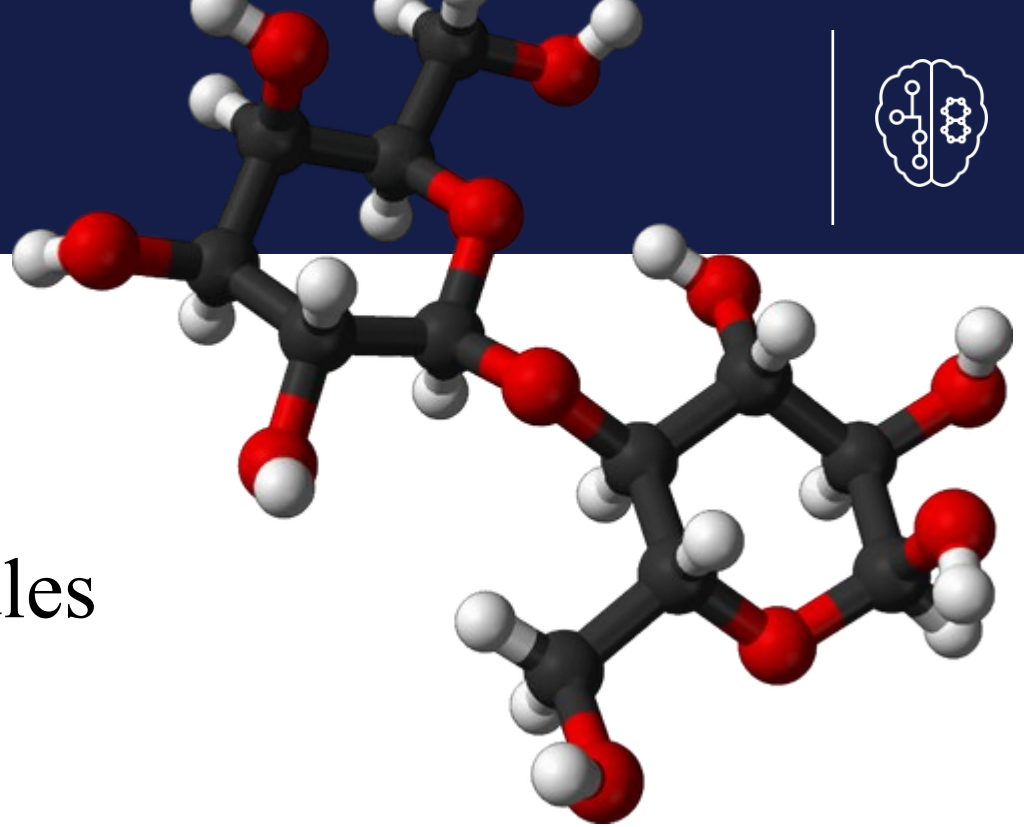
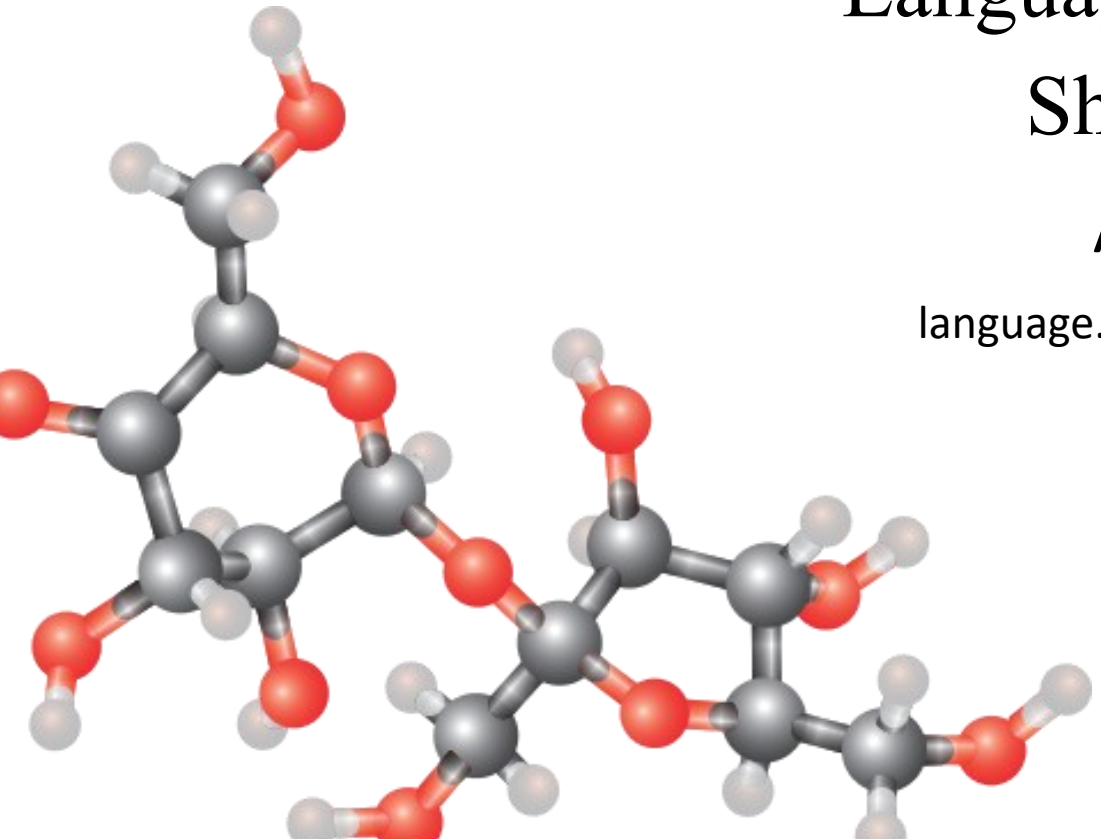
THRUST 1



Language + Molecules Shared Task

ACL 2024

language.molecules@gmail.com



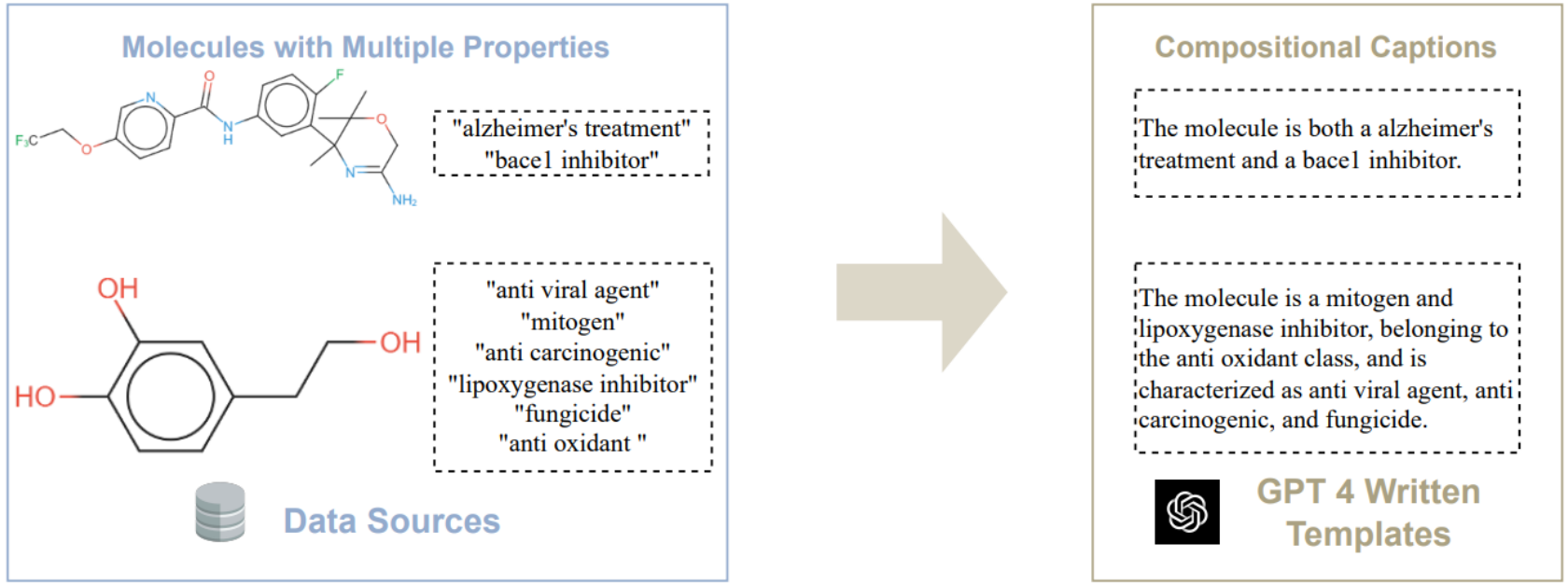
Why design a new shared task

L+M-24: Building a Dataset for Language+Molecules @ ACL 2024

- Existing datasets were:
  1. Small and scraped from existing databases
  2. Large but noisy and constructed by performing entity linking on the scientific literature
  3. Built by converting property prediction datasets to natural language using templates
- The goal was to focus on 3 key benefits of natural language:
  - Compositionality
  - Functionality
  - Abstraction
- We grouped properties into four key categories of interest:
  - Biomedical
  - Light and Electricity
  - Human Interaction and Organoleptics
  - Agriculture and Industry

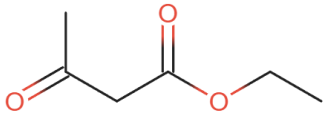
Compositionality, Functionality, and Abstraction

- To focus on functionality and abstraction, we used specific source datasets containing (mostly) natural language annotations rather than numerical properties.
- To add compositionality, we composed multiple properties together using natural language templates.
  - Some property combinations were held-out of the training set. This may have caused certain properties to be poorly represented in the training data.

