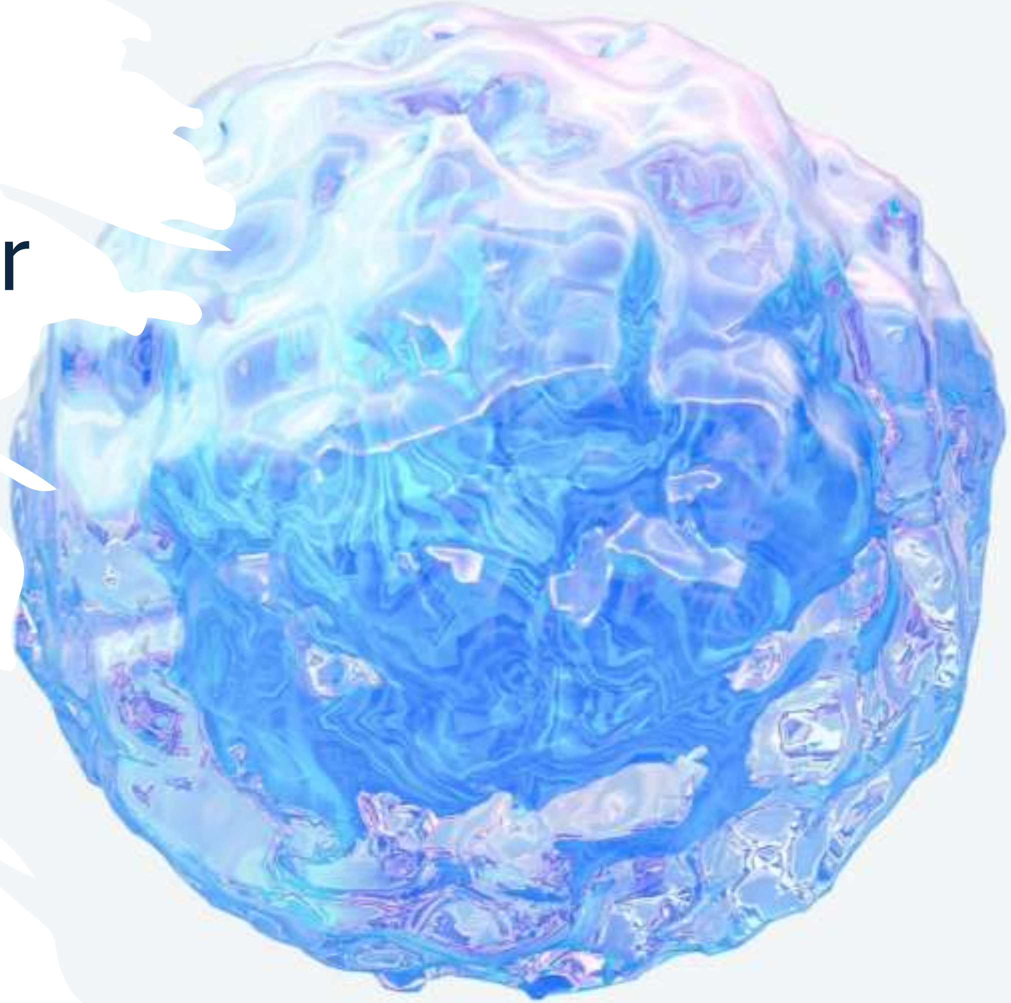
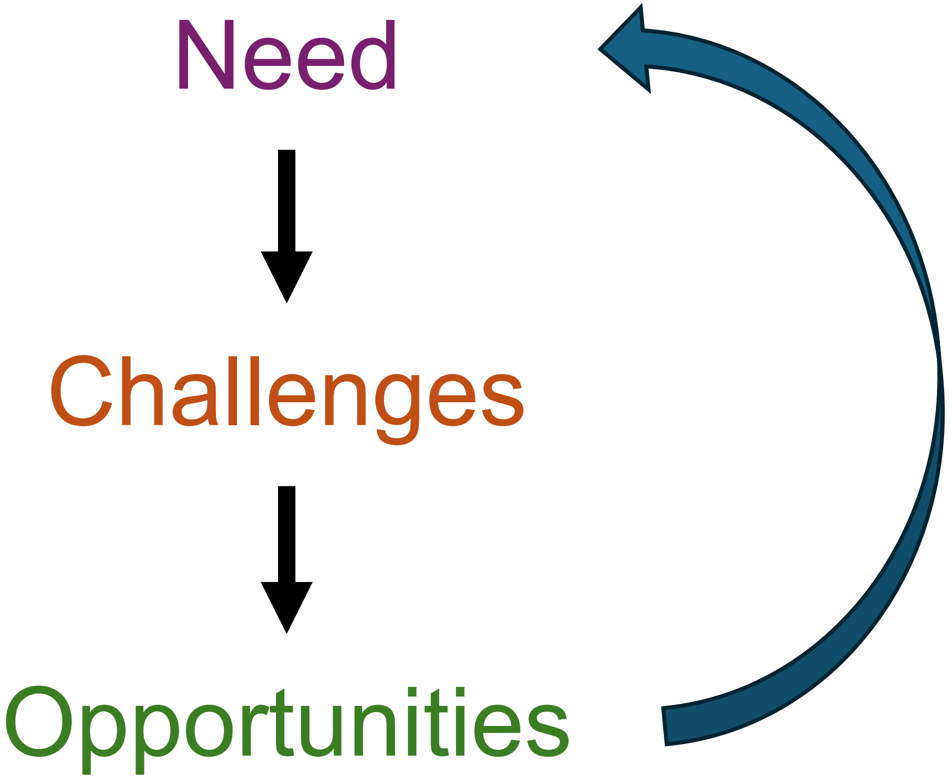
WISJ Machine Learning Summer School 2025

LLMs and Foundation Models for Material Science

Indra Priyadarsini, Ph.D.
Lisa Hamada, Ph.D.

