MOLECULE MAKER LAB INSTITUTE

Language-Guided Scientific Discovery for Chemistry

Carl Edwards

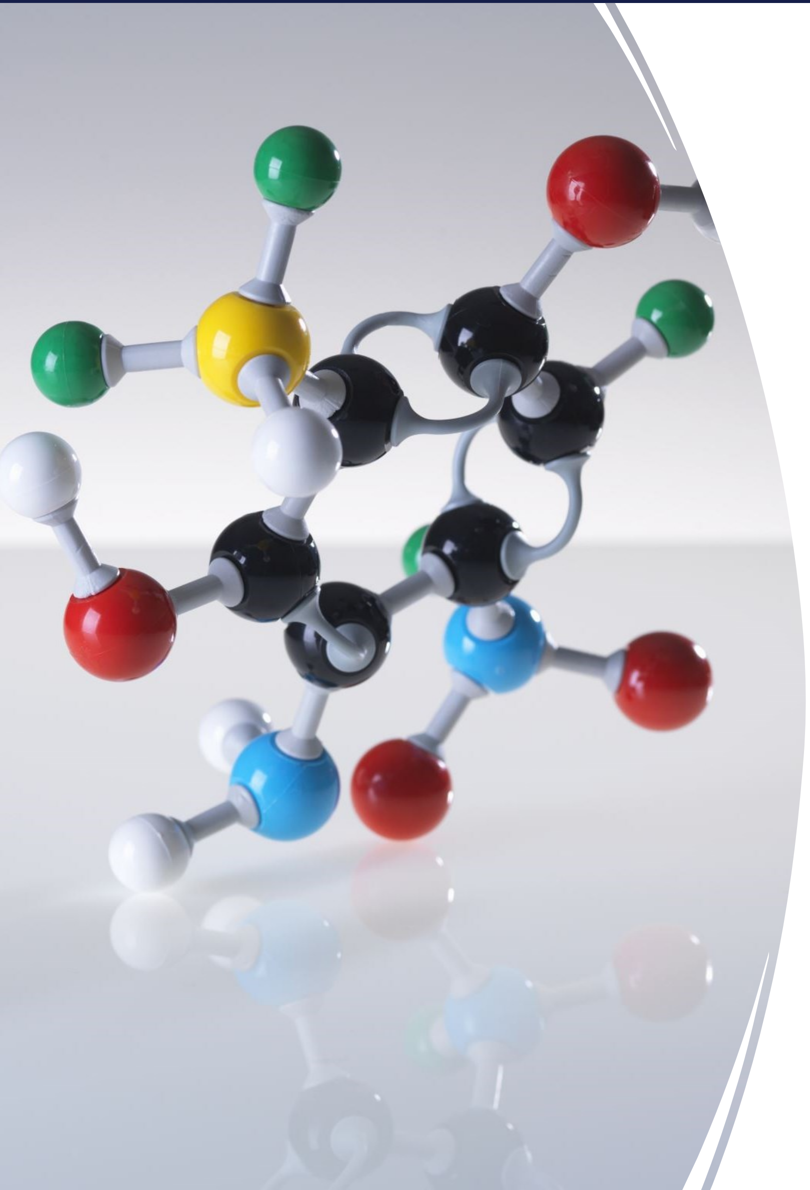
University of Illinois Urbana-Champaign

cne2@illinois.edu

BLENDER | Cross-source Information Extraction Lab

THRUST 1

HOW CAN WE INCREASE AI MODEL UNDERSTANDING OF CHEMISTRY?



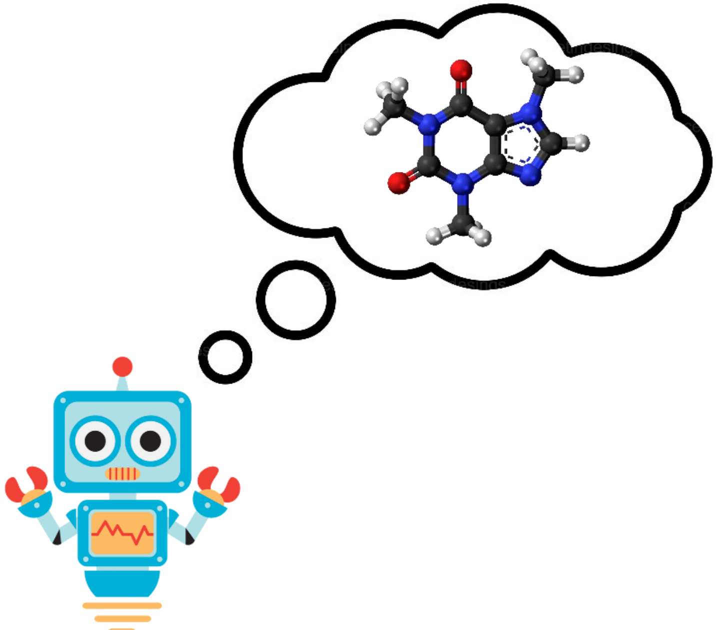
1. The inherent properties of molecules arising from their structure, composition, and interaction.
2. Centuries of work that has been collected into corpora by generations of scientists.

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 discovery
   a) Here, language model “hallucination” is actually a strength!

I’ll focus on 2, which is a newly developing field.

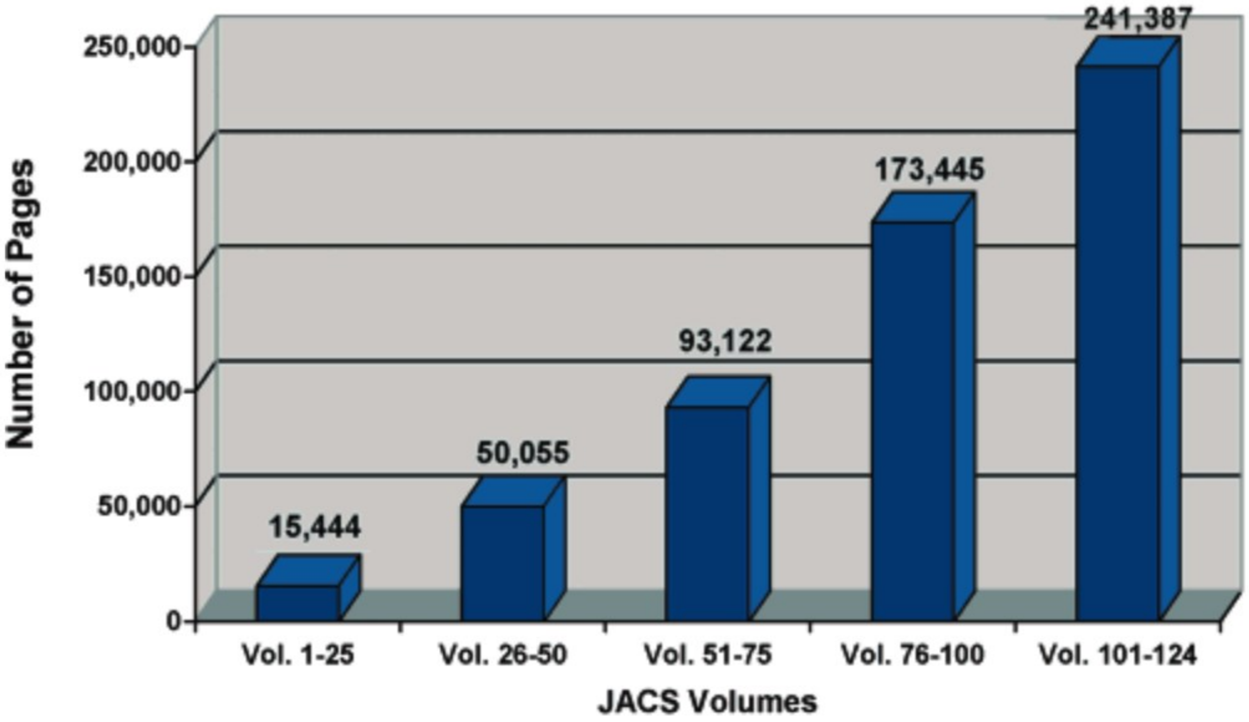


THRUST 1

1. Information extraction and retrieval for chemistry -- Motivation

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

Papers in JACS journals to 2003



J. Am. Chem. Soc. 2003, 125, 1, 1–8

THRUST 1

2. Language for controlling and interfacing with chemistry (the goal/my current working framework)

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.

Connecting Language and Molecules: Tasks

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?

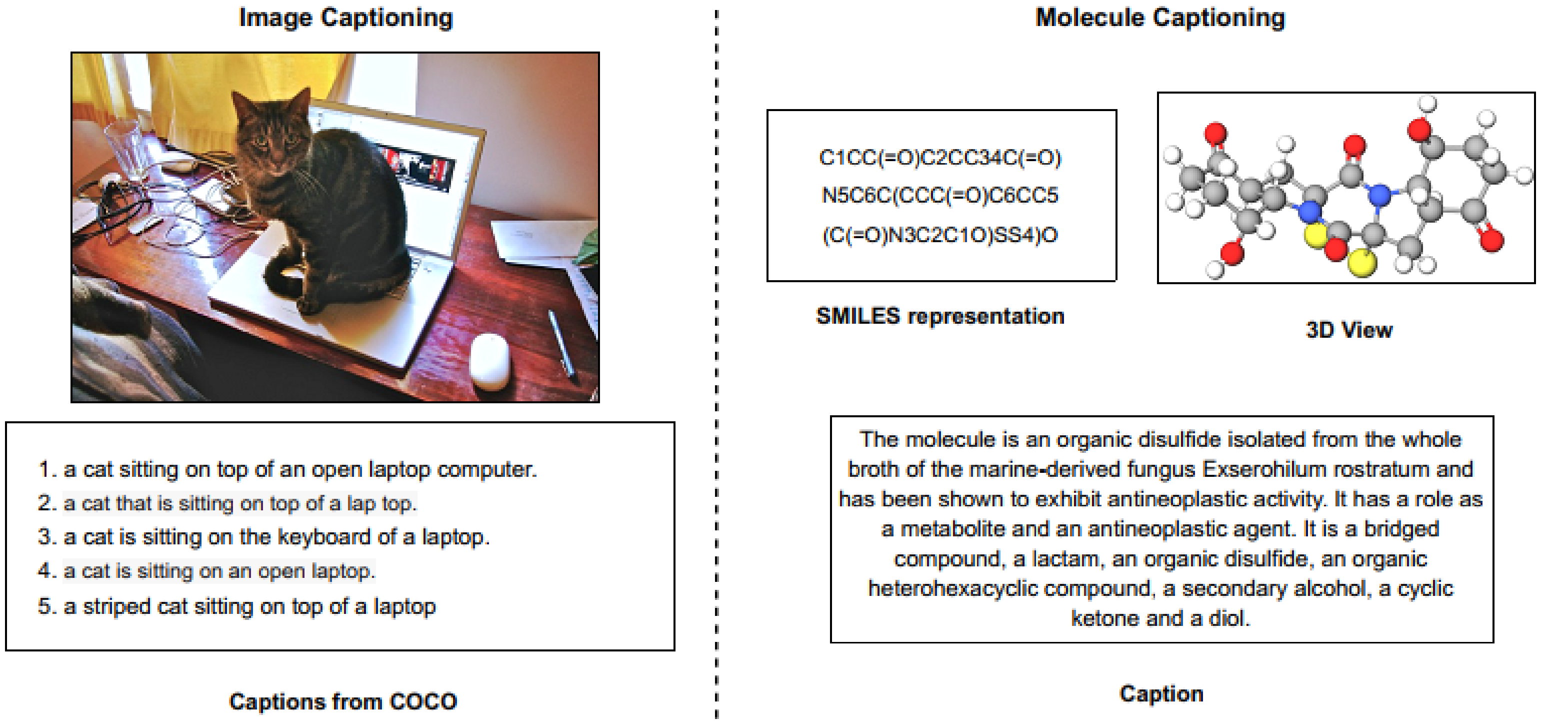


Image Captioning

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.