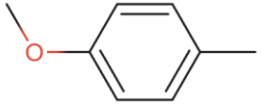


Data Sources -- PubChem

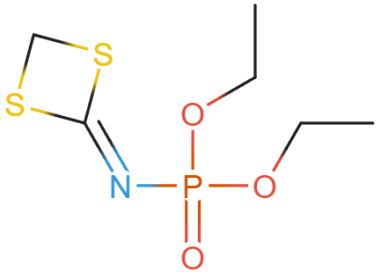
- Properties including odor, taste, polymerization, and decomposition were taken from PubChem.
  - Descriptions are written free-form and do not fit into categories
- PubChem Examples:
  - PLEASANT GREEN, FRUITY, RUM ODOR



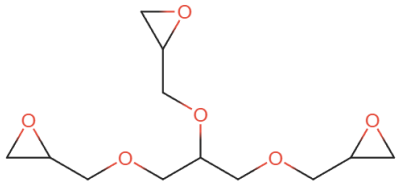
  - PUNGENT ODOR SUGGESTIVE OF YLANG-YLANG



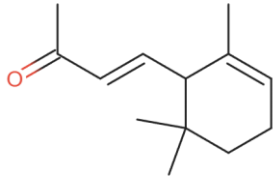
  - Mercaptan-like odor



  - When subjected to high temperatures alone or in the presence of catalysts or strong oxidizing agents it is possible that violent polymerization will take place.



  - Woody, floral, berry, fruity with powdery nuances

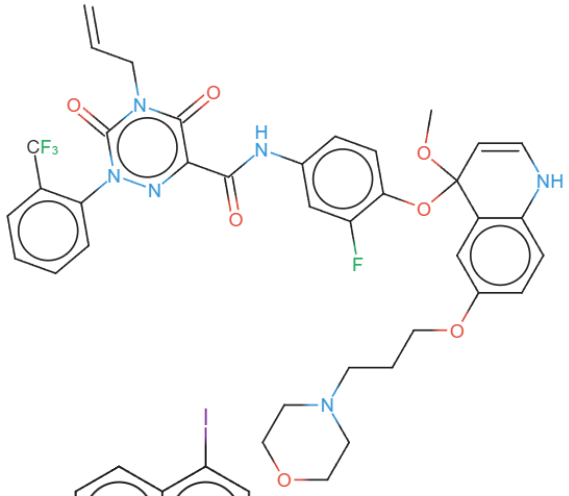


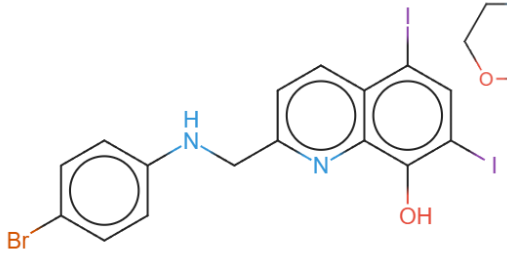
Data Sources – Chemical Function

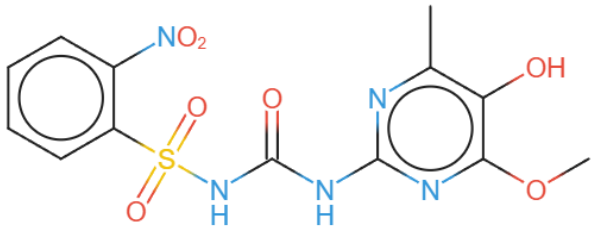
- Properties were extracted from the patent literature and standardized. We performed additional processing and kept properties in the following categories:
  - “X-icide”
  - “anti-X”
  - “X treatment”
  - “X modulators”
  - “X inhibitors”
  - “X agonists”
  - “X antagonists”
  - “light”,
  - “electricity.

CheF Examples:

- Inhibiting c-Met kinase
- Antimicrobial / Iron(II)-dependent / Biofilm
- herbicide / antidiabetic / urea





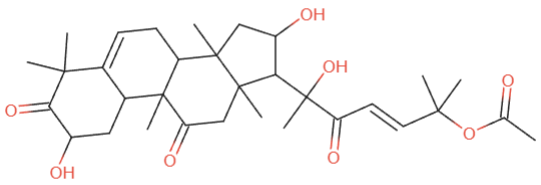


Data Sources – ChemFOnt: Chemical Function Ontology

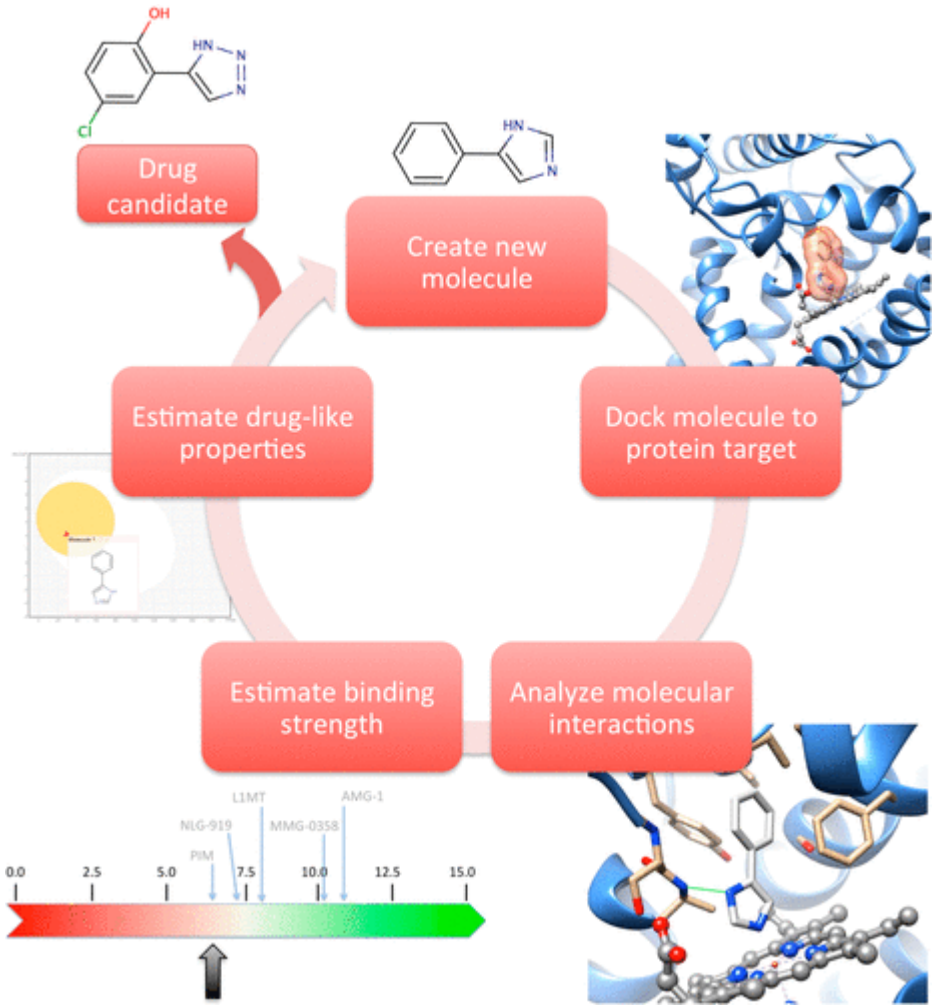
- We considered organoleptic effects, compound roles, and health effects

ChemFOnt Examples:

Functional Ontology	
Physiological effect	**Organoleptic effect** Taste ∘ Bitter
Process	**Naturally occurring process** Biological process Chemical reaction ∘ Lipid Peroxidation (HMDB: HMDB0034927) Biochemical pathway Metabolic pathway ∘ Fatty Acid Metabolism (HMDB: HMDB0034927) ∘ Lipid metabolism pathway (HMDB: HMDB0034927) Cellular process ∘ Cell signaling (HMDB: HMDB0034927) Biochemical process ∘ Lipid transport (HMDB: HMDB0034927)
Role	**Biological role** ∘ Membrane stabilizer (HMDB: HMDB0034927) ∘ Anti gibberellin (HMDB: HMDB0034927) ∘ Anti hepatotoxic (HMDB: HMDB0034927) ∘ Anti-inflammatory (HMDB: HMDB0034927) ∘ Cytotoxic (HMDB: HMDB0034927) ∘ Insectifuge (HMDB: HMDB0034927) ∘ Insectiphile (HMDB: HMDB0034927)