Research Scientist
IBM Research - Tokyo





Need

Challenges

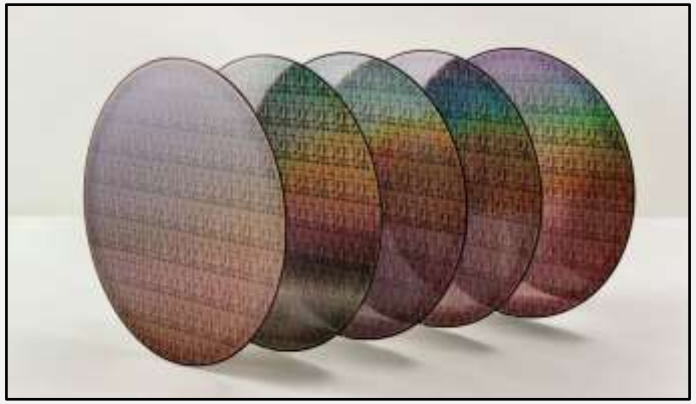
Opportunities

New materials are potentially high impact

but discovery is very difficult …

Energy and climate solutions

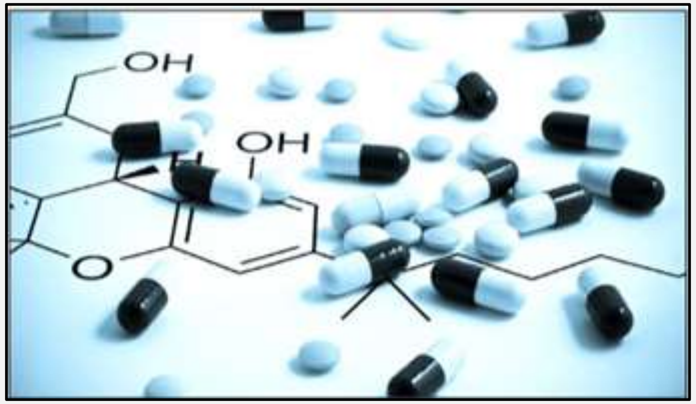
Greener feedstocks



Sustainable semiconductor processing

Clean water & food supplies

Plastics recycling



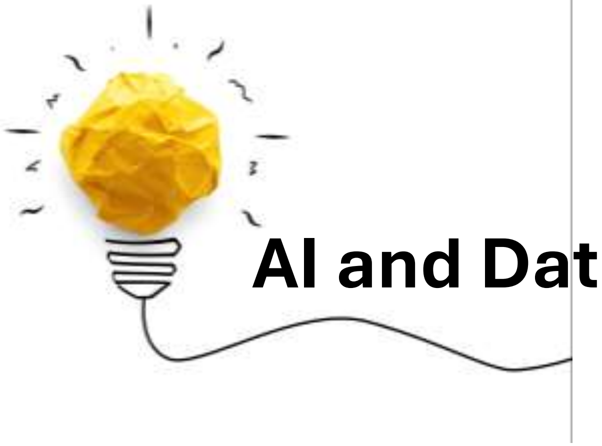
Novel therapies

Finding a good suitable material is like finding a needle in a haystack

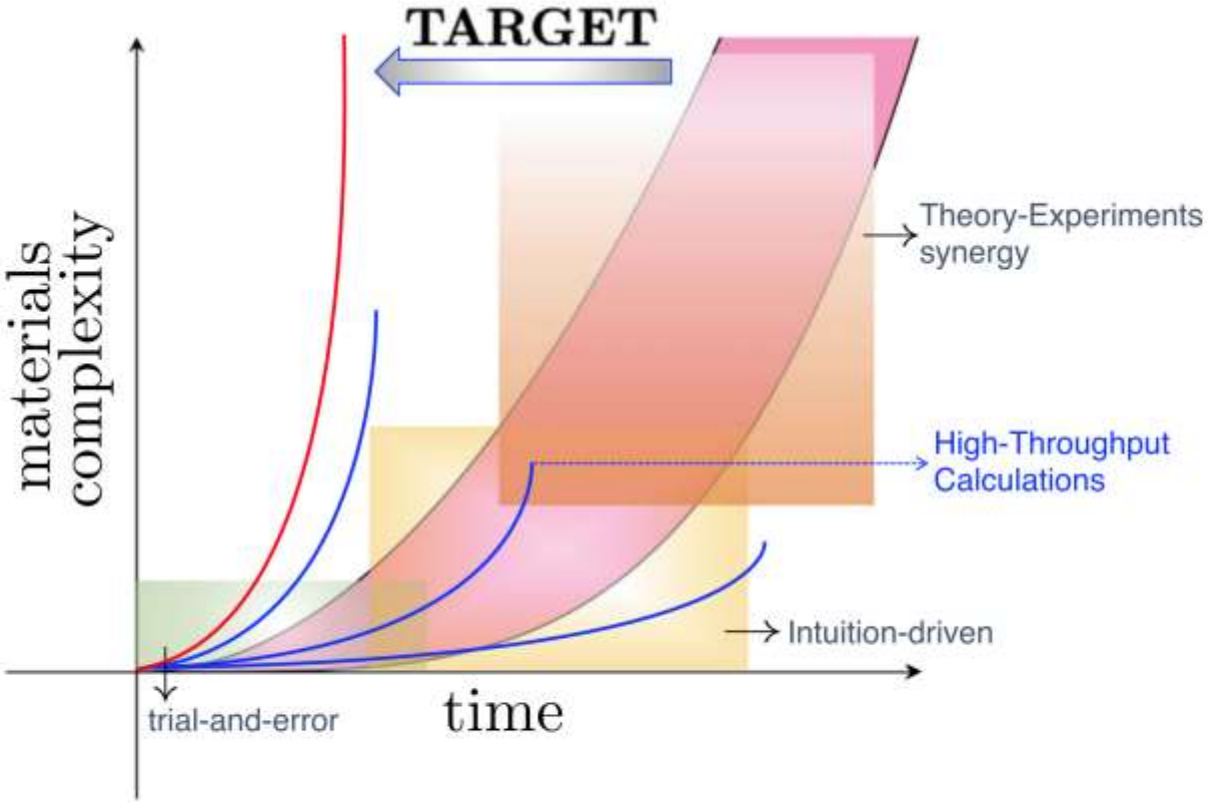
How do we guide experiments towards materials with desired properties?

Given desired material properties,

10^{60} possible candidates!



AI and Data-driven Modelling



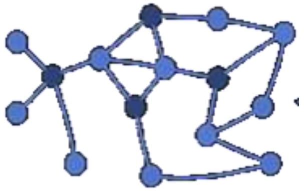
T. Lookman, et al. *npj Computational Materials* 5.1 (2019): 21.

Accelerated Discovery – Discovery Technology Foundations

Accelerating scientific workflows with AI foundation models, toolkits and discovery platform services

Deep Search

Deep Search Toolkit for Scientific Discovery

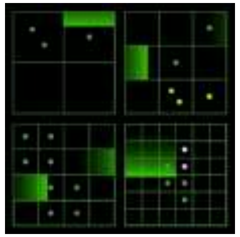


Collection and integration of knowledge

ds4sd.github.io

Simulation (ST4SD)

Simulation Toolkit for Scientific Discovery

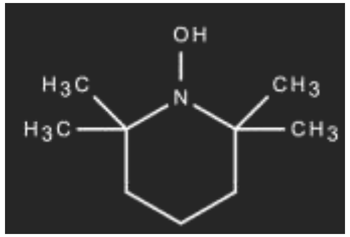


Acceleration of simulation

github.com/st4sd

Generative (GT4SD)

Generative Toolkit for Scientific Discovery



Generation of new chemicals

github.com/gt4sd

Lab Automation (RXN)

