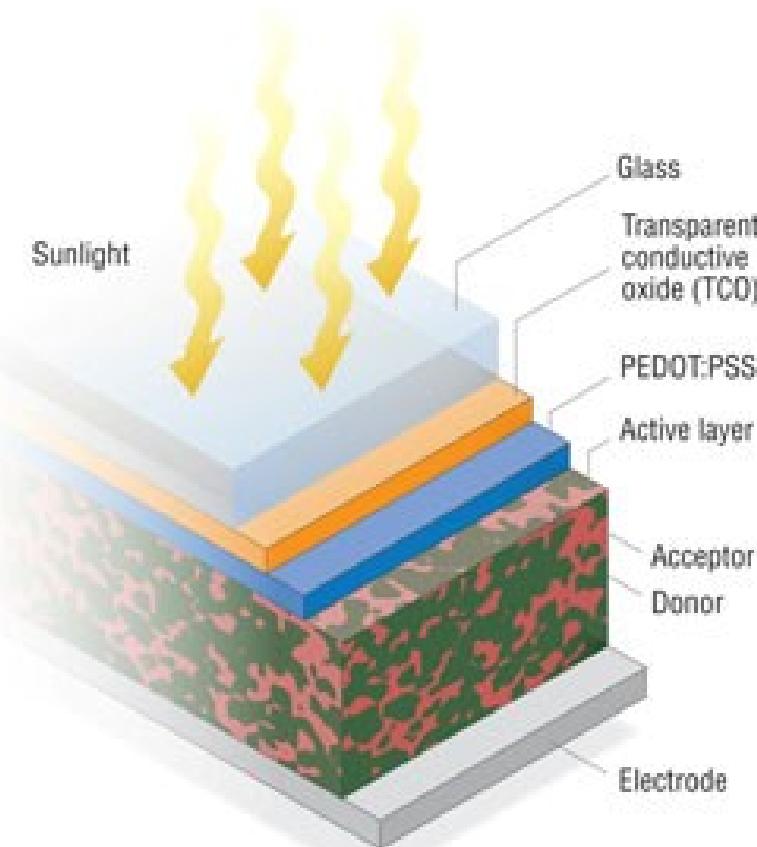


THRUST 1

Light and Electricity: Organic Photovoltaics



Organic solar cells are lightweight, flexible, and cheap, but current cells are still inefficient and unstable.



Department of Energy

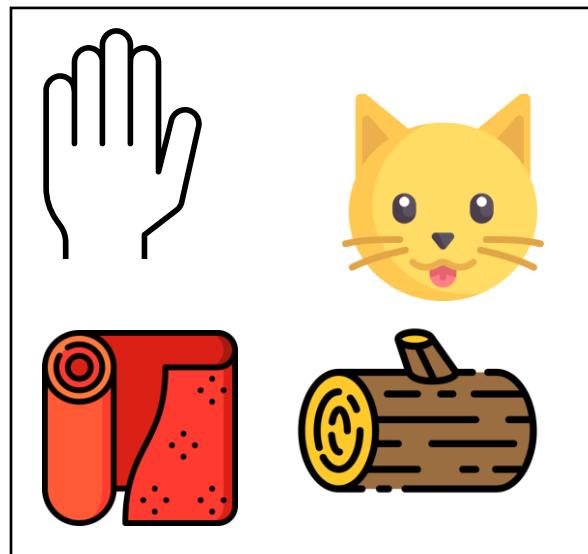
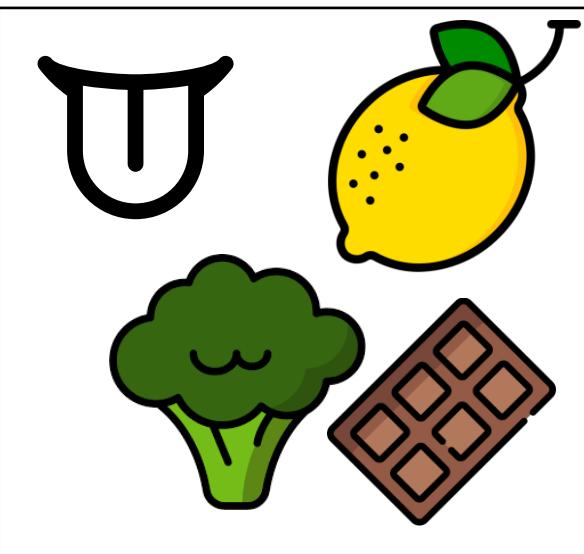
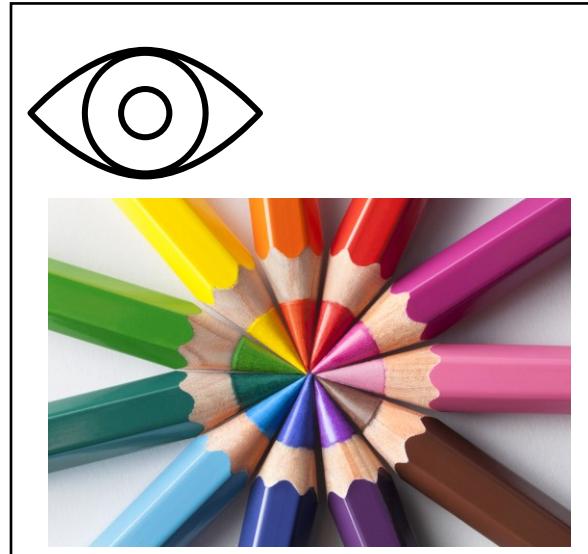
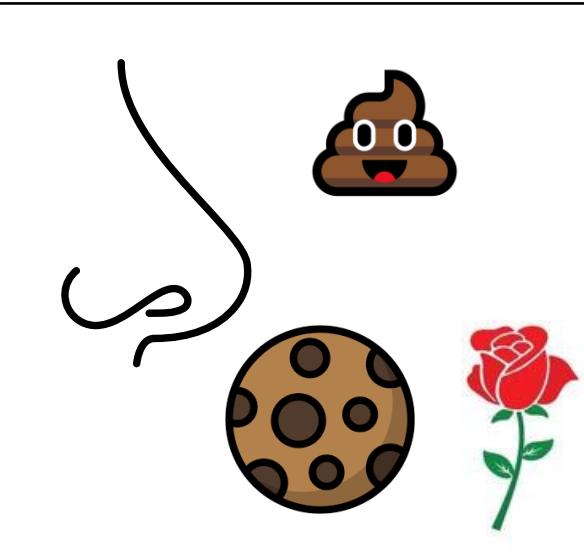
Salient Properties (a sample):

- Organic electroluminescent compound – A material which emits light in the presence of a electric current
- Fluorescent – Emission of light by a substance that has absorbed light
- Phosphorescent – Delayed emission of the absorbed light
- Photoacid generator – Compounds that produce acids upon exposure to light
- Photopolymerization – A chemical reaction linking small monomers when exposed to light
- Photochromic – Undergoes a reversible change in color upon exposure to light



THRUST 1

Human Interaction and Organoleptics



Salient Properties (a sample):

- Bitter
- Nephrotoxic agent – Toxic to the kidneys
- Artichoke
- Red cedar
- Cucumber seed

Interesting reading: A principal odor map unifies diverse tasks in olfactory perception



THRUST 1

Agriculture and Industry



Salient Properties (a sample):

- Pesticide – An agent which kills pests
- Acidity regulator – Food additives which change or maintain acidity
- Crustacicide – An agent which kills crustaceans
- Fertilizer – improves growth and productiveness of plants
- Lubricant – A substance which helps reduce friction between surfaces



THRUST 1

Language + Molecules Shared Task @ ACL



Abstraction, Functionality, and Composition in Complex Molecular Properties

Molecules with Multiple Properties

"alzheimer's treatment"
"bace1 inhibitor"

"anti viral agent"
"mitogen"
"anti carcinogenic"
"lipoxygenase inhibitor"
"fungicide"
"anti oxidant "

Data Sources



Compositional Captions

The molecule is both a alzheimer's treatment and a bace1 inhibitor.

The molecule is a mitogen and lipoxygenase inhibitor, belonging to the anti oxidant class, and is characterized as anti viral agent, anti carcinogenic, and fungicide.

GPT 4 Written Templates

Language
+
Molecules

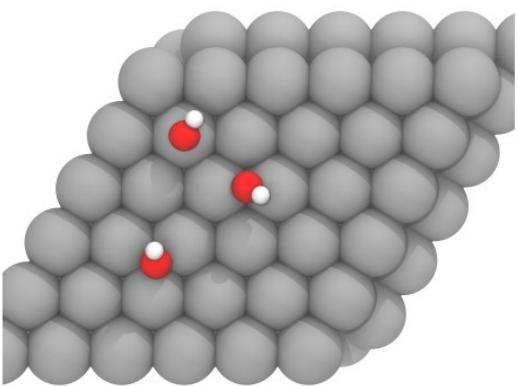
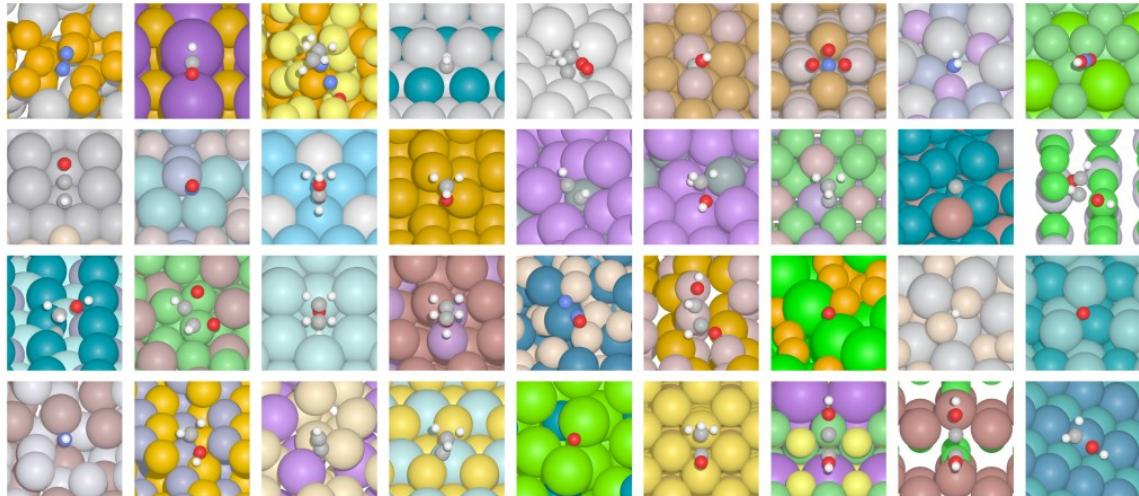
ACL
2024

