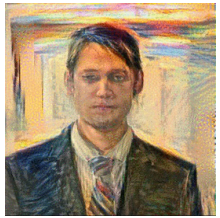
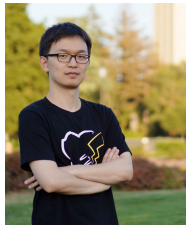
AIScientist: Knowledge Empowered Joint Large and Molecule Modeling to Accelerate Drug Discovery

Heng Ji
Computer Science Professor@UIUC, Amazon Scholar
Founding Director of Amazon-Illinois Center on AI for Interactive Conversational Experiences (AICE)

Based on the wonderful work done by my PhD students and postdocs:
Carl Edwards, Hongwei Wang, Tuan Lai and Zixuan Zhang
Collaborations with Martin Burke (UIUC)

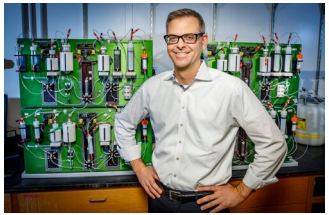
hengji@illinois.edu
University of Illinois Urbana-Champaign


Carl Edwards

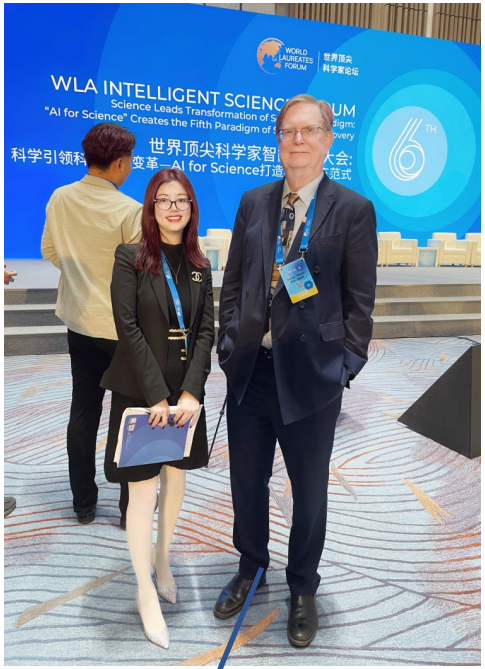

Hongwei Wang


Tuan Lai


Zixuan Zhang


Martin Burke
Prof. of Chemistry, UIUC

Who are Most Excited about Chat-GPT?

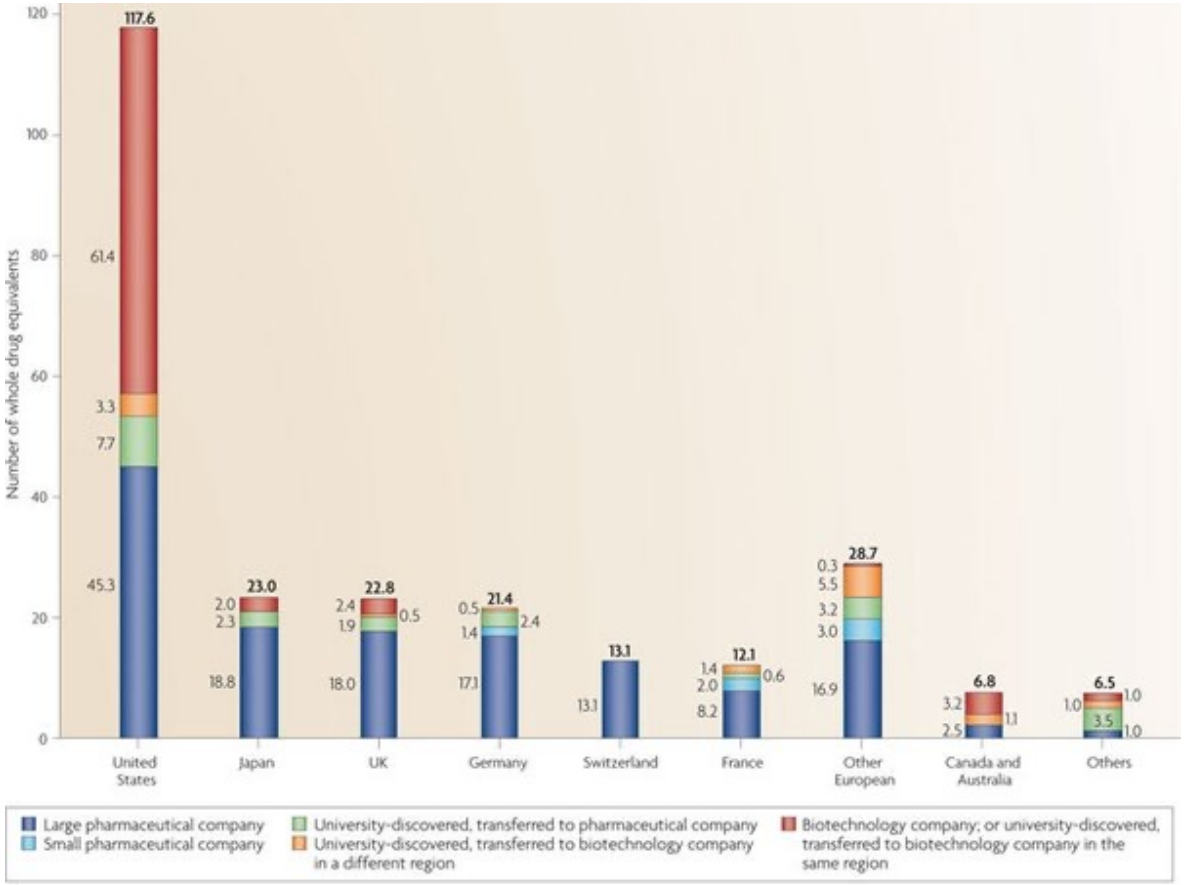


Dr. Michael Levitt (Nobel Laureate in Chemistry 2013)

- But they are searching gold in sand land: there are approximately **166 billion** small molecules, and **970 million** of them are druglike, discovering each drug costs $1.3 billion

Dr. George Smoot (Nobel Laureate in Physics 2006)

Drug Discovery is Currently a Very Luxury Research Area



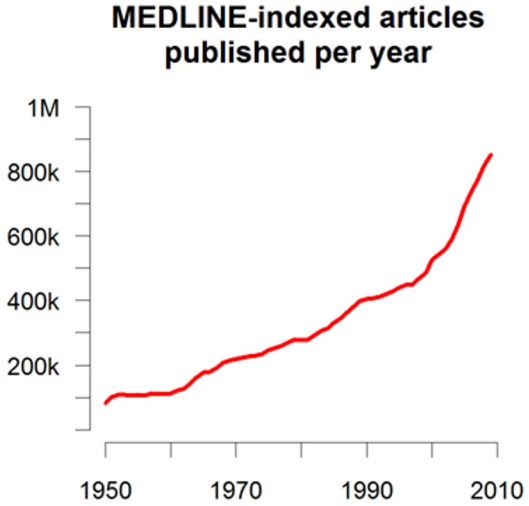
- We need to democratize drug discovery with AI
- A lot of repetitive work too
- Lab is quite unstructured

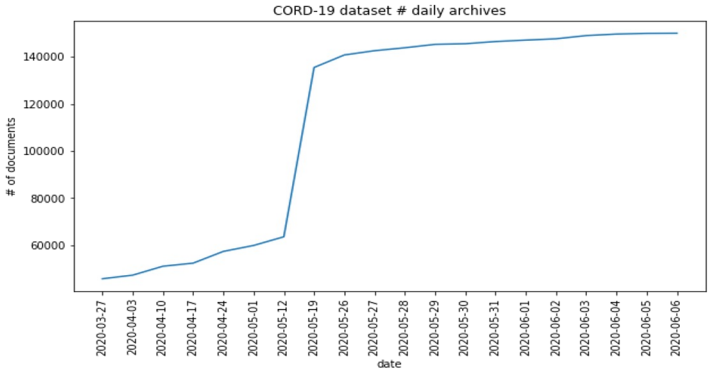
Who Are Most Excited about Scientific Discovery?

- Dominated by computer scientists, almost all papers are about numbers using training/dev/test split of known molecules, so we are not discovering anything
- 95% accuracy is exciting to computer scientists, but not to chemists and doctors
- Most work is overly simplified (e.g., 2D instead of 3D molecule modeling)

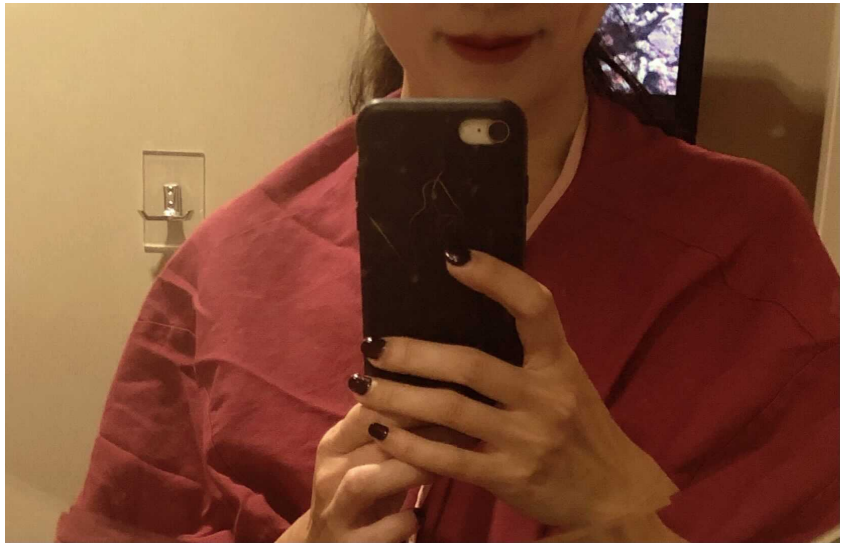
A Professional Motivation: Too Many papers

- More than 500K papers are published at PubMed every year, and more than 1.2 million new papers are published in 2016 alone, bringing the total number of papers to over 26 million (Van Noorden, 2014)
- As of June 13, 2020, there are at least 140K papers about coronavirus
- Quality: Given the rapid publications of preprints without peer reviews, many research results are redundant, complementary or even conflicting with each other
- Human's reading ability keeps almost the same across years: US scientists estimated that they read, on average, only 264 papers per year (1 out of 5000 available papers, the same across years)





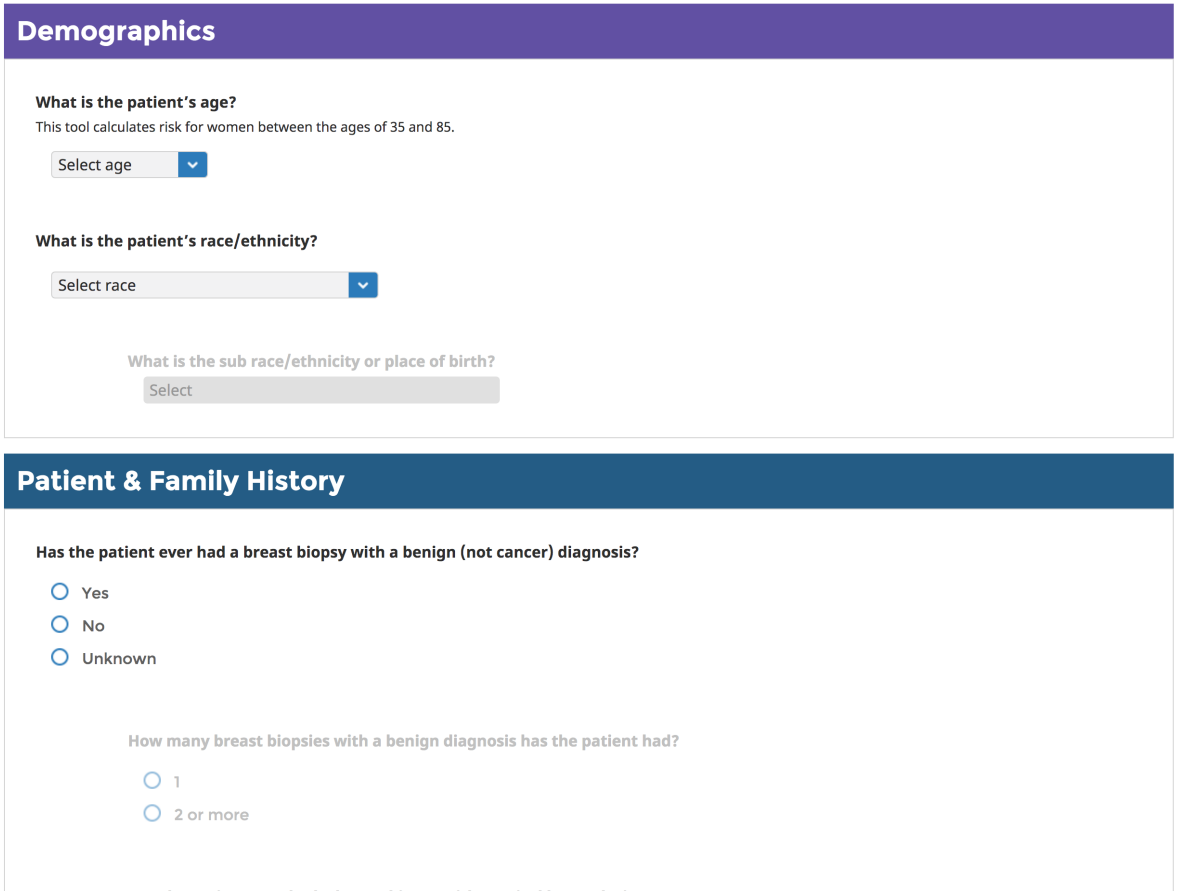
A Personal Motivation: Democratizing Scientific Discovery



My Darkest Day During Pandemic:

Trust a prediction model trained from seven features, or a surgeon who treats you like his sister?