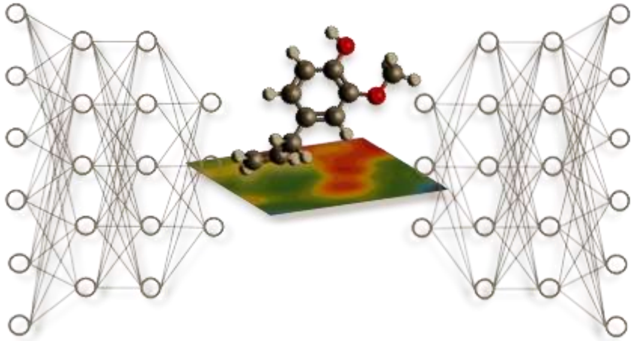
AI for Chemistry

Automated synthesis

rxn.app.accelerate.science

Foundation Model for Materials

Foundation Model for materials, chemistry, biology



Global Presence

3000

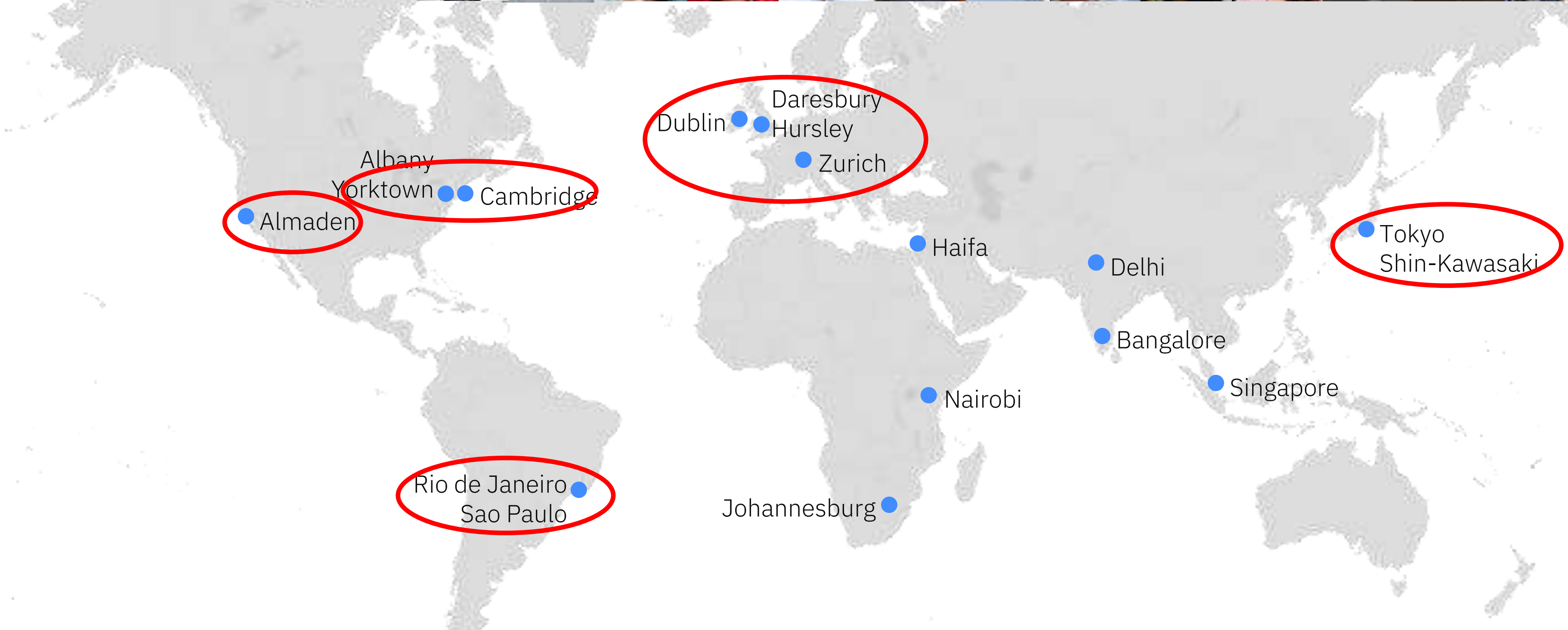
Researchers

79

Years

18

Sites



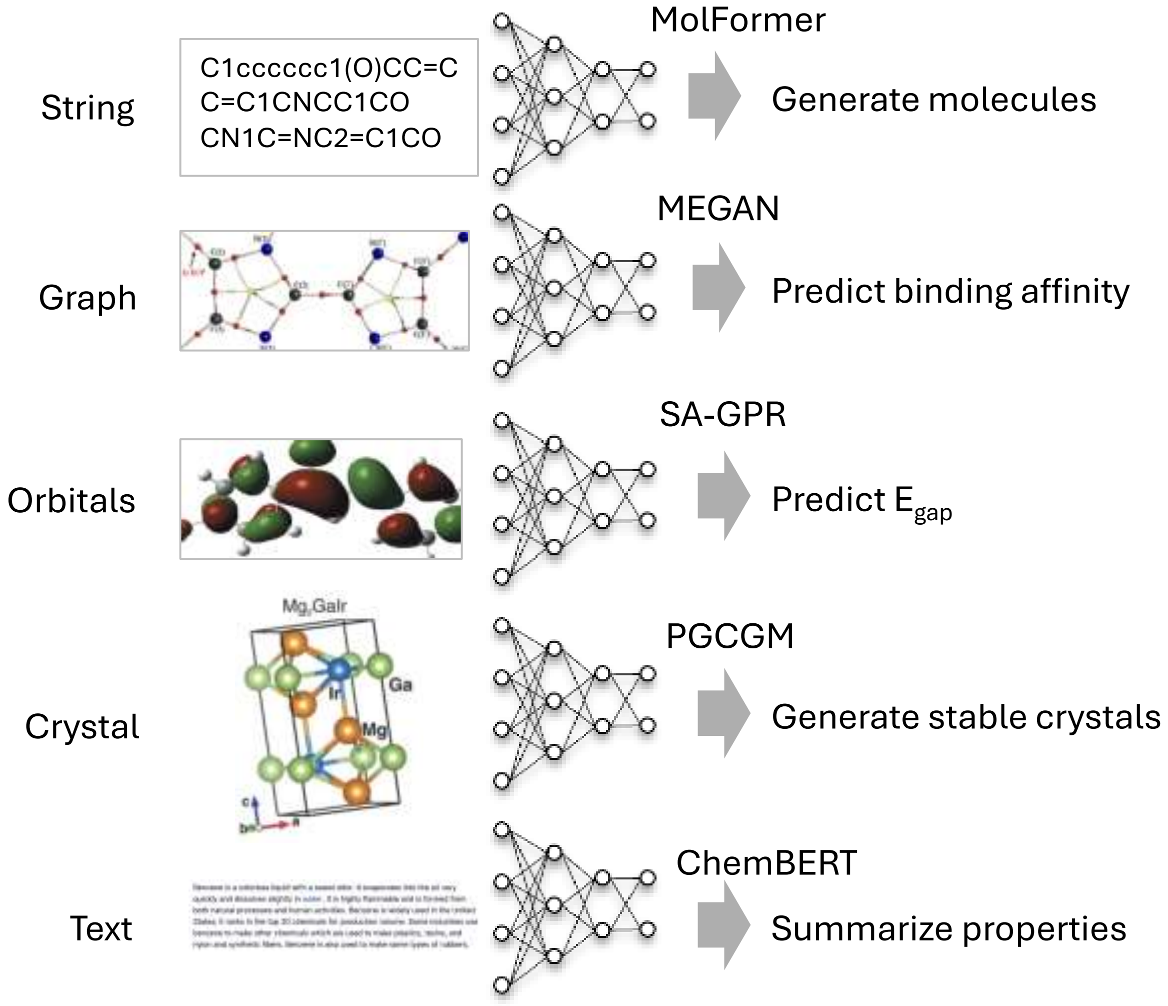
Members working on Foundation Model for Materials and Chemistry

AI is gaining focus in materials, but efforts are fragmented and limited

Uni-modal models for discrete chemical classes focus on limited tasks such as prediction of individual properties, de novo generation of molecules, or prediction of synthesis pathways, etc.

Limitations:

- Scale & data – **today's models are small & limited**
  - Small parameter size : ~100M
  - Small ground-truth data: ~1M
  - Single modality
  - Few chemical classes
  - Limited tasks
- Performance – **today's models perform poorly**
  - Insufficient accuracies
  - Heavy emphasis still on human-workflow
  - No synergies between models
- Many redundant efforts reinventing similar models



String

C1cccccc1(O)CC=C
C=C1CNCC1CO
CN1C=NC2=C1CO

MolFormer

Generate molecules

Graph

MEGAN

Predict binding affinity

Orbitals

SA-GPR

Predict E_{gap}

Crystal

PGCGM

Generate stable crystals

Text

ChemBERT

Summarize properties

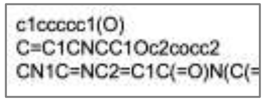
Greater impact can be achieved by centralizing global efforts to build a foundation model

A foundation model trained with multi-modal data sets can be applied to diverse classes of downstream application tasks.