- Physiological effect
  - Health effect
    - Cancer
    - Circulatory system disorder
    - Digestive system disorder
    - Endocrine system disorder
    - Exposure
    - Genetic disorder
    - Immune system disorder
    - Infection or infestation
    - Injury
    - Intervention
    - Metabolic disorder
    - Musculoskeletal system disorder
    - Nervous system disorder
    - Nonspecific condition
    - Nutritional disorder
    - Observation
    - Poisoning
    - Renal system and urinary system
    - Reproductive system disorder
    - Respiratory system disorder
    - Skin and subcutaneous tissue disorder
  - Organoleptic effect
    - Appearance
    - Odor
    - Taste
    - Touch

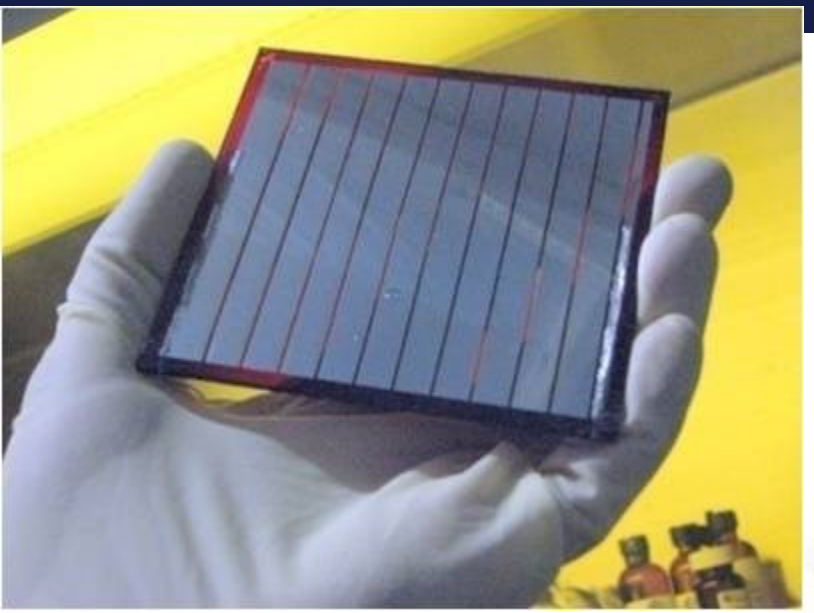


Drug Design
Daina et al. 2017 *J. Chem. Educ.*

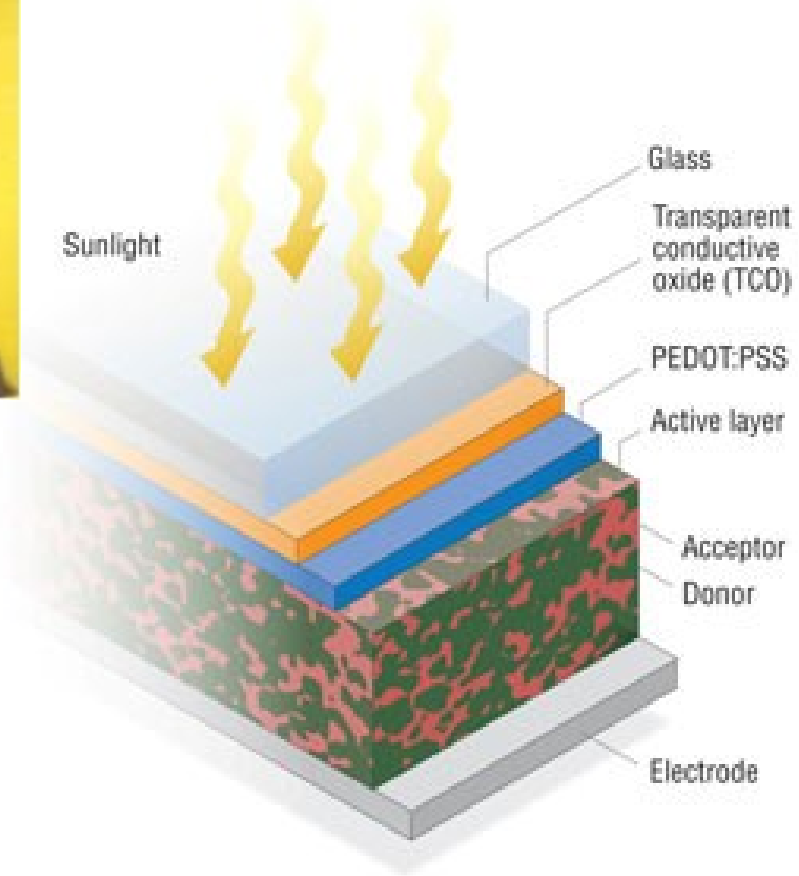
Salient Properties (a sample):

- 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

Light and Electricity: Organic Photovoltaics



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

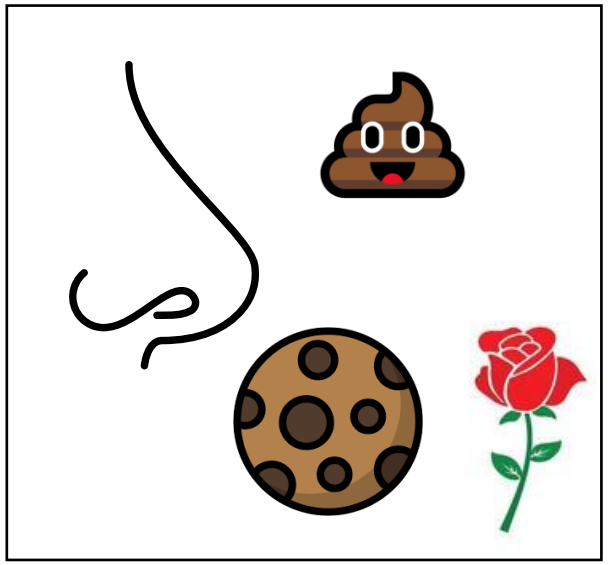


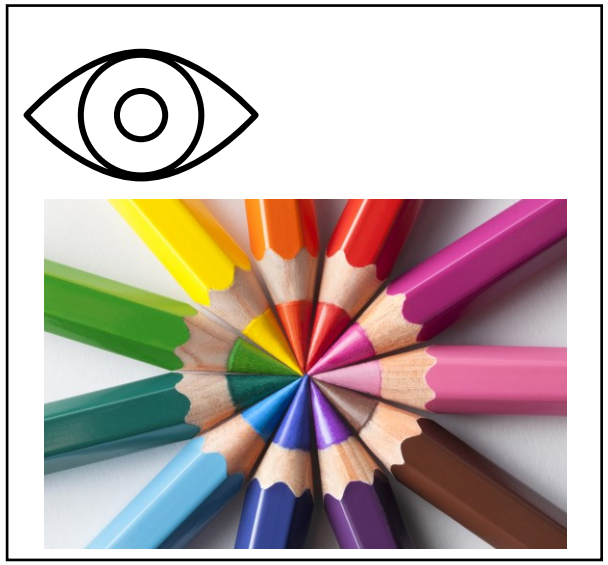
Department of Energy

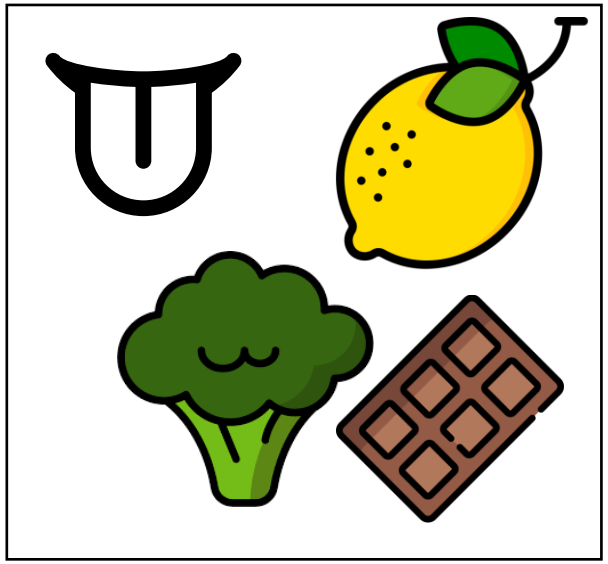
Salient Properties (a sample):