How Doctors Predict Cancer Today



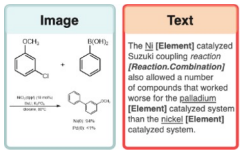
- The classification features are extremely coarse-grained, generic and fragile
  - Changing the number of biopsies from 1 to 2 will change the cancer risk level from 17% to 37%, despite of the positive/negative results of biopsies
- Precision Medicine is only affordable for a tiny population
- Development cost is about $2.6 billion

Our Road Map: Converting Unstructured Scientific Data to Structured Knowledge

DATA

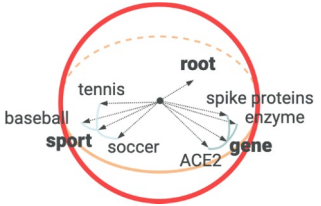
SEMANTICS

KNOWLEDGE BASE

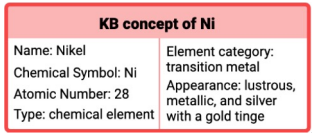


Scientific Literature

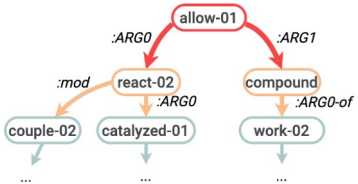
Chemical Ontology & Existing Databases



Hierarchical Spherical Embedding

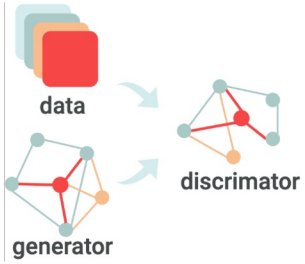


Ontology Enriched Text Embedding



Cross-media Structured Semantic Representation

Graph neural networks
Joint entity/relation/event extraction and ontology construction



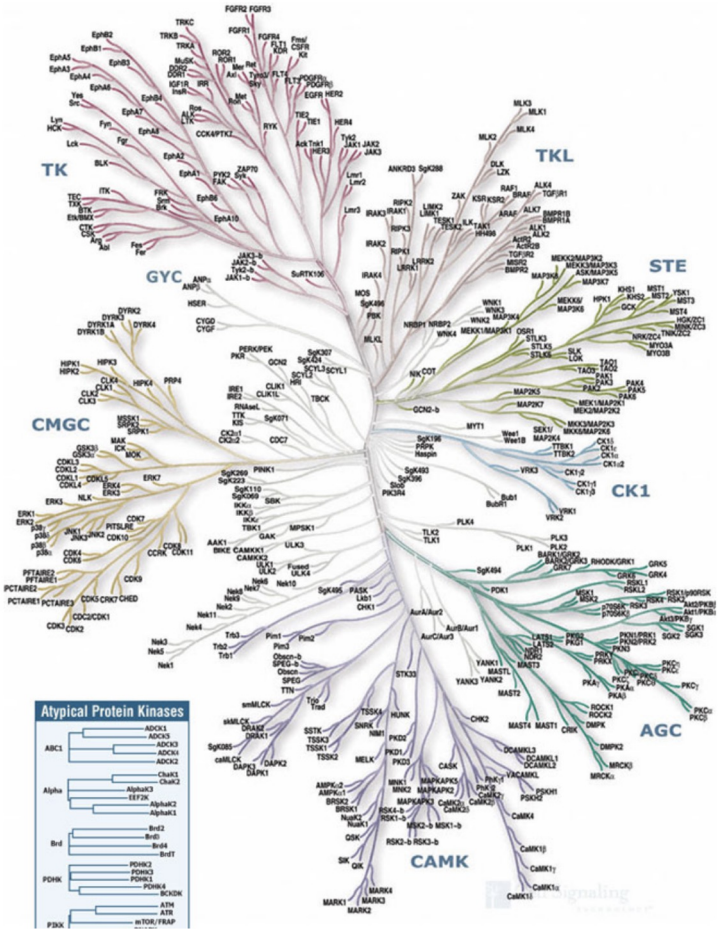
Generative Adversarial Networks for Data Augmentation and Distant Supervision

Multimedia Knowledge Base

Multimedia Search and Summarization

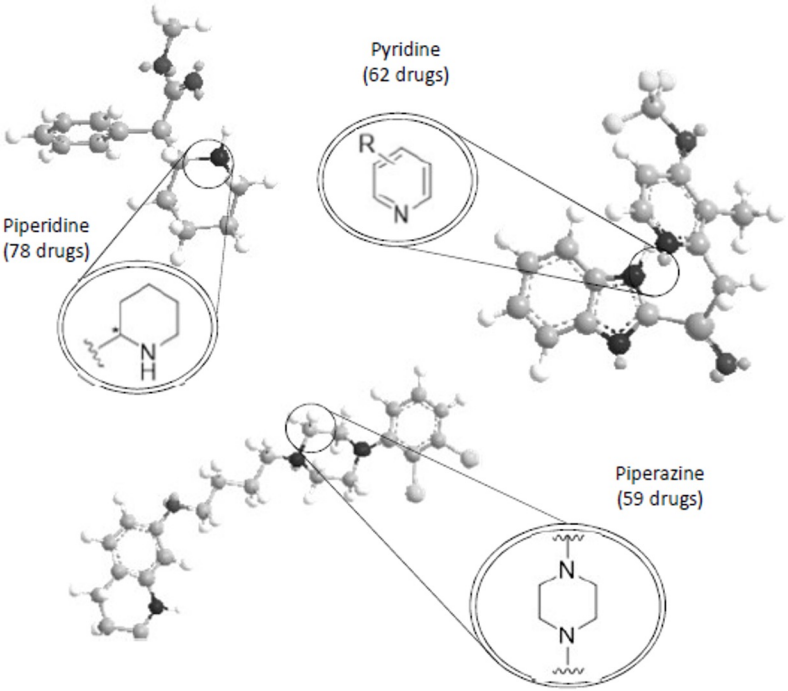
The power of small molecule drugs: kinase inhibitors

- A substance that blocks a type of enzyme called a kinase.
- Human cells have many different kinases, and they help control important functions, such as cell signaling, metabolism, division, and survival.
- Certain kinases are more active in some types of cancer cells and blocking them may help keep the cancer cells from growing.
- Kinase inhibitors may also block the growth of new blood vessels that tumors need to grow
- Some kinase inhibitors are used to treat cancer; but there are too many of them, so we should try to use AI to discover them automatically!



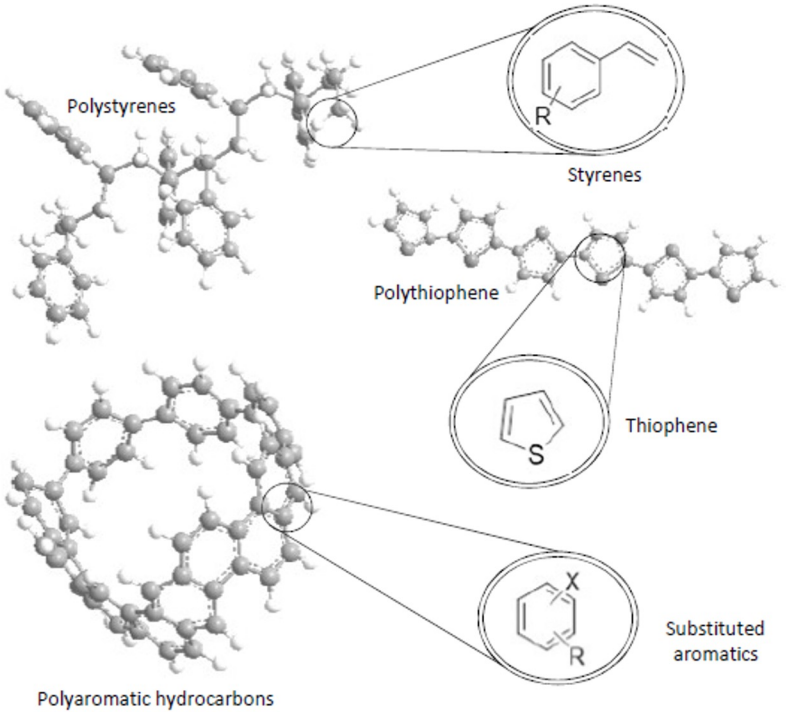
Finding Building Blocks in Drugs

Drugs



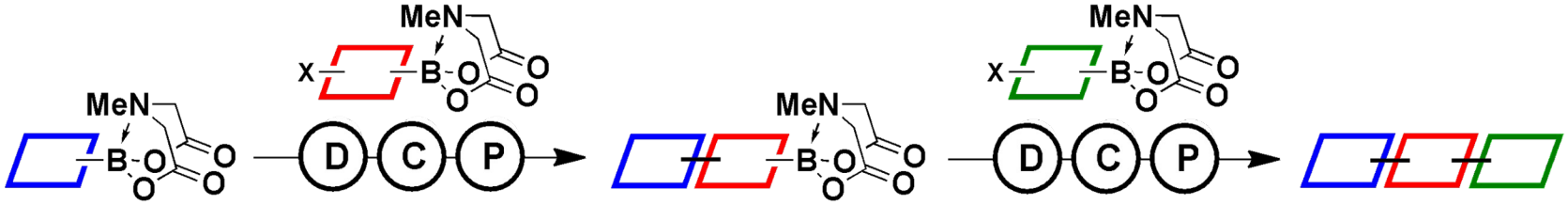
Njardarson *J Med Chem* **2014,** *57* (24), 10257-74

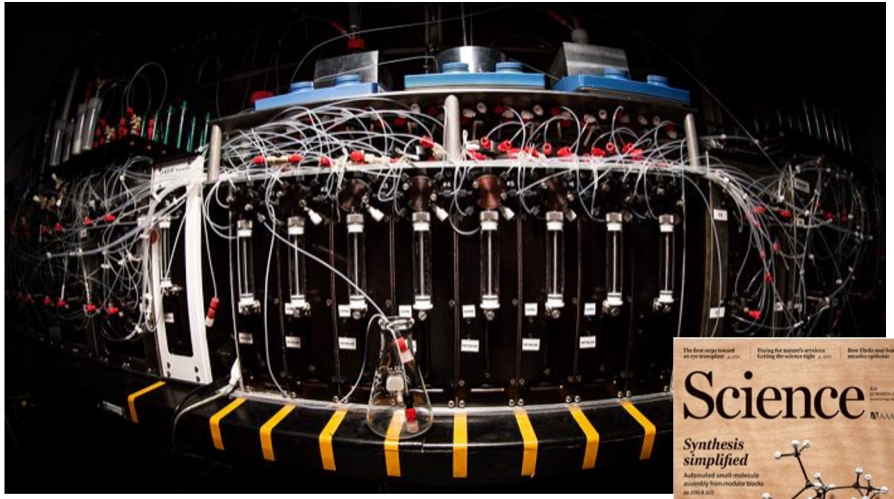
Materials



Johnathan Lehmann and Daniel Blair *Nature Review Chemistry* **2018**, *2*, 0115

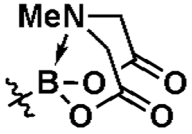
Molecule-Making Machines at UIUC

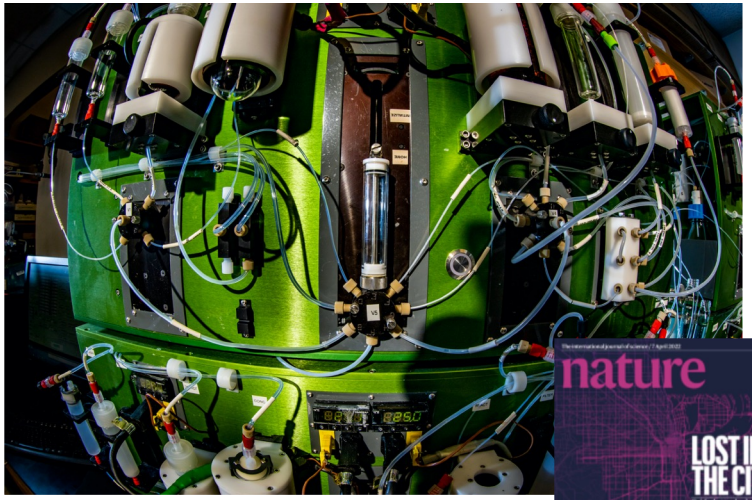




MIDA

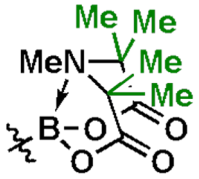
201 5



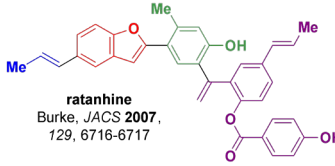


TIDA

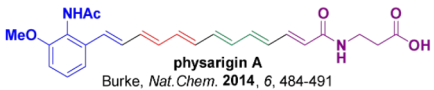
2022



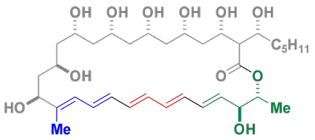
Small Molecules Synthesized via Lego Platform



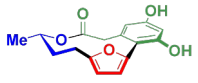
ratanhine
Burke, *JACS* **2007**, *129*, 6716-6717



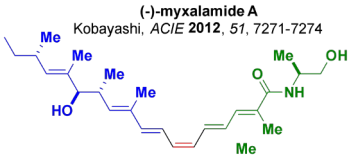
physarigin A
Burke, *Nat.Chem.* **2014**, *6*, 484-491



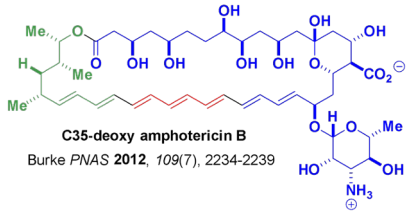
polyene core of filipin III
Cossy, *JOC* **2016**, *81*, 8206-8221



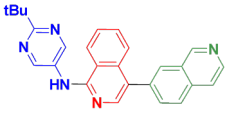
Citreofuran
Burke, *Science* **2015**, *347*, 1221-1226



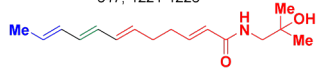
(-)-myxalamide A
Kobayashi, *ACIE* **2012**, *51*, 7271-7274



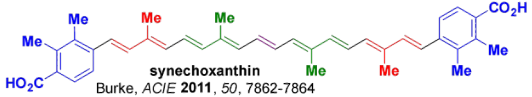
C35-deoxy amphotericin B
Burke *PNAS* **2012**, *109*(7), 2234-2239



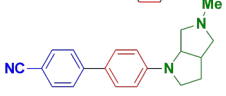
B-Raf Kinase Inhibitor
Burke, *Science* **2015**, *347*, 1221-1226



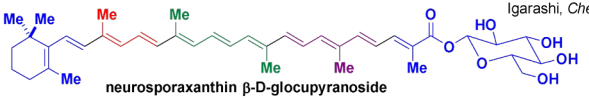
hydroxy-β-sanshool
Igarashi, *Chem. Pharm. Bull.* **2012**, *60*, 1088-1092



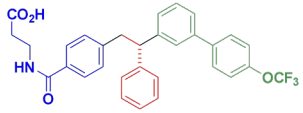
synechoxanthin
Burke, *ACIE* **2011**, *50*, 7862-7864



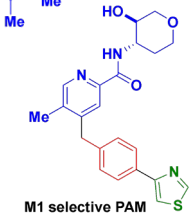
Histamine H3 antagonist
Grob, *Org. Lett.* **2012**, 14, 5578-5581



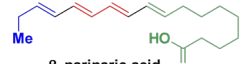
neurosporaxanthin β-D-glocupyranoside
Burke, *Nat. Chem.* **2014**, *6*, 484-491



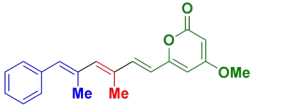
glucagon receptor inhibitor
Burke, *JACS* **2011**, *133*, 13774-13777



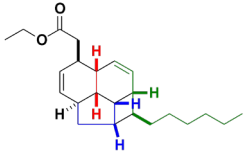
M1 selective PAM
Davoren, *J. Med. Chem.* **2016**, *59*, 6313-6328



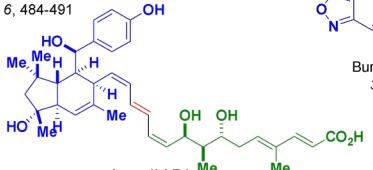
β–parinaric acid
Burke, JACS **2008**, *130*, 466-468



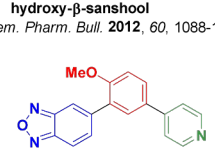
asnipyrone B
Burke, *Nat. Chem.* **2014**, *6*, 484-491



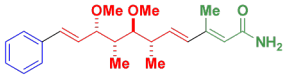
cryptobeilic acid D ethyl ester
Vosburg, *Tetrahedron* **2016**, *72*, 3790-3794



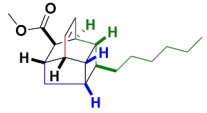
elansolid B1
Kirschning, *Org. Lett.* **2014**, *16*, 568-571



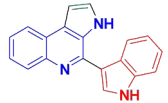
PDE472
Burke, *Science* **2015**, *347*, 1221-1226