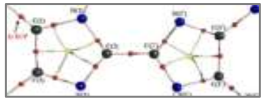
- **Increase scale and grow data:**
  - Larger parameter models: 10B+
  - More diverse data
  - Multi-modality & multi-representation

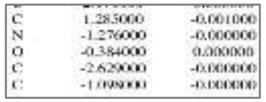
- **Improve performance and impact:**
  - Richer representation by fusing data
  - Higher accuracy in predictions
  - Higher fidelity generation
  - Integrated knowledge

More modalities & representations

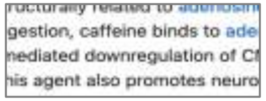


```
c1ccccc1(O)
C=C1CNCC1Oc2cocc2
CN1C=NC2=C1C(=O)N(C(=
```



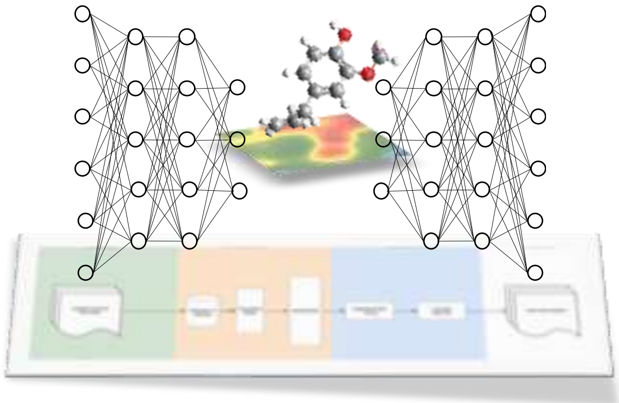


```
C        1.285000      -0.001000
N       -1.276000      -0.000000
O       -0.384000       0.000000
C       -2.629000      -0.000000
C       -1.098000      -0.000000
```



ructurally related to adenosin
gestion, caffeine binds to ade
nediated downregulation of C
nis agent also promotes neuro