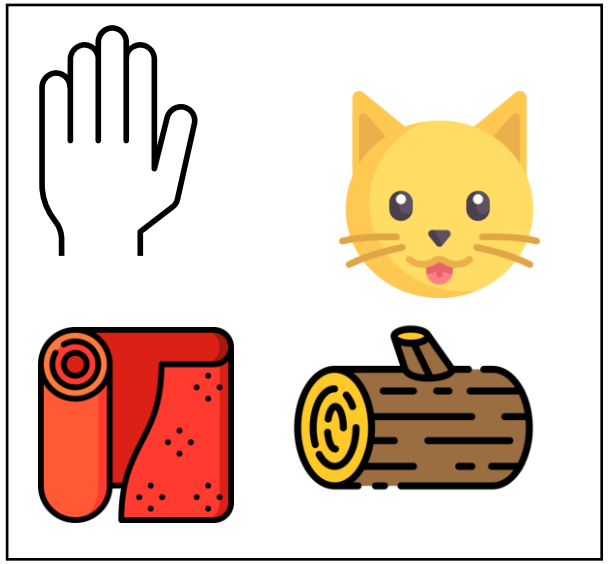
- Organic electroluminescent compound – A material which emits light in the presence of a electric current
- Fluorescent – Emission of light by a substance that has absorbed light
- Phosphorescent – Delayed emission of the absorbed light
- Photoacid generator – Compounds that produce acids upon exposure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

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

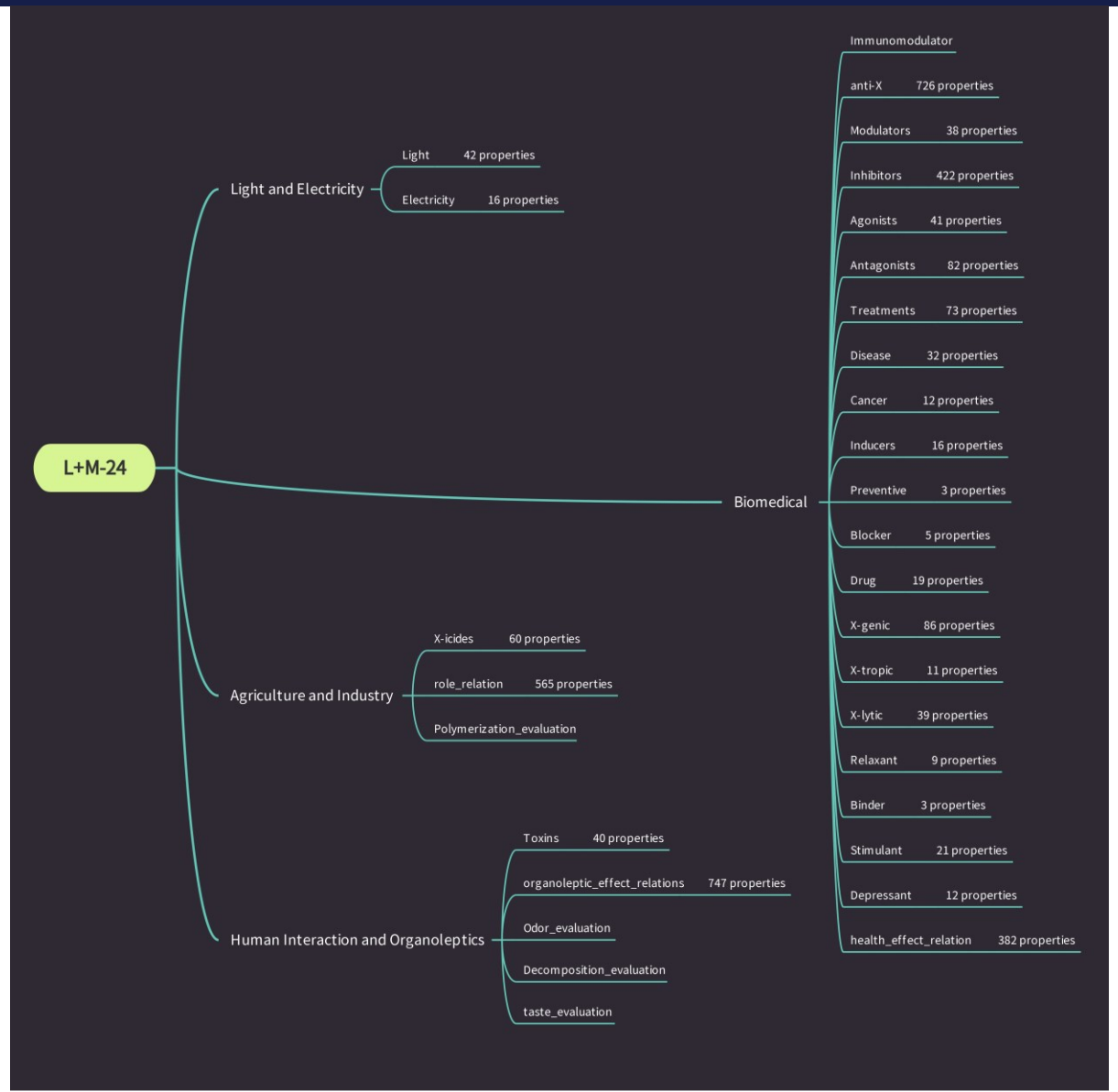
Agriculture and Industry

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

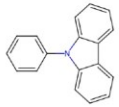
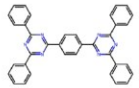
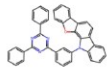
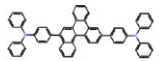
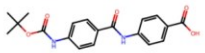
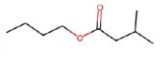
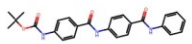
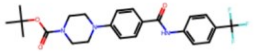
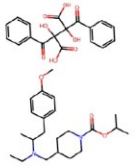
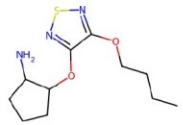
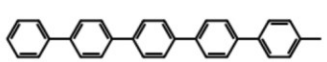
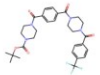
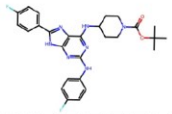
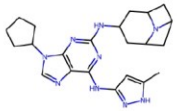
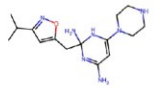
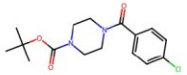
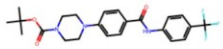
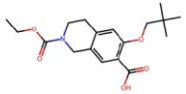
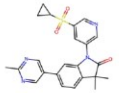
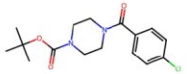
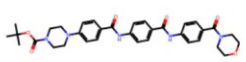
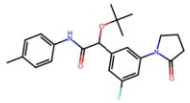
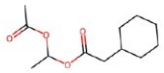
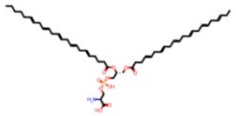
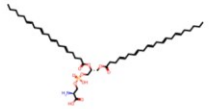
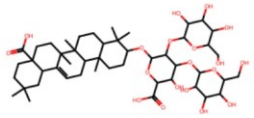
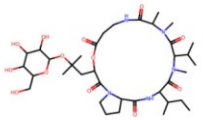
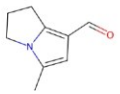
Property Breakdown



Group	Property-Molecule Pair Count
Total	1512865
Biomedical	776712
anti-X	24884
Modulators	2787
Inhibitors	23257
Agonists	1161
Antagonists	3172
Treatments	53070
Disease	316380
Cancer	41456
Inducers	31
Preventive	0
Blocker	47
Drug	260
X-genic	172
X-tropic	17
X-lytic	84
Relaxant	40
Binder	4
Stimulant	60
Depressant	52
health_effect_relations	309532
Light and Electricity	14077
Light	11069
Electricity	3008
Human Interaction	27457
Toxins	1070
organoleptic_effect_relations	20501
Agric. and Industry	694619
X-icides	809
role_relation	693648

THRUST 1

Baseline Model Examples

Input	MolT5-small	MolT5-base	MolT5-large	Meditron	Ground Truth
The molecule is a luminescent member of the organic light-emitting class.	Cc1ccc(-c2ccc(-c3ccc(-c4ccc(-c5ccccc5)cc5)cc4)cc3)cc2)cc1 **Invalid**				
The molecule is both a platelet aggregation inhibitor and a cell adhesion inhibitor.	Cc1ccc(-c2ccc(-c3ccc(-c4ccc(-c5ccccc5)cc5)cc4)cc3)cc2)cc1 **Invalid**		CC(C)(C)OC(=O)NC(C(=O)N1CCCC1C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)NC(Cc1ccccc1)C(=O)O' **Invalid**		
The molecule is a muscarinic agonist that impacts pain treatment and is both alzheimer's treatment and anxiety treatment.	Cc1ccc(-c2ccc(-c3ccc(-c4ccc(-c5ccccc5)cc5)cc4)cc3)cc2)cc1 **Invalid**				
The molecule is a jak2 inhibitor and is cancer treatment.					
The molecule is both a anti psychotic and a nmda antagonist.	Cc1ccc(-c2ccc(-c3ccc(-c4ccc(-c5ccccc5)cc4)cc3)cc2)cc1 **Invalid**				
The molecule is a factor ixa inhibitor, a factor xa inhibitor, and anti thrombotic.	Cc1ccc(-c2ccc(-c3ccc(-c4ccc(-c5ccc(-c6ccccc6)cc6)cc4)cc4)cc3)cc2)cc1 **Invalid**				
The molecule is a flavoring agent and a nutrient, as well as nutty and green.					