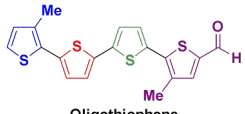
Crocacin C
Burke, *JACS* **2008**, *130*, 14084-14085



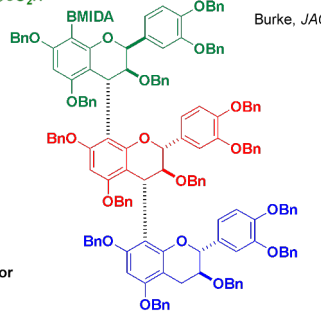
endiandric acid K
Vosburg, *Tetrahedron* **2016**, 72, 3790-3794



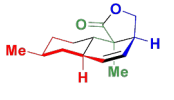
marinoquinoline E
Sperry, *Synlett* **2013**, *24*, 461-464



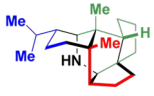
Oligothiophene
Burke, *Science* **2015**, *347*, 1221-1226



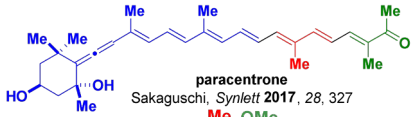
procyanidin derivative
Smith, *Tetrahedron* **2012**, *68*, 340-348



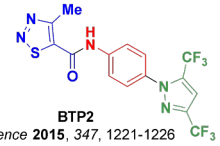
Oblongolide
Burke, *Science* **2015**, *347*, 1221-1226



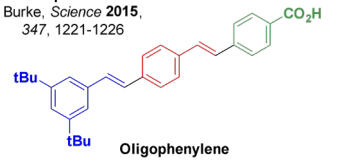
secodaphnane core
Burke, *Science* **2015**, *347*, 1221-1226



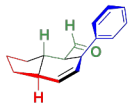
paracentrone
Sakaguschi, *Synlett* **2017**, *28*, 327



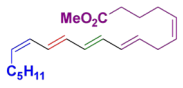
BTP2
Science **2015**, *347*, 1221-1226



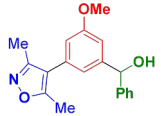
Oligophenylene
Burke, *Science* **2015**, *347*, 1221-1226



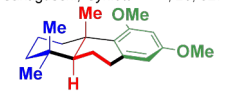
Hexahydroindene core
Burke, *Science* **2015**, *347*, 1221-1226



methyl eicosapentaenoate
Hansen, *Tetrahedron Lett.* **2012**, *52*, 1057-1059



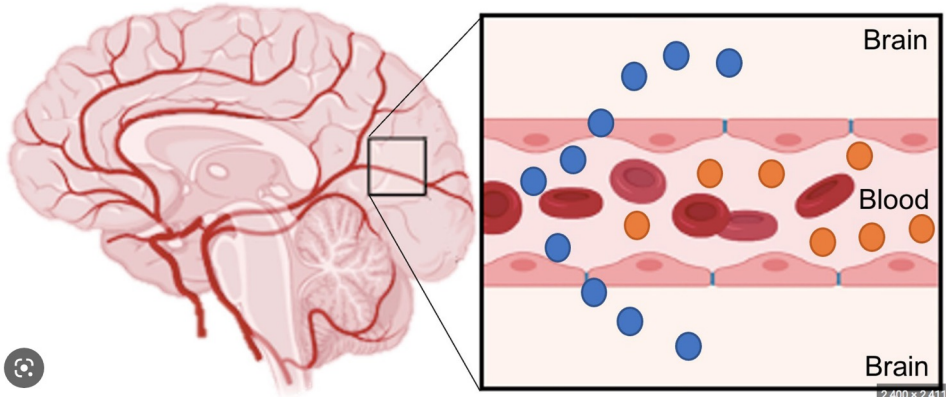
BET bromodomain inhibitor
Watson, *ACIE* **2015**, *54*, 9976-9979



Steroid-like core
Burke, *Science* **2015**, *347*, 1221-1226

What are the Problems with This Beautiful Manual Approach?

- Only covers the tip of an iceberg
  - When it was first introduced on the market, Imatinib was heralded as a breakthrough for treating certain types of cancers using molecularly targeted oncotherapy.
  - Since the FDA-approval of Imatinib in 2001 for leukemia, the class of Kinase inhibitors has quickly become an important tool for treating various cancers, as well as other diseases.
  - There are now **89** approved tyrosine kinase inhibitors in various healthcare systems worldwide, and there are **184** blocks based on these
  - But there are approximately **166 billion** small molecules, and **970 million** of them are druglike
- The functions for the 89 FDA-approved kinase inhibitors are also often limited
  - e.g., Imatinib has widely been reported to not penetrate the blood brain barrier
  - This means that Imatinib may successfully treat a cancer at its origin, but the patient may still pass away due to a brain metastasis.



Joint Molecule and Language Modeling to Discover Building Blocks for Improved Drug Design

- FDA approved drugs consist of substituent **building blocks** which are particularly relevant for beneficial drug properties
- Our goal: Using blood brain barrier penetration (BBBP) as a case study to produce better candidate kinase inhibitor variants to pass BBB.
- **Existing knowledge bases cannot solve the problem**: 0 of the 89 kinase inhibitors we care about are in the popular human constructed BBBP dataset.
- **GPT-4 cannot solve the problem**: LLMs, by design, make up (hallucinate) some false claims in a confident tone:

HE

Why does Imatinib effectively cross the blood brain barrier?

Imatinib, also known by its trade name Gleevec or Glivec, is a medication primarily used to treat certain types of cancer, such as chronic myelogenous leukemia (CML) and gastrointestinal stromal tumors (GIST). While Imatinib is not specifically designed to cross the blood-brain barrier (BBB), it has been found to have some ability to penetrate the central nervous system (CNS). The exact mechanism by which Imatinib crosses the BBB is not fully understood, but several factors likely contribute to its ability to reach the brain effectively.

HE

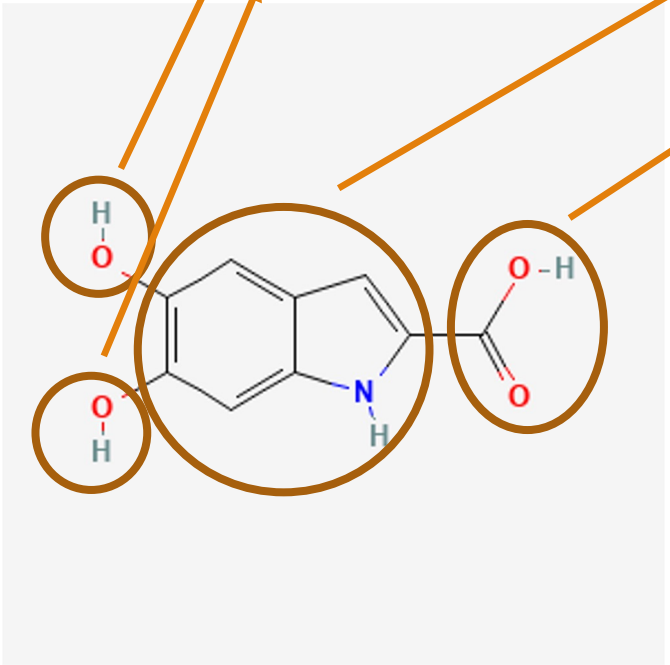
Why doesn't Imatinib effectively cross the blood brain barrier?

I apologize for the incorrect information in my previous response. Imatinib does not effectively cross the blood-brain barrier (BBB). In fact, the penetration of Imatinib into the central nervous system (CNS) is generally limited.

The main reason for the limited penetration of Imatinib across the BBB is its physicochemical properties. Imatinib is a relatively large molecule with a molecular weight of around 589 Daltons, which makes it more challenging for it to passively diffuse through the tight junctions of the BBB.

Challenge 1: Chemistry Language != Regular Natural Language

5,6-DIHYDROXY-1H-INDOLE-2-CARBOXYLIC ACID



Definition: *5,6-dihydroxyindole-2-carboxylic acid is a dihydroxyindole that is indole-2-carboxylic acid substituted by hydroxy groups at positions 5 and 6. It has a role as a mouse metabolite. It is a conjugate acid of a 5,6-dihydroxyindole-2-carboxylate. It is a tautomer of a dopachrome.*

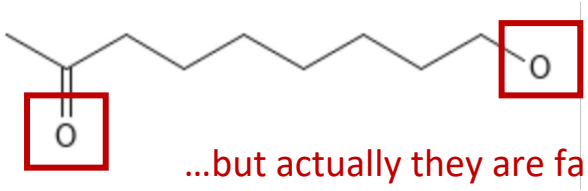
Property Name	Property Value
Molecular Weight	**193.16**
XLogP3-AA	1.2
Hydrogen Bond Donor Count	4
Hydrogen Bond Acceptor Count	4
Rotatable Bond Count	1
Exact Mass	**193.03750770**
Monoisotopic Mass	**193.03750770**
Topological Polar Surface Area	93.6 Å²
Heavy Atom Count	14
Formal Charge	0
Complexity	245

Limitation of SMILES-based Methods

- Smiles are 1D linearization of molecule structures, which makes them hard to learn the original structural information of molecules

CC(CCCCCCCO)=O

These two O's are close in SMILES string…

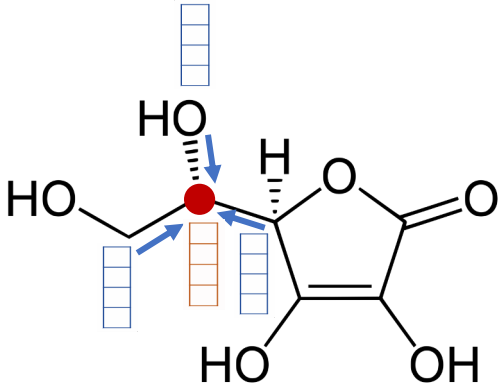


…but actually they are far from each other

Graph Neural Networks (GNN)-based Methods

- Each atom has an initial feature vector consisting of information from literature:
  - Element type, Charge, Whether the atom is in an aromatic ring, the count of attached hydrogen atom(s)
  - Bond type can be inferred by the features of its two associated atoms

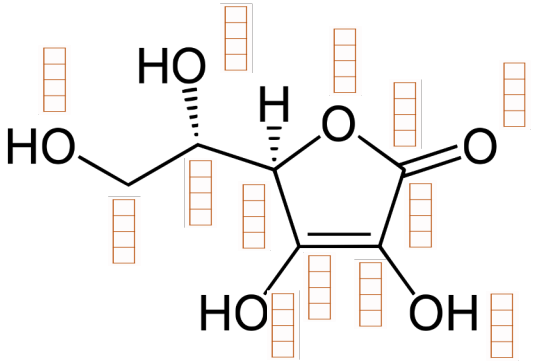
1. Propagating messages over the graph



In the k-th iteration, aggregating neighborhood information and update the embedding for atom i:

$$h_i^k = AGGREGATE\left(\{h_j^{k-1}\}_{j\in\mathcal{N}_i\cup\{i\}}\right)$$

2. Read out the molecule graph embedding



After K iteration, using a readout function to aggregate all aom embeddings and return the whole graph embedding:

$$h_G = READOUT\left(\{h_i^K\}_{i\in G}\right)$$

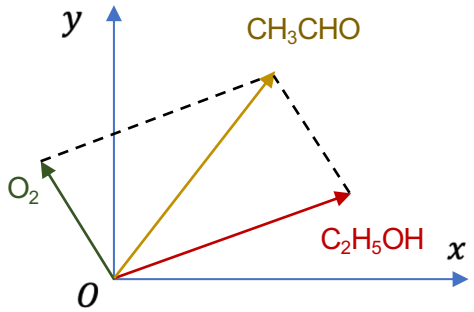
Taking Advantage of Chemical Reaction Equivalence for Embedding Composition [Wang et al., ICLR2022]

- A chemical reaction defines a particular relation "→" between reactant set $R = \{r_1, r_2, \dots\}$ and product set $P = \{p_1, p_2, \dots\}$:

$$r_1 + r_2 + \cdots \rightarrow p_1 + p_2 + \cdots$$

- Several physical quantities retain constant before and after the reaction
  - Mass, energy, charge, etc.
- We aim to preserve such equivalence in the molecule embedding space:

$$\sum_{r \in R} h_r = \sum_{p \in P} h_p$$



$$C_2H_5OH + O_2 \rightarrow CH_3CHO$$

$$h_{C_2H_5OH} + h_{O_2} = h_{CH_3CHO}$$

Chemical reaction space

Molecule embedding space

Experiments: Chemical Reaction Prediction

- USPTO Dataset - Training set: 409k, validation set: 30k, test set: 40k

Metrics	MRR	MR	Hit@1	Hit@3	Hit@5	Hit@10
Mol2vec	0.681	483.7	0.614	0.725	0.759	0.798
Mol2vec-FT1	0.688 ± 0.000	417.6 ± 0.1	0.620 ± 0.000	0.734 ± 0.000	0.767 ± 0.000	0.806 ± 0.000
MolBERT	0.708	460.7	0.623	0.768	0.811	0.858
MolBERT-FT1	0.731 ± 0.000	457.9 ± 0.0	0.649 ± 0.000	0.790 ± 0.000	0.831 ± 0.000	0.873 ± 0.000
MolBERT-FT2	0.776 ± 0.000	459.6 ± 0.2	0.708 ± 0.000	0.827 ± 0.000	0.859 ± 0.000	0.891 ± 0.000
MolR-GCN	0.905 ± 0.001	34.5 ± 2.4	0.867 ± 0.001	0.938 ± 0.001	0.950 ± 0.001	0.961 ± 0.002
MolR-GAT	0.903 ± 0.002	35.3 ± 2.8	0.864 ± 0.002	0.935 ± 0.003	0.948 ± 0.003	0.961 ± 0.003
MolR-SAGE	0.903 ± 0.004	53.0 ± 4.6	0.865 ± 0.005	0.935 ± 0.004	0.948 ± 0.004	0.961 ± 0.002
MolR-TAG	**0.918** ± 0.000	**27.4** ± 0.4	**0.882** ± 0.000	**0.949** ± 0.001	**0.960** ± 0.001	**0.970** ± 0.000
MolR-TAG (1% training data)	0.904 ± 0.002	33.0 ± 3.7	0.865 ± 0.003	0.937 ± 0.003	0.951 ± 0.002	0.963 ± 0.002
Improvement of MolR-TAG over the best baseline	**14.2%**	**390.2**	**17.4%**	**12.2%**	**10.1%**	**7.9%**

Can We Translate between Molecules and Natural Language? [Edwards et al., 2022arxiv]

Image Captioning

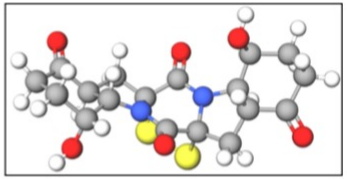
1. a cat sitting on top of an open laptop computer.
2. a cat that is sitting on top of a lap top.
3. a cat is sitting on the keyboard of a laptop.
4. a cat is sitting on an open laptop.
5. a striped cat sitting on top of a laptop