<https://github.com/language-plus-molecules/LPM-24-Dataset>



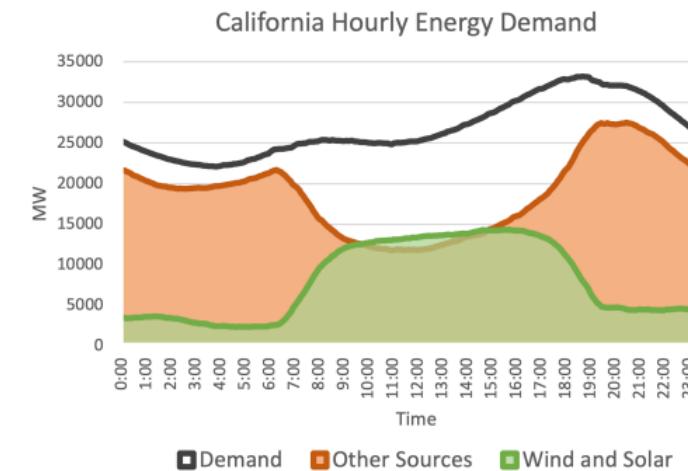
THRUST 1

Catalyst Design – Reducing Chemical Reaction Costs

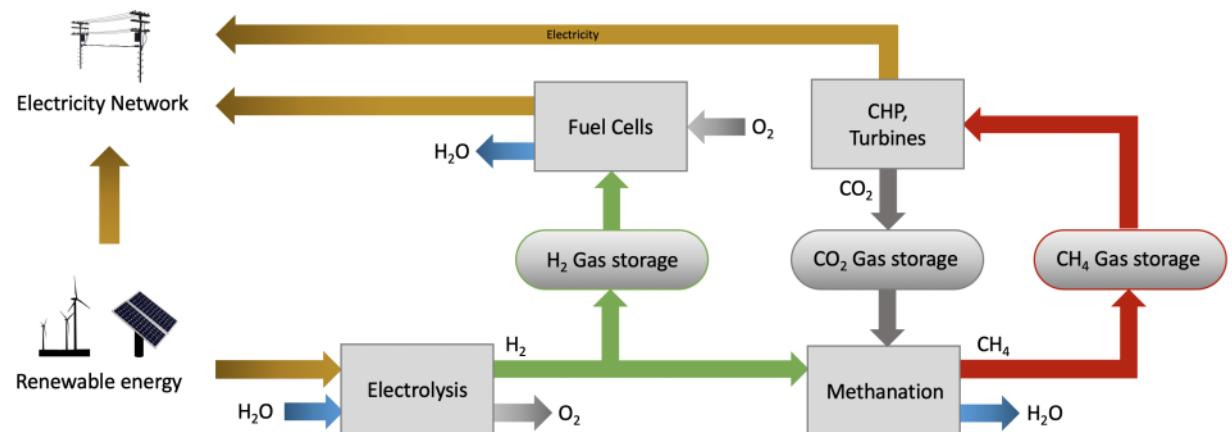


OH binding to a metallic surface

Catalysts (the big atoms) allow energy-efficient reactions to occur in the small atoms via their interactions.



Energy demand exceeds renewable storage and capacity!



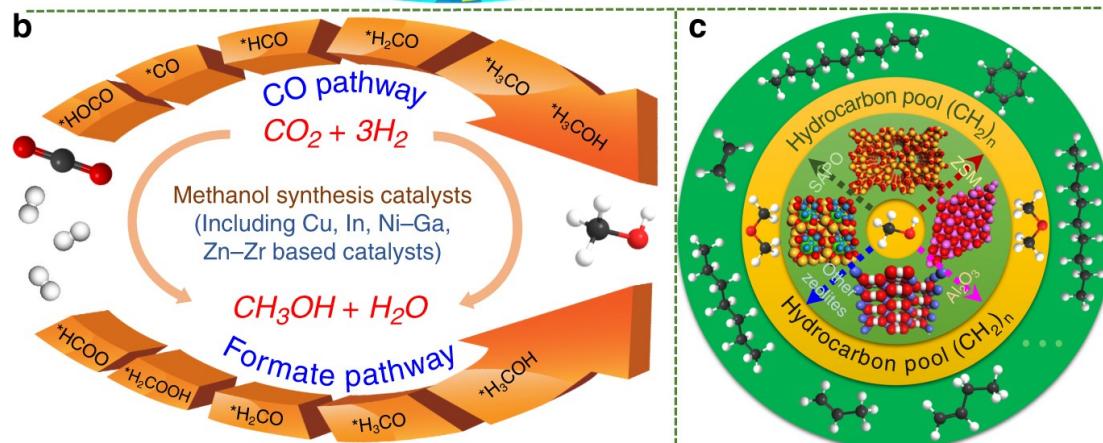
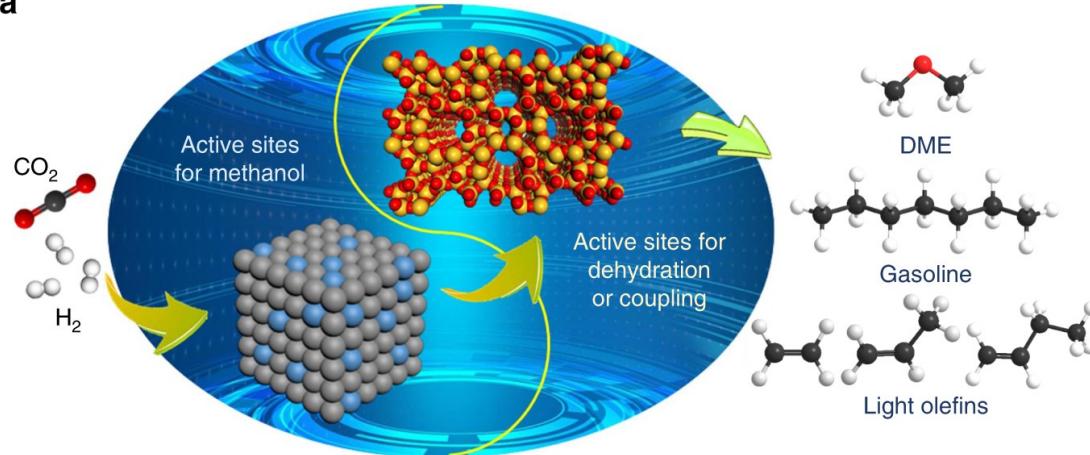
Zitnick et al. 2020

Improved catalysts can help mitigate the storage issue!