THRUST 1

Results – Molecule Captioning

Team	Model	Overall Increase	Translation Metric Increase	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
avaliev	RAG_SIM_098	27.08	6.37	73.81	53.04	80.06	60.17	57.5	77.45
qizhipei	BioT5+_large_voting	14.66	6.45	75.58	54.77	79.41	59.89	57.46	75.43
protonunfold	SciMind	12.39	5.77	75.66	54.98	78.24	58.42	56.34	74.76
NLPeople	Ensembled	12.3	5.68	75.54	54.83	78.1	58.47	56.37	74.57
hecao	bioagent	10.95	5.57	74.11	53.84	78.73	59.38	57.06	74.08
xwk89	mistral_4b9_e1	10.8	5.56	74.38	54.08	78.49	59.49	56.82	73.91
mengmeng	Mistral	10.37	5.48	75.04	54.56	78.1	58.81	56.42	73.73
langmolecules	Meditron	10.34	5.47	75.16	54.72	77.97	58.75	56.33	73.69
NLPeople	Rank_model_1	9.94	4.8	74.73	54.1	77.3	57.7	55.53	73.26
dimitris	ALMol~10%DataTrained	9.61	4.27	74.71	54.34	76.43	56.54	54.66	72.76
xygui	MDEG	9.43	2.96	73.98	53.33	75.08	54.39	52.58	72.21
danielshao	SMol+LPM	7.83	2.7	72.2	52.02	74.76	56.1	53.34	71.57
xwk89	mistral_e1	7.45	3.46	73.18	53.24	75.29	56.74	54.39	71.75
duongttr	Mol2Lang-VLM	4.52	4.11	73.43	53.19	76.72	57.67	55.43	72.05
langmolecules	MolT5-Large	4.22	3.64	73.63	53.2	75.79	56.47	54.42	72.16
bluesky333	phi3-knowchem-sft-beam1	2.87	0.81	70.56	50.83	72.61	53.61	52.01	69.01
langmolecules	MolT5-Base	1.06	1.2	69.83	50.56	73.34	54.55	52.86	69.86
langmolecules	MolT5-Small	0	0	66.82	48.29	72.8	54.44	53.33	68.14
guiyike	yike	-16.64	-43.45	12.8	6.37	25.9	13.83	24.1	20.11

THRUST 1

Results – Molecule Captioning – Property F1 (%)

Team	Model	Overall Increase	Property Metric Increase	Overall Property F-1 Increase	Biomedical	Human Interaction	Agr. + Industry	Light + Electro	X-icides	Toxins	Light	Electricity	Inhibitors	anti-X	Modulators	Antagonist	Treatments	Agonists	Cancer	Disease	Held-out Combos
avaliev	RAG_SIM_098	27.08	33.99	26.99	27.9	4.18	3.55	72.32	0	6.74	71.92	72.71	44.52	11.07	68.09	55.7	65.49	41.53	52.07	66.15	69.09
qizhipei	BioT5+_large_voting	14.66	17.39	13.76	19.76	4.01	3.07	28.2	0	6.25	31.28	25.12	20.55	3.64	37.02	30.77	23.89	16.5	59.47	67.98	70.05
protonunfold	SciMind	12.39	14.6	11.51	18.17	3.94	2.93	21	0.04	6.2	23.95	18.05	18.06	2.55	30.14	25.36	19.42	14.81	57.54	67.43	69.92
NLPeople	Ensembled	12.3	14.5	11.63	17.86	4.02	2.97	21.67	0.04	6.33	24.74	18.59	17.02	2.77	30.96	25.89	18.61	15.1	53.69	67.44	69.99
hecao	bioagent	10.95	12.74	9.94	16.86	3.9	2.76	16.24	0	6.43	18.12	14.37	18.84	2.48	26.78	23.41	16.33	14.02	50.28	67.22	69.64
xwk89	mistral_4b9_e1	10.8	12.54	9.93	16.98	3.85	2.81	16.1	0	6.26	17.15	15.05	15.2	2.12	31.67	19.11	14.34	14.59	51.17	67.69	70.02
mengmeng	Mistral	10.37	12	9.73	16.72	3.81	2.83	15.57	0.04	6.03	18.49	12.64	14.74	2.12	24.43	21.21	14.34	10.55	53.61	67.45	69.99
langmolecules	Meditron	10.34	11.96	9.7	16.87	3.75	2.83	15.36	0.04	6.07	18.19	12.53	14.71	2.17	24.72	20.61	14.35	10.54	53.74	67.4	69.99
NLPeople	Rank_model_1	9.94	11.66	9.88	16.5	3.85	2.87	16.28	0.08	5.97	19.29	13.28	13.91	2.14	23.71	17.88	13.87	12.27	50.82	66.01	69.45
dimitris	ALMol~10%DataTrained	9.61	11.39	10.05	15.75	3.74	2.78	17.94	0.21	6	17.11	18.77	12.39	1.83	17.34	18.53	11.14	11.16	51.66	67.09	69.76
xygui	MDEG	9.43	11.59	8.74	16.08	3.57	2.72	12.6	0	5.86	12.68	12.52	13.2	1.87	29.96	19.61	12.47	14.08	54.55	67.1	69.32
danielshao	SMol+LPM	7.83	9.54	8.55	14.97	2.39	2.62	14.2	0	3.19	16.38	12.01	10.9	1.61	22.13	17.43	9.66	9.93	40.41	64.84	68.74
xwk89	mistral_e1	7.45	8.78	8.23	14.44	2.69	2.74	13.05	0	4.33	13.43	12.67	14.23	2	26.09	18.6	12.99	11.75	20.44	59.5	69.19
duongttr	Mol2Lang-VLM	4.52	4.66	5.76	10.73	3.18	2.36	6.78	0	5.5	9.39	4.18	1.08	0.3	0	2.49	1.72	0.69	40.62	67.61	69.67
langmolecules	MolT5-Large	4.22	4.42	5.81	10.33	3.2	2.36	7.36	0	5.58	7.77	6.94	0.65	0.13	0.39	0.11	2.02	0.58	38.34	66.97	69.21
bluesky333	phi3-knowchem-sft-beam1	2.87	3.56	5.36	10.05	3.15	2.13	6.11	0	5.45	8.01	4.21	0.93	0.23	0	0.29	2.04	0.21	35.39	63.71	64.96
langmolecules	MolT5-Base	1.06	1.02	3.99	8.27	2.68	2	3	0	4.63	5.52	0.48	0.08	0.04	0	0	1.79	0	18.04	47.64	68.46
langmolecules	MolT5-Small	0	0	3.23	7.87	0.27	1.65	3.12	0	0	6.24	0	0.06	0	0	0	1.66	0	17.99	41.09	65.06

THRUST 1

Results – Molecule Generation

Team	Model	Overall Increase	Test Metric Increase	BLEU	Exact Match	Levenshtein	Validity	MACCS FTS	RDK FTS	Morgan FTS	FCD
qizhipei	BioT5+_large	12.97	13.2	73.17	0.01	41.05	100	76.05	68.7	50.05	3.13
protonunfold	SciMind	12.68	12.76	73.44	0	40.35	99.78	75.06	67.05	47.82	2.54
avaliev	PLAIN	12.41	12.39	71.82	0.01	43.91	98.98	74.97	66.85	48.92	0.28
mengmeng	Mistral	12.26	12.22	70.56	0	43.75	99.4	75.6	67.57	48.62	2.01
langmolecules	Meditron	11.81	11.66	68.84	0.01	46.47	99.54	75.59	67.66	48.72	2.44
dimitris	ALMol~10%Data	10.41	9.26	69.74	0.01	43.24	92.84	70.22	62.79	42.96	3.05
hecao	bioagent_epoch5	10.21	10.67	61.98	0.02	47.12	99.67	75.94	68.38	46.92	2.17
langmolecules	MolT5-Base	10.07	10	67.04	0	45.71	99.89	74.61	63.7	46.29	nan
danielshao	SMol+LPM	8.74	8.45	59.74	0.01	55.09	97.66	74.35	66.79	46.6	4.25
hecao	bioagent	6.39	5.36	51.5	0.01	70.67	98.05	74	66.43	45.56	3.74
langmolecules	MolT5-Large	3.65	4.78	55.31	0	56.47	99.12	74.14	63.4	38.54	17.63
erikxiong	PUF	0	0	55.44	0	57.21	81.03	63.06	56.83	36.69	nan
langmolecules	MolT5-Small	0	0	55.44	0	57.21	81.03	63.06	56.83	36.69	nan
ndhieunguyen	Lang2mol-diff	-1.21	-0.75	54.15	0	55.26	100	59.6	32.44	31.98	10.71
guiyike	Nano	-7.39	-6.64	43.51	0	83.38	100	49.25	37.82	23.52	5.64