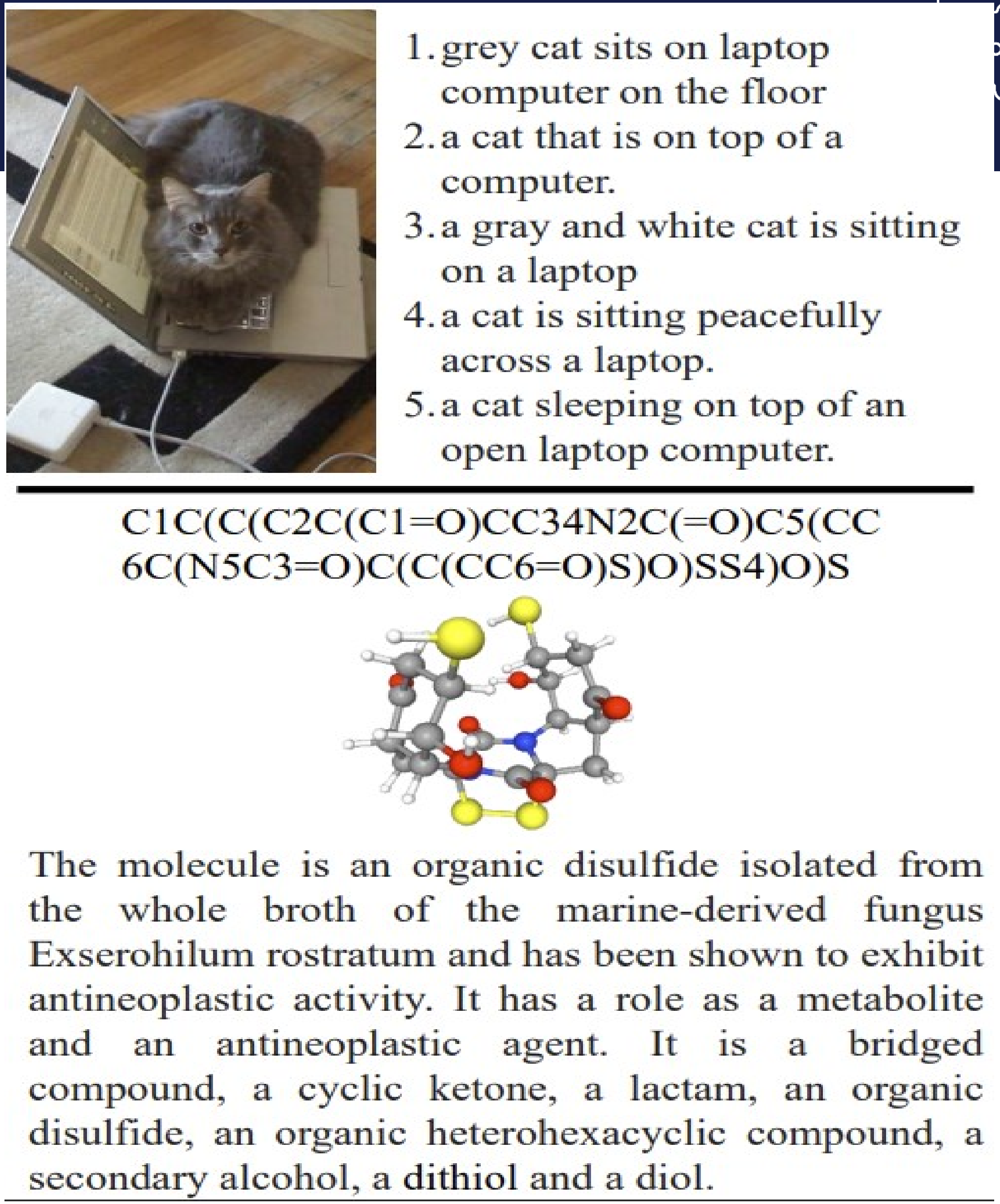
Caption

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.

C1C(C(C2C(C1=O)CC34N2C(=O)C5(CC6C(N5C3=O)C(C(CC6=O)S)O)SS4)O)S

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.

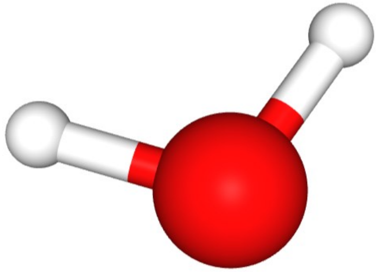
THRUST 1

Finding the right molecule: cross-modal retrieval

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

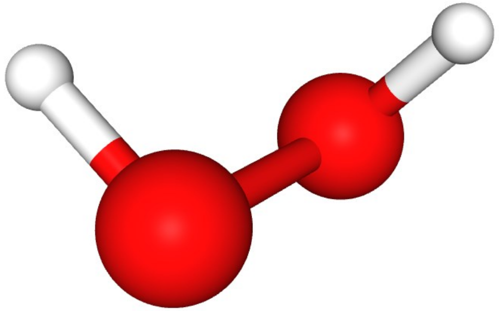
Query

1.



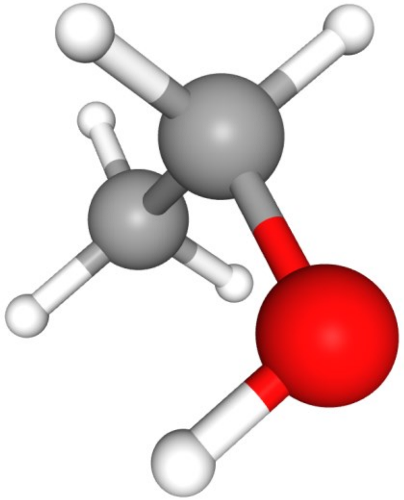
H_2O

2.



H_2O_2

3.

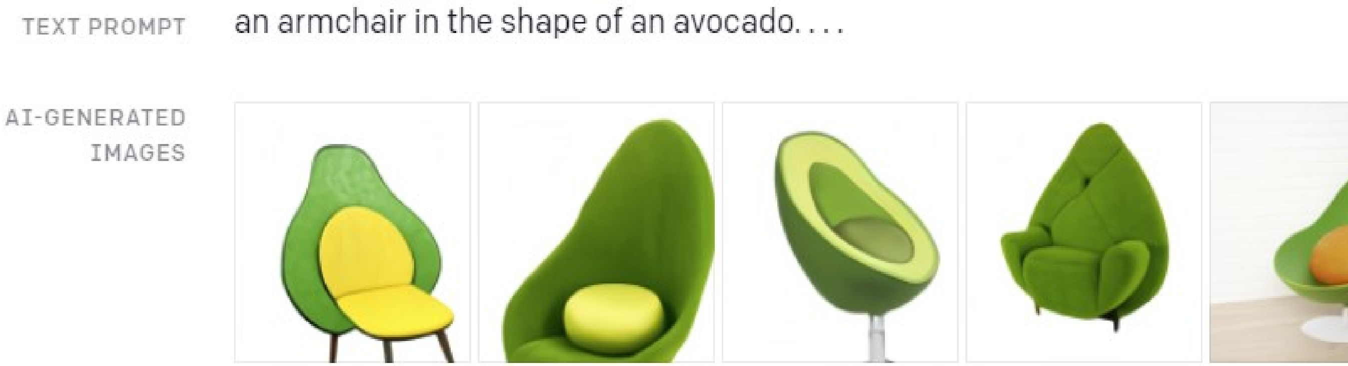


C_2H_6O

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?



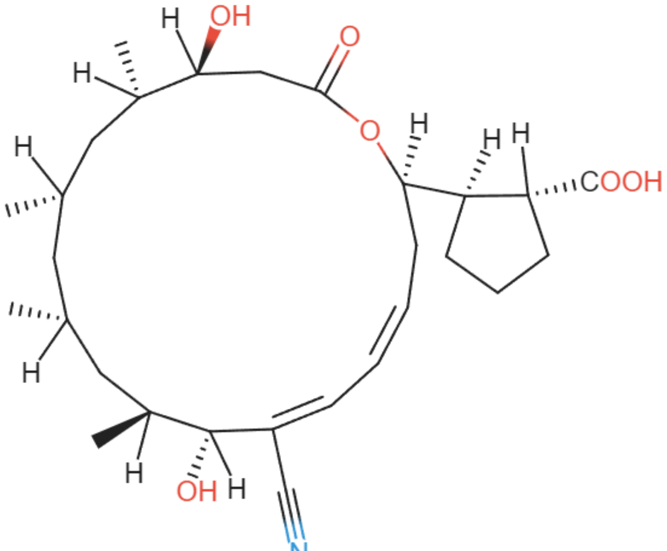
TEXT PROMPT an armchair in the shape of an avocado. . . .

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 antimalarial properties. It has a role as a bacterial metabolite, an antimicrobial agent, an antifungal agent, an antineoplastic agent, an apoptosis inducer and an antimalarial. It is a macrolide, a monocarboxylic acid, a secondary alcohol, a diol and an aliphatic nitrile.

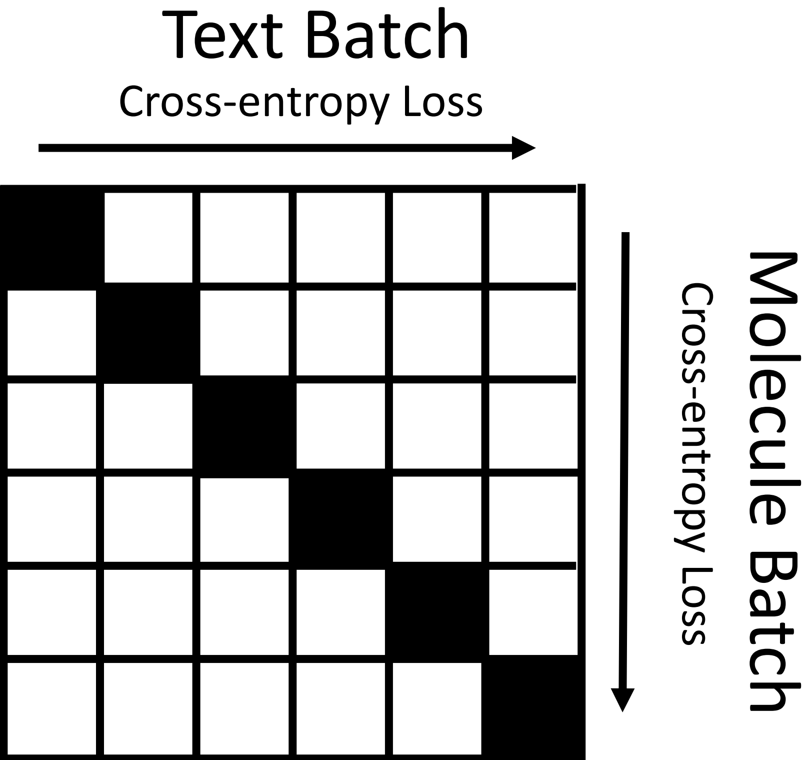


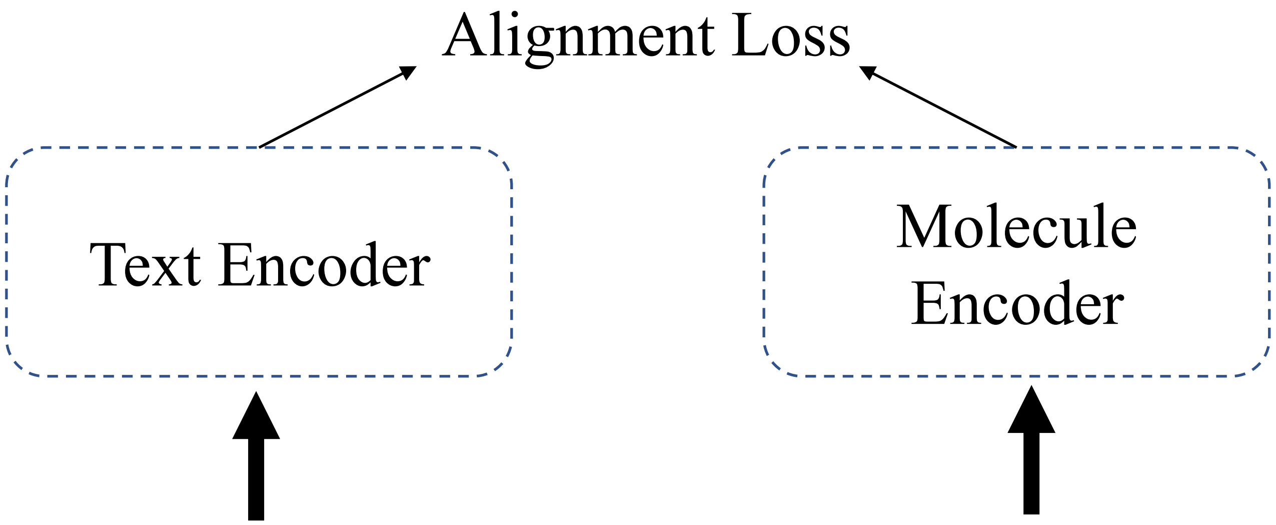
Connecting Language and Molecules: Bi-encoder models