THRUST 1

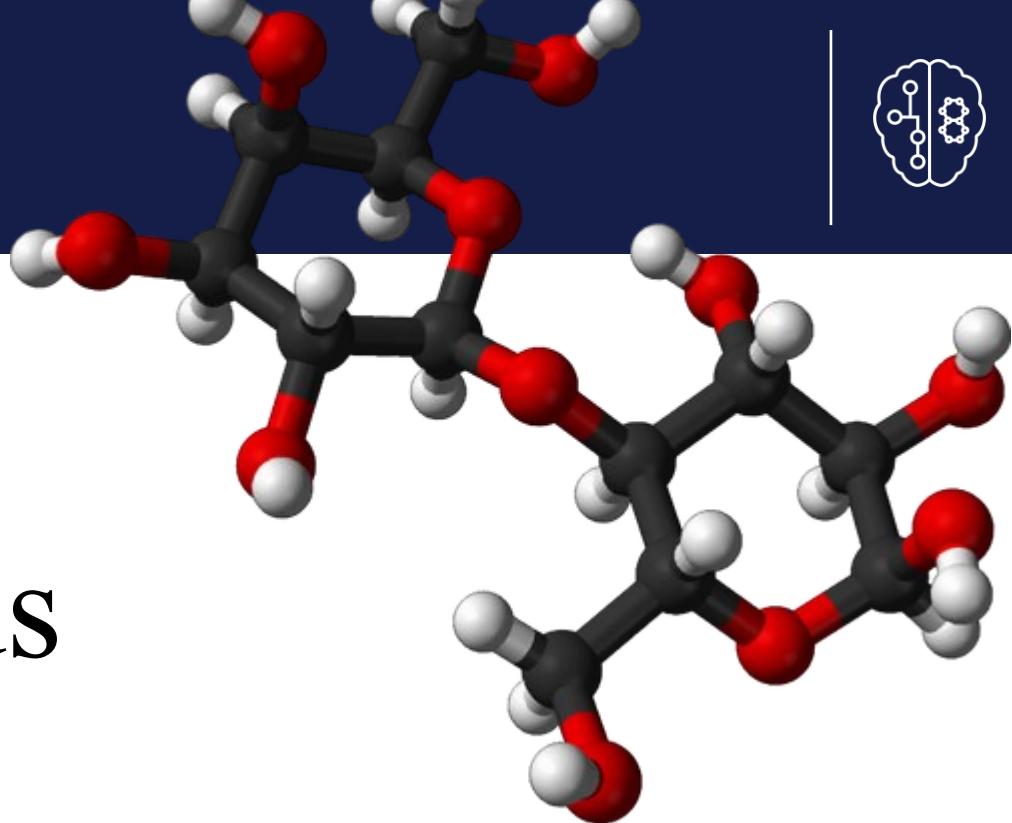
Carbon Capture and Biofuel Production

**a**

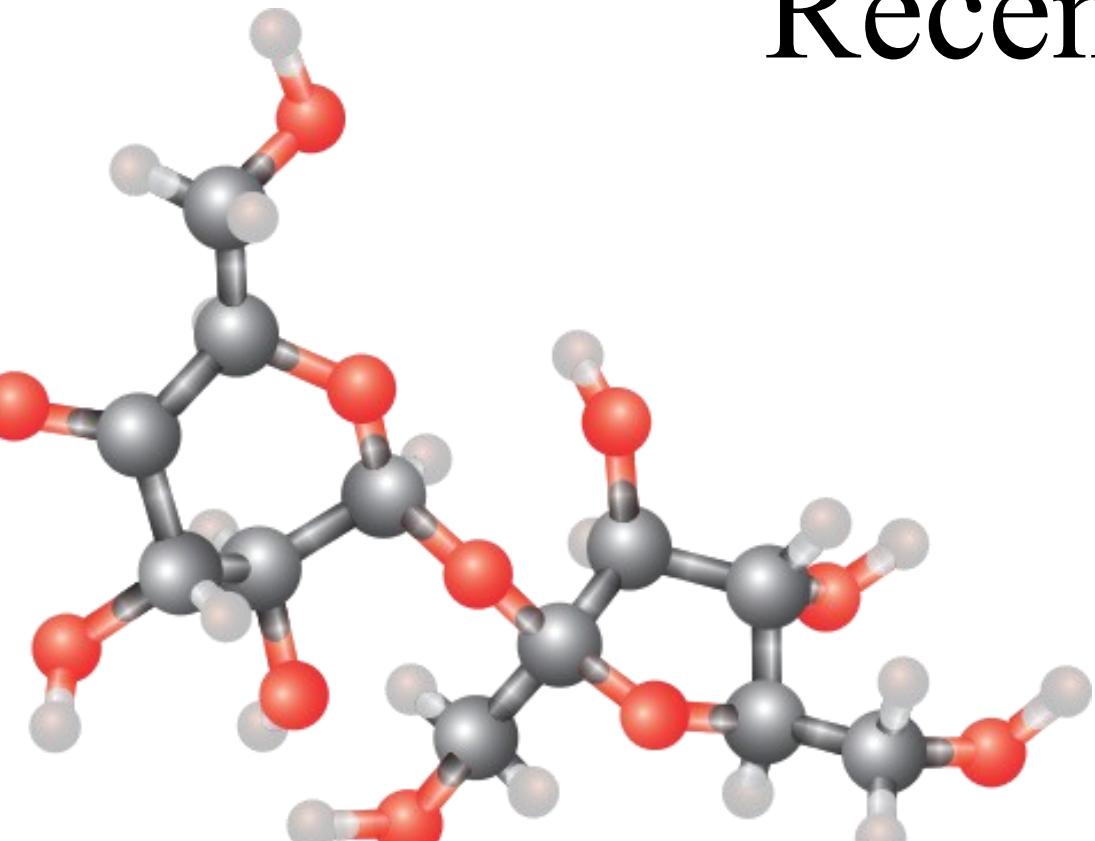


THRUST 1

Recent Trends and Conclusion



Recent Trends

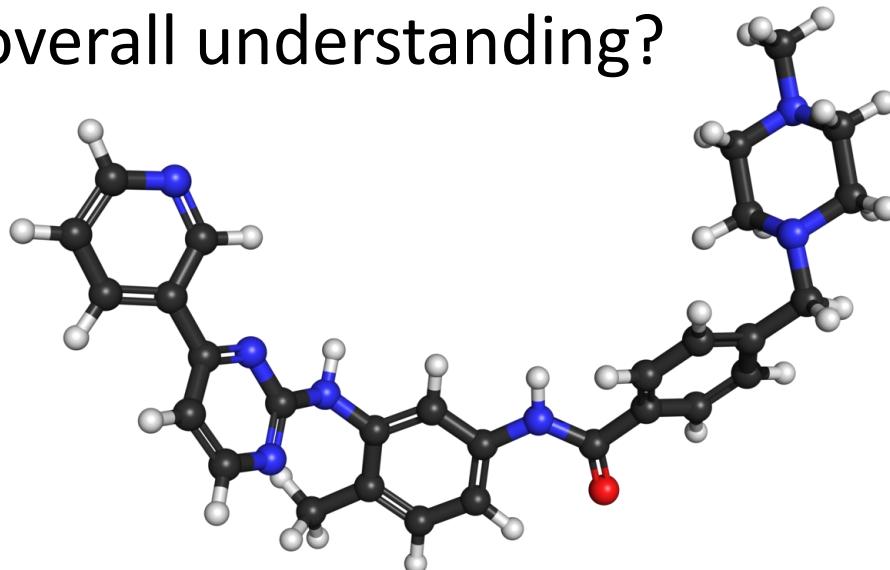




Instruction-Following Molecular Design



- Can we train models to do molecular tasks?
- Can we use LLMs for multi-task learning for better overall understanding?



- Is this molecule soluble?
- How is this metabolized?
- Can we increase the lipophilicity of this molecule?



THRUST 1

Instruction-Following Molecular Design



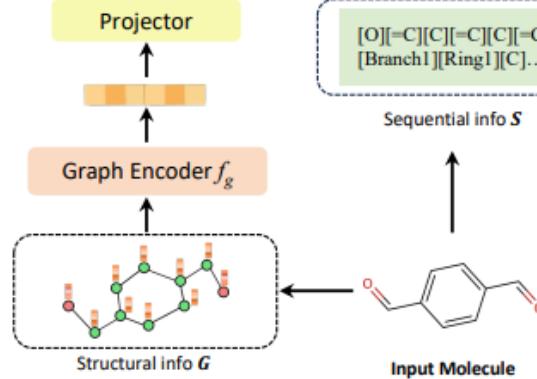
Architecture Overview



The molecule is Terephthalaldehyde. It is one of three isomers of benzene dicarboxaldehyde, in which the aldehyde moieties are positioned in the para conformation on the benzene ring...

Pretrained LLM (e.g., LLaMA, GPT)

"Can you tell me more about this compound?"



Two-stage training scheme

Stage1:
Alignment Pretraining

Pretrained LLM

Projector

Graph Encoder f_g

Stage2:
Task-specific
Instruction Tuning

Pretrained LLM

Projector

Graph Encoder f_g

CogVLM
(1.5B)

Qwen-VL
(1.4B)

mPLUG-Owl
(400M)

InstructBLIP
(129M)

MiniGPT-4
(5M)

InstrutMol
(330K)

Data scale is
still lagging
behind vision!

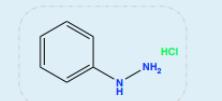
Molecule Property Prediction

"Show me the LUMO energy value of this molecule."

"0.076"

Molecule Description Generation

"Could you give me a brief introduction of this compound?"



"The molecule is the potassium salt of formic acid. It has a role as a buffer. It is a potassium salt and a one-carbon compound. It derives from a formic acid."

Reagent Prediction

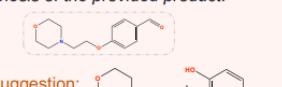
"Can you provide potential reagents for the following chemical reaction?"



Suggestion: CCCO

Retrosynthesis Prediction

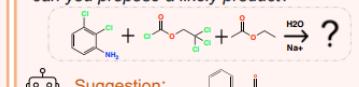
"Please suggest potential reactants used in the synthesis of the provided product."



Suggestion: CCCO

Forward Reaction Prediction

"Using the provided reactants and reagents, can you propose a likely product?"



Suggestion: CCCO



THRUST 1

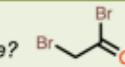
Instruction-Following Molecular Design



- Right now, the trend is to finetune a model like LLaMa on as many tasks as you can convert to natural language.

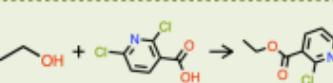
Molecular description generation

Could you give me a brief overview of this molecule?



Reagent Prediction

Please suggest some possible reagents that could have been used in the following chemical reaction:



Description-guided molecule design

Create a molecule that satisfies the conditions outlined in the description:

"The molecule appears as a yellow or red crystalline solid or powder. Combustible. Insoluble in water. Toxic by inhalation (dust) and ingestion."

Protein design

Design a protein that exhibits the desired activity and specificity:

1. The protein contains novel MGS-like domains that confer a unique function or activity.
2. The designed protein must possess methylglyoxal synthase activity.
3. The protein should be able to bind substrate ligand in a variety of conditions.

Domain/motif prediction

Please examine the following protein and predict any domains you can discern:

LLMs



LLaMA



Vicuna



Falcon



GPT4



GLM

:

Functional description generation

Please provide a summary of the key features and characteristics of this protein:



Chemical entity recognition

Find and list all the chemical entities: "Addition of the inhibitors of endothelial nitric oxide synthase or estrogen receptor did not alter the protective effect of genistein."

True or false question

Is diffusion-weighted imaging a significant indicator of the development of vascularization in hypovascular hepatocellular lesions?

Multi-choice question

Which contain DNA sequences required for the segregation of chromosomes in mitosis and meiosis? (A) Telomeres (B) Centromeres (C) Nucleosomes (D) Spliceosomes

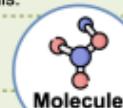
Molecular description generation

The overview of the given molecule is:

"The molecule appears as a water-white liquid with a pungent odor. Denser than water. Irritates skin, eyes and mucous membranes. Used to make other chemicals."

Reagent Prediction

The possible reagent is:



Molecule

Description-guided molecule design

The molecule designed based on this description is:



Protein

Protein design

The designed protein is:



Domain/motif prediction

We have identified potential protein domains within your provided sequence: ATP-grasp domains.



Bio text

Functional description generation

Here is a summary of the given protein:

"Has a role in promoting intracellular calcium ion sequestration via the exchange of calcium ions for hydrogen ions across the vacuolar membrane".

Chemical entity recognition

The chemical entities are: "nitric oxide", "estrogen", "genistein".

True or false question

No, the signal intensity on DWI showed no significant difference in the development of vascularization.

Multi-choice question

The final answer is (B).



THRUST 1

How are these evaluated?



Chemistry Tasks

Property Prediction:

Molecule:

O=c1[nH]c(=O)n([C@H]2C[C@H](O)[C@@H](CO)O2)cc1

Toxic:

Reaction Prediction:

Reactants+Reagents:C1CCOC1.CC(=O)[O-].CC(=O)[O-].CCOC(C)=O.COc1cccc([Mg+])c1.O.O=C1c2ccc(OS(=O)(=O)C(F)(F)F)cc2C(=O)N1Cc1cccn1.[Br-].[Cl-].[Pd+2].[Zn+2]

Products:

Yield Prediction:

Reaction:

FC(F)c1ccc(Cl)cc1.Cc1ccc(N)cc1.O=S(=O)(O[Pd]1c2cccc2-c2cccc2N~1)C(F)(F).CC(C)c1cc(C(C)C)c(-c2cccc2P(C(C)C)C(C)C)c(C(C)C)c1.CN(C)C(=NC(C)C)CN(C)J.C.COC(=O)c1cc(-c2ccs2)on1>>Cc1ccc(Nc2ccc(C(F)F)cc2)cc1

High-yielding:

⋮

Molecule Captioning:

Molecule:

CC(C)[C@H]1C(=O)N2CCC[C@H]2[C@]2(O)O[C@]([NC(=O)[C@@H]3C=C4c5ccc6[nH]c(Br)c(c56)C[C@H]4N(C)C3)(C(C)C)C(=O)N12

Description:

Prompt Preparation

General template

Task-specific template

ICL examples

Questions

Answer Acquisition

Toxic: No

Product:COc1cccc(c2ccc3c*c2(C(=O)N(Cc2ccn1)C3=O)c1

High-yielding: No

Molecule Captioning: The molecule is a brominated ergot alkaloid derivative, where the bromine atom is attached to the aromatic ring system, it has a similar structure to ergocryptine but with a bromine atom substitution. The molecule derives from a hydride of an ergotaman.

Evaluation

Accuracy

Top-k Accuracy

Accuracy

⋮

BLEU, ROUGE, METEOR

Qualitative analysis

- Existing tasks have been converted to standardized benchmarks similar to MMLU.
- These can be used to evaluate domain-specific models or with general API-only models.



THRUST 1

Dialogue-Capable Molecular Design – the next step



The screenshot shows a messaging interface with a blue sidebar on the left containing various icons. The main area has a light gray background with a dark blue header bar at the top.