Captions from COCO

Molecule Captioning

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

SMILES representation

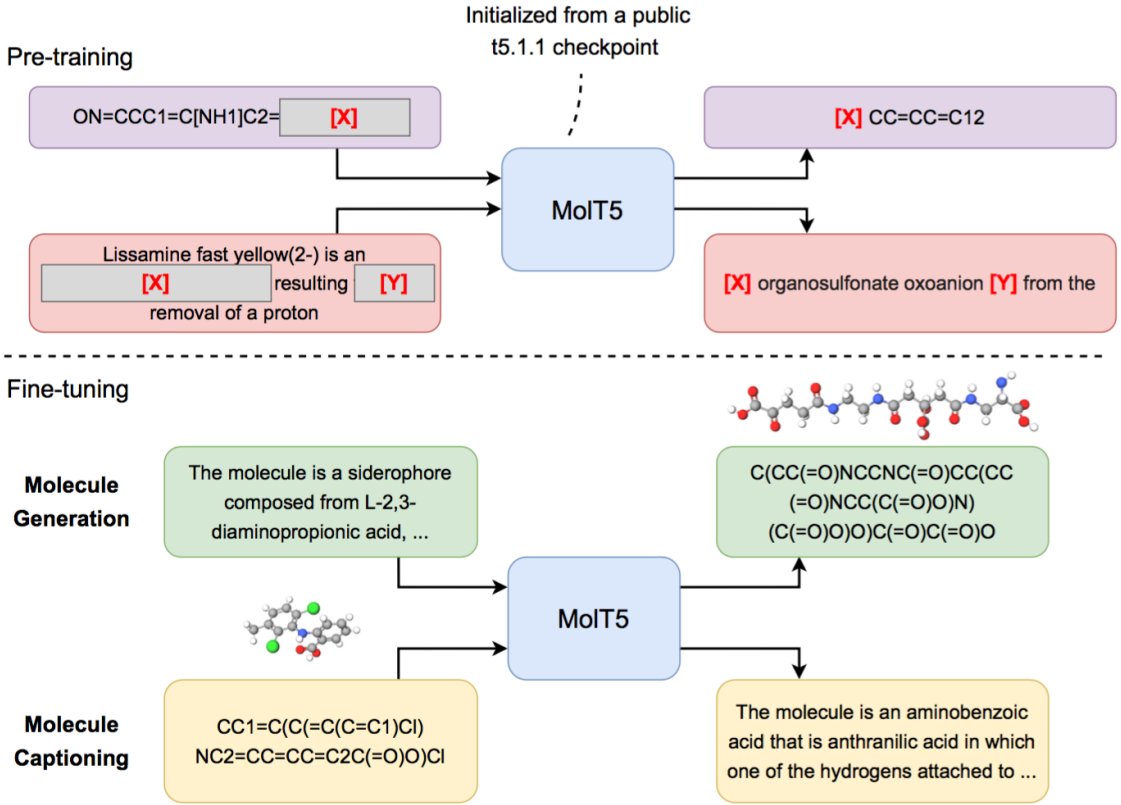


3D View

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

Now We can Enable Translation between Natural Language and Molecules [Edwards et al., EMNLP2022]



Can We Translate between Molecules and Natural Language? [Edwards et al., 2022arxiv]

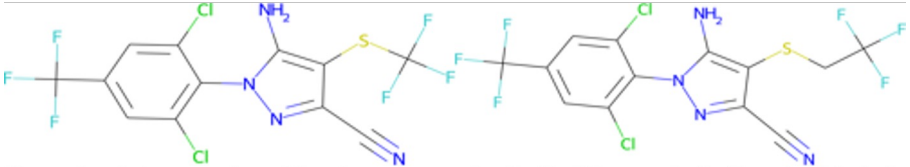
- Molecule Captioning Performance

Model	BLEU-2	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	Text2Mol
Ground Truth							0.609
RNN	0.303	0.213	0.347	0.191	0.303	0.337	0.426
Transformer	0.061	0.027	0.188	0.0597	0.165	0.126	0.0575
T5-Small	0.525	0.414	0.612	0.457	0.568	0.533	0.526
MolT5-Small	0.520	0.436	0.624	0.475	0.581	0.549	0.540
T5-Base	0.533	0.423	0.614	0.460	0.571	0.538	0.522
MolT5-Base	0.540	0.457	0.636	0.489	0.594	0.563	0.547
T5-Large	0.558	0.467	0.631	0.482	0.584	0.570	0.563
MolT5-Large	**0.594**	**0.508**	**0.650**	**0.509**	**0.605**	**0.591**	**0.582**

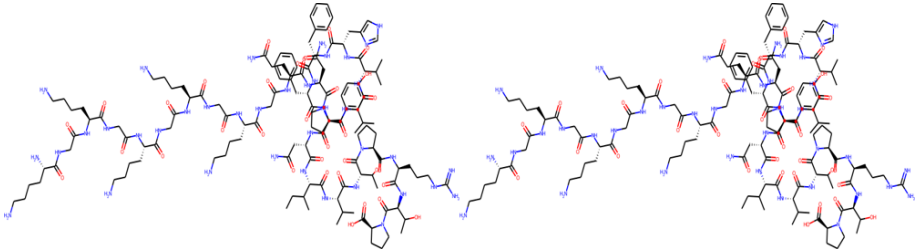
- Molecule Generation Performance

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	0.004	38.09	0.591	0.400	0.362	0.223	0.409	0.542
Transformer	0.499	0.000	57.66	0.480	0.320	0.217	0.379	0.277	**0.906**
T5-Small	0.741	0.063	27.7	0.704	0.578	0.525	0.213	0.479	0.608
MolT5-Small	0.755	0.076	25.99	0.704	0.568	0.517	0.198	0.482	0.721
T5-Base	0.762	0.067	24.95	0.731	0.605	0.545	0.177	0.499	0.66
MolT5-Base	0.769	0.080	24.46	0.721	0.588	0.529	0.185	0.496	0.772
T5-Large	0.854	0.272	16.721	0.823	0.731	0.670	0.117	0.552	0.902
MolT5-Large	**0.854**	**0.302**	**16.07**	**0.834**	**0.746**	**0.684**	**0.116**	**0.554**	0.905

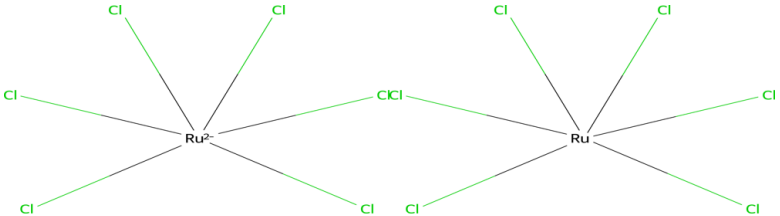
Molecule Generation Results



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.



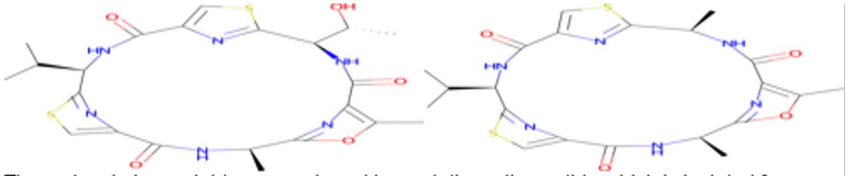
The molecule is a linear 27-membered polypeptide comprising the sequence Lys-Gly-Lys-Gly-Lys-Gly-Lys-Gly-Lys-Gly-Glu-Asn-Pro-Val-Val-His-Phe-Phe-Tyr-Asn-Ile-Val-Th Corresponds to the sequence of the myelin basic protein 83-99 (MBP83-99) immunodominant epitope with the lysyl residue at position 91 replaced by tyrosyl [MBP83-99(Y(91))] and with an (L-lysylglycyl)5 [(KG5)] linker attached to the glutamine(83) (E(83)) residue.



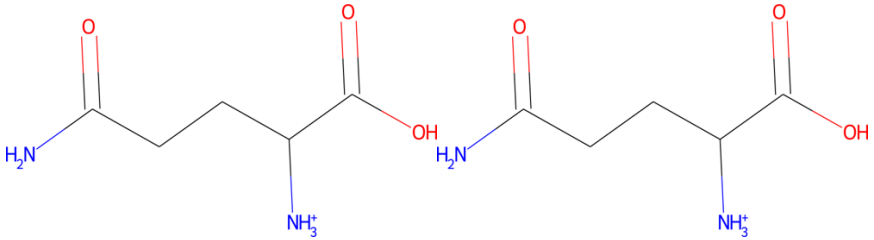
The molecule is a perchlorometallate anion having six chlorines and ruthenium(IV) as the metal component. It is a perchlorometallate anion and a ruthenium coordination entity.

Left - Ground truth, Right - Predicted

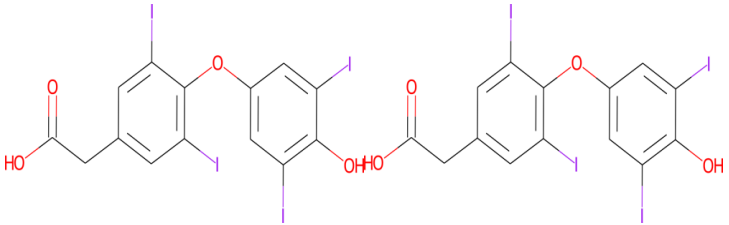
Molecule Generation Results



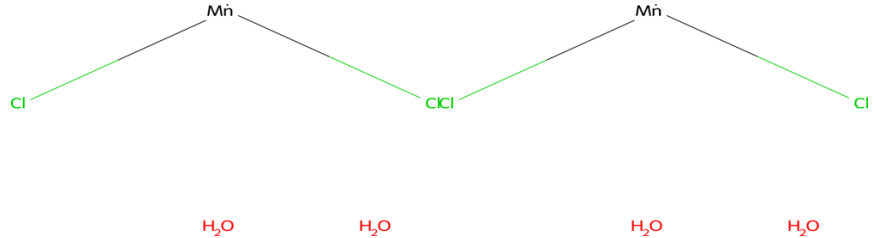
The molecule is an eighteen-membered homodetic cyclic peptide which is isolated from Oscillatoria sp. and exhibits antimalarial activity against the W2 chloroquine-resistant strain of the malarial parasite, Plasmodium falciparum. It has a role as a metabolite and an antimalarial. It is a homodetic cyclic peptide, a member of 1,3-oxazoles, a member of 1,3-thiazoles and a macrocycle.



The molecule is an alpha-amino-acid cation that is the conjugate acid of glutamine, arising from protonation of the amino group. It is a conjugate acid of a glutamine.



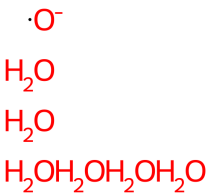
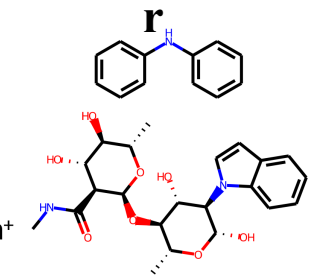
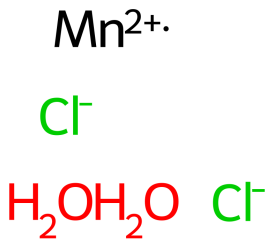
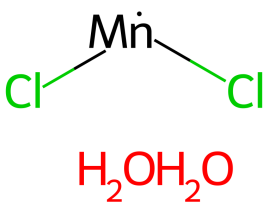
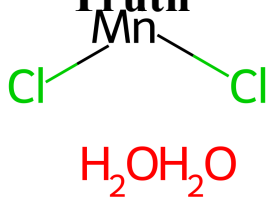
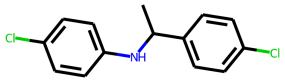
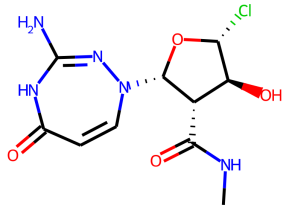
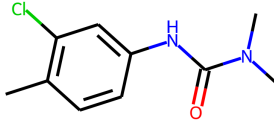
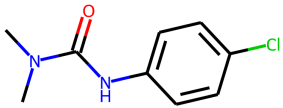
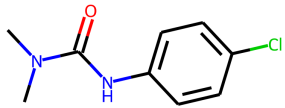
The molecule is a monocarboxylic acid that is thyroacetic acid carrying four iodo substituents at positions 3, 3', 5 and 5'. It has a role as a thyroid hormone, a human metabolite and an apoptosis inducer. It is an iodophenol, a 2-halophenol, a monocarboxylic acid and an aromatic ether.



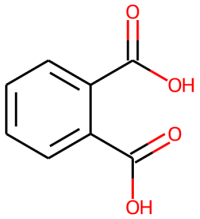
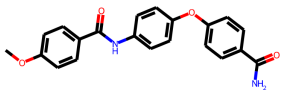
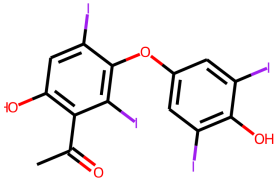
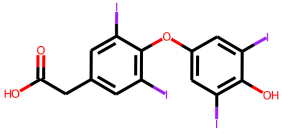
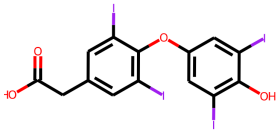
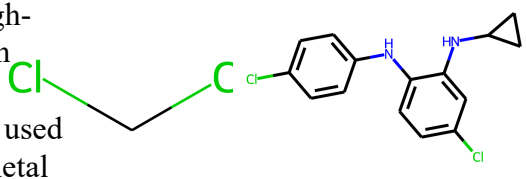
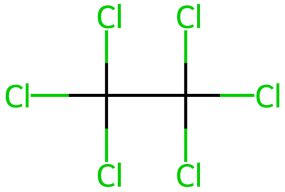
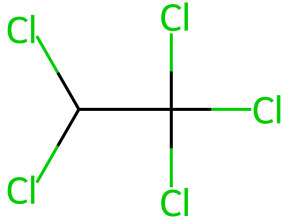
The molecule is a hydrate that is the dihydrate form of manganese(II) chloride. It has a role as a MRI contrast agent and a nutraceutical. It is a hydrate, an inorganic chloride and a manganese coordination entity. It contains a manganese(II) chloride.

Left - Ground truth, Right - Predicted

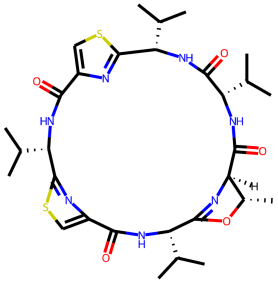
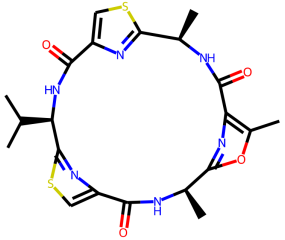
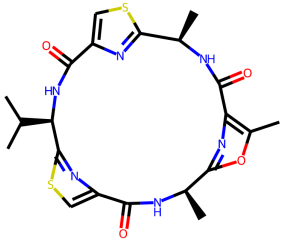
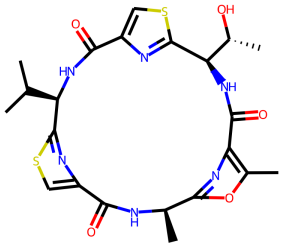
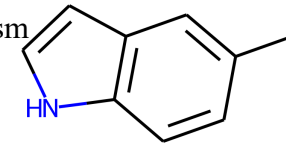
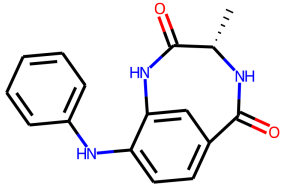
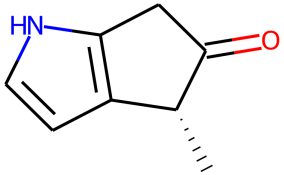
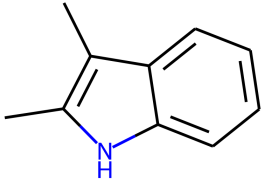
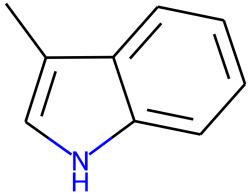
Molecule Generation Results: Different Models

Input	RNN	Transforme r	T5	MolT5	Ground Truth
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.					
The molecule is a member of the class of phhenylureas that is urea in which one of the nitrogens is substituted by a p-chlorophenyl group while the other is substituted by two methyl groups. It has a role as a herbicide, a xenobiotic and an environmental contaminant. It is a member of monochlorobenzenes and a member of phenylureas.					