THRUST 1

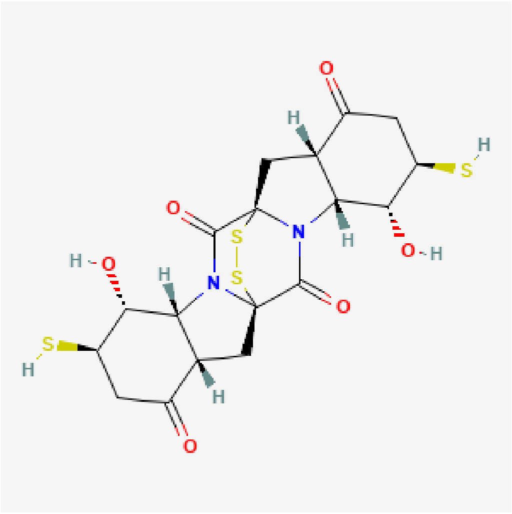
CONNECTING LANGUAGE AND MOLECULE: BI-ENCODER MODELS

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





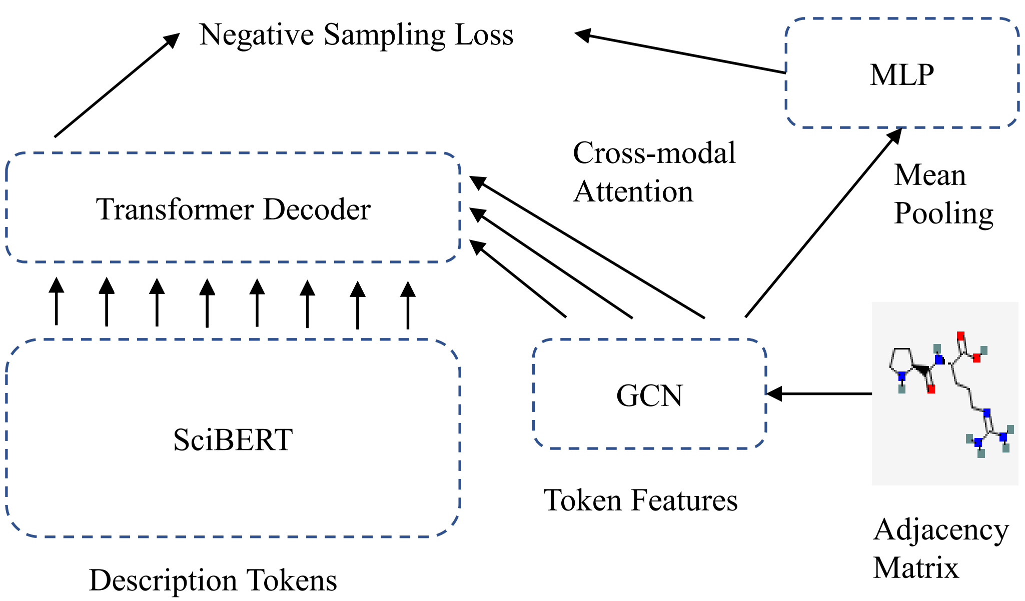
Rostratin D is an organic disulfide isolated from the whole broth of the marine-derived fungus Exserohilum rostratum and has been shown to exhibit antineoplastic activity. [...] It is a bridged compound, a cyclic ketone, a lactam, an organic disulfide, an organic heterohexacyclic compound, a secondary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



Edwards et al. EMNLP 2021

THRUST 1

LEARNING ASSOCIATIONS FROM TEXT-MOLECULE DATA



Negative Sampling Loss

MLP

Cross-modal Attention

Mean Pooling

Transformer Decoder

GCN

SciBERT

Token Features

Adjacency Matrix

Description Tokens

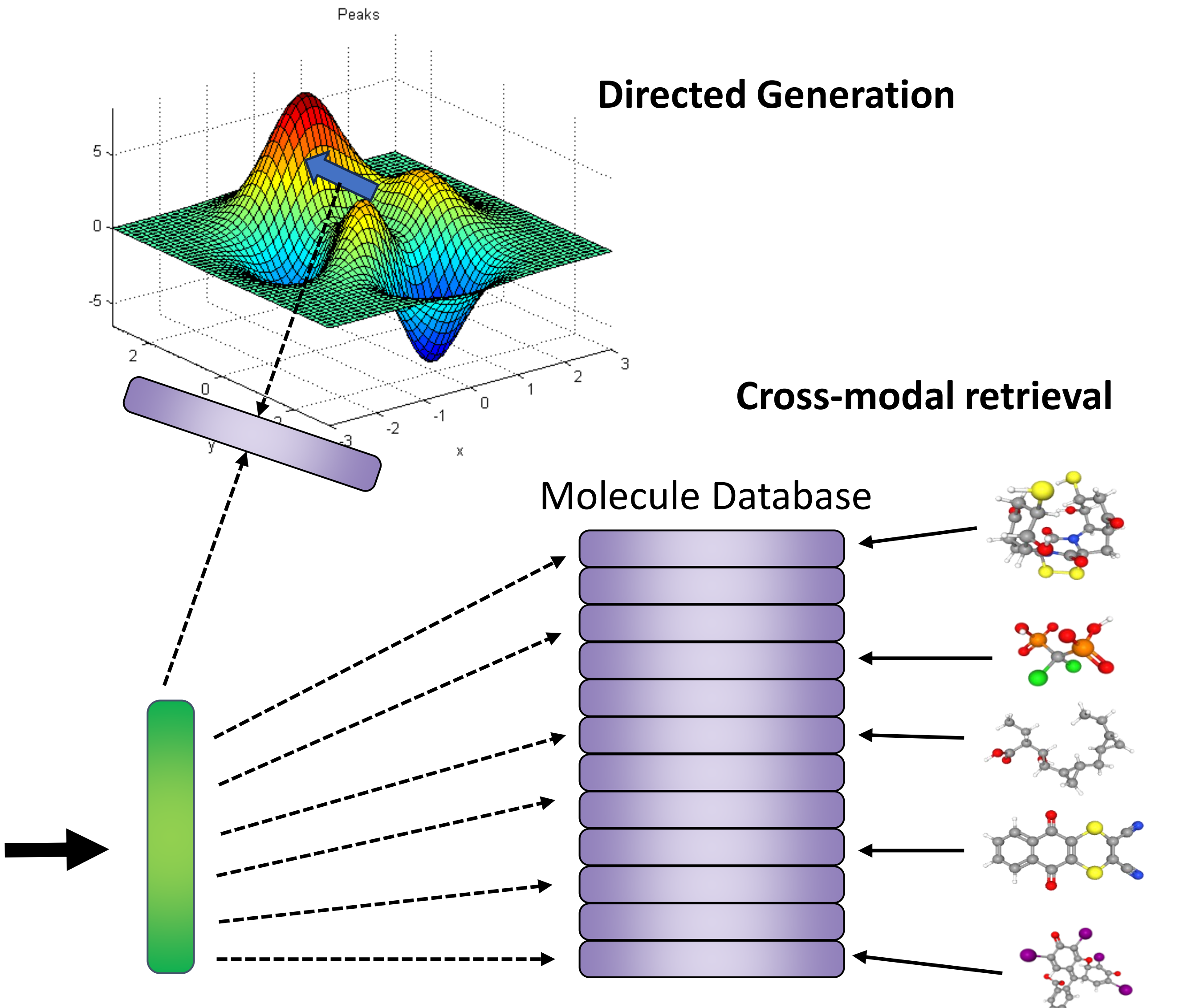
Token	Substructure	Supp	Conf
Titanium	$Ti{=}O$	1.29	0.65
Aluminium	Al^{3+}	4.31	0.23
Manganese	Mn^{2+}	10.08	0.30
Toluene	$C - C{=}C$	12.93	0.231
Toluene	C_7H_8	23.79	0.425
##chloro	$Cl - C$	18.81	0.207
pollutant	$F - C$	3.097	0.208
chromatography	$C - Si$	2.976	0.271
acid	$C - O - H$	2398.7	0.078
crown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

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. 2023)
- 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 […]

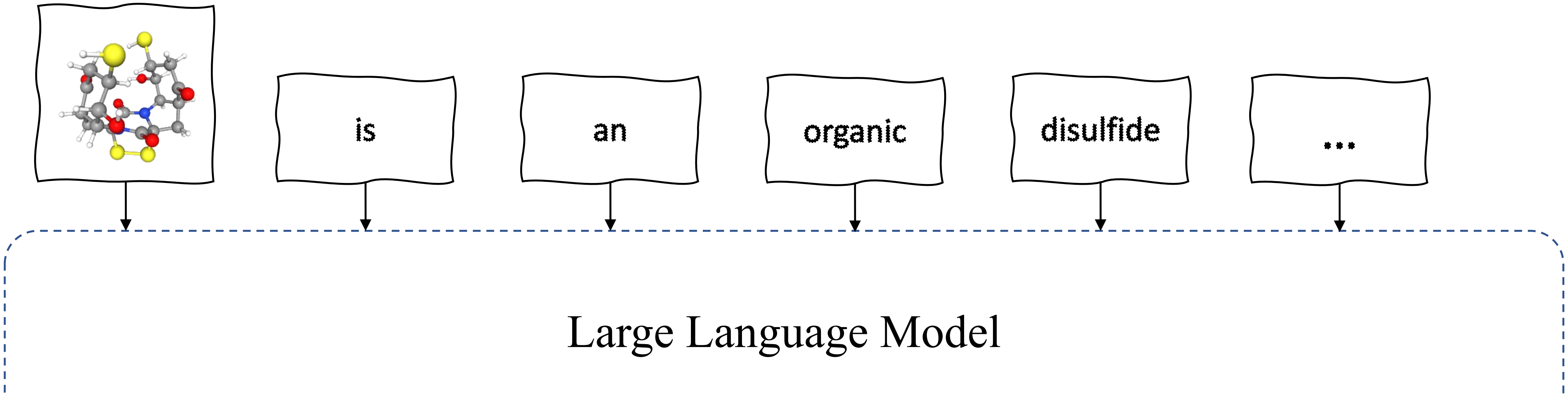


Directed Generation

Cross-modal retrieval

Molecule Database

Fine-Grained Connections between Language and Molecules: Joint-Representation Models



is

an

organic

disulfide

...