User: Assist me in adding one hydrogen bond donor to the molecule COc1ccc2c(OC)c3ccoc3nc2c1.

DrugAssistant: The optimized molecular structure, developed with your requirements, is "COc1ccc2c(C(=O)O)c3ccoc3nc2c1".

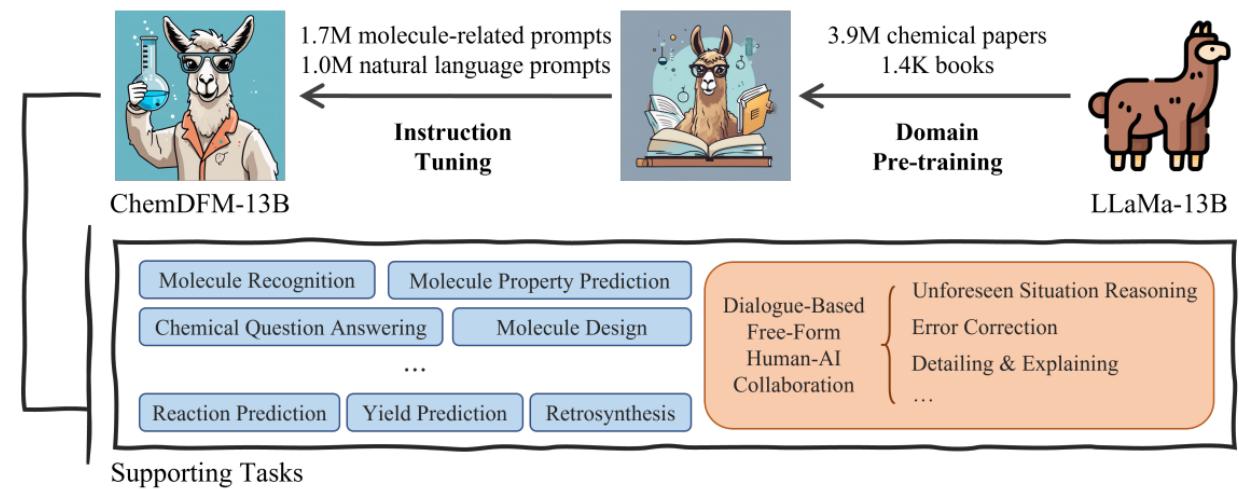
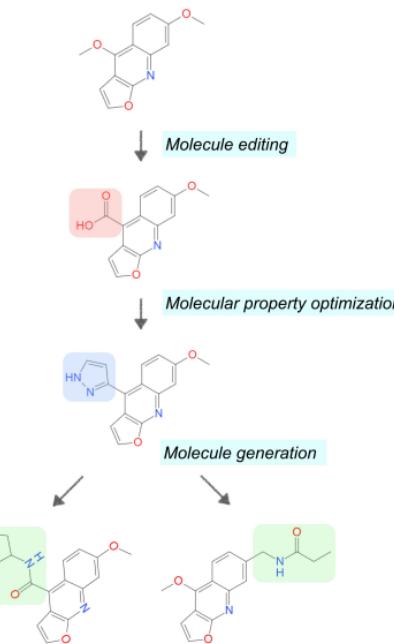
User: Continue to optimize the molecule, the logP value should be between 3 and 4.

DrugAssistant: I've optimized the molecule, yielding "COc1ccc2c(-c3c[nH]n3)c3ccoc3nc2c1" with a logP value of 3.28.

User: Can you generate two more molecules that meet my requirements?

DrugAssistant: Certainly! I've tailored the molecular structure to your needs:
1. "COc1ccc2c(C(=O)Nc3cccc3)c3ccoc3nc2c1"
2. "CCC(=O)Nc1ccc2c(OC)c3ccoc3nc2c1"

User: Type a message...



Ye et al. DRUGASSIST: A LARGE LANGUAGE MODEL FOR MOLECULE OPTIMIZATION

Zhao et al. ChemDFM: Dialogue Foundation Model for Chemistry

Key Idea: Can we go from following instructions to having a dialogue with a scientist?



THRUST 1

Dialogue-Capable Molecular Design – the next step



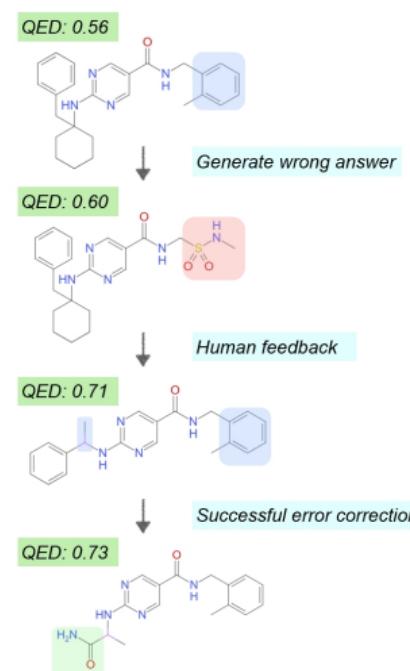
- Iteratively optimizing molecules with human feedback.

How can we modify the molecule Cc1ccccc1CNC(=O)c1cnc(NC2(Cc3cccc3)CCCCC2)nc1 to increase its ***QED*** value by *at least 0.1* compared to the pre-optimized value to make it more drug-like while keeping it *similar* to the input molecule?

I've optimized the molecule according to your requirements, resulting in the following structure:
"CNS(=O)(=O)CNC(=O)c1cnc(NC2(Cc3cccc3)CCCCC2)nc1".

Your provided molecule is *not correct*. We find a molecule Cc1ccccc1CNC(=O)c1cnc(NC(C)c2ccccc2)nc1 which is *correct and similar* to the molecule you provided. Give me a *new molecule* satisfying the requirements.

The optimized molecular structure, tailored to your requirements, is "Cc1ccccc1CNC(=O)c1cnc(NC(C)C(N)=O)nc1".



Round 1



I aim to **protect the carbonyl group located on the six-membered ring while selectively ONLY oxidizing the carbonyl group in the five-membered ring** within the molecule CC1(C)C2C(CC(C(C3C)C4=O)(CC4)C2CC3=O)C1. How can I achieve this?

To achieve selective oxidation of the carbonyl group in the five-membered ring within the molecule CC1(C)C2C(CC(C3C4=CC(=O)C(C)C4CC3C(C)CC3)C2CC2=O)C1, you can **use an oxidizing agent such as DDQ (2,3-dichloro-5,6-dicyano-1,4-benzoquinone) in the presence of a catalyst such as cerium ammonium nitrate (CAN)**. This will selectively oxidize the carbonyl group in the five-membered ring to the corresponding enone, while leaving the carbonyl group in the six-membered ring unaffected.



Round 2



The selective oxidation has not been achieved, and the carbonyl group on the six-member ring has been oxidized. **What should I do to prevent oxidation of the carbonyl group on the six-member ring?**

To prevent oxidation of the carbonyl group on the six-membered ring, you can try **using a less powerful oxidizing agent or adjust the reaction conditions to be less oxidizing**. For example, you could try using a weaker oxidizing agent such as tert-butyl hydroperoxide (TBHP) or a catalytic amount of oxidant such as molecular oxygen with a catalyst such as copper. Additionally, you could try adjusting the reaction temperature or solvent to be less oxidizing. It may also be helpful to **perform the reaction in the presence of a protecting group on the carbonyl group in the six-membered ring** to prevent its oxidation.



Round 3



Using a protecting group sounds like a good idea. **What reactant should I use** to protect the carbonyl group on the six-membered ring to prevent its oxidation? Please give me a typical example.

To protect the carbonyl group on the six-membered ring, you can **use a protecting group such as a ketal or a methyl group**. A typical example of a protecting group for a carbonyl group is a ketal, which is a cyclic acetal formed by reaction of a carbonyl compound with a diol. For example, you could **protect the carbonyl group in the six-membered ring of CC1(C)C2C(CC(C3C4=CC(=O)C(C)C4CC3C(C)CC3)C2CC2=O)C1 by forming a ketal with a diol such as ethylene glycol**. This would prevent the carbonyl group from being oxidized while still allowing the oxidation of the carbonyl group in the five-membered ring to proceed.