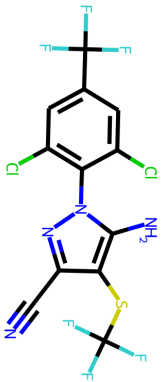
Molecule Generation Results: Different Models

Input	RNN	Transformer	T5	MolT5	Ground
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. It is an iodophenol, a 2-halophenol, a monocarboxylic acid and an aromatic ether.					
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, it is used as a solvent for oil and grease, in metal cleaning, and in the separation of coal from impurities. It has a role as a non-polar solvent.			Invalid, fixed 		

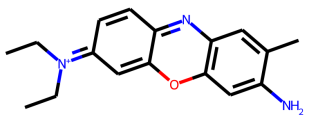
Molecule Generation Results: Different Models

Input	RNN	Transforme	T5	MolT5	Ground
The molecule is an eighteen-membered homodetic cyclic peptide which is isolated from Oscillatoria sp. and exhibits antimalarial activity against the W2 chloroquine-resistant strain of the malarial parasite, Plasmodium falciparum. It has a role as a metabolite and an antimalarial. It is a homodetic cyclic peptide, a member of 1,3-oxazoles, a member of 1,3-thiazoles and a macrocycle.	Invalid				
The molecule is a methylindole carrying a methyl substituent at position 3. It is produced during the anoxic metabolism of L-tryptophan in the mammalian digestive tract. It has a role as a mammalian metabolite and a human metabolite.					

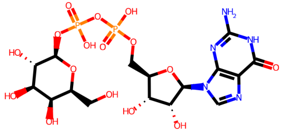
Molecule Captioning Results: Different Models

Input	RNN	Transformer	T5	MolT5	Ground Truth
	the molecule is an organofluorine compound that is 1, 2, 3, 4 - triazol - 1h - 1, 2, 4 - triazole which is substituted at positions 2, 3, and 5 by a 2, 3, 5 - triazol - 1 - yl group and at position 5 by a 2 - (trifluoromethyl) - 1, 3, 5 - triazol - 1 - yl group. it is an organofluorine compound, an organofluorine compound, an organofluorine compound, an organofluorine compound, an organofluorine compound, an organofluorine compound, an organofluorine compound, an organofluorine compound and a member of monochlorobenzenes.	the molecule is a deuterated compound that is is is is is an isotopologue of chloroform in which the four hydrogen atoms have been replaced by deuterium. it is a deuterated compound, a gamma - lactam and an aliphatic sulfide.	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	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)sulfin yl, 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	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.

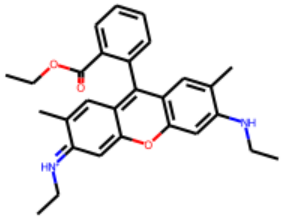
Molecule Captioning Results: Different Models

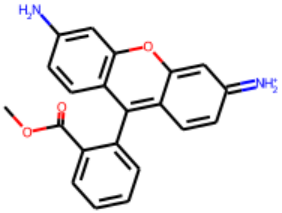
Input	RNN	Transformer	T5	MolT5	Ground Truth
	the molecule is a cationic fluorescent dye having 2, 3 - dimethyl - 1, 2, 3, 4, 6 - tetrahydro - 1h - 1, 2, 3, 4, 6 - tetrahydropyridin - 1 - yl] amino } amino group, respectively. it has a role as a fluorochrome.	the molecule is a deuterated compound that is is is is is an isotopologue of chloroform in which the four hydrogen atoms have been replaced by deuterium. it is a deuterated compound and an alpha, omega - dicarboxylic acid.	The molecule is a quaternary ammonium ion and a member of phenanthridines. It has a role as an intercalator and a fluorochrome.	The molecule is an organic cation that is phenoxazin-5-ium substituted by amino and methylamino groups at positions 3 and 7 respectively. The chloride salt is the histological dye 'azure C'.	The molecule is an organic cation that is phenoxazin-5-ium substituted by methyl, amino and diethylamino groups at positions 2, 3 and 7 respectively. The tetrachlorozincate salt salt is the histological dye 'brilliant cresyl blue'.

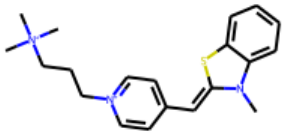
Molecule Captioning Results: Different Models

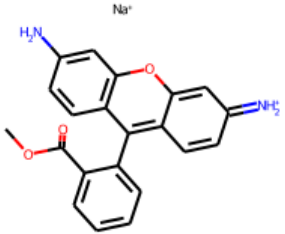
Input	RNN	Transformer	T5	MolT5	Ground Truth
	the molecule is a gdp - d - glucoside - a - […]	the molecule is the stable isotope of helium with relative atomic mass 3. 016029. the least abundant (0. 000137 atom percent) isotope of naturally occurring helium.	The molecule is a GDP-D-glucose in which the anomeric centre of the pyranose fragment has alpha-configuration. It is a GDP-D-glucose and a ribonucleoside 5'-diphosphate-alpha-D-glucose. It is a conjugate acid of a GDP-alpha-D-glucose(2-).	The molecule is a GDP-L-galactose in which the anomeric oxygen is on the same side of the fucose ring as the methyl substituent. It has a role as a plant metabolite and a mouse metabolite. It is a conjugate acid of a GDP-beta-L-galactose(2-).	The molecule is a GDP-L-galactose having beta-configuration at the anomeric centre of the L-galactose fragment. It is a conjugate acid of a GDP-beta-L-galactose(2-).

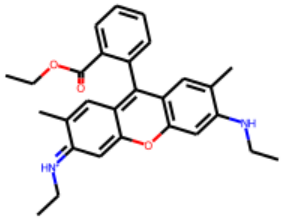
The molecule is a blue dye.

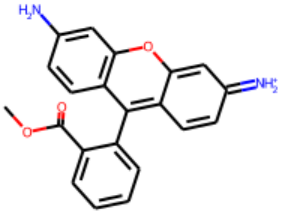


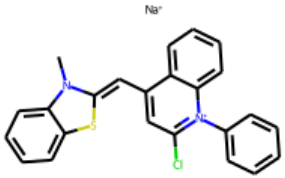




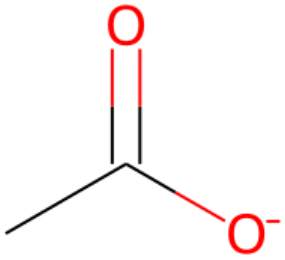


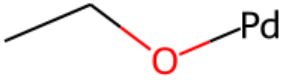


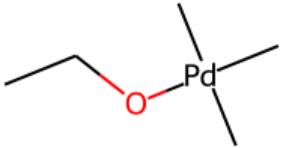


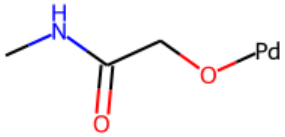


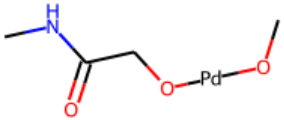
The molecule is an explosive.