Large Language Model

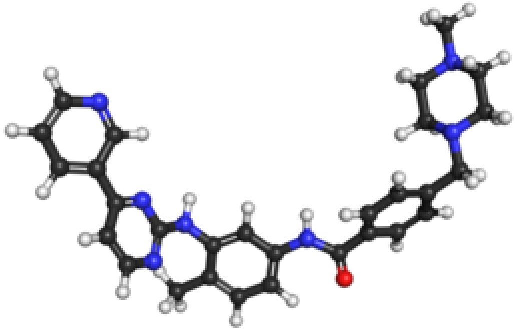
THRUST 1

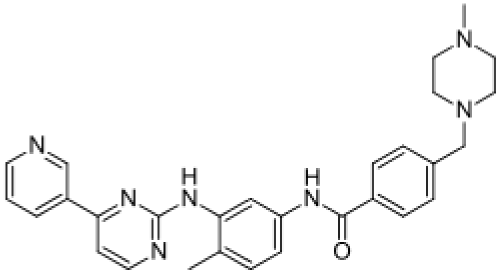
A brief aside

- Chemistry is a language

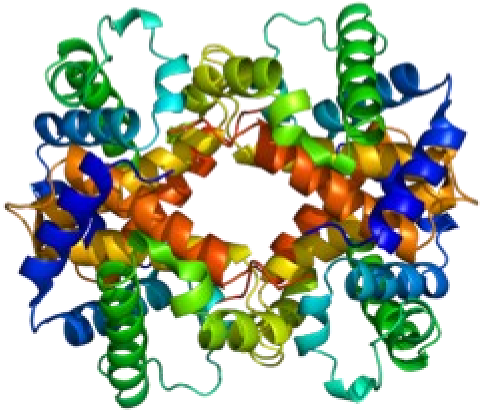
Machine Translation

Chemie ist eine Sprache





Cc1ccc(cc1Nc2nccc(n
2)c3cccnc3)NC(=O)c4
ccc(cc4)CN5CCN(CC
5)C



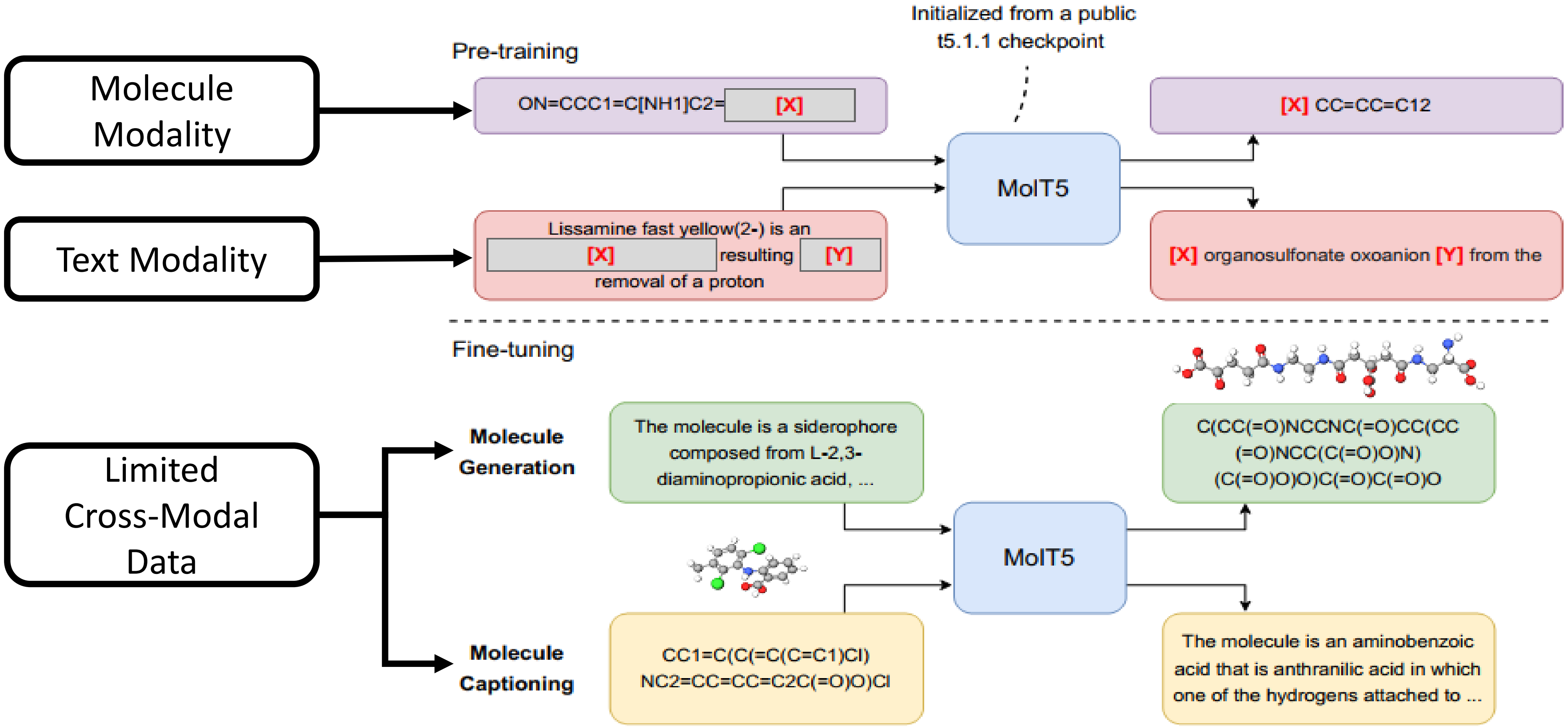
VLSPADKTNVKAAW
GKVGAHAGEYGAE
ALERMFLSFPTTKTY
FPHFDLSHGSAQV...

actcttctggtccccacagactcag
agagaacccaccatggtgctgtct
cctgccgacaa gaccaacgtca
aggccgcctggggtaaggt ...

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?



Molecule Modality

Text Modality

Limited Cross-Modal Data

Pre-training

Initialized from a public t5.1.1 checkpoint

ON=CCC1=C[NH1]C2= [X]

Lissamine fast yellow(2-) is an [X] resulting [Y] removal of a proton

MolT5

[X] CC=CC=C12

[X] organosulfonate oxoanion [Y] from the

Fine-tuning

Molecule Generation

The molecule is a siderophore composed from L-2,3-diaminopropionic acid, ...

MolT5

C(CC(=O)NCCNC(=O)CC(CC(=O)NCC(C(=O)O)N)(C(=O)O)O)C(=O)C(=O)O

Molecule Captioning

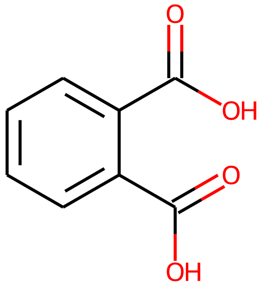
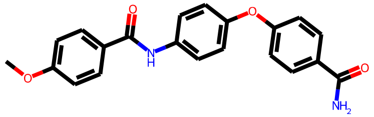
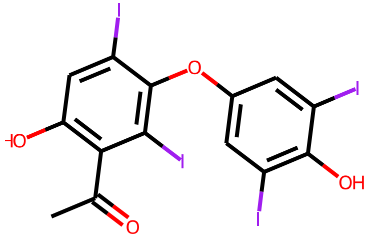
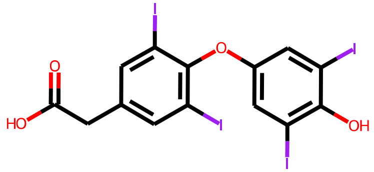
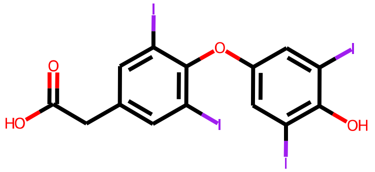
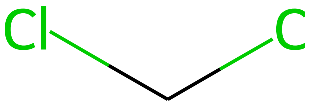
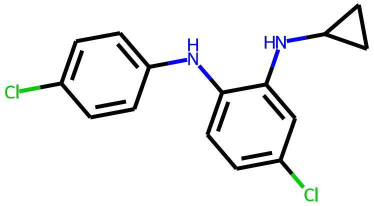
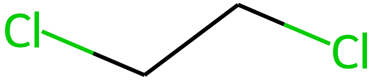
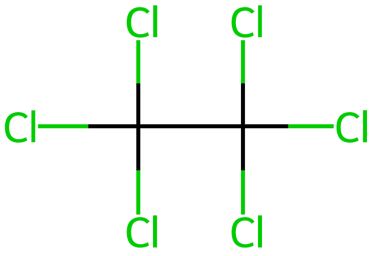
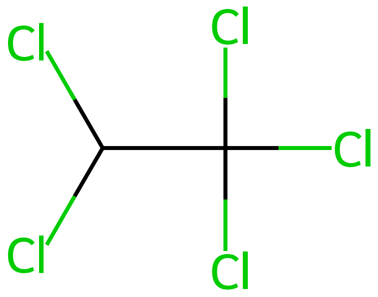
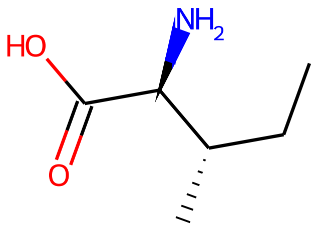
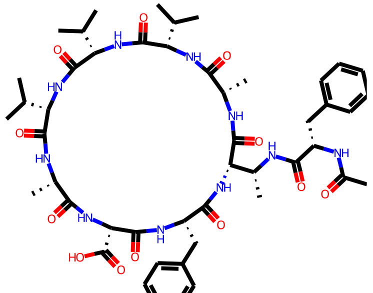
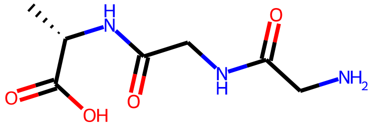
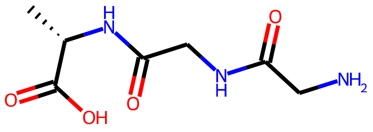
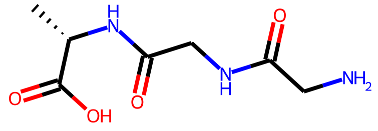
CC1=C(C(=C(C=C1)Cl)NC2=CC=CC=C2C(=O)O)Cl

The molecule is an aminobenzoic acid that is anthranilic acid in which one of the hydrogens attached to ...

Edwards et al. EMNLP 2022

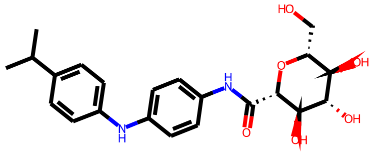
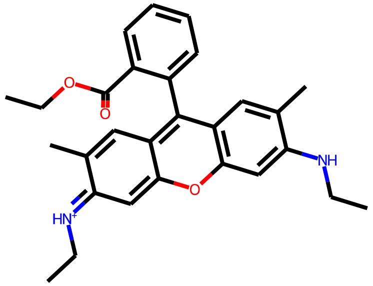
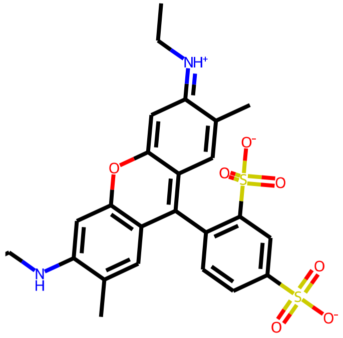
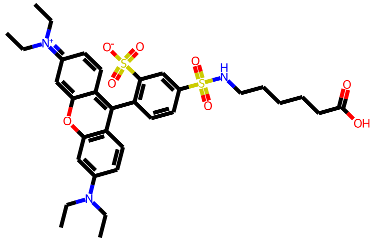
THRUST 1

MOLECULE GENERATION RESULTS – CONCRETE DESCRIPTIONS

Input	RNN	Transformer	T5	MolT5	Ground Truth
The molecule is a monocarboxylic acid that is thyroacetic acid carrying four iodo substituents at positions 3, 3', 5 and 5'. It has a role as a thyroid hormone, a human metabolite and an apoptosis inducer […]					
The molecule is a member of the class of chloroethanes that is ethane in which five of the six hydrogens are replaced by chlorines. A non-flammable, high-boiling liquid (b.p. 161-162°C) with relative density 1.67 and an odour resembling that of chloroform, […]			Invalid, fixed 		
The molecule is a tripeptide composed of glycine, glycine and L-alanine residues joined in sequence. It has a role as a metabolite.					

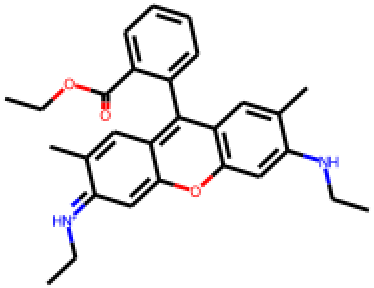
THRUST 1

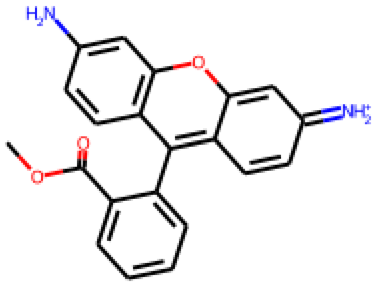
MOLECULE GENERATION RESULTS – LESS SPECIFIC DESCRIPTIONS

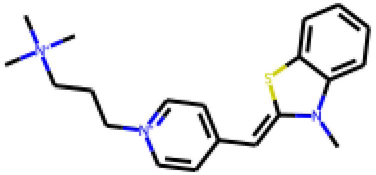
Input	RNN	Transformer	T5	MolT5	Ground Truth
The molecule is a sulfonated xanthene dye of absorption wavelength 573 nm and emission wavelength 591 nm. It has a role as a fluorochrome.	Invalid				

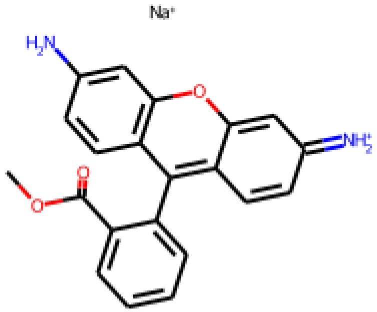
THRUST 1

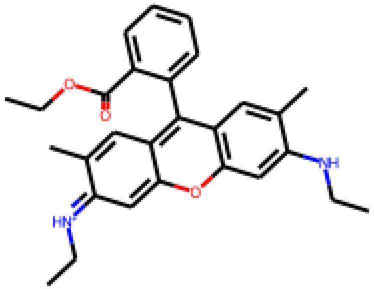
Looking for properties: “The molecule is a blue dye.”