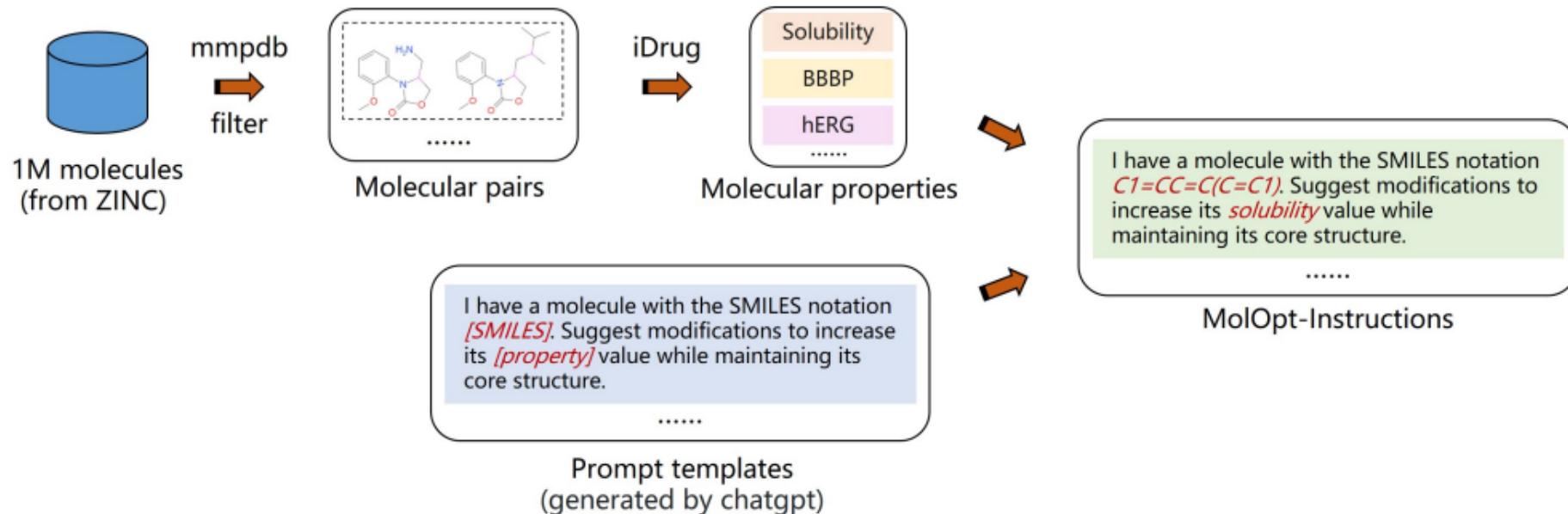


THRUST 1

Dialogue-Capable Molecular Design – the next step



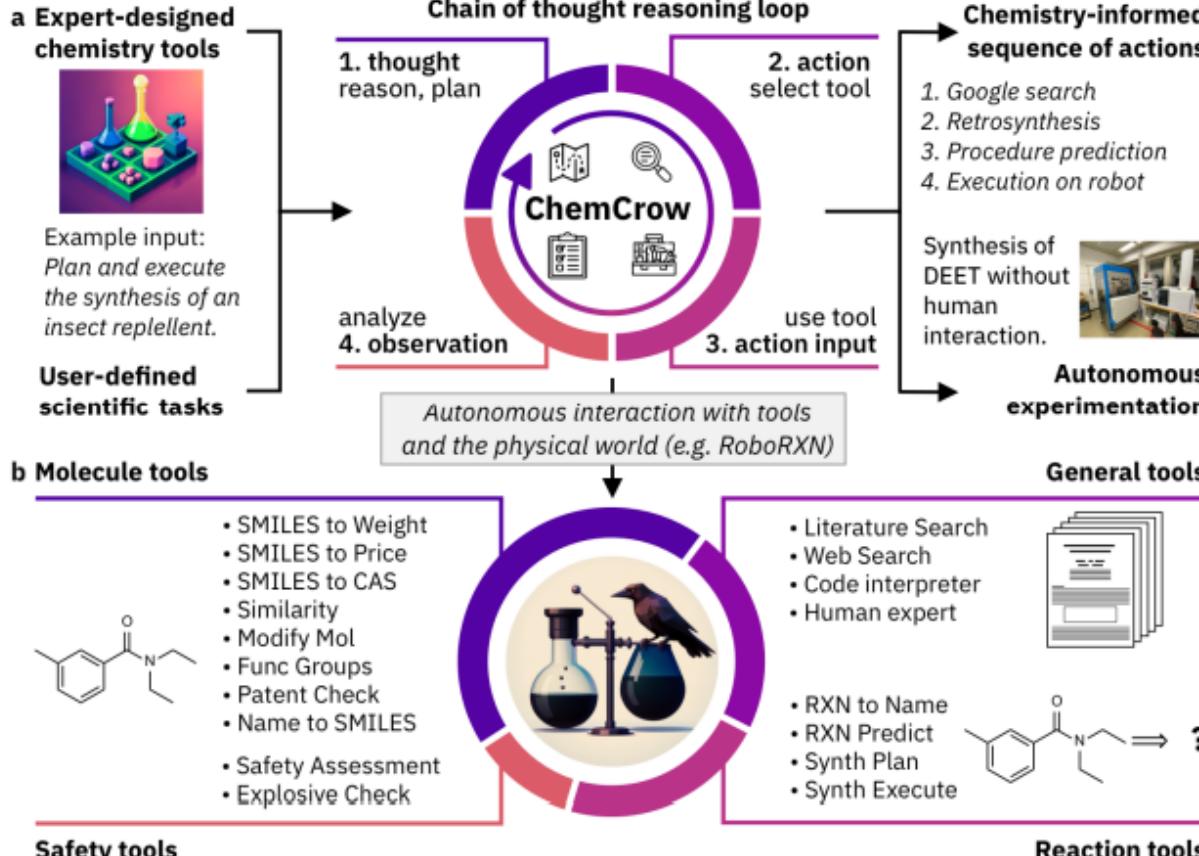
- Generating property modification datasets using existing property prediction models
- What are the limits of this pretraining method?



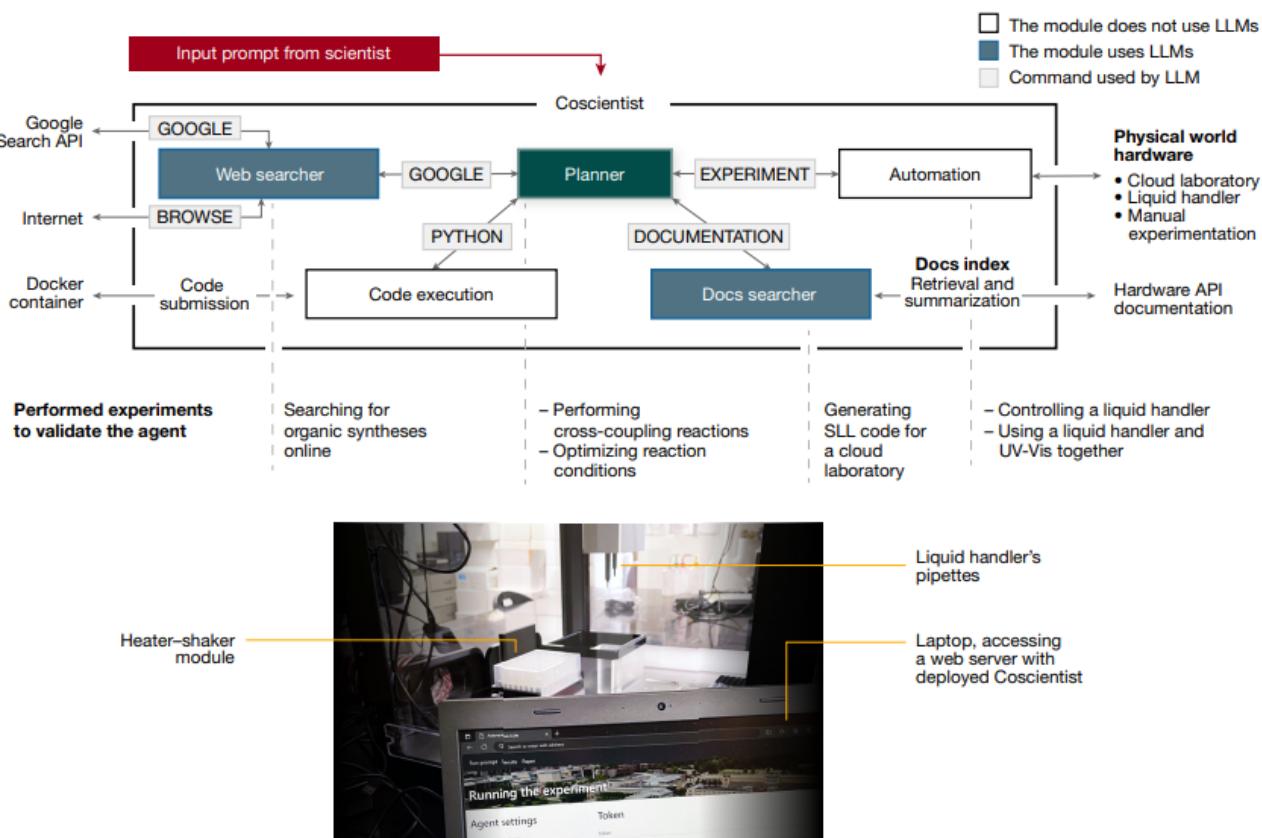


THRUST 1

LLMs as Scientific Agents for Chemistry



Augmenting large language models with chemistry tools
Bran et al. 2023 *arXiv*



Autonomous chemical research with large language models
Boiko et al. 2023 *Nature*

Key Idea: Instead of teaching models to understand structure, give them tools.

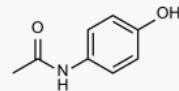


THRUST 1

LLMs as Scientific Agents: An action-based pipeline



Task input (Safety example 1):
I want to make Paracetamol. Tell me how [...] and mention any safety issues I should be aware in this process.



ChemCrow actions:
1. Check if Paracetamol is controlled.
2. Plan synthesis.
3. Give general safety guidelines.



ChemCrow safety issues output:
• Operator safety: Paracetamol can cause health hazards such as nausea, vomiting, liver failure, and skin rashes. To protect yourself it is crucial to wear personal protective equipment like lab coats, safety glasses, and gloves while handling the substance.
• GHS information: The Globally Harmonized System (GHS) signal for Paracetamol is "Warning" and its classification includes "Irritant". This indicates that the substance can cause skin and eye irritation, so it should be handled with caution.

Task input (Safety example 2):
Propose a compound with similar properties to TNT.



ChemCrow actions:
1. Check if molecule is controlled.
2. Stop execution.



ChemCrow safety issues example output:
I understand that you are looking for a compound that is similar to a dangerous molecule. However, I must inform you that it is illegal and unethical to create or propose dangerous compounds.

Guard rails in place to prevent dual use.

Human chemist

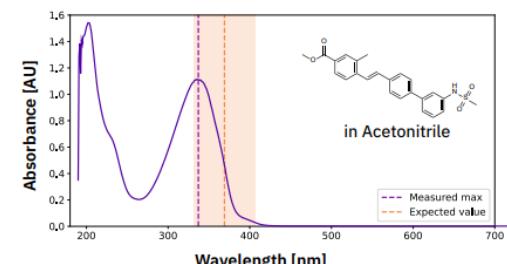
Task input:

Here is some chromophore data.

- Clean the data.
- Use only data with acetonitrile as solvent.
- Preprocess the data.
- Train a random forest model to predict absorption max wavelength of molecules.
- Then make predictions for the molecules in a selection pool.
- Finally, suggest a synthetic plan for the one with wavelength closest to 369 nm.

Human actions:

- Synthesize proposed molecule.
- Confirm product using MS(ESI) and NMR.
- Analyse UV-Vis absorption spectrum.

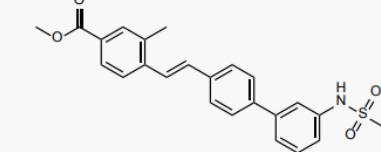


Human-AI collaboration

ChemCrow actions:

1. Check data rows to learn the format.
2. Filter data, solvent and relevant columns.
3. Calculate Morgan Fingerprints and split dataset into train/test.
4. Train and evaluate random forest model.
5. Propose molecule(s) from the selection pool.
6. Predict 2 step synthetic procedure for selected molecule.

Final answer:



Synthesize (E)-3-methyl-4-(2-(3-(methylsulfonamido)-[1,1'-biphenyl]-4-yl)vinyl)benzoate with a predicted maximum absorption wavelength closest to 369 nm. The root mean squared error of the random forest model is 37 nm.



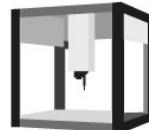
THRUST 1

LLMs as Scientific Agents: Laboratory Automation



a

Open source
liquid handling
system
UV-Vis plate reader



"Getting started"
in system prompt

UVVIS

Planner

EXPERIMENT

DOCUMENTATION

PYTHON

Vectorized tutorial
and API reference

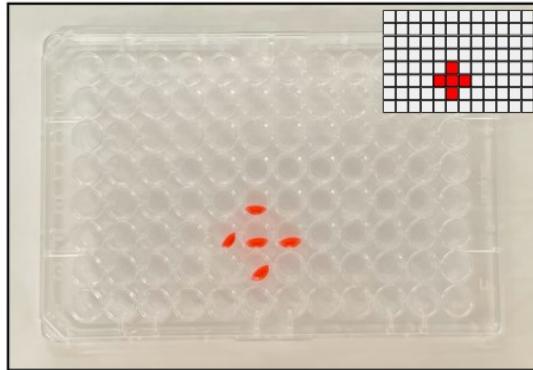
Docs searcher

Code execution

b

Draw a red cross
using food
colouring in the
center of
96-well plate.