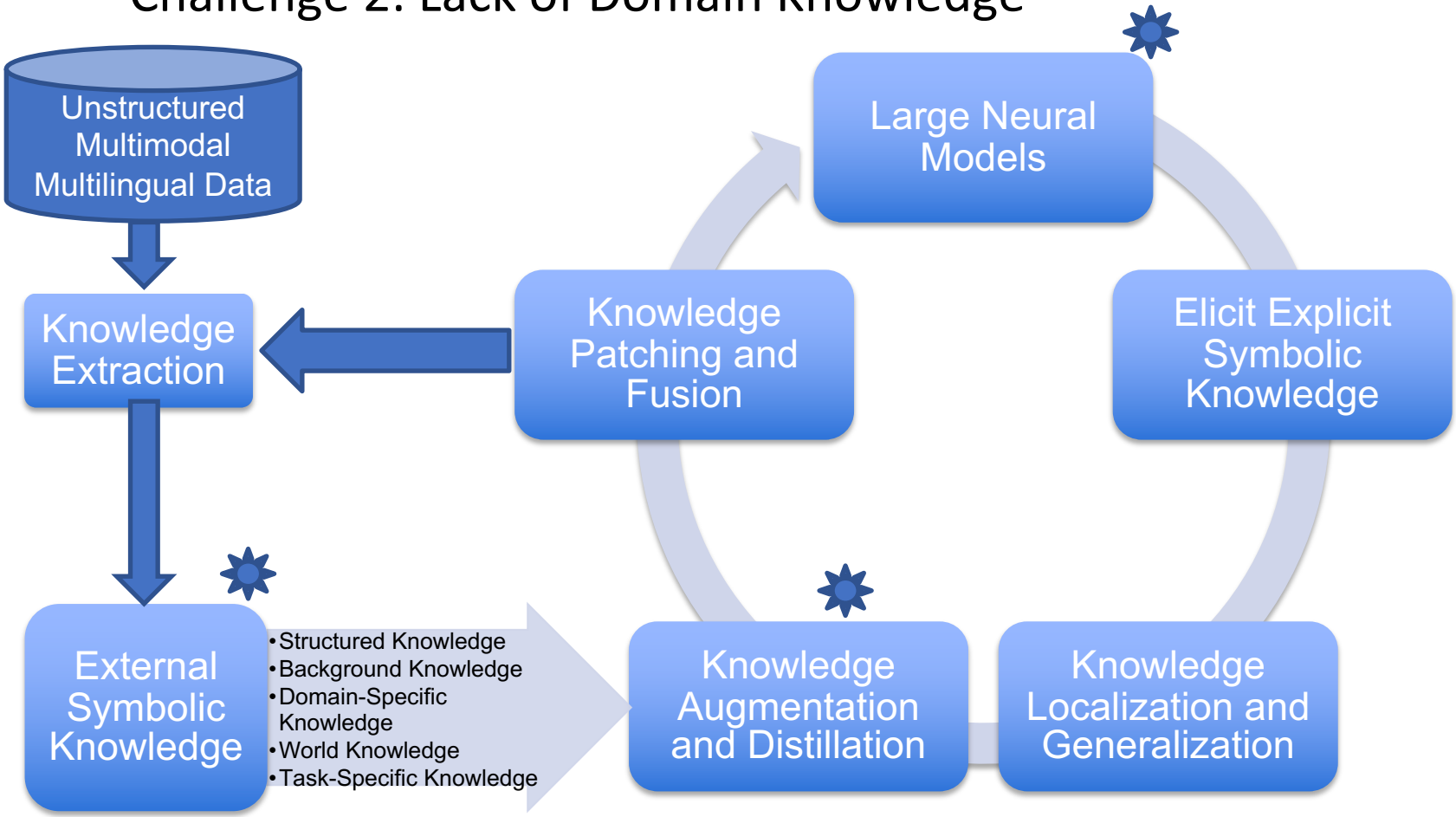






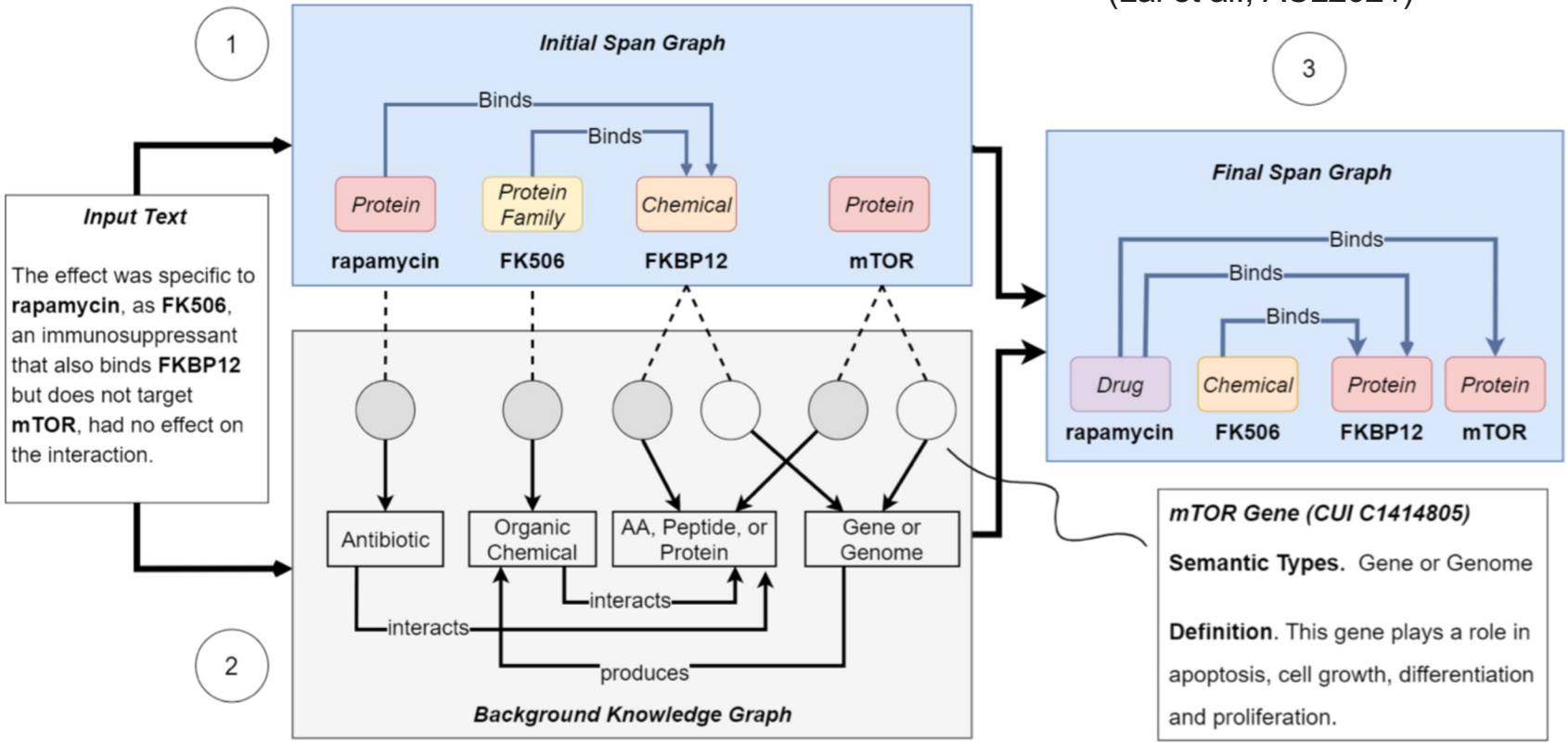


Challenge 2: Lack of Domain Knowledge

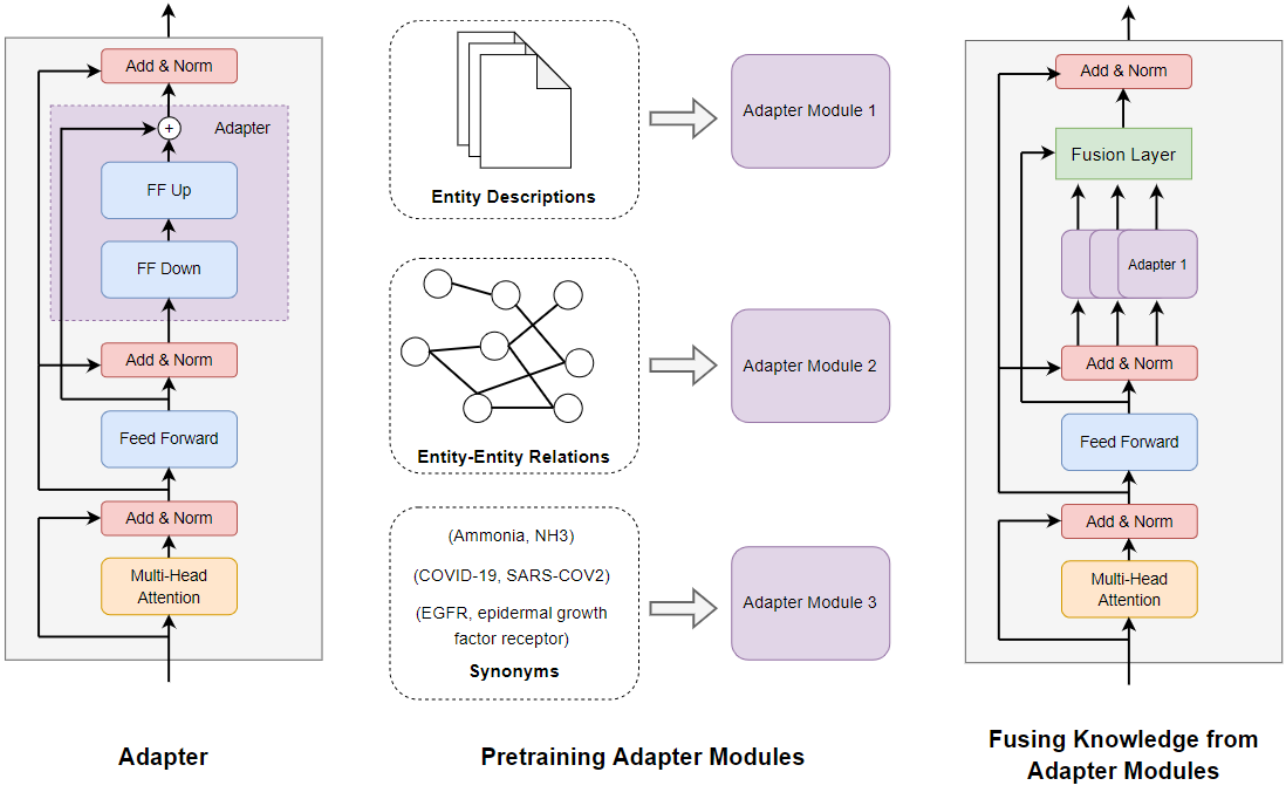


Challenge 2: Constructing an “Extra Brain” to Incorporate Domain Knowledge

- (Lai et al., ACL2021)



Knowledge-Enhanced Scientific Language Model [Lai et al., Journal of Biomedical Informatics 2023]



- (1) Encode domain knowledge using lightweight adapter modules, bottleneck feed-forward networks that are inserted into different locations of a backbone PLM
- (2) Pretrain an adapter module to capture each knowledge source in a self-supervised way
- (3) Employ fusion layers to combine the knowledge encoded within these adapters for downstream tasks

Challenge 3: Very Long Context

- Paper authors tend to write long sentences with *clauses* and *appositions* for better presentations.
- ***Example:*** *Foxp3**[Argument]** contains a proline-rich amino-terminal domain reported to function as a nuclear factor of activated T cells (NF-AT) and nuclear factor-kappaB (NF-kappaB) binding domain, a central region containing a zinc finger and leucine zipper potentially important for protein-protein interactions, and a carboxyl-terminal forkhead (FKH) domain required for nuclear localization**[Trigger]** and DNA-binding activity [14-16] .*
- Distance between event trigger and argument:

Dataset	Average Distance	Maximal Distance
ACE05-E (News)	0.212 sentence	56 words
GENIA-2011 (Papers)	0.330 sentence	77 words

LLM'S Length Generalization Problem [Han et al., 2023 "LMInfinite"]

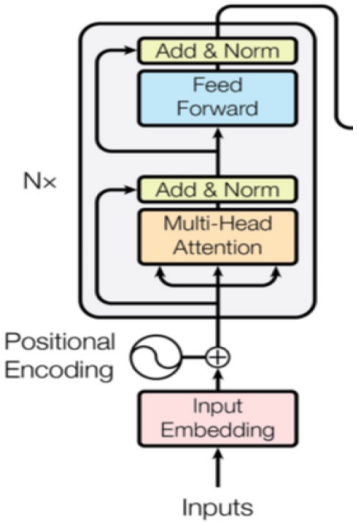
Length Generalization Failure:
After the length exceeds training lengths, LLMs start to generate nonsense texts.

Context:
Previously, Kottwitz [@Kottwitz92] proved the formula (\[eq_intro:Kottwitz_formula\]) in PEL-type cases (of simple Lie type A or C) by a method which is based on the Honda-Tate theory. This method however cannot be applied in general Hodge-type situations, and indeed:

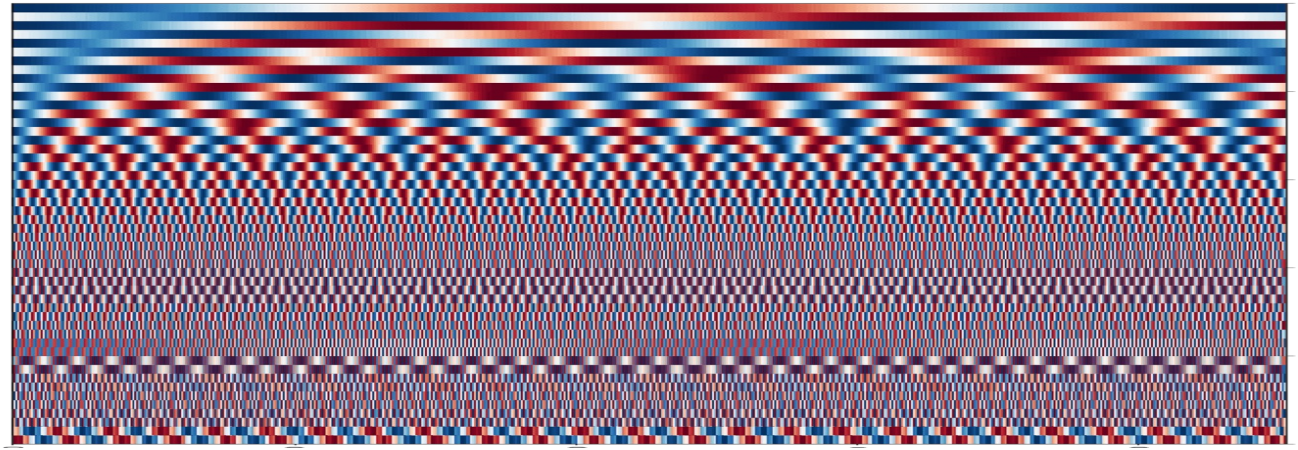
Generation:
*of this (of over-equary-her, and [(and, in the...cister '-- and an of the model to\n↵by. by ..., this, by the. It, and it, 7. --(of an equist (of the.\nand to the [[[WNE (to. and for the (((de in the (for the andistile–c.\n-[de (for in an inc ort, ort (betness in >with (with, based (and (>~~such ((c of a or for the abstract as. of **.

Traditionally: Absolute Position Embeddings

Absolute Position Embeddings:
Adding position-specific vectors onto word embeddings.
Popular in vanilla Transformers, they are not extendable to unseen positions.



Transformer structure

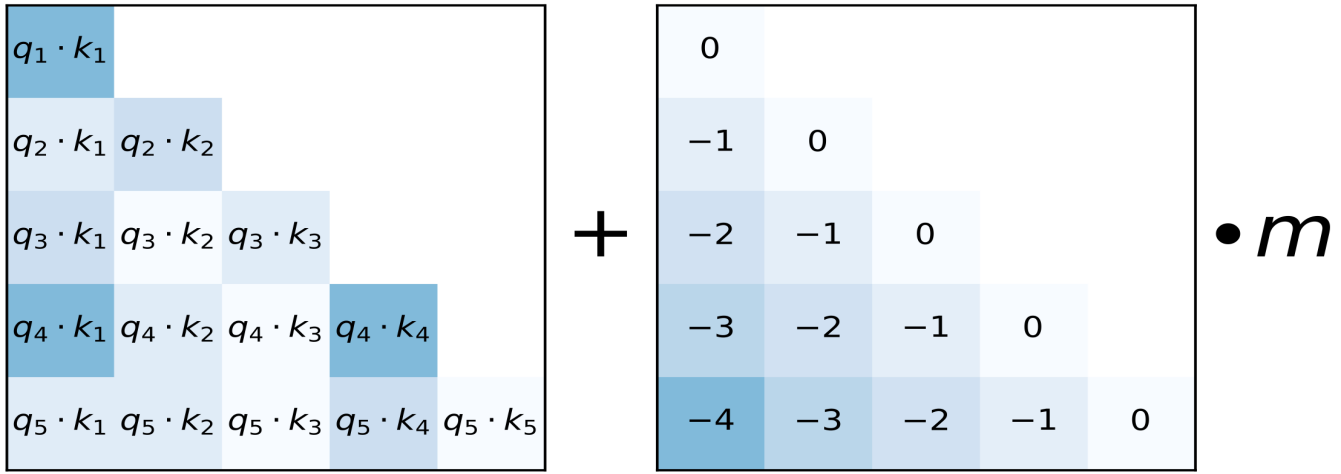


One popular option: sinusoidal position embeddings

https://erdem.pl/2021/05/understanding-positional-encoding-in-transformers

Recently: Relative Position Embeddings

Relative Position Embeddings:
Core idea: determining attentions based on inter-token distance.
Proposed in hope to generalize to unseen lengths.



Representative work: **Alibi**. It adds a linear attention decay (right) onto original attention scores (left)

Press, Ofir, Noah A. Smith, and Mike Lewis. "Train short, test long: Attention with linear biases enables input length extrapolation." arXiv preprint arXiv:2108.12409 (2021).

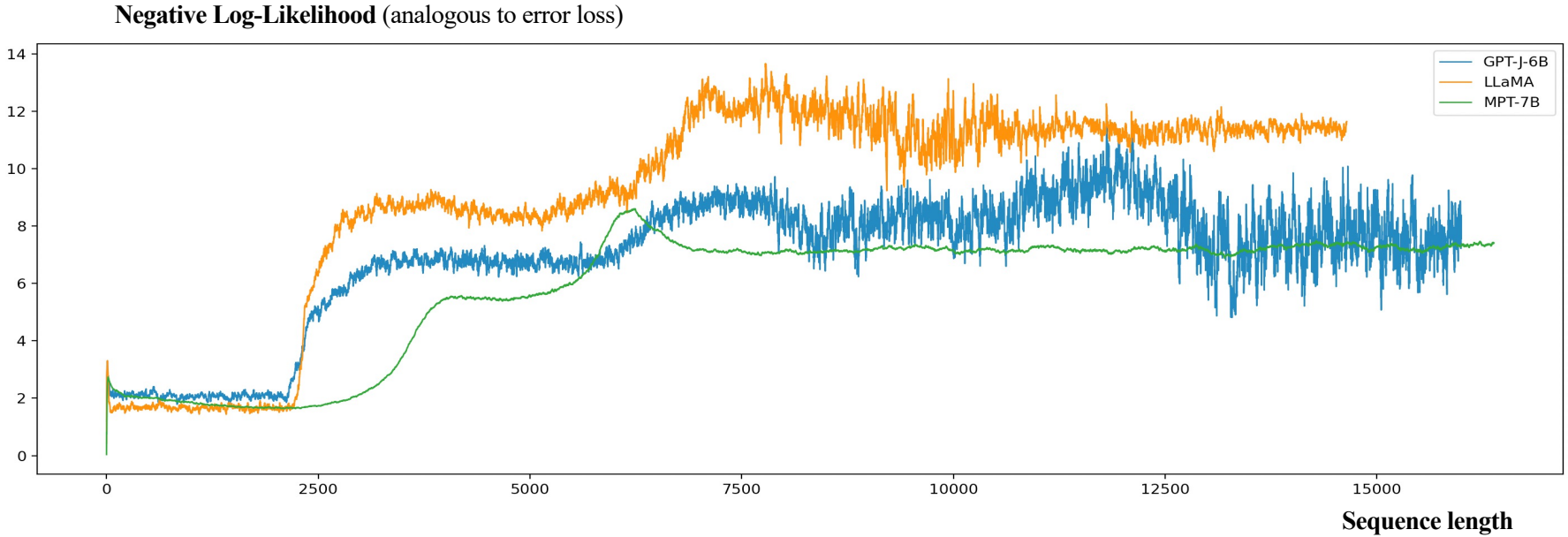
Recently: Relative Position Embeddings

However, length generalization failures are still observed!

Factor 1: Unseen Distance

Factor 2: Too Many Tokens Under Attention

Factor 3: Messing Up Implicitly Encoded Positions



On-the-Fly Length Generalization: LM-Infinite

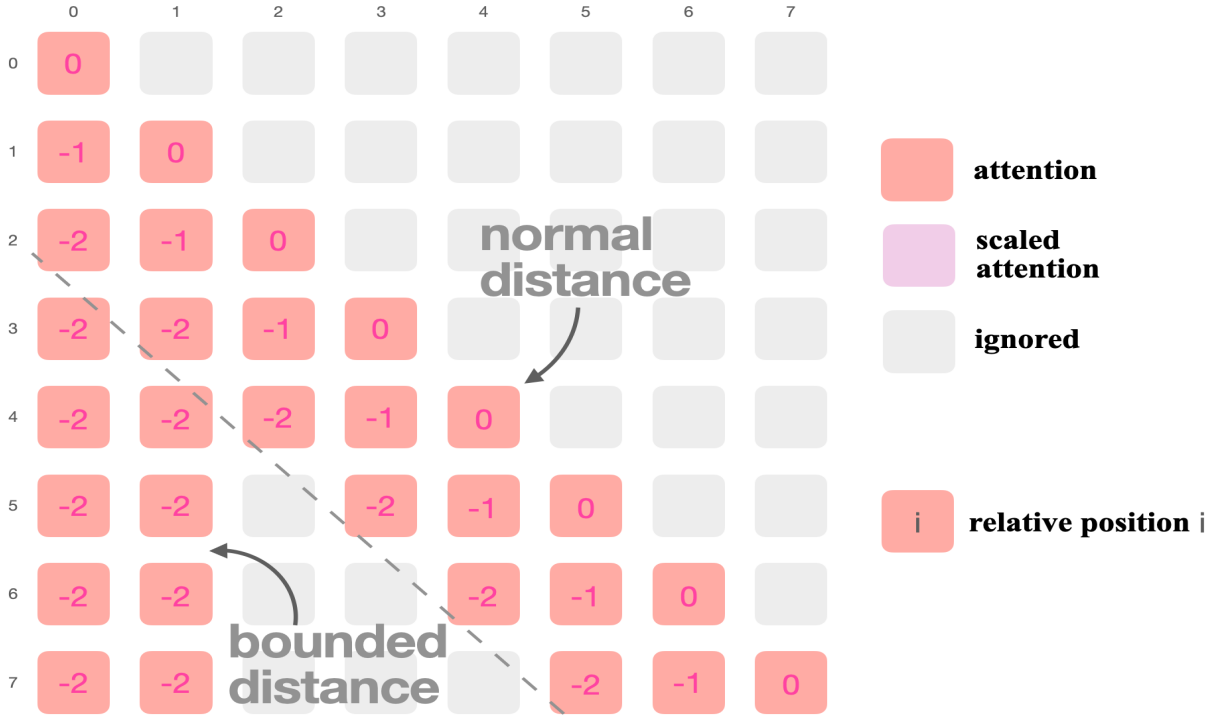
Advantages:

Plug-and-play, no fine-tuning needed.