THRUST 1

Results – Molecule Generation – Held-out Combos

Team	Model	Overall Increase	Withheld Combo Increase	BLEU	Exact Match	Levenshtein	Validity	MACCS FTS	RDK FTS	Morgan FTS	FCD
qizhipei	BioT5+_large	12.97	12.74	78.48	0.02	43.99	100	86.54	80.43	59.54	4.32
protonunfold	SciMind	12.68	12.6	78.99	0	43.39	99.77	86.03	79.16	58.06	3.06
avaliev	PLAIN	12.41	12.42	78.19	0	47.84	99.64	86.3	79.16	59.02	0.35
mengmeng	Mistral	12.26	12.31	78.08	0	46.34	99.48	86.14	78.75	59.39	2.27
langmolecules	Meditron	11.81	11.96	77.11	0	48.29	99.53	86.23	78.91	59.57	2.59
dimitris	ALMol~10%Data	10.41	11.56	76.93	0.01	45.68	98.99	85.9	79.35	55.28	3.51
hecao	bioagent_epoch5	10.21	9.75	65.86	0	52.06	99.83	86.55	80.13	55.68	3.24
langmolecules	MolT5-Base	10.07	10.15	72.48	0	50.9	99.84	86.22	78.23	57.56	nan
danielshao	SMol+LPM	8.74	9.03	65.95	0	56.31	99.1	86.14	79.77	56.75	4.36
hecao	bioagent	6.39	7.42	64.6	0	65.63	99.11	85.86	79.58	54.81	4.18
langmolecules	MolT5-Large	3.65	2.36	57.74	0	66.94	99.17	83.19	74.77	40.97	nan
erikxiong	PUF	0	0	56.96	0	72.44	91.89	81.03	73.36	41.6	nan
langmolecules	MolT5-Small	0	0	56.96	0	72.44	91.89	81.03	73.36	41.6	nan
ndhieunguyen	Lang2mol-diff	-1.21	-1.67	59.38	0	63.17	100	73.26	38.6	39.76	6.38
guiyike	Nano	-7.39	-8.15	44.84	0	85.92	100	59.67	47.95	30.91	7.86

THRUST 1

Mystery Molecules!

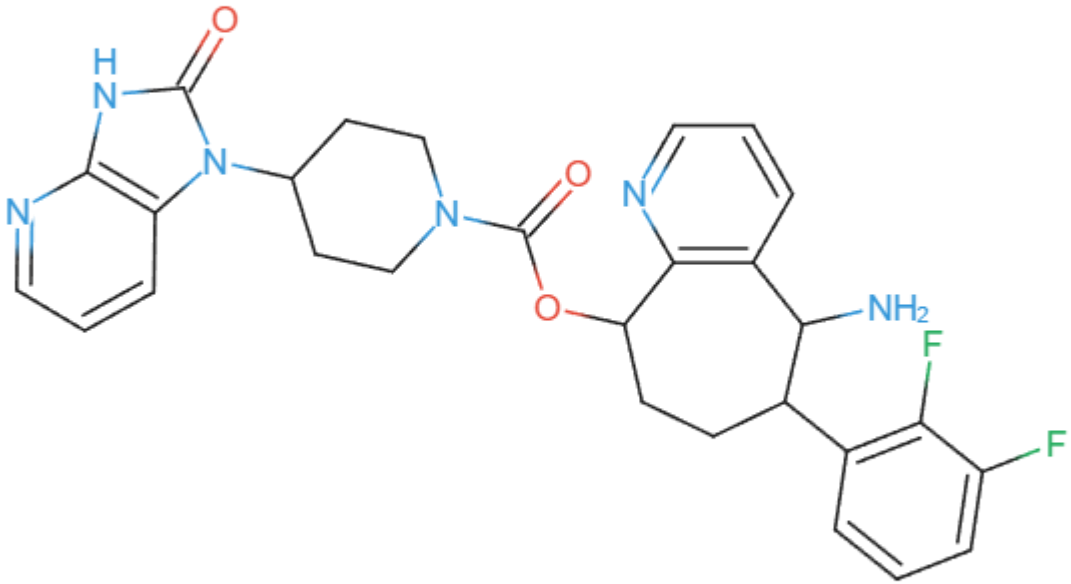
- 137 molecules without ground truth properties were added to the test split.
  - These consisted of:
    - Small molecule drugs approved by the FDA in the 2020s
    - Molecules suggested by the scientific advisory board
- The goal is to create relevant property predictions important molecules by ensembling the different submissions to the task.
  - We plan to make a submitted prediction dataset available for future work to build more sophisticated ensembles.
- In particular, we were interested in looking for “off-label” properties of newly approved drugs
  - We looked at a voting ensemble of properties suggested for these molecules by multiple models.

Mystery Molecules!

Rimegepant

- a medication used for the acute treatment of migraine with or without aura in adults and the prophylactic/ preventive treatment of episodic migraine in adults.

- Suggested to be a cgrp receptor antagonist (correct)

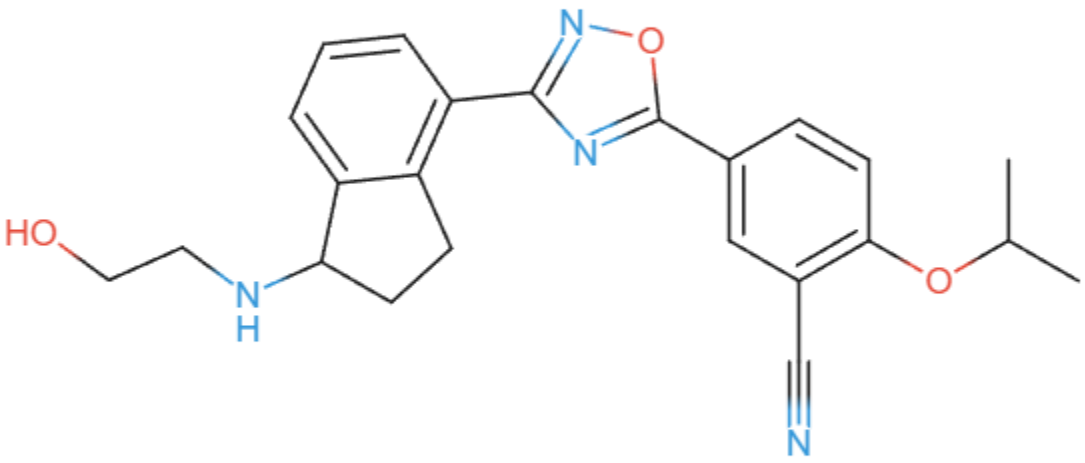


Mystery Molecules!

Ozanimod

- an immunomodulatory medication for the treatment of relapsing multiple sclerosis and ulcerative colitis

- Suggested to be an antifungal by several models

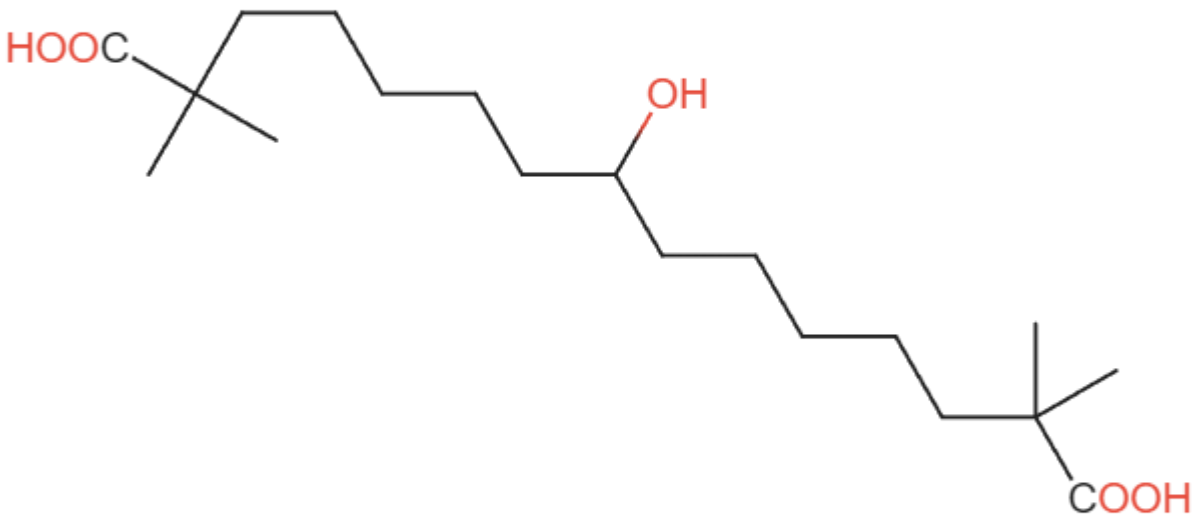


Mystery Molecules!