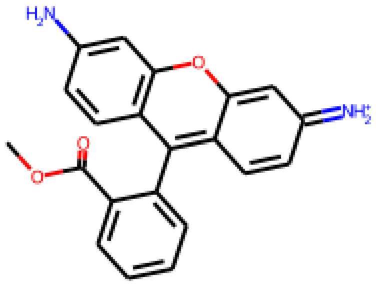


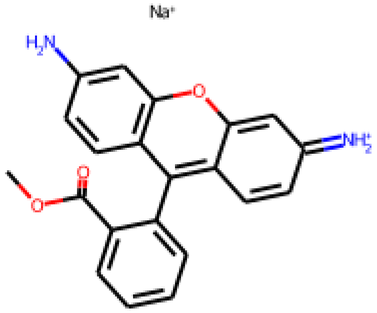


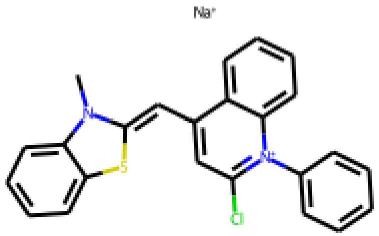






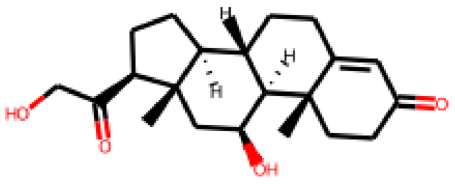


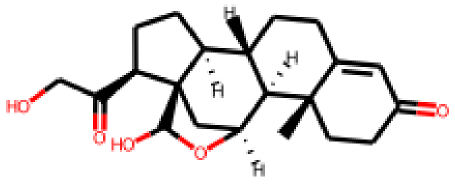


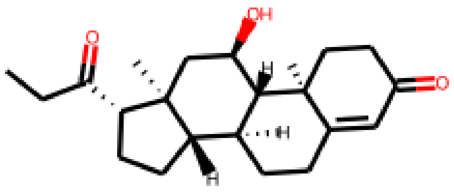


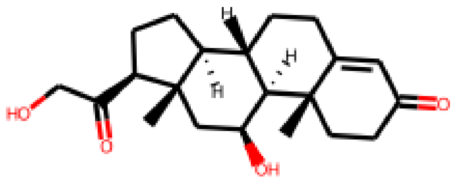
THRUST 1

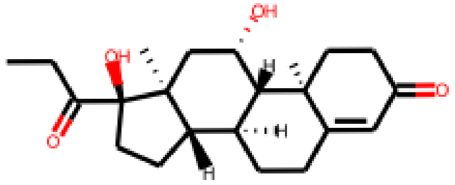
Looking for properties: “The molecule is a corticosteroid.”

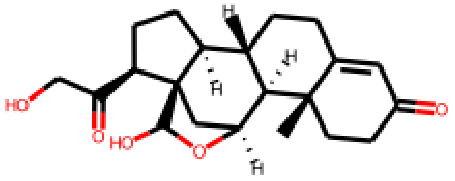


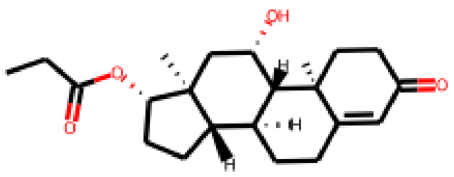


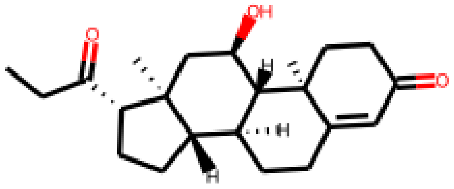


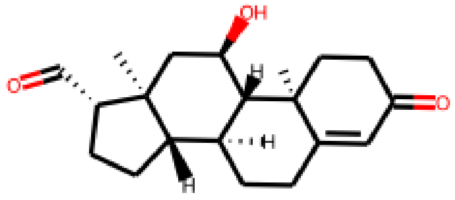


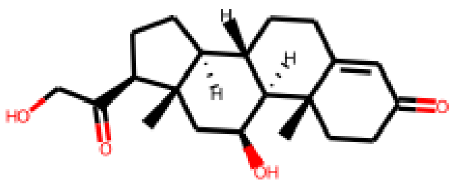








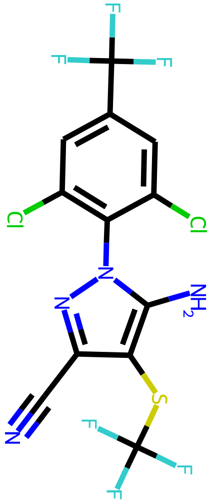




THRUST 1

MOLECULE CAPTIONING RESULTS

Input



MolT5

The molecule is a member of the class of pyrazoles that is 1H-pyrazole that is substituted at positions 1, 3, 4, and 5 by 2,6-dichloro-4-(trifluoro methyl)phenyl, cyano, (trifluoromethyl)sulfinyl, and amino groups, respectively. It is a nitrile, a dichlorobenzene, a primary amino compound, a member of pyrazoles, a sulfoxide and a member of (trifluoromethyl) benzenes

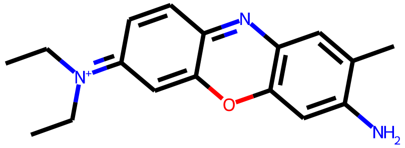
Ground Truth

The molecule is a member of the class of pyrazoles that is 1H-pyrazole that is substituted at positions 1, 3, 4, and 5 by 2,6-dichloro-4-(trifluoromethyl)phenyl, cyano, (trifluoromethyl) sulfanyl, and amino groups, respectively. It is a metabolite of the agrochemical fipronil. It has a role as a marine xenobiotic metabolite. It is a member of pyrazoles, a dichlorobenzene, a member of (trifluoromethyl)benzenes, an organic sulfide and a nitrile.

THRUST 1

Molecule Captioning Results

Input



T5

The molecule is a quaternary ammonium ion and a member of phenanthridines. It has a role as an intercalator and a fluorochrome.

MolT5

The molecule is an organic cation that is phenoxazin-5-ium substituted by amino and methylamino groups at positions 3 and 7 respectively. The chloride salt is the histological dye 'azure C'.

Ground Truth

The molecule is an organic cation that is phenoxazin-5-ium substituted by methyl, amino and diethylamino groups at positions 2, 3 and 7 respectively. The tetrachlorozincate salt salt is the histological dye 'brilliant cresyl blue'.

THRUST 1

Large models are surprisingly good?

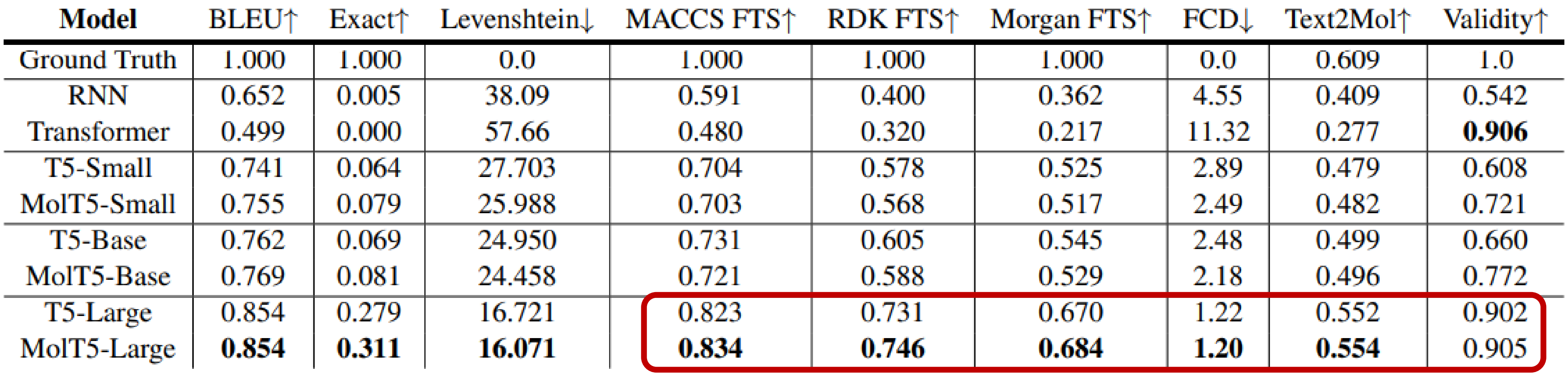
String Metrics

Fingerprint metrics

77M params

250M params

800M params

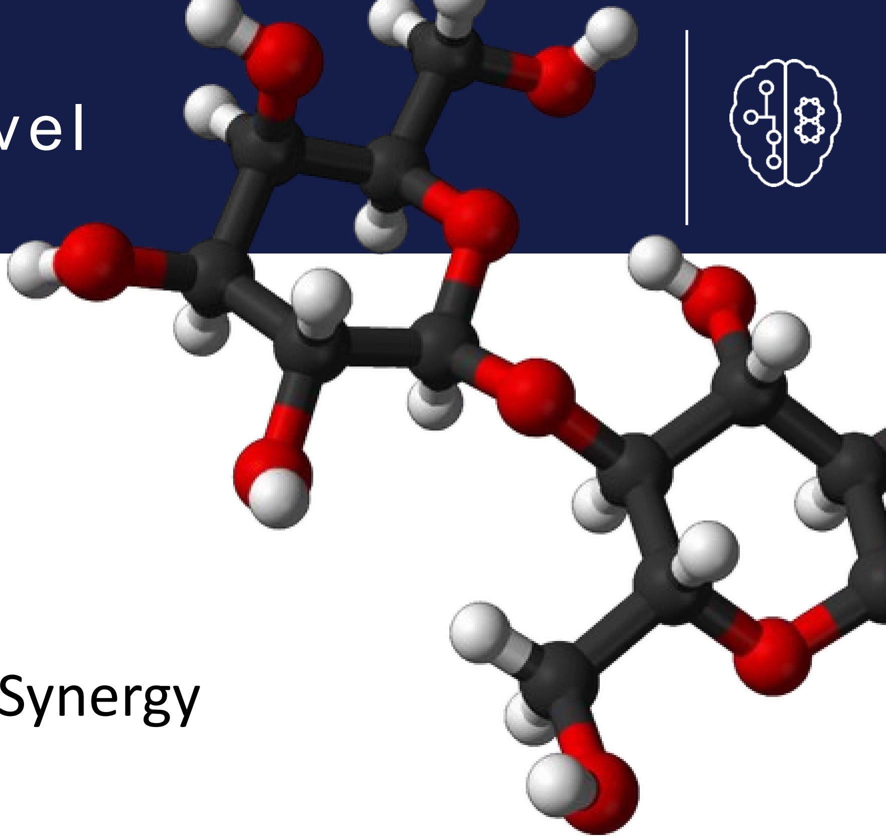


Model	BLEU↑	Exact↑	Levenshtein↓	MACCS FTS↑	RDK FTS↑	Morgan FTS↑	FCD↓	Text2Mol↑	Validity↑
Ground Truth	1.000	1.000	0.0	1.000	1.000	1.000	0.0	0.609	1.0
RNN	0.652	0.005	38.09	0.591	0.400	0.362	4.55	0.409	0.542
Transformer	0.499	0.000	57.66	0.480	0.320	0.217	11.32	0.277	**0.906**
T5-Small	0.741	0.064	27.703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

What is going on inside the language model?

THRUST 1

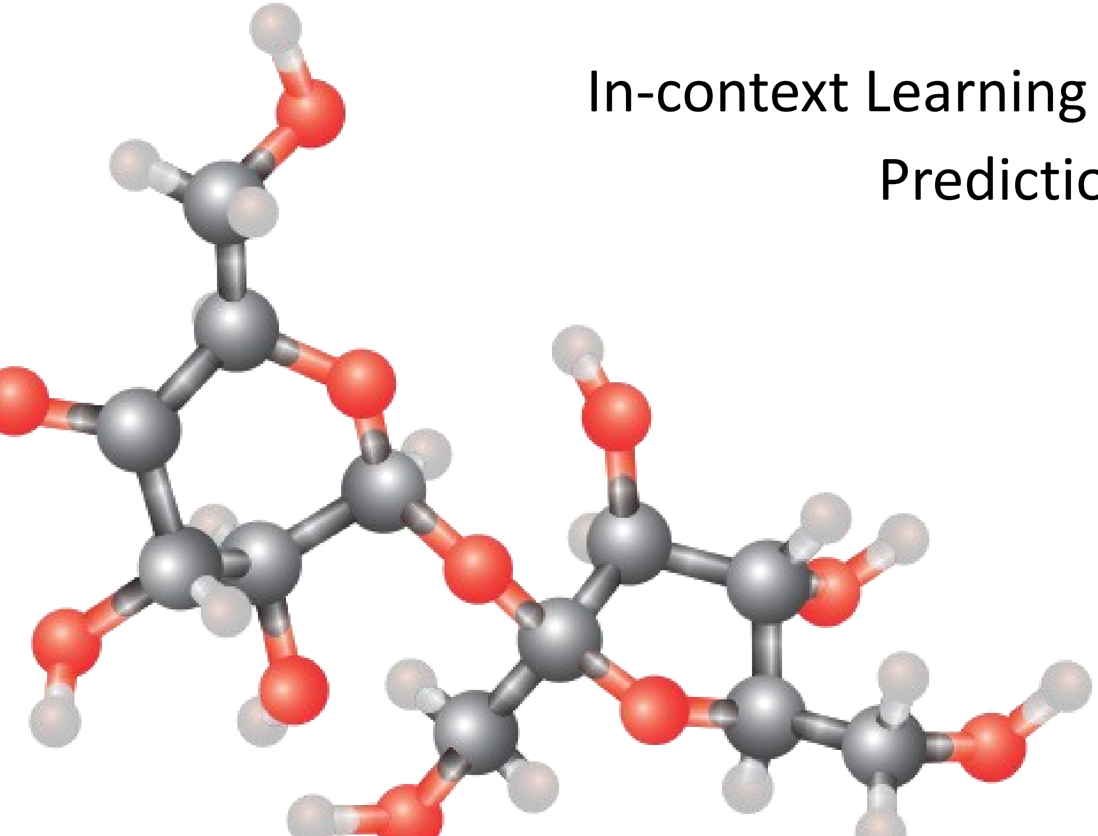
Language models on higher-level molecule interaction