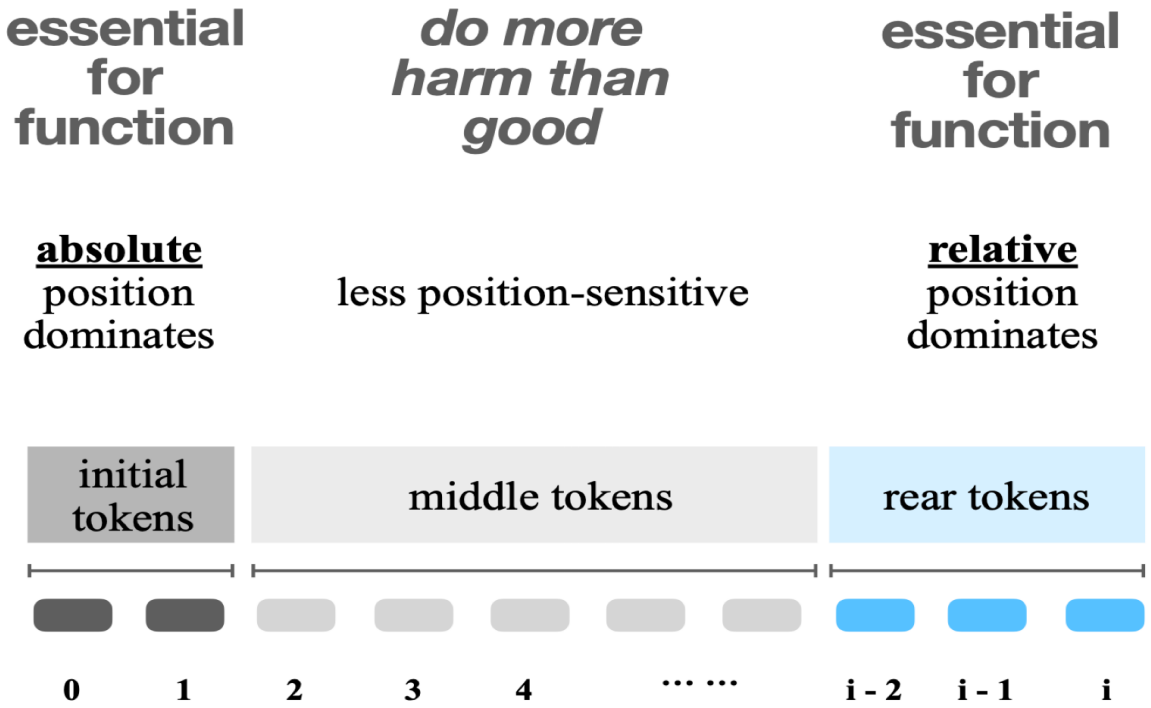
Easy to implement

Compatible with various LLMs.

O(n) efficiency.

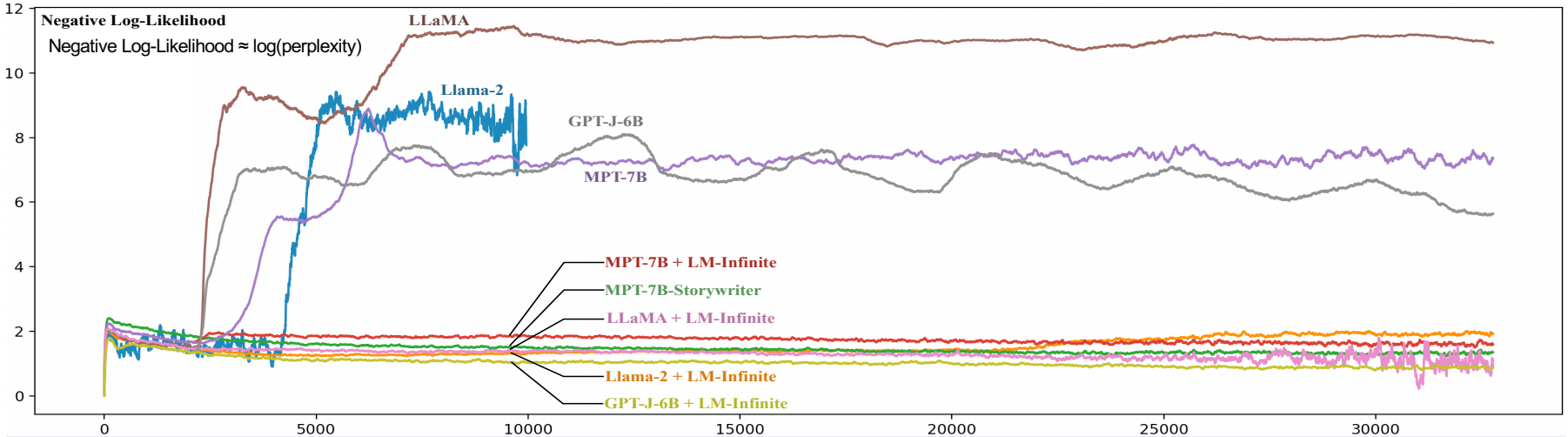


Conceptually Understanding Positions in LLMs



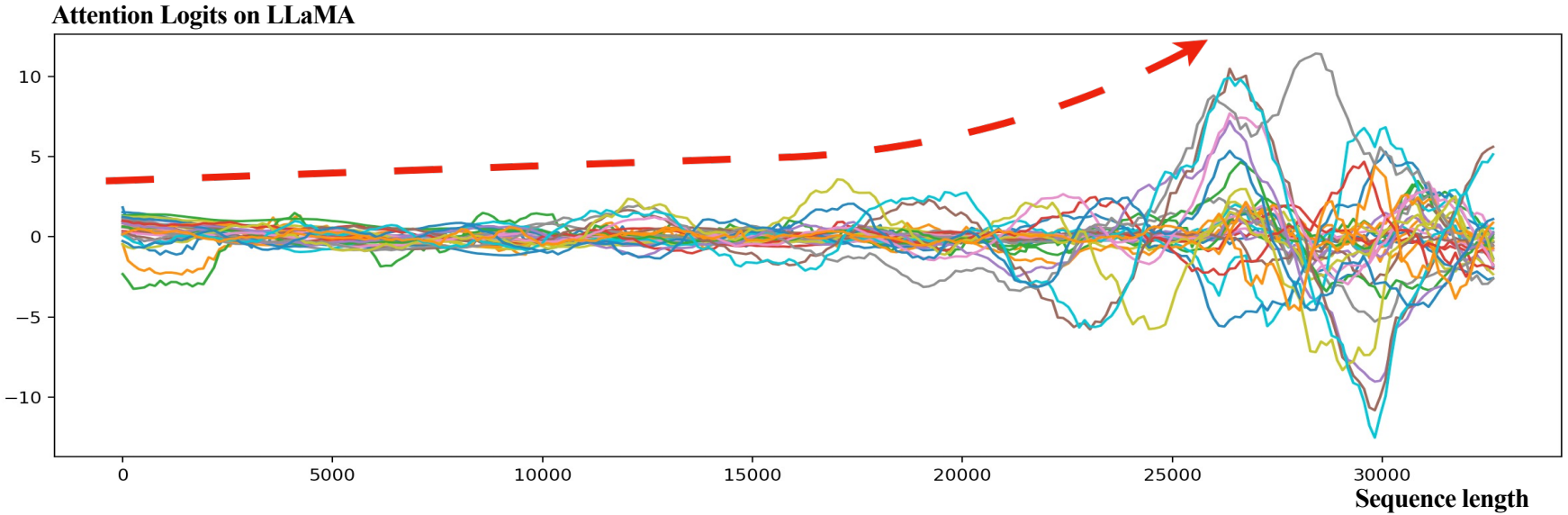
Evaluation: Perplexity

LM-Infinite flattens the perplexity curves of various LLMs.



What are “unfamiliarity” (out-of-distribution) factors to LLMs?

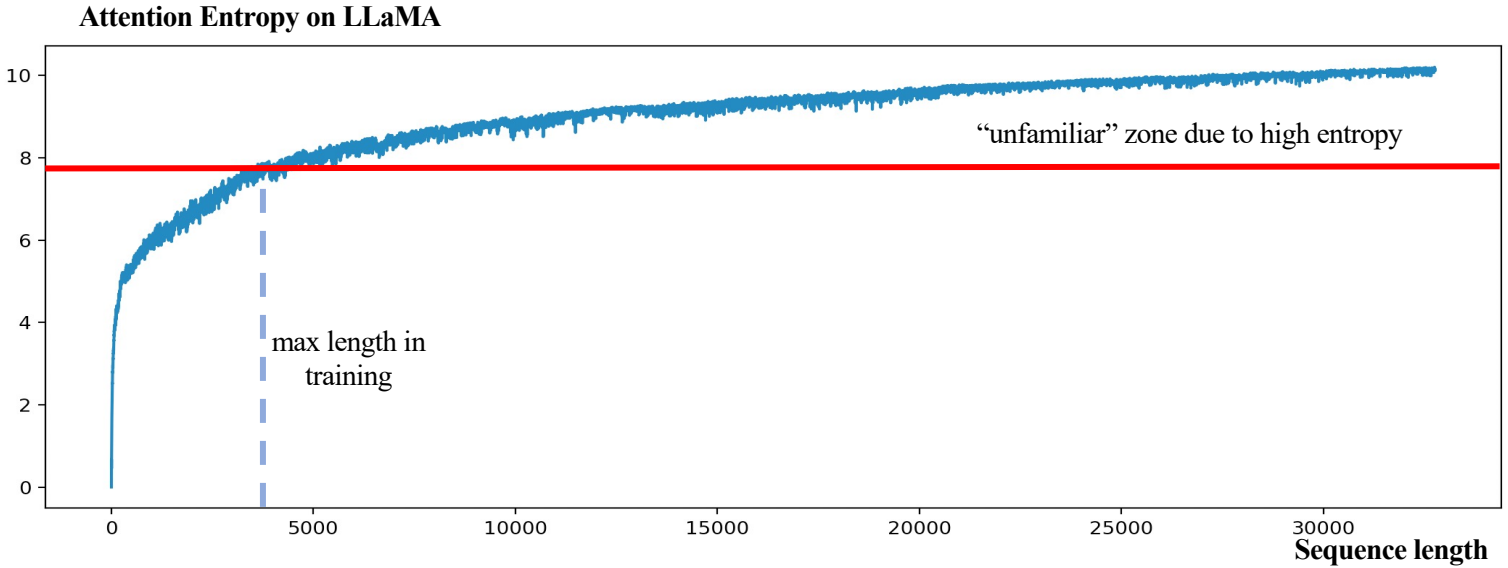
Factor 1: Unseen Distance



Theoretically, when sequence increases, **attention logits tend to “explode”** to unseen magnitude.

What are “unfamiliarity” (out-of-distribution) factors to LLMs?

Factor 2: Too Many Tokens Under Attention



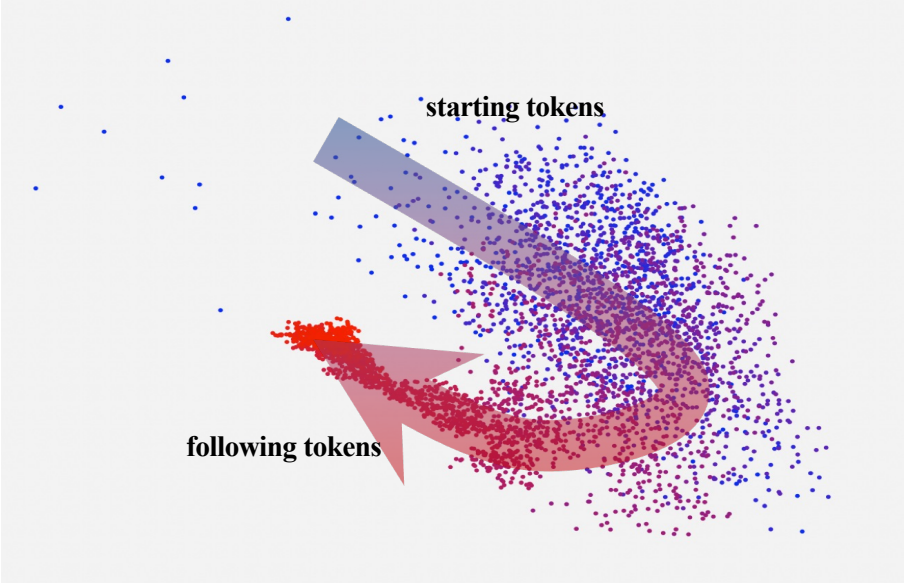
Theoretically, attention on too many tokens will enlarge its entropy, **making itself increasingly chaotic**.

What are “unfamiliarity” (out-of-distribution) factors to LLMs?

Factor 3: Messing Up Implicitly Encoded Positions

Theoretically, absolute position information might be implicitly encoded [Kazemnejad et al, 2023].

If the starting tokens are processed by unseen-distance attention functions, it can mess up the sub-space.



LLaMA features implicitly encode positions

Kazemnejad, Amirhossein, et al. "The Impact of Positional Encoding on Length Generalization in Transformers." arXiv preprint arXiv:2305.19466 (2023).

Evaluation: Generation Quality

LM-Infinite lets LLMs generate texts with higher quality after long context, superior or similar to their finetuned counterpart MPT-7B-SW.