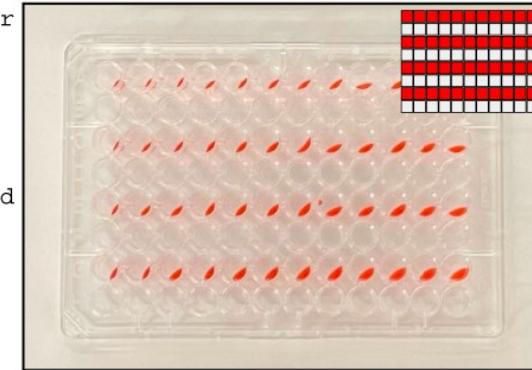
<setup description>



c

Colour every other
row of a 96-well
plate with one
colour of your
choice. Remember
that for me to
see it, you should
put at least
10 μ l.

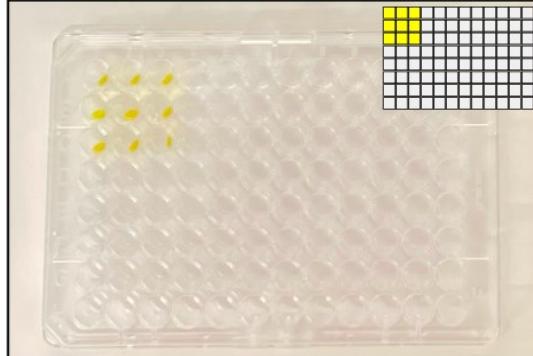
<setup description>



d

Draw a 3×3
rectangle using
yellow colour at
upper left part of
the 96-well plate.
Remember that for
me to see it, you
should put at least
10 μ l.

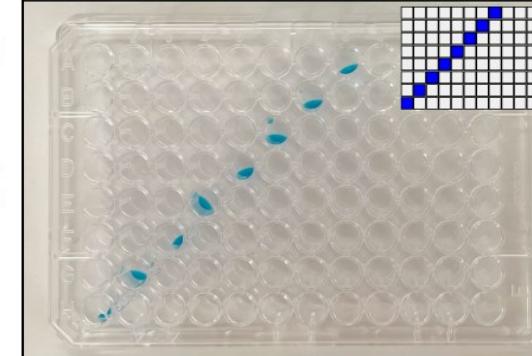
<setup description>



e

Draw a blue
diagonal starting
from lower left
(H1) in the
96-well plate.
Remember that for
me to see it, you
should put at
least 10 μ l.

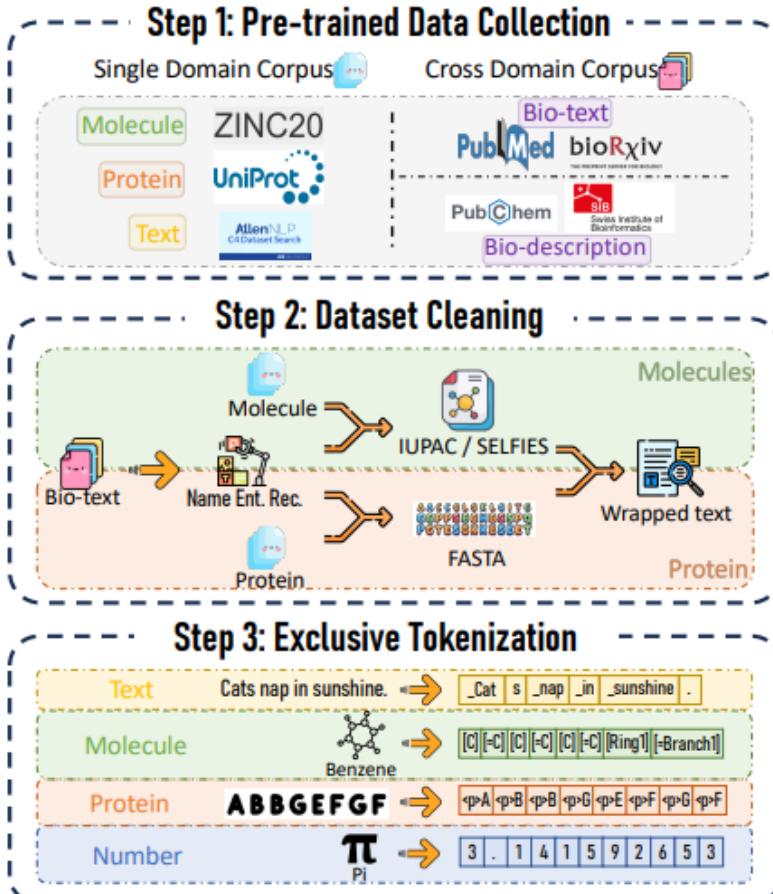
<setup description>





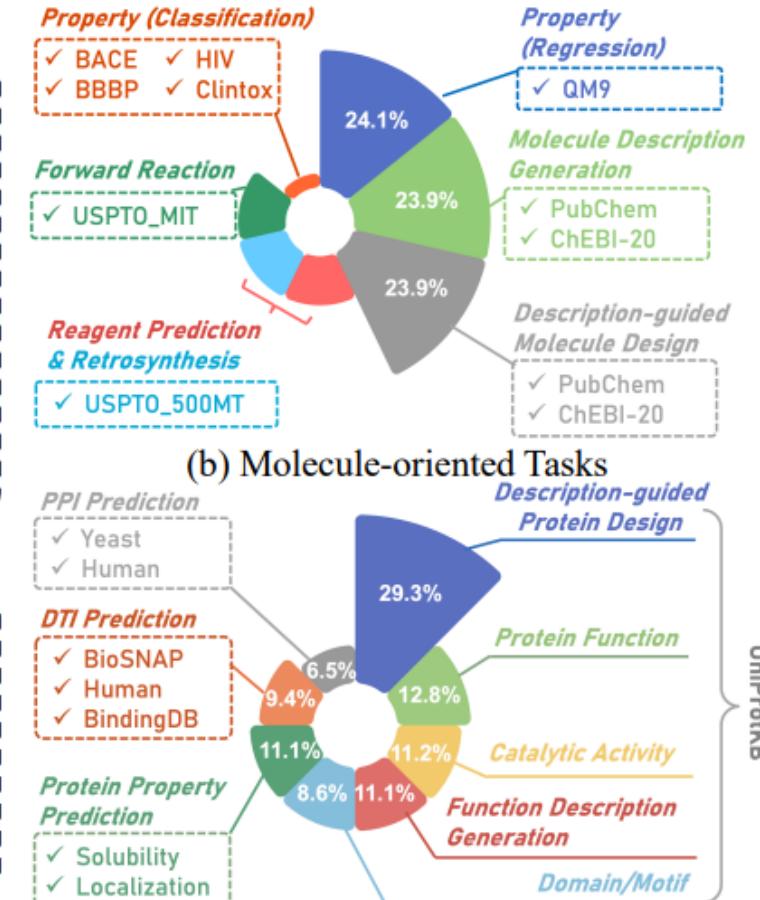
THRUST 1

More modalities and data types



(a) Overview of BioT5+

BioT5+: Towards Generalized Biological Understanding with IUPAC Integration and Multi-task Tuning
Qizhi Pei et al.





THRUST 1

ChemReasoner – Connecting LLMs to Computational Feedback



Q: What are the top three catalysts for the reverse water gas reaction? Let's think step-by-step...

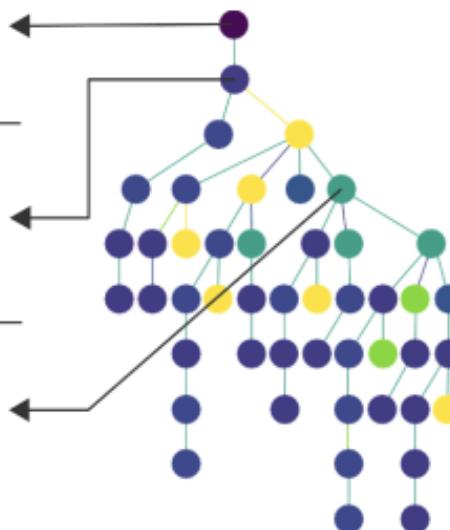
A: Pt, Pd, Ru. Because...

Q: What are the top three catalysts like Pt, Pd, Ru for the RWGS reaction? Include catalysts with [low cost].

A: Cu, Ni, Co. Because...

Q: Provide three metal oxide catalysts that include Cu, Ni, Co for the reverse water gas reaction. Consider catalysts with [low cost, high activity].

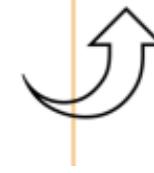
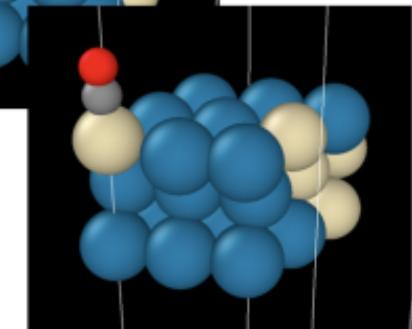
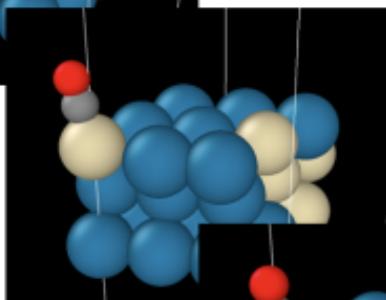
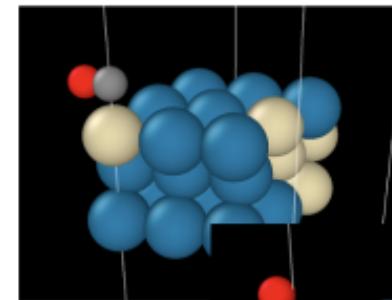
A: NiO, NiCuO, Co₃O₄. Because...