Fused multi-modal FMs and framework



Downstream tasks

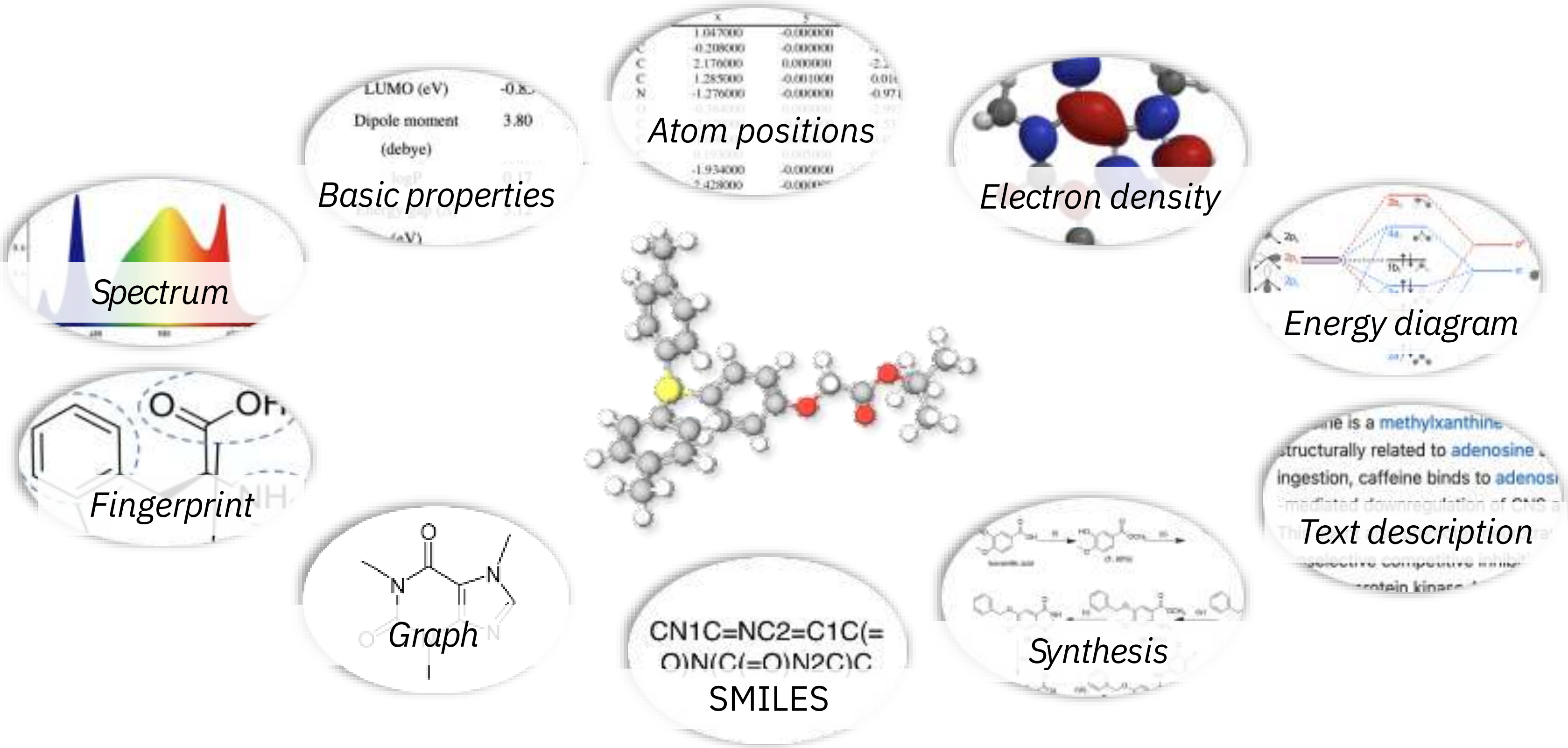
Prediction of comprehensive properties

Generation of molecules, reaction pathways, materials characterization

Prediction of experimental outcomes

Conversational user interface

Multi-modal representations enhance modeling capabilities



Atom positions

Basic properties

Electron density

Spectrum

Energy diagram

Fingerprint

Text description

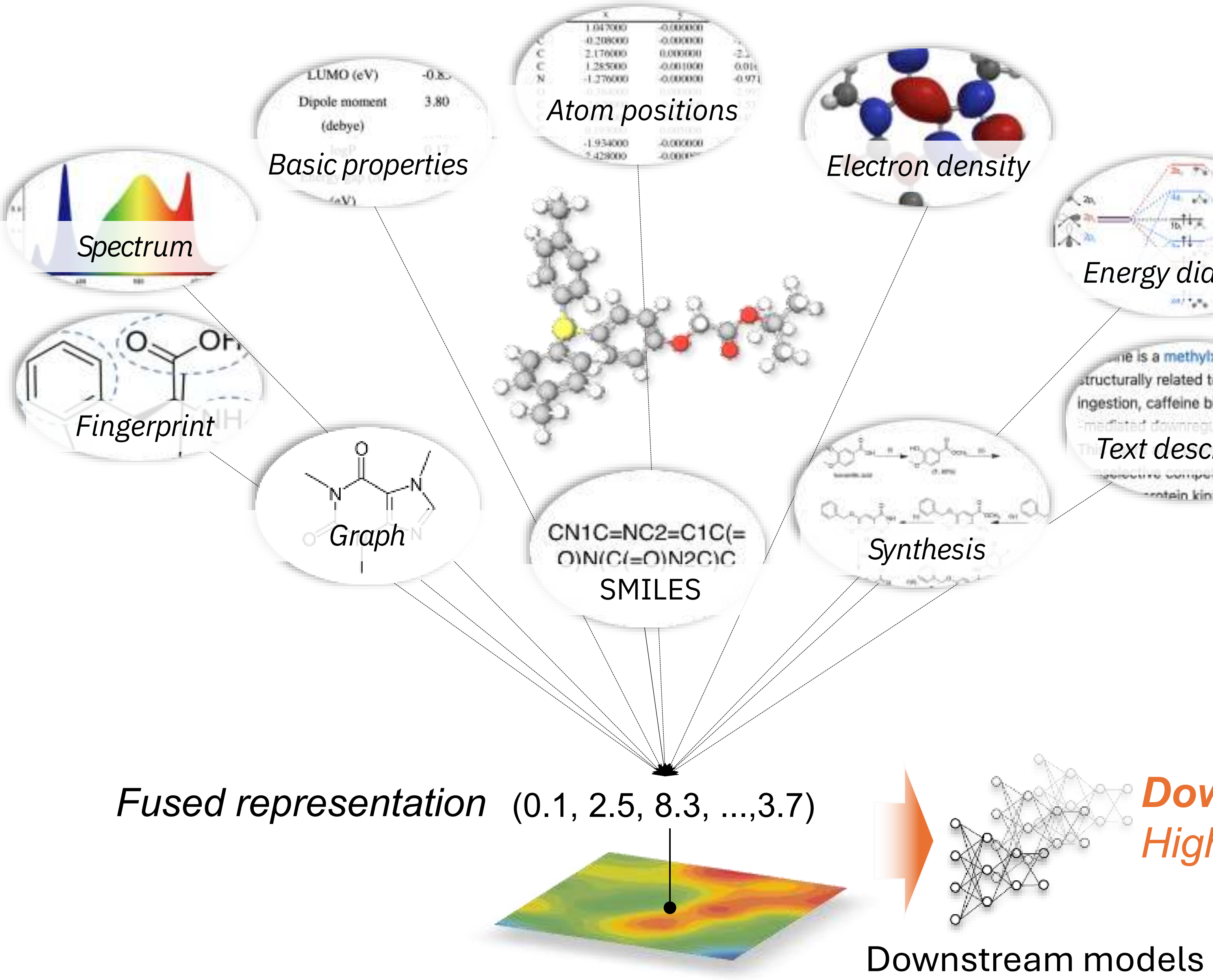
Graph

Synthesis

SMILES

LUMO (eV)
Dipole moment (debye)
3.80
CN1C=NC2=C1C(=
O)N(C(=O)N2C)C
ingestion, caffeine binds to adenosi
structurally related to adenosine

Multi-modal representations enhance modeling capabilities



LUMO (eV)
Dipole moment (debye)
3.80
methylxanthine
structurally related to adenosine
ingestion, caffeine binds to adenosi

Spectrum

Basic properties

Atom positions

Electron density

Energy diagram

Fingerprint

Graph

CN1C=NC2=C1C(=O)N(C(=O)N2C)C

SMILES

Synthesis

Text description

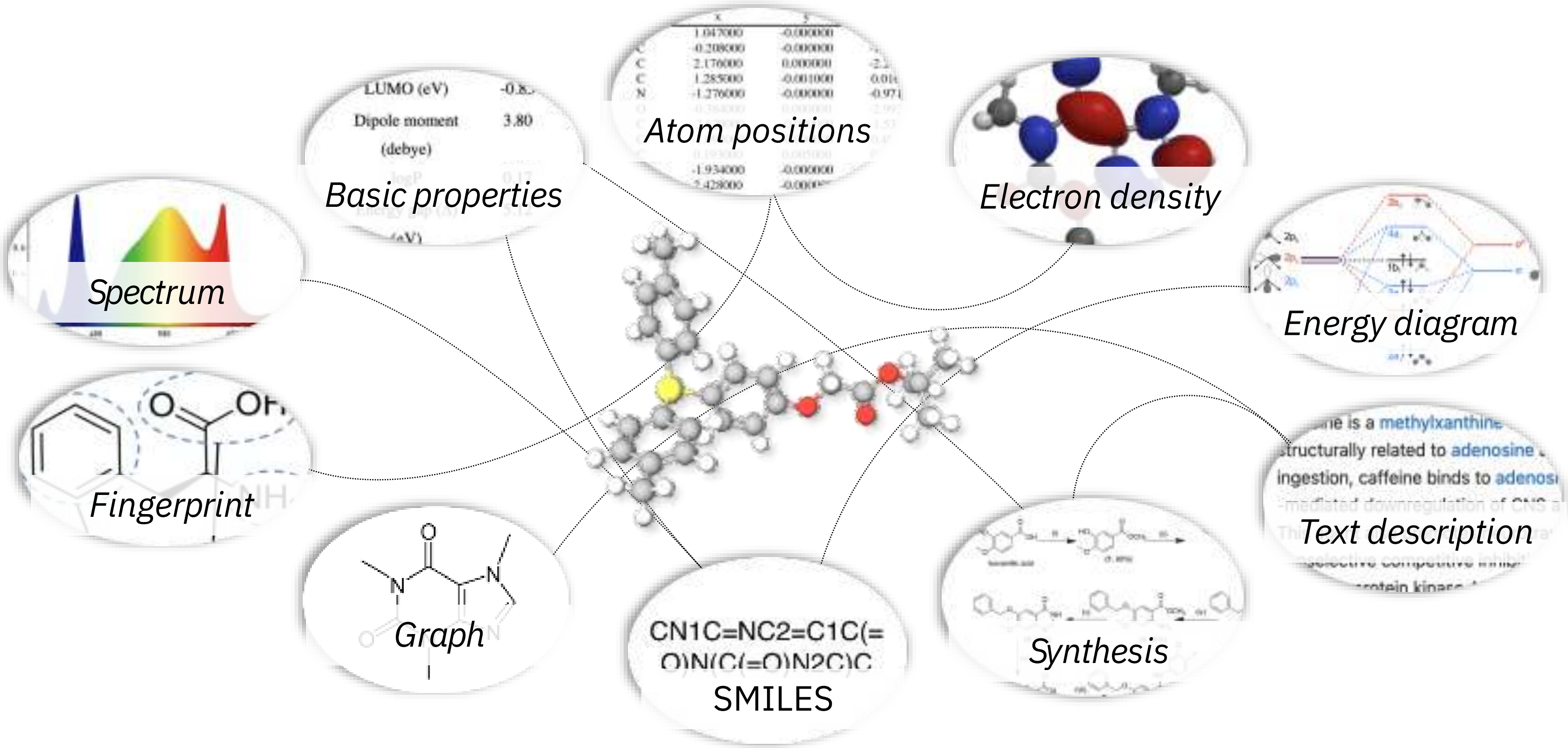
Fused representation (0.1, 2.5, 8.3, ...,3.7)