SynerGPT:

In-context Learning for Personalized Drug Synergy Prediction and Drug Design

Preprint

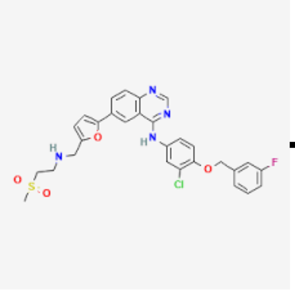


THRUST 1

Drug Synergy Prediction

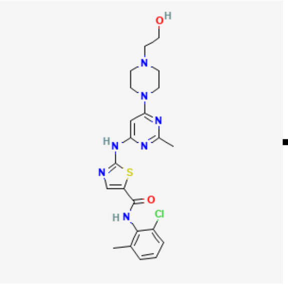
Drug 2

Lapatinib

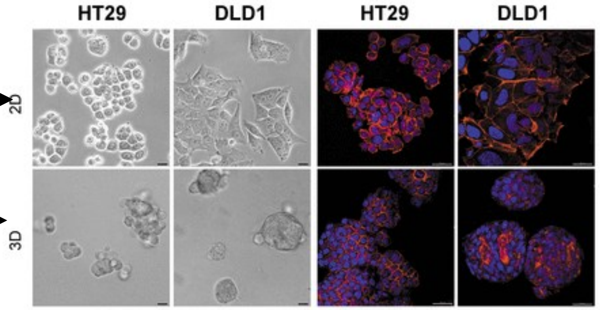


Drug 1

Dasatinib



Cell Type



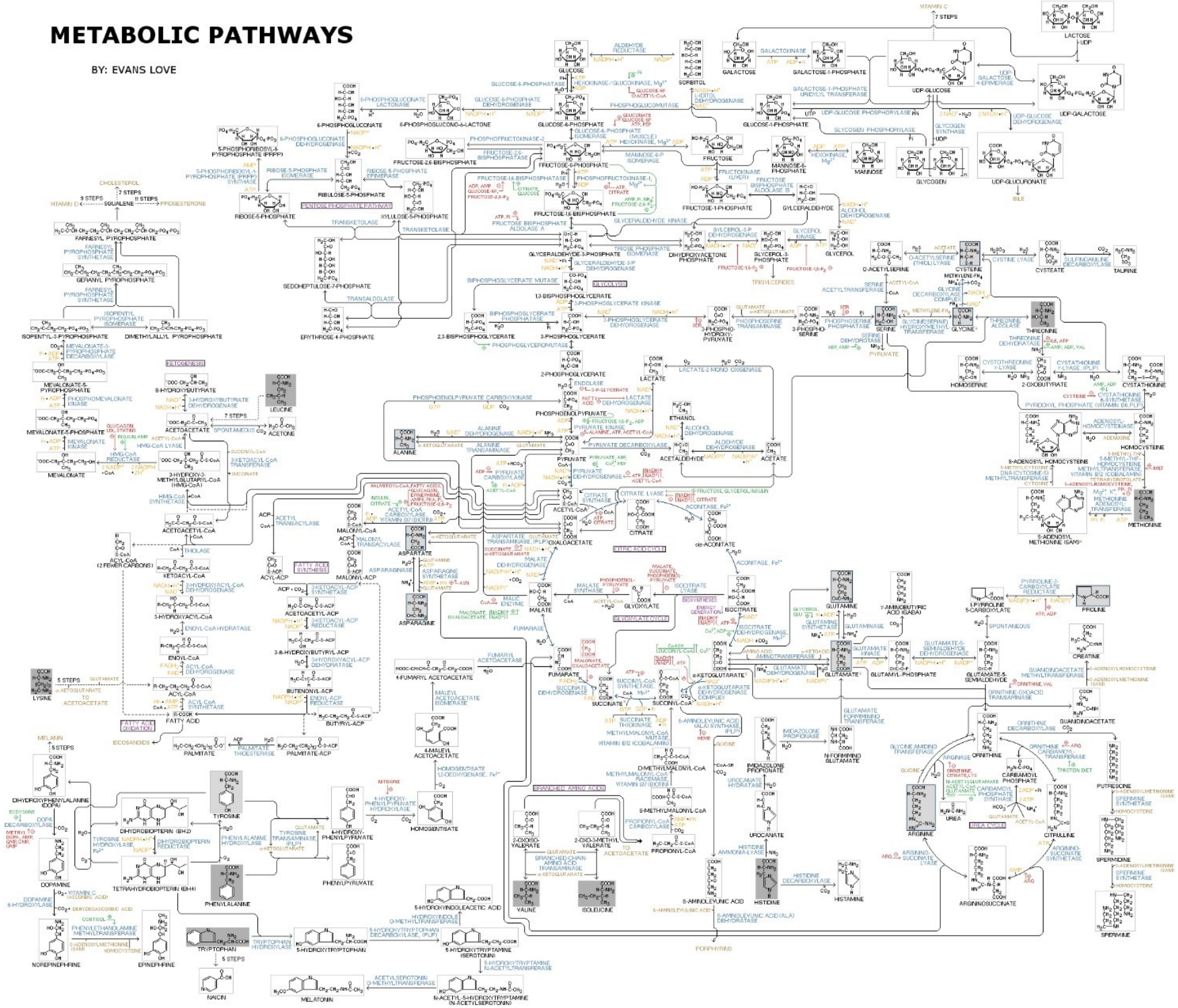
(Stankevicius et al. ‘16, Anticancer Research)

Synergy?

Goal: addressing drug resistance and increasing efficacy.

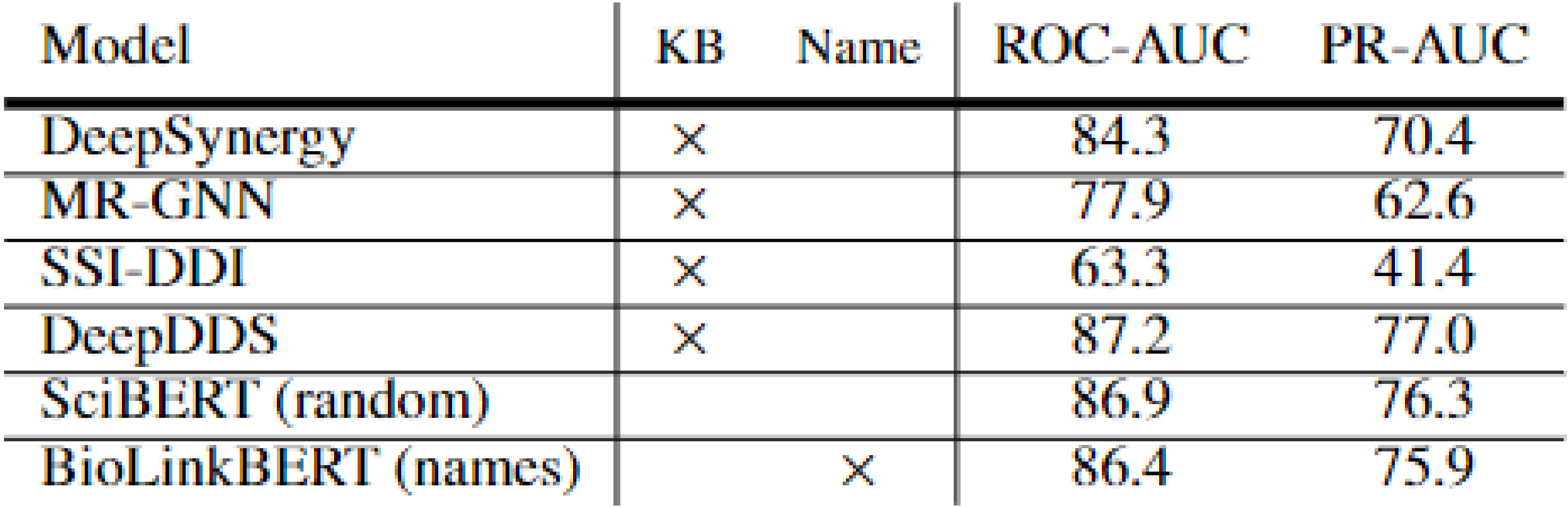
THRUST 1

Predicting drug synergies is complex



THRUST 1

BERT language models *without* information on the drug structure perform similarly !?



Model	KB	Name	ROC-AUC	PR-AUC
DeepSynergy	×		84.3	70.4
MR-GNN	×		77.9	62.6
SSI-DDI	×		63.3	41.4
DeepDDS	×		87.2	77.0
SciBERT (random)			86.9	76.3
BioLinkBERT (names)		×	86.4	75.9

- Language models can achieve results on par with several baselines.
- Is this because of pre-trained knowledge about the drugs?
  - No, random tokens as input still achieves strong performance

THRUST 1

CAN WE PERSONALIZE THIS TO A SPECIFIC PATIENT OR DRUG?

Key Motivation: Can we evaluate a limited number of drug synergies in biopsied patient tumor cells and use it to predict synergies for that patient?

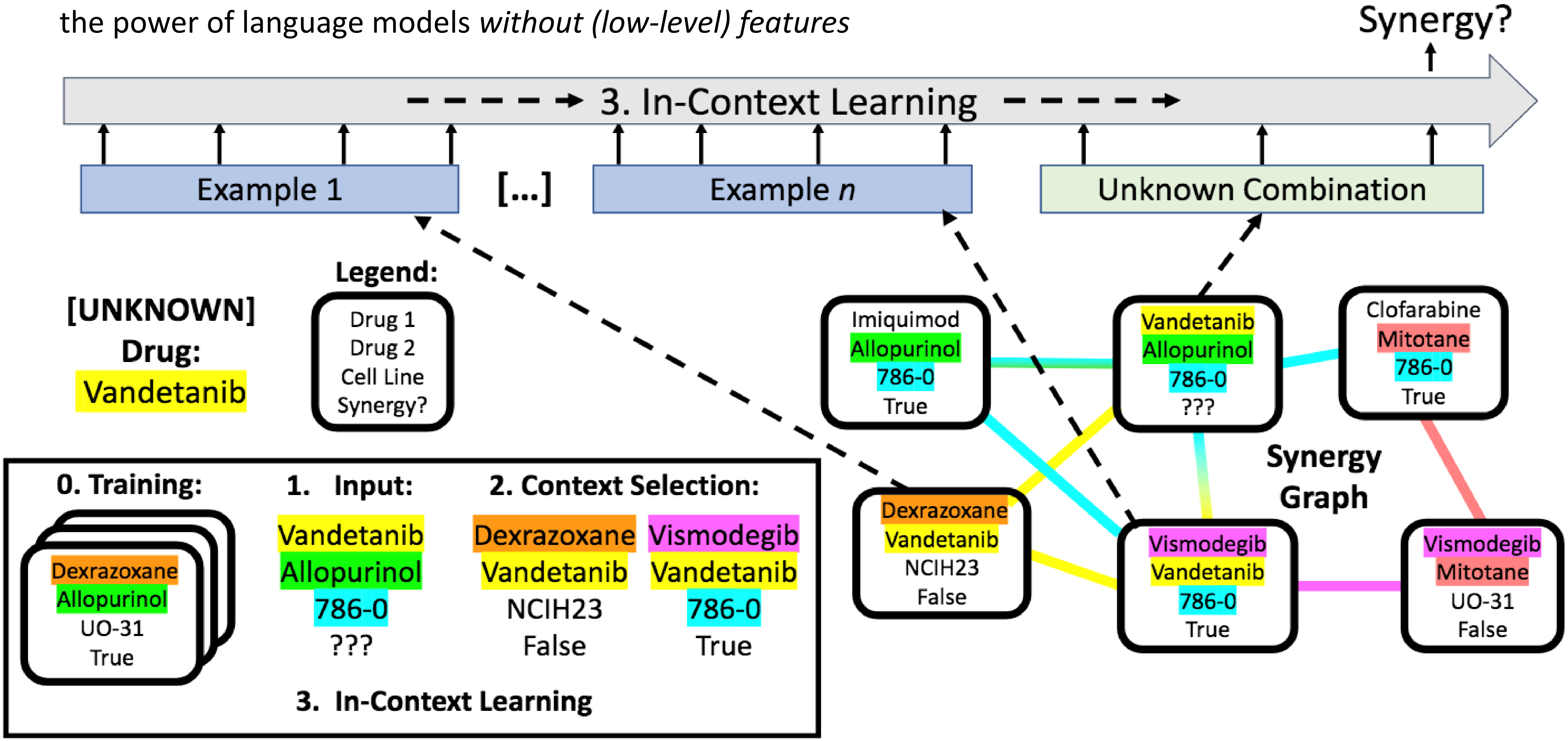
How do we do that?