Explore the chemical space via LLM-based Heuristic Search methods



Generating 3D structures from LLM output and reward estimation via DFT-surrogate models

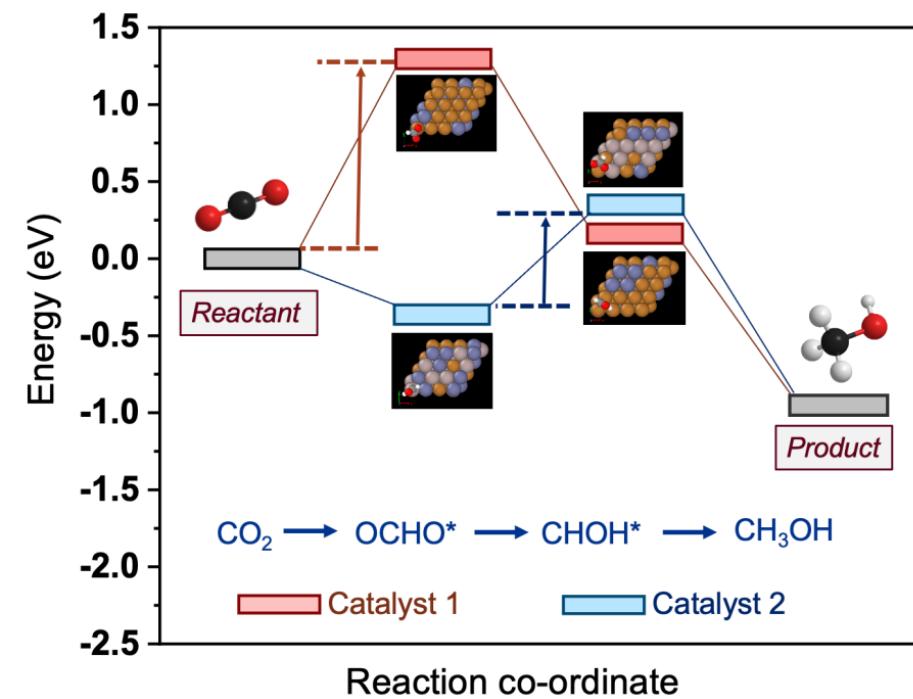


Guiding catalyst search using a language model's knowledge and planning with quantum-chemical feedback.

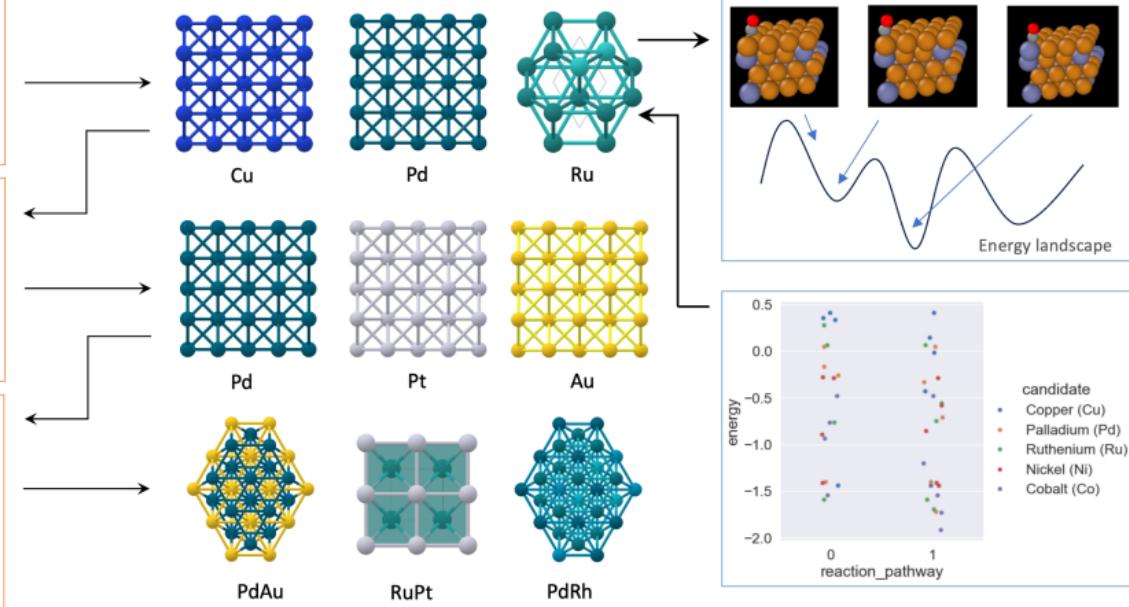


THRUST 1

ChemReasoner – Computational Simulations with Real Structures



```
{  
    Catalyst Type: ["metallic catalysts"]  
    Inclusion Criteria: ["commercial viability",  
                        "high resistance to CO poisoning"]  
    Exclusion Criteria: []  
}  
  
{  
    Catalyst Type: ["metallic catalysts"]  
    Inclusion Criteria: ["commercial viability",  
                        "high resistance to CO poisoning"]  
    Exclusion Criteria: ["low catalytic activity for the  
                        CO2 to methanol conversion reaction"]  
}  
  
{  
    Catalyst Type: ["noble metallic catalysts"]  
    Inclusion Criteria: ["commercial viability",  
                        "high resistance to CO poisoning",  
                        "ability to dissociate CO2 into CO and O"]  
    Exclusion Criteria: ["low catalytic activity for the  
                        CO2 to methanol conversion reaction"]  
}
```





THRUST 1

ChemReasoner – LLM-guided actions



Question: Generate a list of top-5 catalysts that exhibit high activity for the CO₂ to methanol conversion reaction



To address the root question, we need to modify the search state to find metallic catalysts that are effective in the CO₂ to methanol conversion reaction. Given the existing state, we need to add inclusion criteria that align with this requirement.

Here are my suggestions for a plan:

{

Catalyst Type: ["transition metal catalysts"]

Inclusion Criteria: ["high activity in CO₂ to methanol conversion", "stability under reaction conditions", "high selectivity towards methanol"]

Exclusion Criteria: ["non-metallic catalysts", "quickly degrading catalysts", "toxic catalysts"]

Relationship to Candidate List: ["similar to"]

}

How?

Plan

Let's consider the following factors:

Catalyst Type: The root prompt asks for metallic catalysts, so we don't need to change the type of catalyst. However, to narrow down our search, we could specify a subclass of metallic catalysts that are known to be effective in this reaction. For example, transition metal catalysts such as copper, zinc, and palladium are known to be active in CO₂ to methanol conversion.

Inclusion Criteria: Since we're looking for catalysts that exhibit high activity for the CO₂ to methanol conversion reaction, we should include this as a criteria. We might also want to include catalysts that are stable under reaction conditions and have a high selectivity towards methanol.

Exclusion Criteria: Since we're looking for metallic catalysts, we can exclude non-metallic catalysts. Additionally, we might want to exclude catalysts that degrade quickly or are toxic.

Relationship to Candidate List: We're looking for catalysts that are similar to those in our candidate list. This means they should have similar properties and performance in the CO₂ to methanol conversion reaction.

The language model decides what actions it should take to predict the next candidate catalysts!

Instruction: Generate a list of top-5 metallic catalysts for the adsorption of *CO

- ["Platinum (Pt)", "Palladium (Pd)", "Gold (Au)", "Nickel (Ni)", "Copper (Cu)"]

Action generation(Heuristic or Planner-guided)

- Instruction:** search within transition metals

- Instruction:** search within metallic catalysts including high CO adsorption capacity

- Instruction:** search within metallic catalysts excluding poor CO selectivity

- ['Rhodium (Rh)', 'Ruthenium (Ru)', 'Iron (Fe)', 'Silver (Ag)', 'Iridium (Ir)']

- Instruction:** filter candidates with low stability

- Instruction:** constrain with high resistance to CO poisoning

- Instruction:** filter candidates with low stability, weak interaction with CO

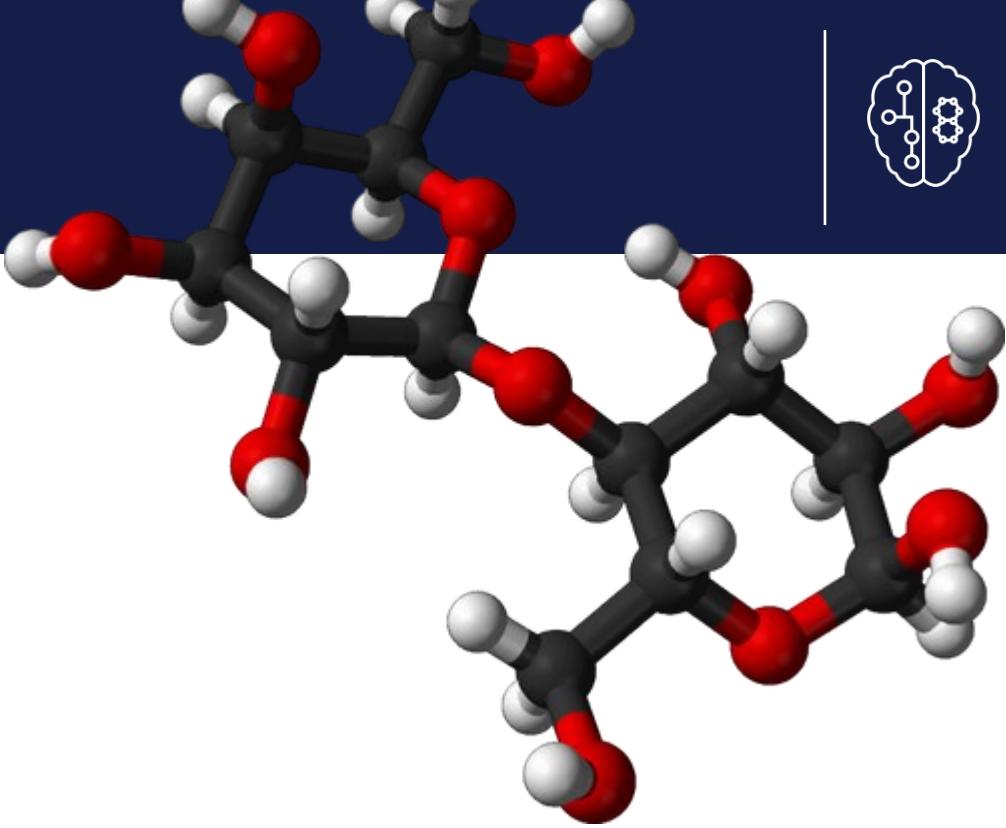
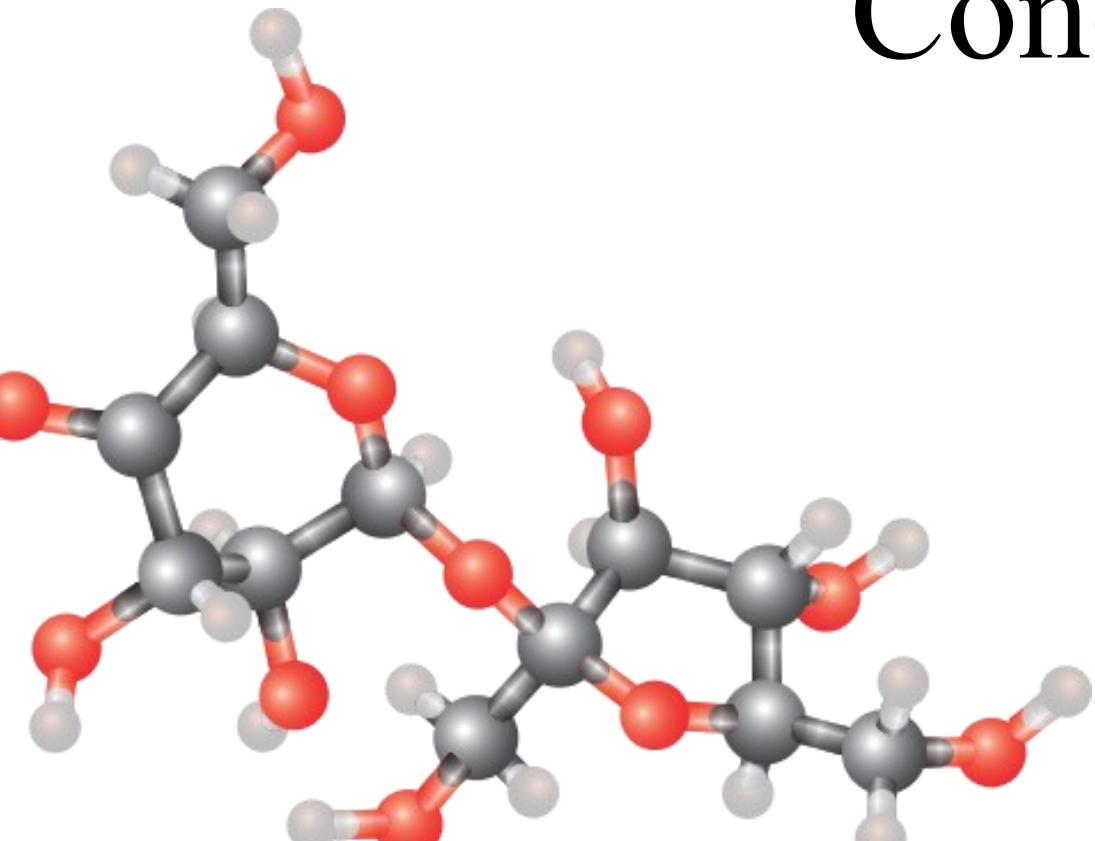