Downstream models

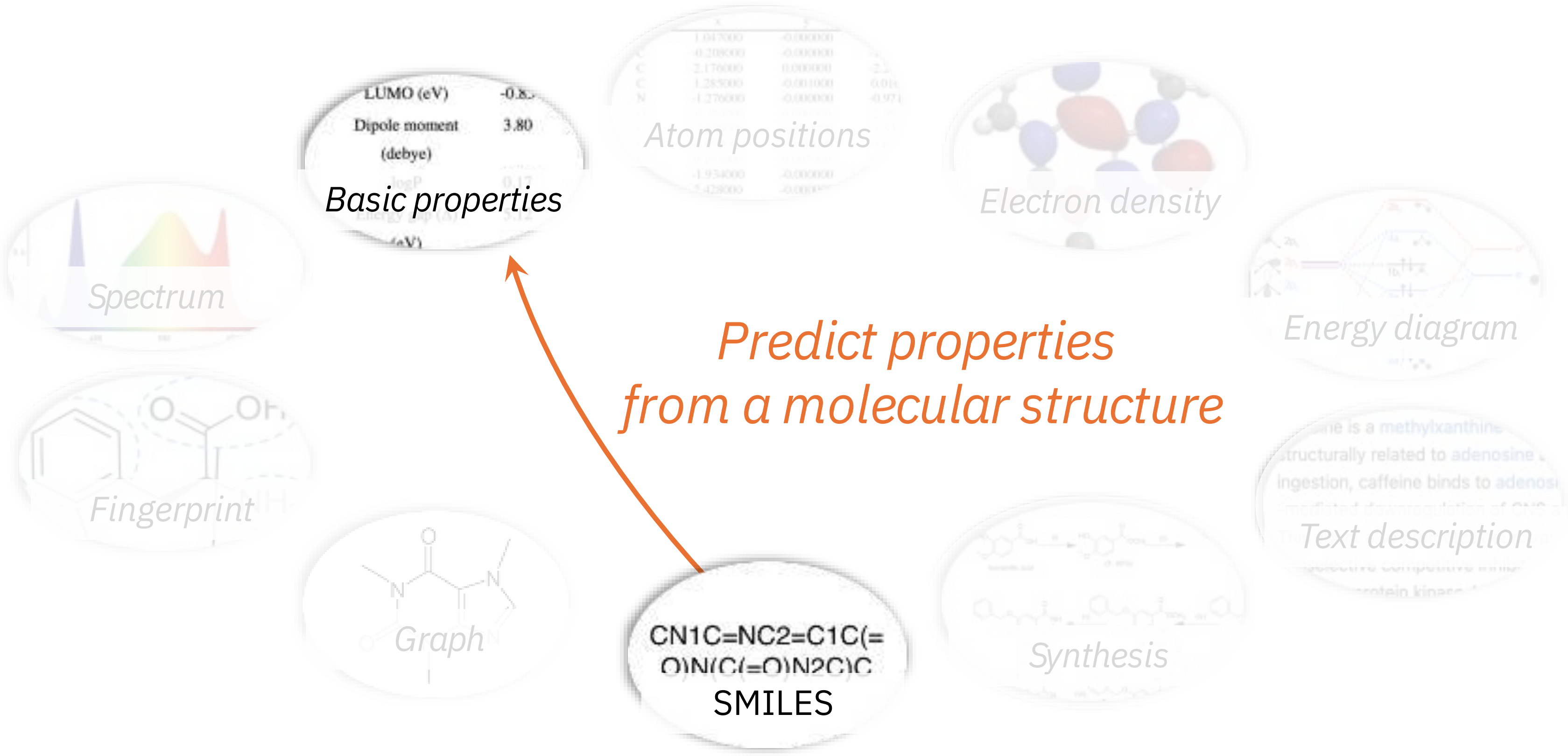
Downstream Tasks #1

High-accuracy predictions

Multi-modal representations enhance modeling capabilities



Multi-modal representations enhance modeling capabilities



Basic properties

Atom positions

Electron density

Spectrum

Energy diagram

Predict properties from a molecular structure

Fingerprint

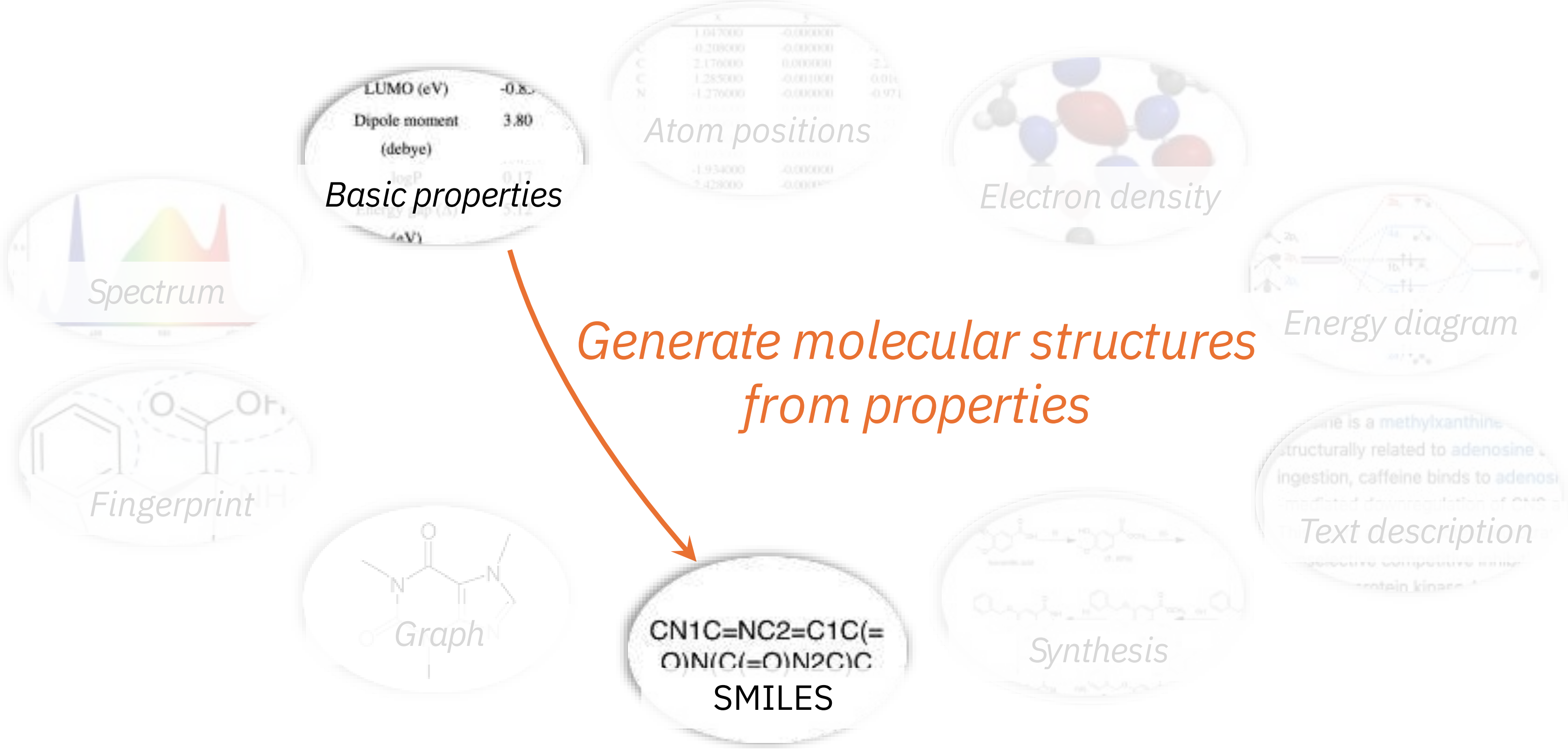
Text description

Graph

SMILES

Synthesis

Multi-modal representations enhance modeling capabilities



Atom positions

Basic properties

Electron density

Spectrum

Energy diagram

Generate molecular structures from properties

Fingerprint

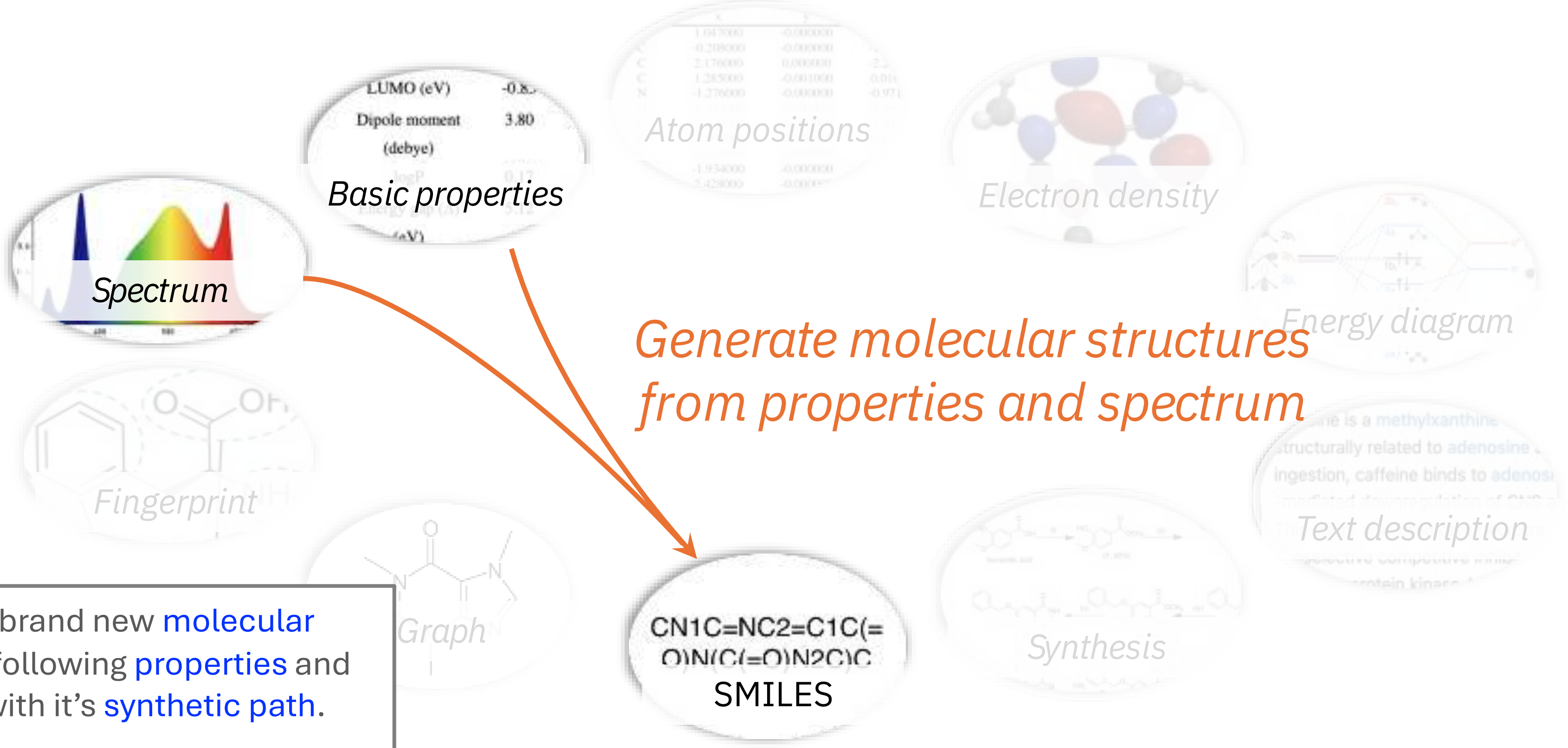
Text description

Graph

Synthesis

SMILES

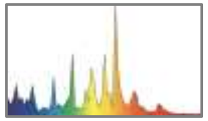
Multi-modal representations enhance modeling capabilities



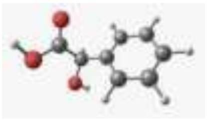