- Can we extend our results on language models to in-context learning?
- In-context learning would enable us to learn from only a few examples of a given drug or patient tumor biopsy.

THRUST 1

What about new drugs and cell lines?

the power of language models *without (low-level) features*



Synergy?

3. In-Context Learning

Example 1

[...]

Example *n*

Unknown Combination

[UNKNOWN]
Drug:
Vandetanib

Legend:
Drug 1
Drug 2
Cell Line
Synergy?

Imiquimod
Allopurinol
786-0
True

Vandetanib
Allopurinol
786-0
???

Clofarabine
Mitotane
786-0
True

Synergy Graph

Dexrazoxane
Vandetanib
NCIH23
False

Vismodegib
Vandetanib
786-0
True

Vismodegib
Mitotane
UO-31
False

0. Training:

Dexrazoxane
Allopurinol
UO-31
True

1. Input:

Vandetanib
Allopurinol
786-0
???

2. Context Selection:

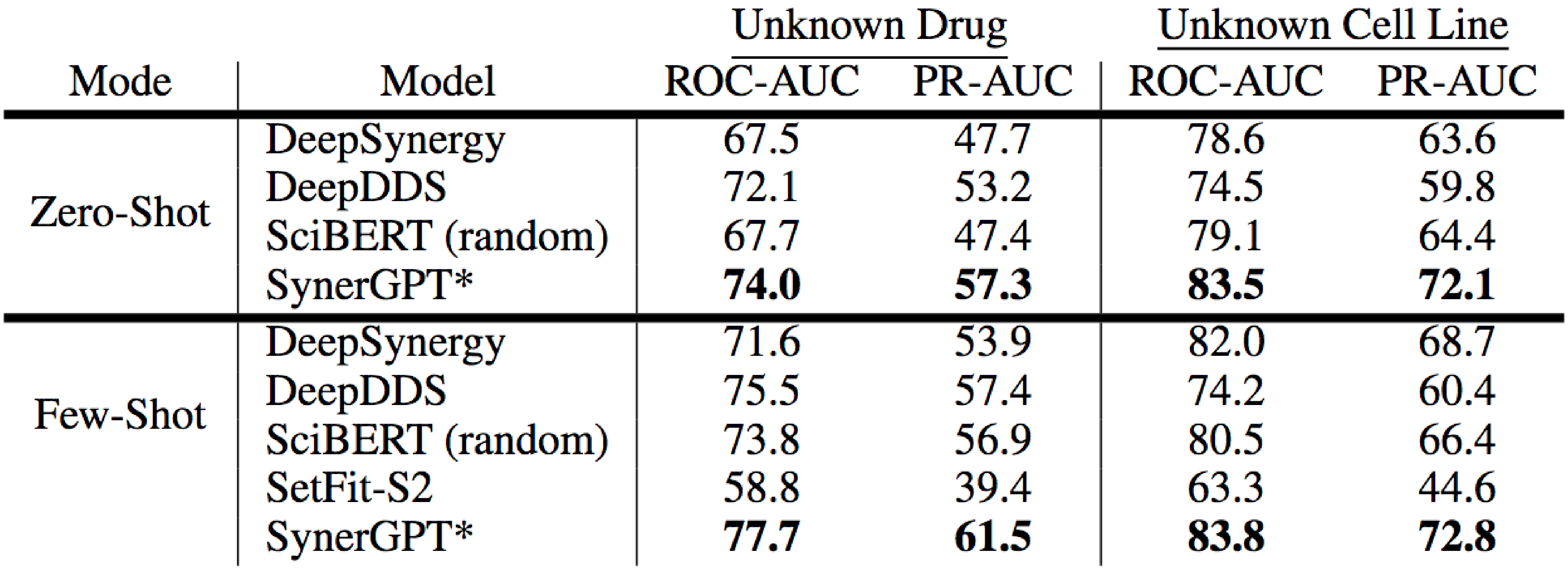
Dexrazoxane
Vandetanib
NCIH23
False

Vismodegib
Vandetanib
786-0
True

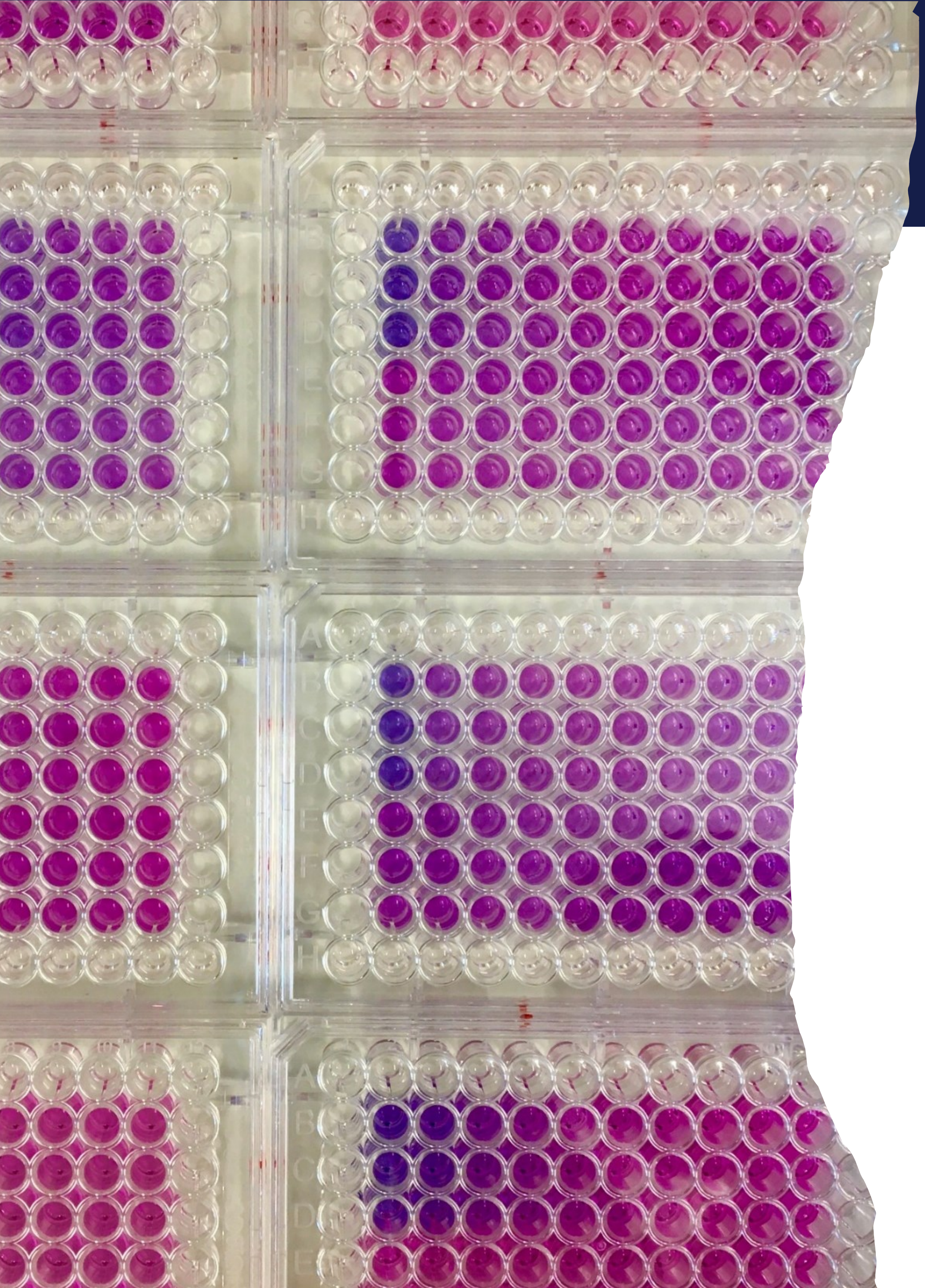
3. In-Context Learning

THRUST 1

RESULTS

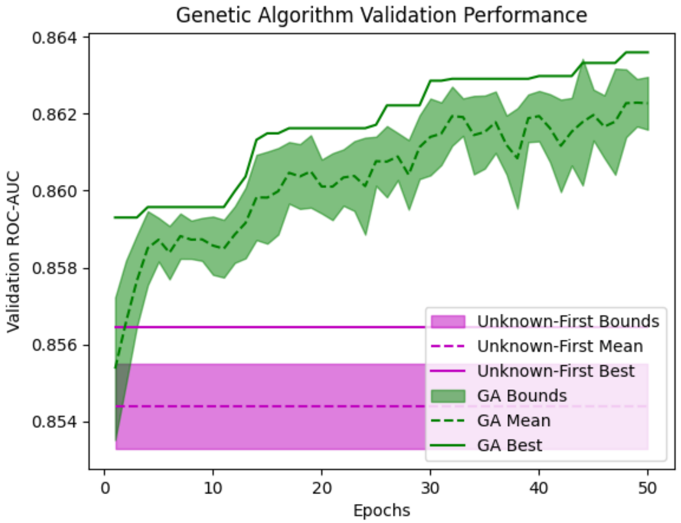


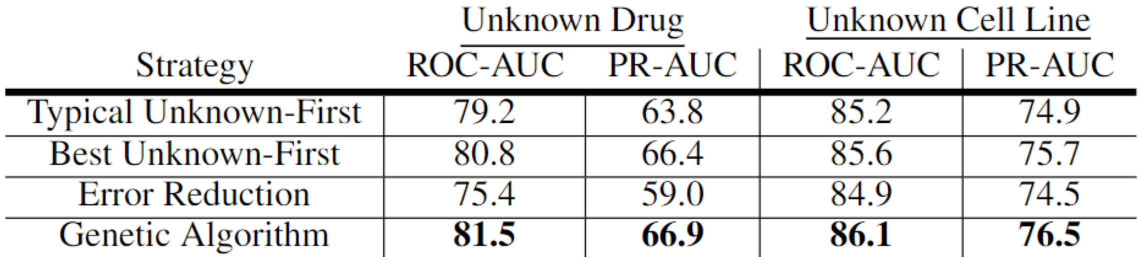
Mode	Model	Unknown Drug		Unknown Cell Line	
		ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
Zero-Shot	DeepSynergy	67.5	47.7	78.6	63.6
	DeepDDS	72.1	53.2	74.5	59.8
	SciBERT (random)	67.7	47.4	79.1	64.4
	SynerGPT*	**74.0**	**57.3**	**83.5**	**72.1**
Few-Shot	DeepSynergy	71.6	53.9	82.0	68.7
	DeepDDS	75.5	57.4	74.2	60.4
	SciBERT (random)	73.8	56.9	80.5	66.4
	SetFit-S2	58.8	39.4	63.3	44.6
	SynerGPT*	**77.7**	**61.5**	**83.8**	**72.8**



How would this look in the real world?

- We need a quick and easy one time test – a standardized assay
- How can we identify which synergies are most informative for such a panel?
    - Optimization of the language model context!