	2k		4k		8k		16k		32k	
ArXiv	bleu	rouge	bleu	rouge	bleu	rouge	bleu	rouge	bleu	rouge
MPT-7B-SW	16.6	26.5	21.5	30.1	15.2	26.6	18.9	27.4	14.8	**27.0**
MPT-7B	0.0	5.6	0.2	3.6	0.0	5.9	0.0	1.7	0.4	1.4
MPT-7B + LM-Infinite	16.1	23.8	20.2	24.9	12.6	24.1	23.9	29.0	**19.7**	26.6
Llama-2	26.6	31.4	0	0.2	0.0	0.0	0.0	0.0	0.0	0.0
Llama-2 + LM-Infinite	**26.9**	**31.8**	**23.6**	**30.9**	**23.9**	**28.2**	**24.8**	**29.2**	18.4	20.4
OpenWebText2	bleu	rouge	bleu	rouge	bleu	rouge	bleu	rouge		
MPT-7B-SW	8.4	21.0	6.1	19.3	7.5	18.5	8.4	**22.0**		
MPT-7B	0.9	7.5	0.9	6.6	1.0	6.4	1.0	6.8		
MPT-7B + LM-Infinite	5.0	16.6	4.1	15.4	5.1	16.2	2.8	16.0		
Llama-2	8.8	**22.4**	0.0	0.2	0.0	0.0	0.0	0.0		
Llama-2 + LM-Infinite	**9.0**	21.9	**7.2**	**21.2**	**9.7**	**19.6**	**9.6**	19.6		

Generation Comparison Examples

ArXiv Dataset (preprints)

Vanilla LLaMA:
*of this (of over-equary-her, and [(and, in the...cister '-- and an of the model to\n↵by. by ..., this, by the. It, and it, 7. --(of an equist (of the.\nand to the [[[WNE (to. and for the (((de in the (for the andistile–c.\n-[de (for in an inc ort, ort (betness in >with (with, based (and (>~~such ((c of a or for the abstract as. of *.*

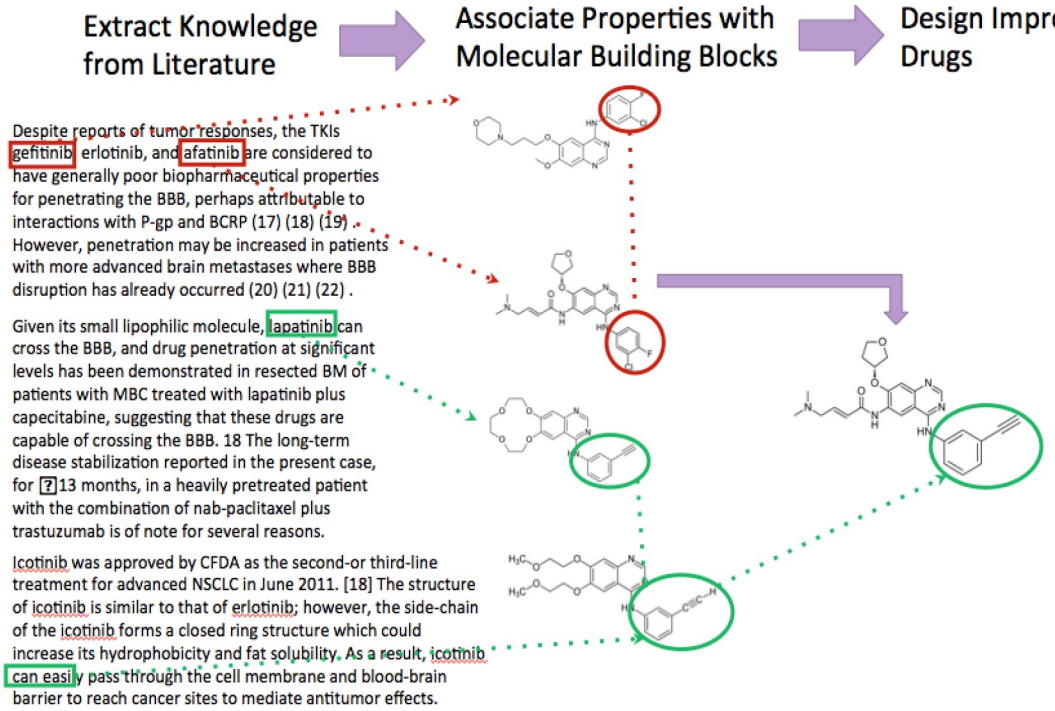
LLaMA + LM-Infinite:
Kottwitz [@Kottwitz84a] had formulated the conjecture (\\[eq_intro:Kottwitz_formula\\]), but at the time no idea had yet been established on how to prove it.
This conjecture (\\[eq_intro:Kottwitz_formula\\]) involves some algebraic statements of automorphic forms which have an algebraic nature, but are only valid for automorphic forms (which are not holomorphic automorphic forms [@Kaplansky1966]), whose (abstract) objects are Galois extensions of F and $\\bar{F}$.

OpenWebText2 Dataset (Reddit posts)

Vanilla LLaMA:
the in that, of it and and \', " of the. of " of " of the pr of of of the of the of the [\n. of “ – s (1 (of s of the " of quends of it &nd of trail of the, the &, 1. – a and at-and and and the " for on the is, core by " (as (:: a. -\n. (on its to “ is a, (on, th is of all but". ".. in the, in the the the and in the for the.

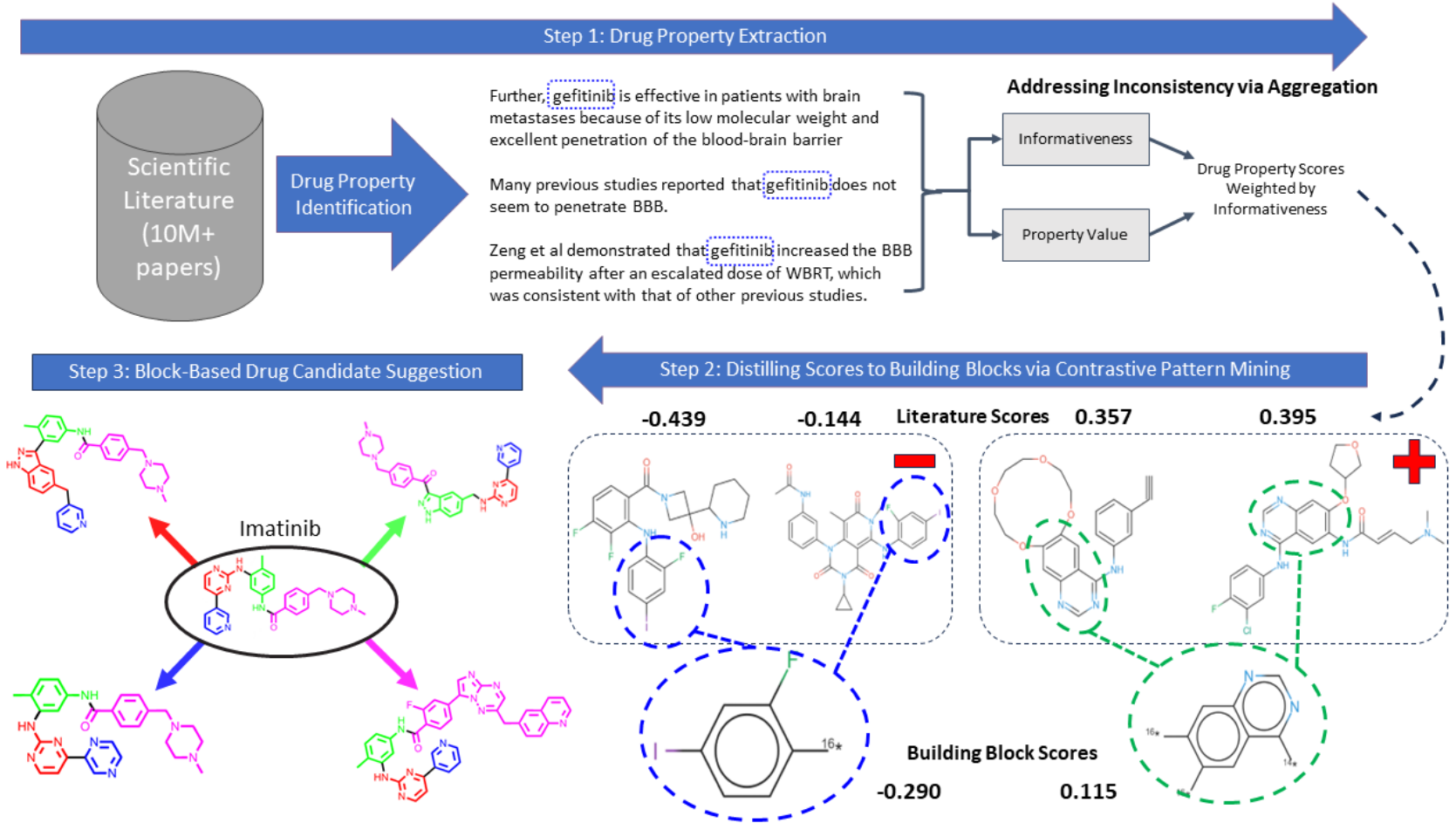
LLaMA + LM-Infinite:
It develops gradually. For Piaget, the cognitive originates in experience as it is subjected to structure : i.e. as it is structured, ordered and transformed. the first stage, the "sensorial stage", structures are formed and integrated based on perception. These structures are built by organic, biological and psychic transformations, which lead to experience.

Joint Molecule and Language Modeling to Discover Building Blocks for Improved Drug Design

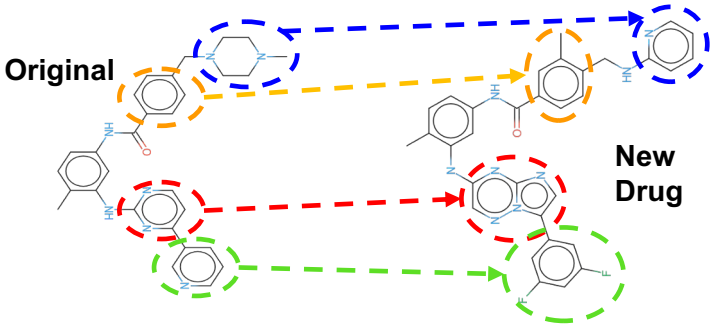


1. Finetune LLM on a few annotated sentences to predict informativeness and BBBP scores of 5000+ sentences.
2. Aggregate scores across all papers because knowledge reported in literature is inconsistent.
3. Apply a graph frequent pattern mining algorithm to identify structures which contribute to BBBP and those which do not.
4. Propose building block replacement based on substructure scores, bond type, docking score, and molecule weight.

Put Everything Together

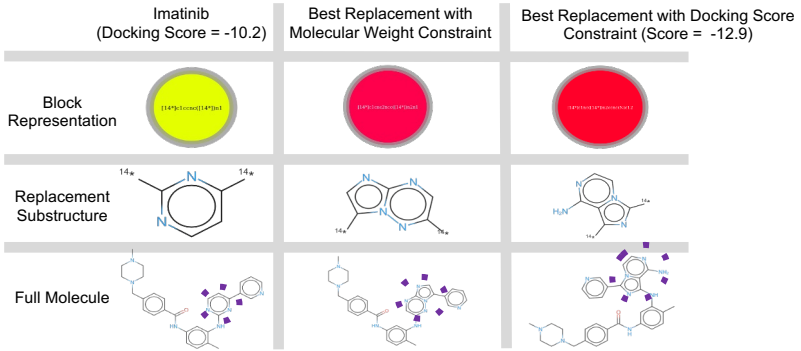


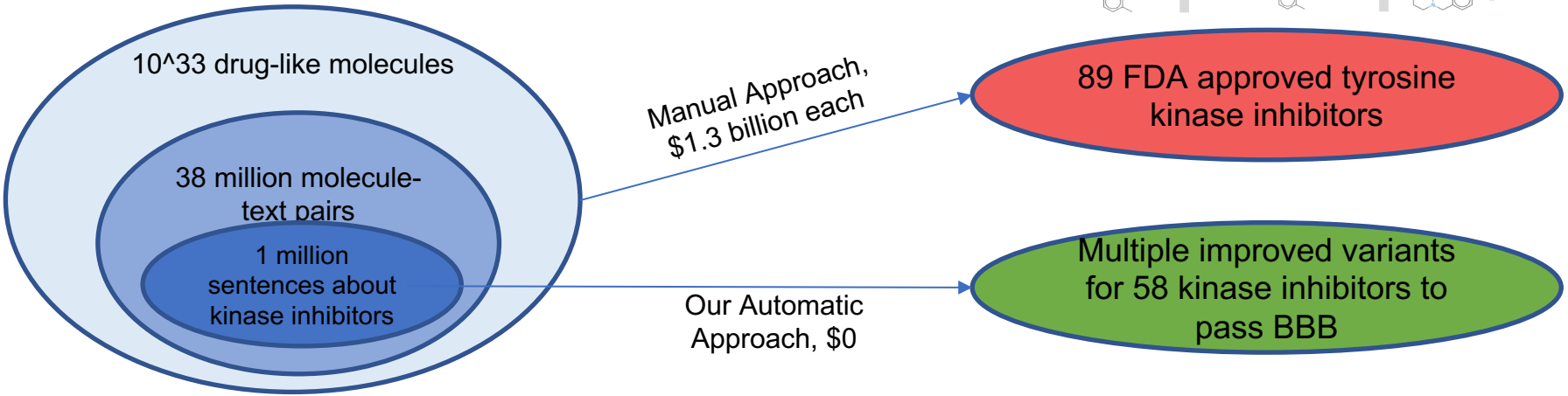
Joint Molecule and Language Modeling to Discover Building Blocks for Improved Drug Design



$$S_b^w = \frac{D_{M(b)}}{D_I} S_b$$

Imatinib Replacement

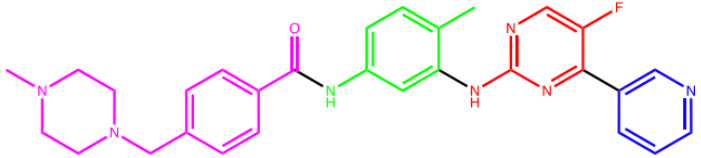




- BBBP Scores Improved Positive molecules from 83.7% to 93.6%, and Negative molecules from 28.9% to 64.0%; currently conducting physical validation (animal screening on mice) on the new improved molecules

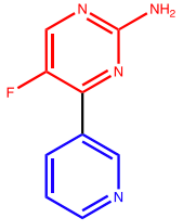
That’s not the end of the story: gap between theory and practice

Red Block Derivative Projected to Pass BBB

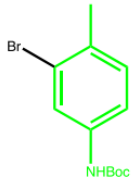


0.6798

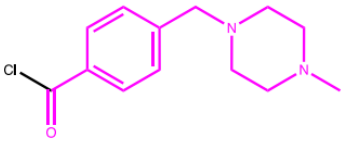
(1)



$1509/ 1 g
eMolcules
Chemspace

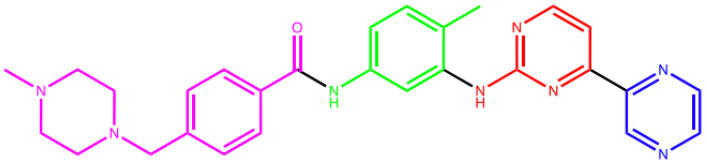


$16/ 1 g
Combi-Blocks



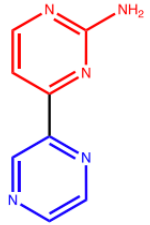
Already have

Blue Block Derivative Projected to Pass BBB

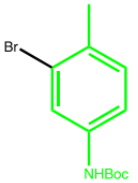


0.6597

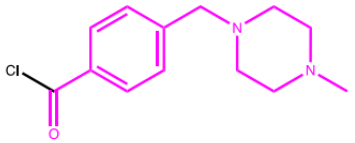
(1)



$363/ 1 g
AA Blocks
USA



$16/ 1 g
Combi-Blocks



Already have

Conclusions and What We Need

- We need a new conference on AI for Science and Science for AI so two communities really marry each other
- Near Future: AI Automated Lab – AI advisor, AI researcher, AI coordinator and AI technician work together with human scientists
- Limitations of existing scientific large language models (e.g., Galatica)
  - Used poor-quality data (publicly available arxiv paper abstracts instead of Nature/Science papers)
  - Design issues: consider structured knowledge bases as short tuple sentences, poor knowledge representation
- For Computer Scientists: More open-minded to close collaboration with researchers from other fields
- For Chemists: please share exciting problems and datasets with us!

Publicly Available Demos, Systems and Resources

- Biomedical Information Extraction System
  - https://github.com/zhangzx-uiuc/Knowledge-AMR
  - https://github.com/laituan245/bio_relex
- COVID Knowledge Graph
  - http://blender.cs.illinois.edu/covid19/ download: 20K+
- ClaimRadar for COVID19: https://blenderdemo.com/covid-list
- Real Time Claim Extraction for COVID19: https://blenderdemo.com/covid-extract
- ClaimRadar source code: https://github.com/uiucnlp/covid-claim-radar
- ClaimRadar docker: https://hub.docker.com/repository/docker/blendernlp/covid-claim-radar
- ClaimRadar demo video: http://blender.cs.illinois.edu/aida/covid_claim_radar.mp4