Generate molecular structures from properties and spectrum

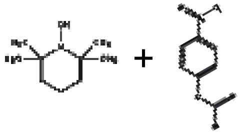
(Query) Generate a brand new molecular structure having the following properties and emission spectrum with it's synthetic path.

- ΔE = 2.3 eV
- T_{melt} = 150 ℃
- …

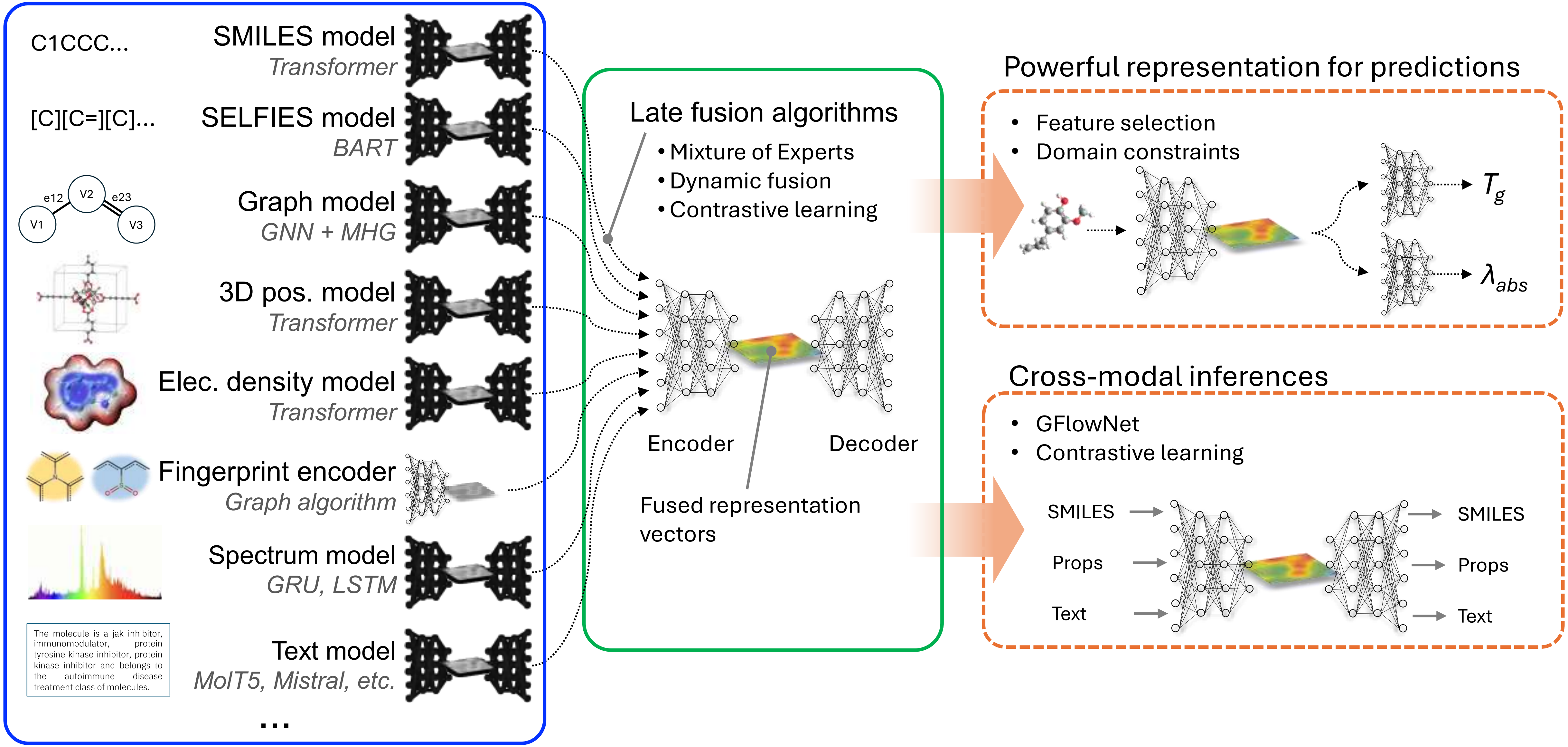


(Answer) Here is the 1st candidate.





Overview of Model Architecture



Modality-specific Foundation Models

C1CCC...
SMILES model
Transformer

[C][C=][C]...
SELFIES model
BART

Graph model
GNN + MHG

3D pos. model
Transformer

Elec. density model
Transformer

Fingerprint encoder
Graph algorithm

Spectrum model
GRU, LSTM

The molecule is a jak inhibitor, immunomodulator, protein tyrosine kinase inhibitor, protein kinase inhibitor and belongs to the autoimmune disease treatment class of molecules.

Text model
MolT5, Mistral, etc.

...