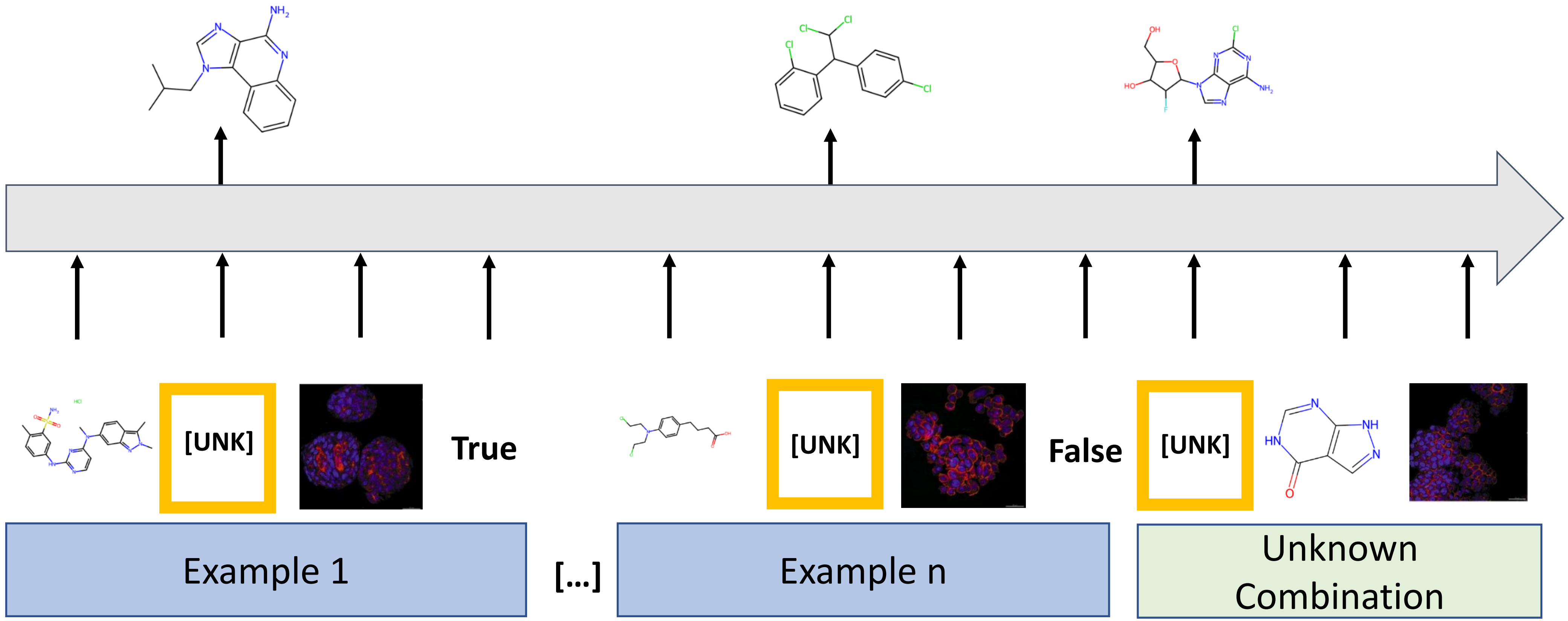


Strategy	Unknown Drug ROC-AUC	Unknown Drug PR-AUC	Unknown Cell Line ROC-AUC	Unknown Cell Line PR-AUC
Typical Unknown-First	79.2	63.8	85.2	74.9
Best Unknown-First	80.8	66.4	85.6	75.7
Error Reduction	75.4	59.0	84.9	74.5
Genetic Algorithm	**81.5**	**66.9**	**86.1**	**76.5**

THRUST 1

WHAT'S GOING ON HERE?

- What kind of latent information is the model learning from drug synergy tuples?
  - Drug structure?
  - Mechanism of action?
- We propose a new type of inverse design based on synergy tuples for explainability.



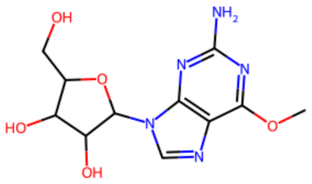
Inverse Design Examples

Retrieved Molecule after *n* Context Examples

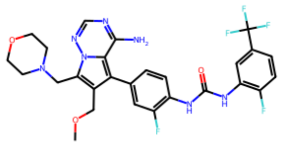
Ground Truth

(2-Amino-6-methoxypurine arabinoside)

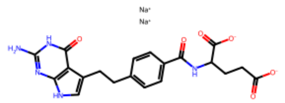
Known Drug



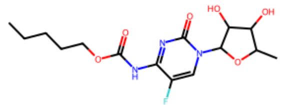
n = 1



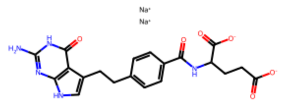
n = 2



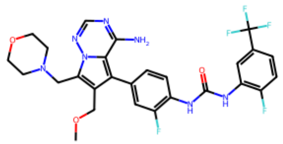
n = 3



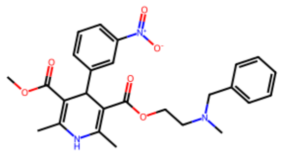
n = 5



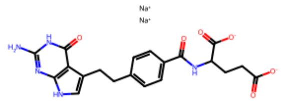
n = 10



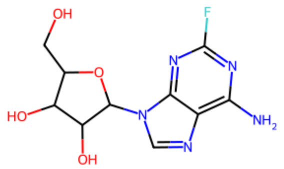
n = 13



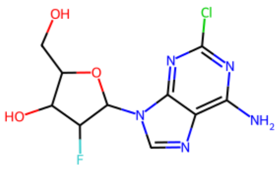
n = 14



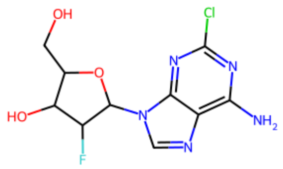
n = 15



n = 19



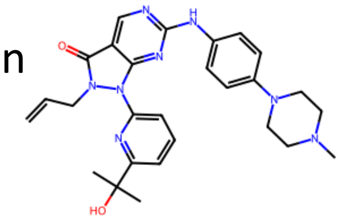
n = 20



Ground Truth

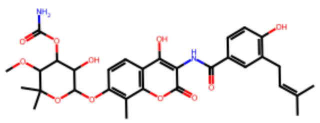
(Adavosertib)

Unknown Drug



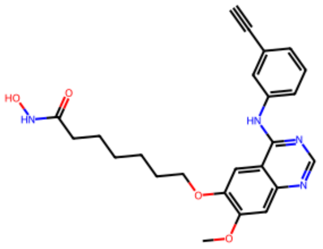
Drug 1
Drug 2
Cell Line
Label

n = 1



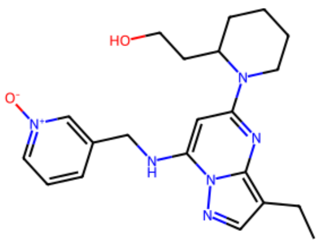
[UNKNOWN]
dactolisib
OV90
True

n = 2



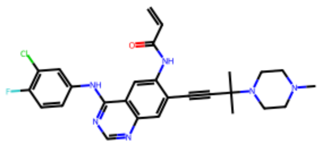
Mit-C
[UNKNOWN]
T-47D
False

n = 5



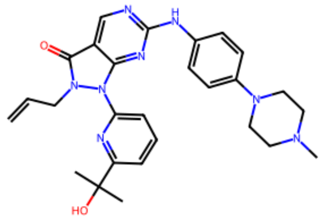
Vorinostat
[UNKNOWN]
NCIH520
True

n = 7



1'-Epi Gemcitabin
[UNKNOWN]
SK-OV-3
True

n = 12



5-Fluorouracil
[UNKNOWN]
MSTO
False

THRUST 1

WHAT IS NEXT?

1. Integration and Application
    1. Integrating LLMs into automated systems
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, …

PARTNERS

NSF

Genentech
A Member of the Roche Group

IBM

janssen

REVOLUTION
MEDICINES

abbvie

AMGEN

NVIDIA.

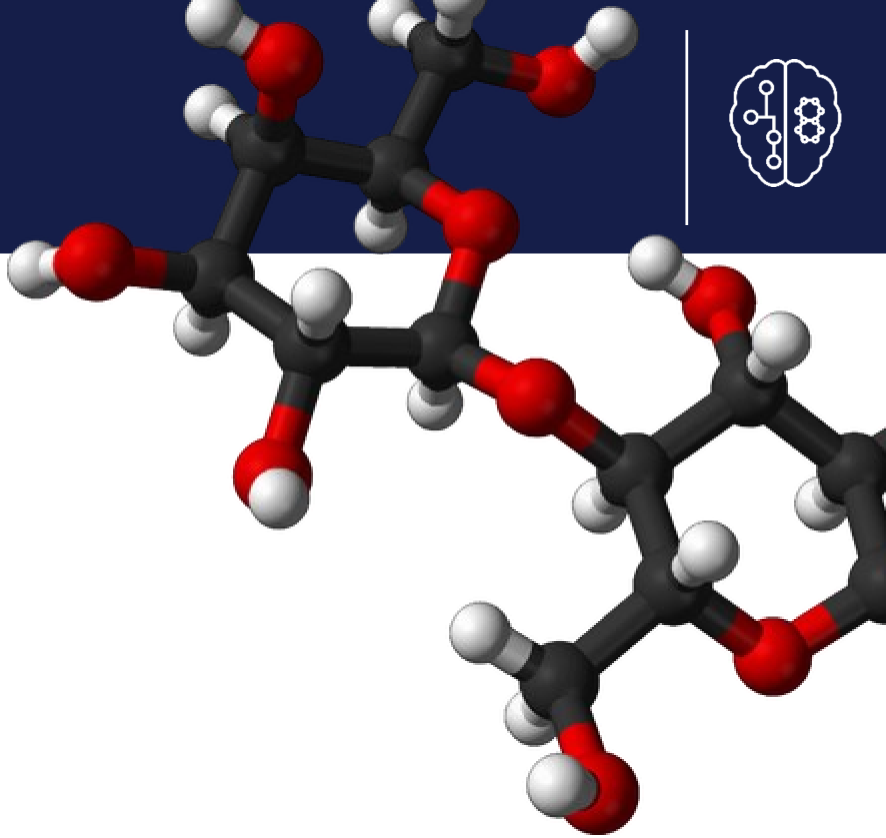
LanzaTech

I UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN

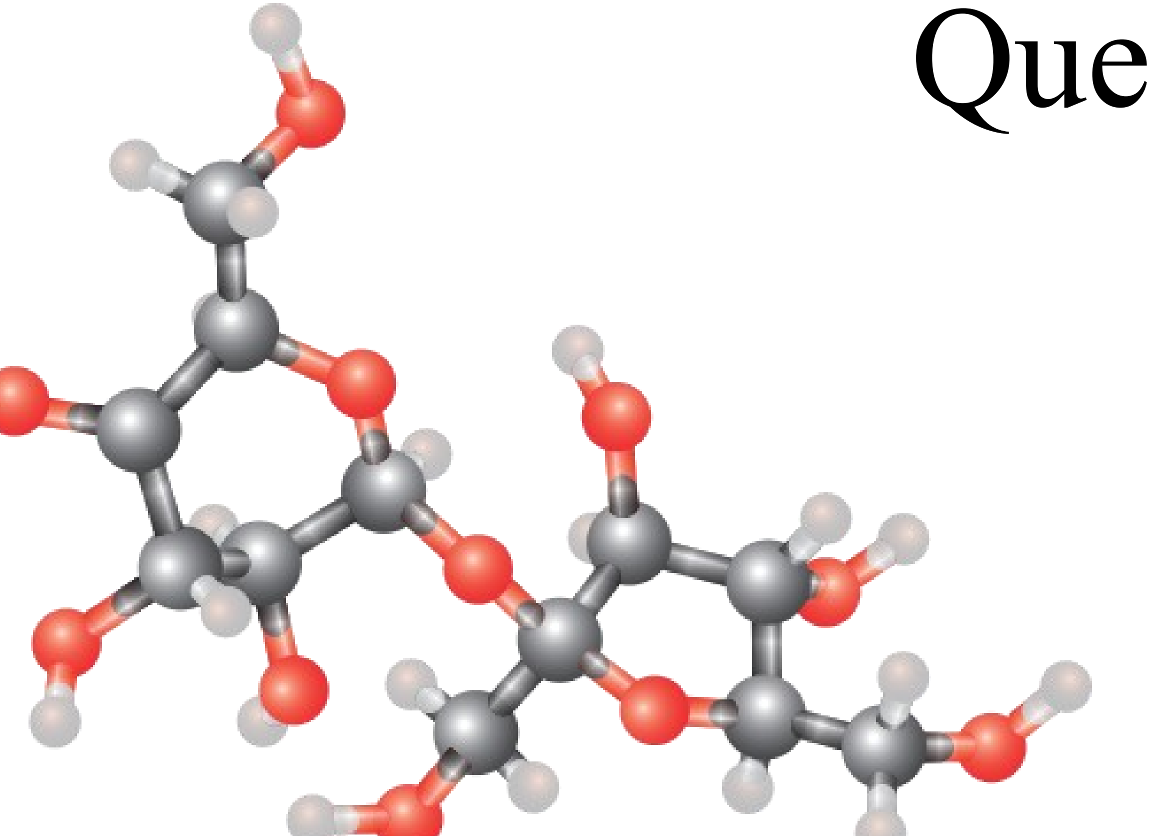
RIT | Rochester Institute
of Technology

PennState

Ai2
Allen Institute for AI



Questions?



UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

MOLECULE MAKER LAB INSTITUTE

NSF