Bempedoic acid

- Used for treatment of high blood cholesterol levels.

- Predicted by several models as a coating for ship hulls to prevent sea life such as algae and mollusks attaching themselves to the hull.

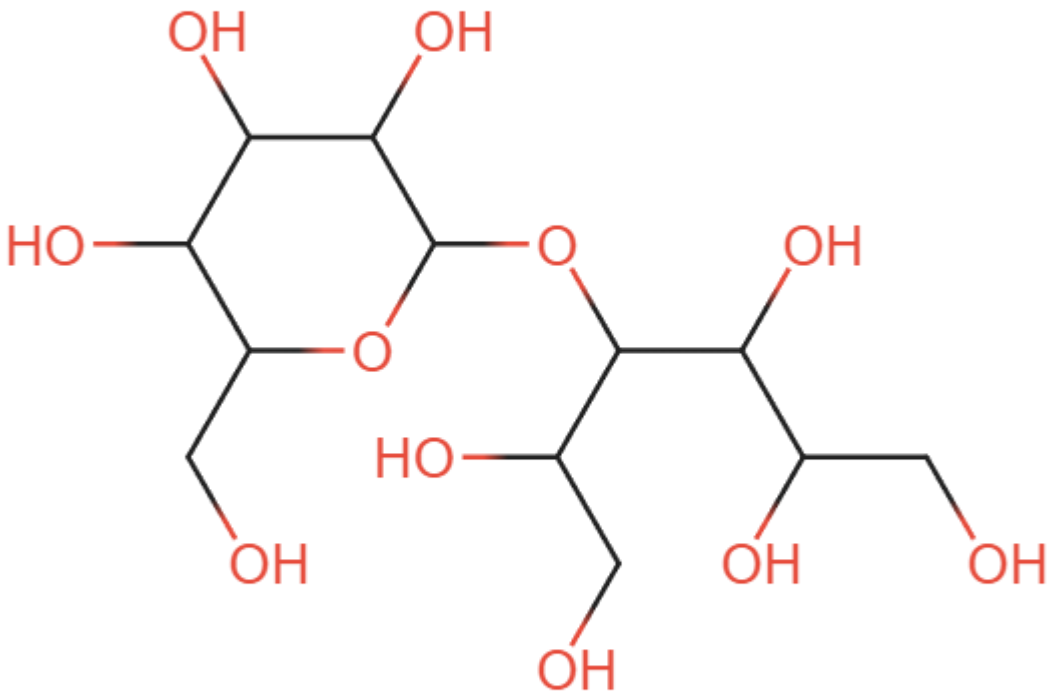


Mystery Molecules!

D-Lactitol monohydrate

- approved by the FDA for use in chronic idiopathic constipation in February 2020.

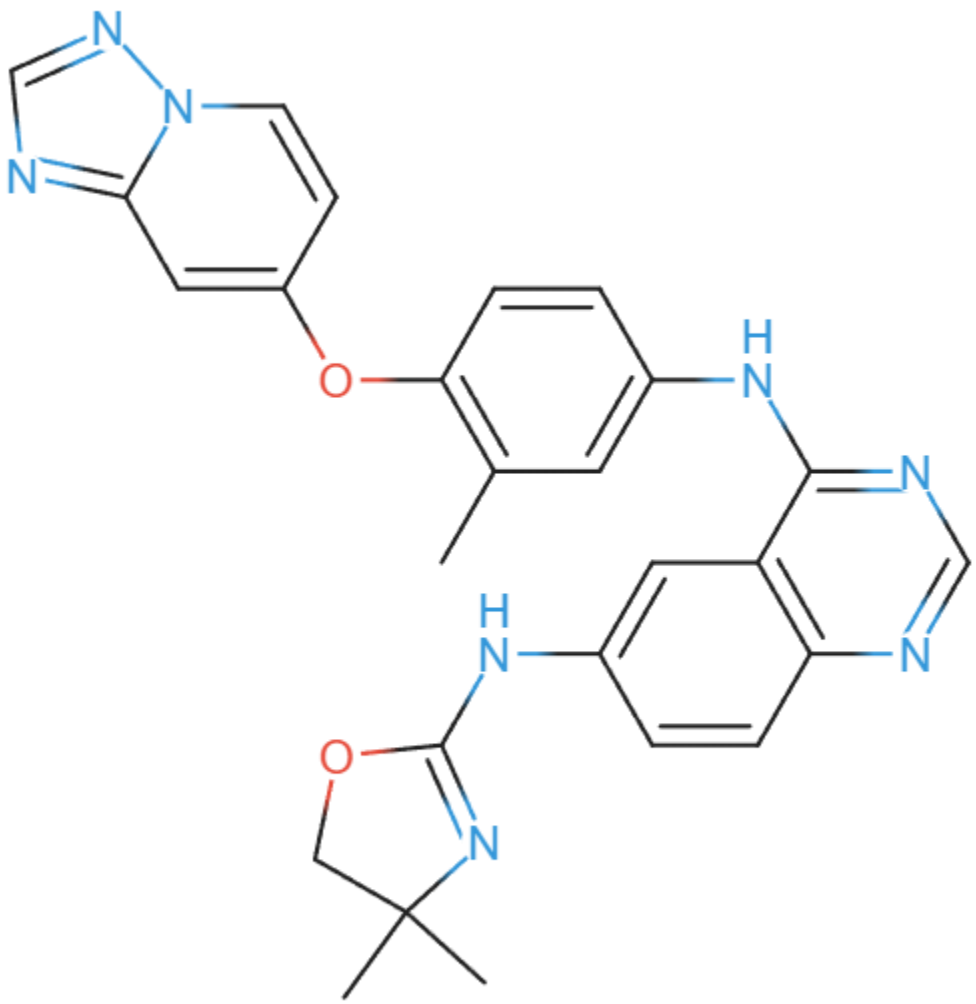
- Predicted as a nutrient and sweet tasting.
- "Sugar alcohol sweet tastant detectable by humans." - Sigma Aldrich



Mystery Molecules!

Tucatinib

- An anticancer medication used for the treatment of HER2-positive breast cancer
- Predicted as a kinase inhibitor by several models.

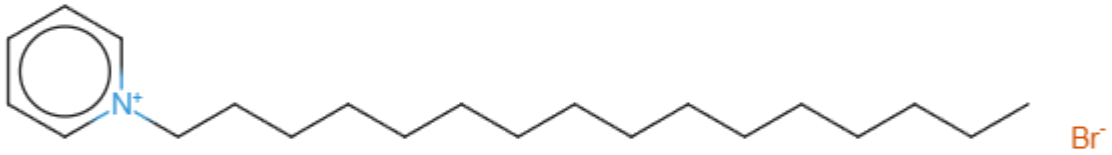


Mystery Molecules!

Cetylpyridinium bromide

- A surfactant, an antiseptic drug and an EC 2.7.11.18 (myosin-light-chain kinase) inhibitor

- Suggested to be toxic.

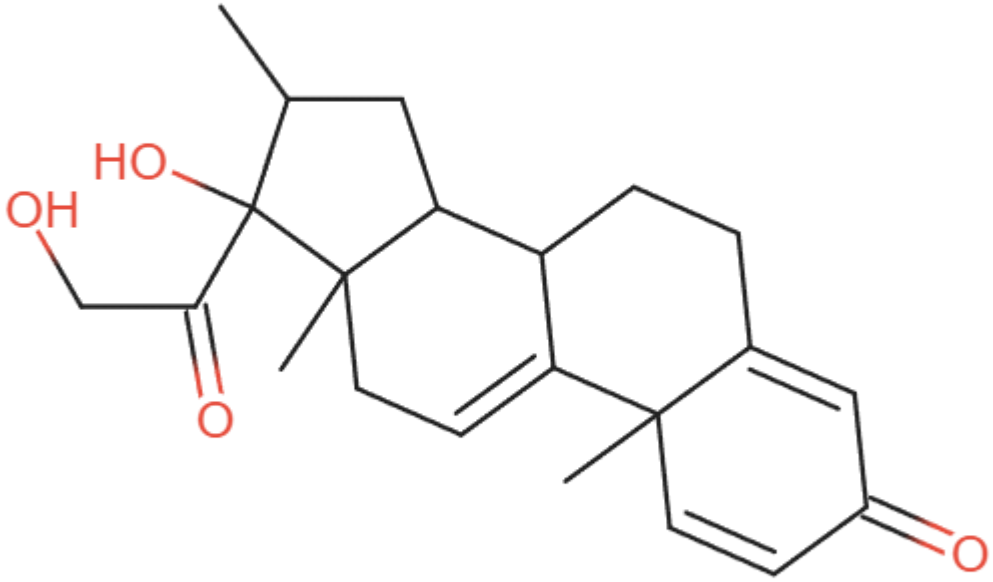


Mystery Molecules!