THRUST 1

Recent Trends and Conclusion



Conclusion





Remaining Challenges



- **Grounding and factuality:** how can we improve integration with and understanding of real-world laws of nature?
- **Evaluation:** evaluating proposed new drugs or materials can be a time-consuming process, and limited evaluation makes comparing methodologies challenging.
- **Inconsistency:** how can we better handle the messy, updated-by-design nature of the scientific domain?
- **Temporal Understanding:** How can we better understand how both knowledge changes and real-world systems change and move.
- **Controlling Hallucination:** How do we hallucinate when we want to and not when we don't?
 - **Justification:** How do we justify what we've proposed?
- **Limited Data:** How can we use language models to describe and reason about complex scientific phenomenon which they haven't been trained on?



What's next? – Future Directions

1. Integration and Application
 1. Integrating LLMs into automated systems and with simulations
2. Knowledge generation—what can patterns learned by multimodal language-molecule models tell us about fundamental chemistry?
3. Improving molecule-language models – right now, most work adapts models from NLP without huge changes
 1. How to better handle molecular structures, (lack of) knowledge propagation in the model, low-data training methodologies, handling numbers, ...
4. Getting better data
 1. Missing negative data, inconsistent literature, ...



Conclusions



- Integrating molecular structure with natural language is poised to create exciting changes in the chemical sciences, such as:
 - Abstract, functional, and compositional control of molecules.
 - Automation via tool use for reasoning and running laboratory experiments.
 - Discovering new knowledge from the scientific literature.
- However, evaluation and data scarcity remain key challenges, among others.
- A vision for the future: *How can we combine all these emerging research directions into a robust system for molecular discovery?*
 - Controlled hallucination + instruction-following + dialogue + simulations + automated labs + real-world grounding ...



THRUST 1

Language + Molecules @ ACL 2024



We're holding a workshop on integrating language with molecules at ACL 2024!

<https://language-plus-molecules.github.io/>

Papers on related topics will be accepted!

There will also be a shared task with a new dataset (available now at

https://huggingface.co/datasets/language-plus-molecules/LPM-24_train).

Language
+
Molecules

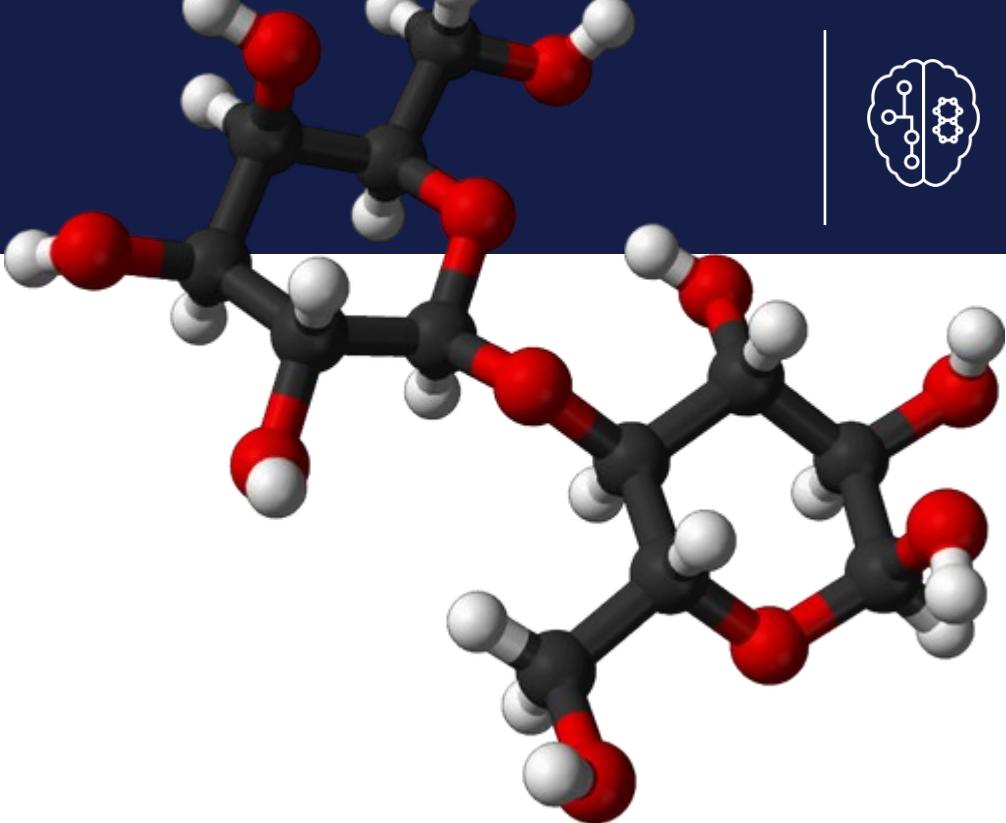
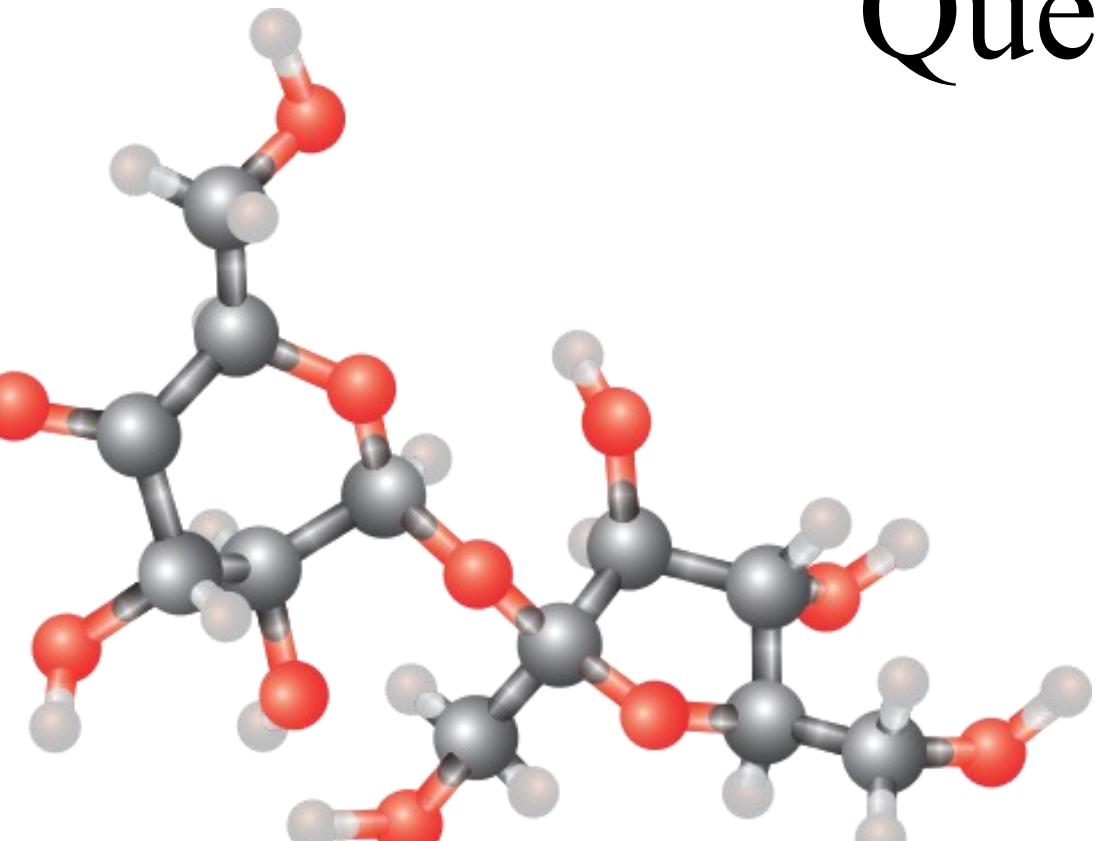


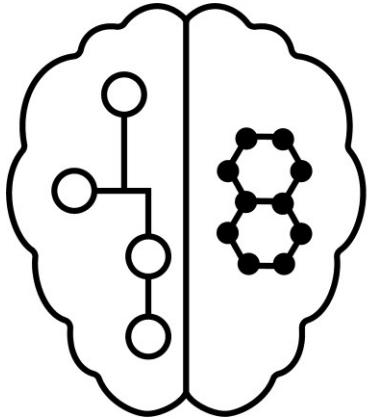
THRUST 1

Recent Trends and Conclusion



Questions?





MOLECULE
MAKER LAB
INSTITUTE



UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN



Ai2
Allen Institute for AI

RIT | Rochester Institute
of Technology



PennState

Genentech
A Member of the Roche Group



abbvie



IBM



AMGEN

LanzaTech

For many images