Vamorolone

- A corticosteroid used for the treatment of Duchenne muscular dystrophy

- Suggested to an anti-inflammatory (correct)
  - Also predicted to be anti-angiogenic

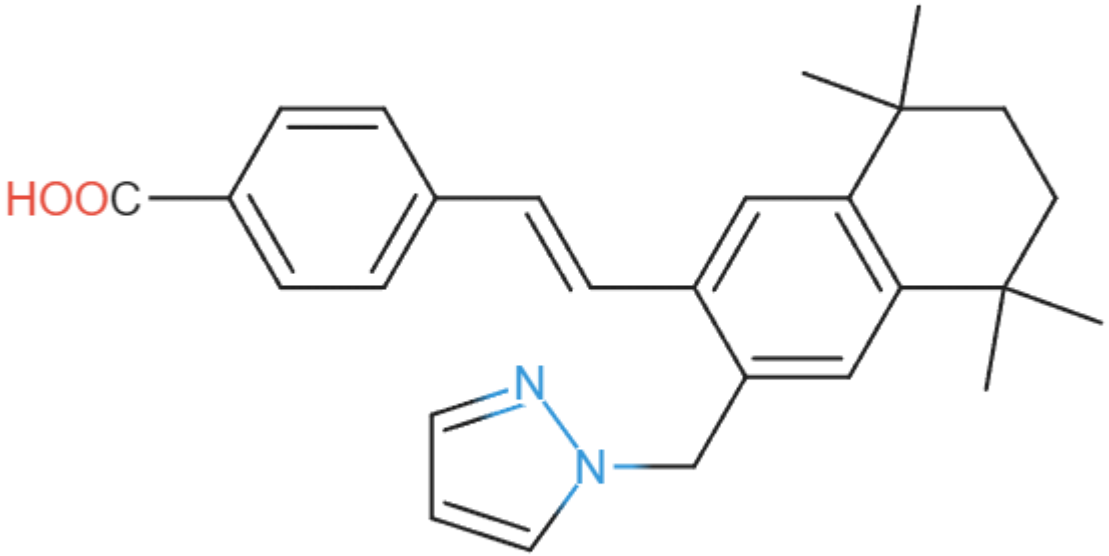


Mystery Molecules!

Palovarotene

- First in class medication for heterotopic ossification

- Suggested to be fluorescent by multiple models

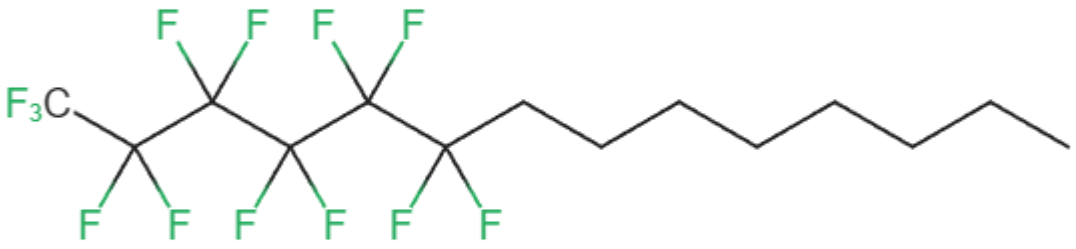


Mystery Molecules!

Perfluorohexyloctane

- Used for the treatment of dry eye disease

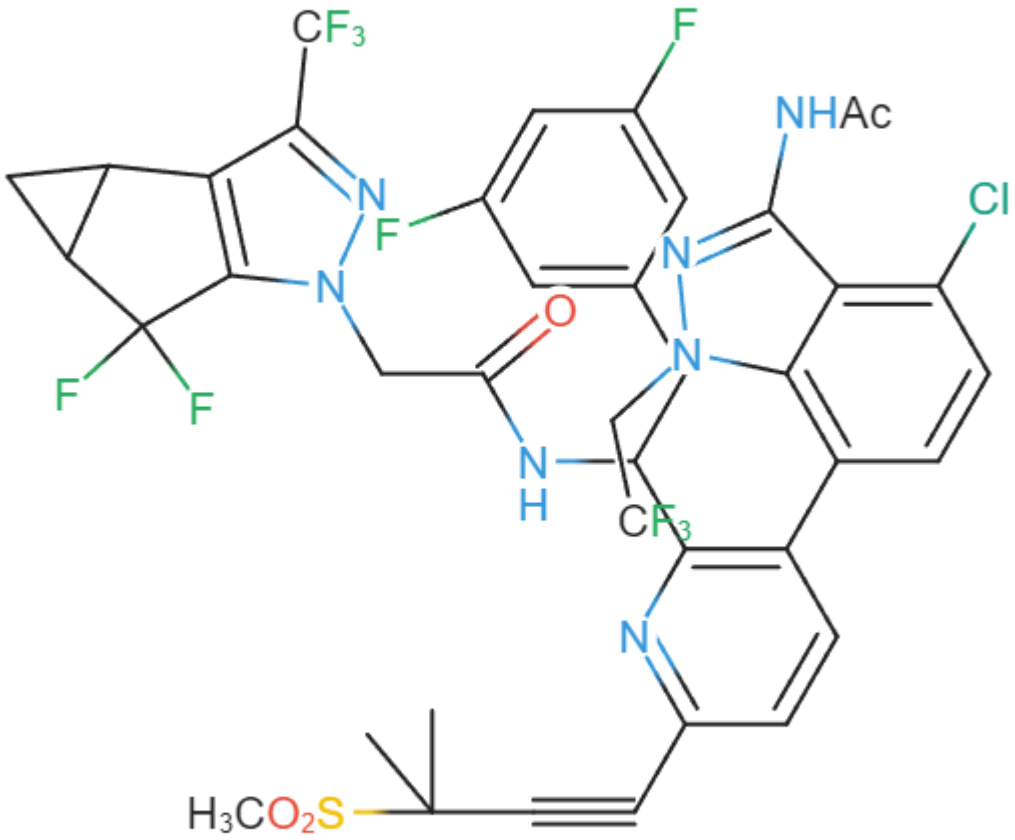
- Suggested to be a dielectric



Mystery Molecules!

Lenacapavir

- An antiretroviral medication used to treat HIV/AIDS
- Predicted as an anti viral and hiv treatment

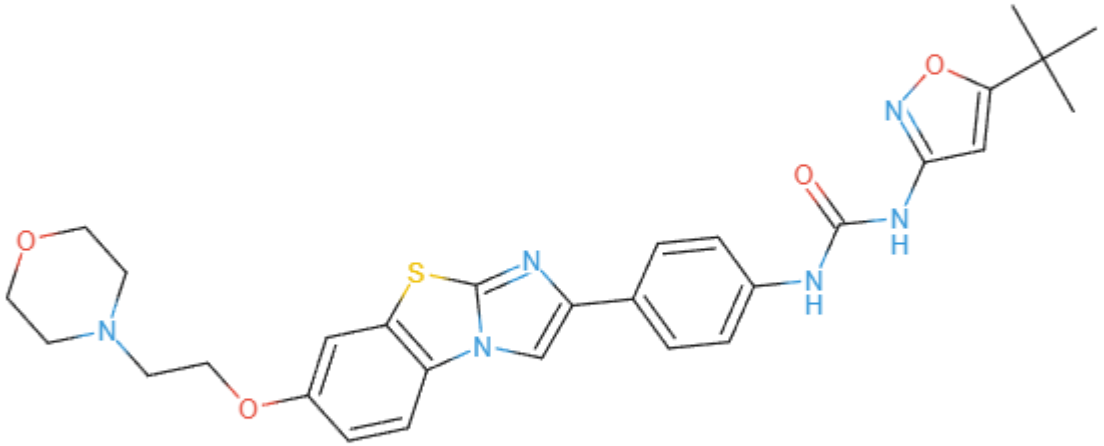


Mystery Molecules!

Quizartinib

- A kinase inhibitor used to treat Acute Myeloid Leukemia

- Predicted as antiviral by several models



Thank you to everyone who participated in the shared task!!

- The task was designed to emphasize benefits of natural language in molecule design.
- It considered new metrics for evaluating property prediction in an open-vocabulary setting (specifically, F-1 score)
  - This revealed that performance varies considerably between different types of properties!
- Your submissions have helped us to reveal a number of exciting points
  - Evaluation is still a big challenge!
  - Data splitting when focusing on abstraction, composition, and functionality is challenging and has room for improvement.
  - Inserting relevant external data, especially on molecular interactions, is critical.
  - Model architectures currently favor building from LLMs, but there's significant opportunity for research on novel ways to interpret different data modalities.

- Shared task submitters will be presenting the details of their methods in the upcoming oral and poster presentations!

MOLECULE MAKER LAB INSTITUTE

NSF

Genentech
A Member of the Roche Group

IBM

janssen

REVOLUTION MEDICINES

abbvie

AMGEN

NVIDIA

LanzaTech

Pacific Northwest NATIONAL LABORATORY

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

RIT Rochester Institute of Technology

PennState

Ai2 Allen Institute for AI

For many images