Fused Foundation Model(s)

Late fusion algorithms

- Mixture of Experts
- Dynamic fusion
- Contrastive learning

Encoder

Decoder

Fused representation vectors

Downstream tasks

Powerful representation for predictions

- Feature selection
- Domain constraints

T_g

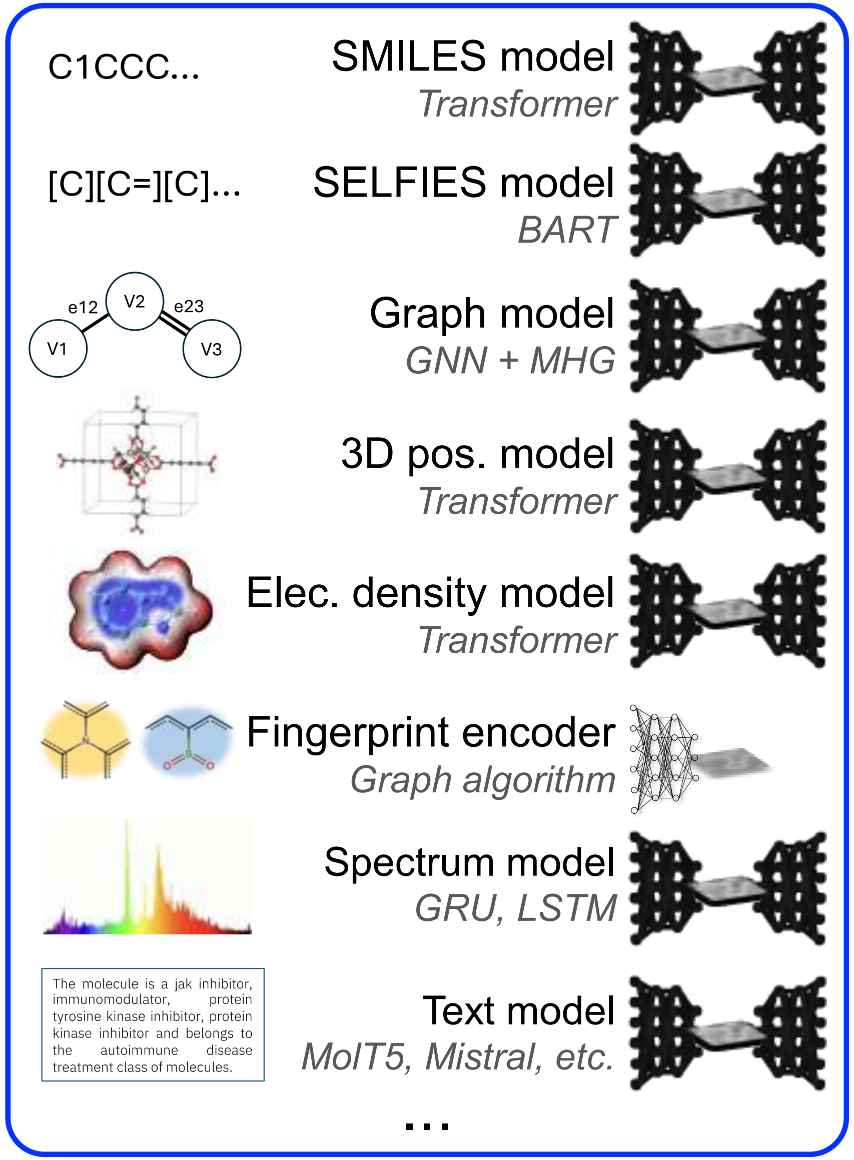
λ_{abs}

Cross-modal inferences

- GFlowNet
- Contrastive learning

Overview of Model Architecture

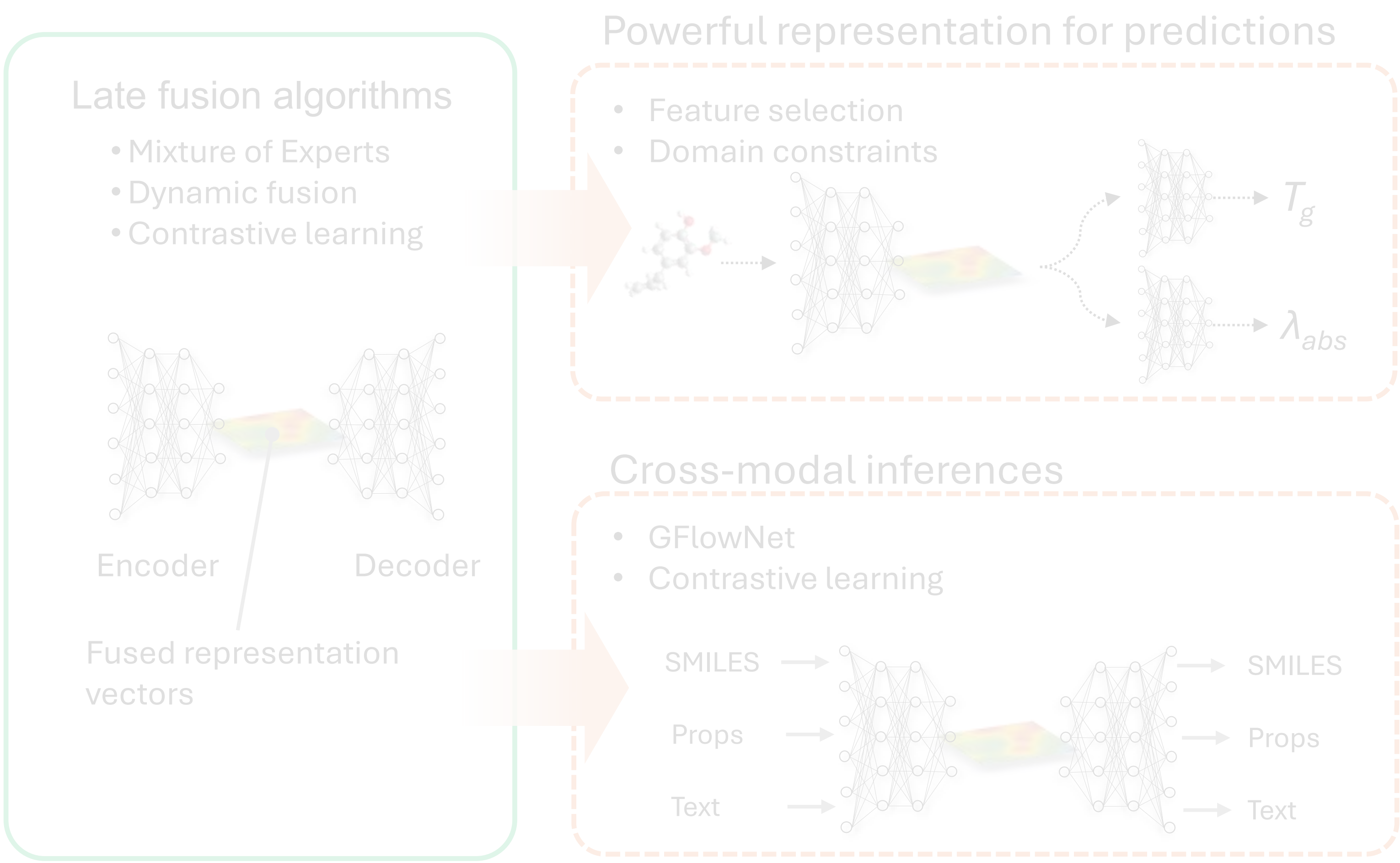
Modality-specific Foundation Models



The molecule is a jak inhibitor, immunomodulator, protein tyrosine kinase inhibitor, protein kinase inhibitor and belongs to the autoimmune disease treatment class of molecules.

Fused Foundation Model(s)

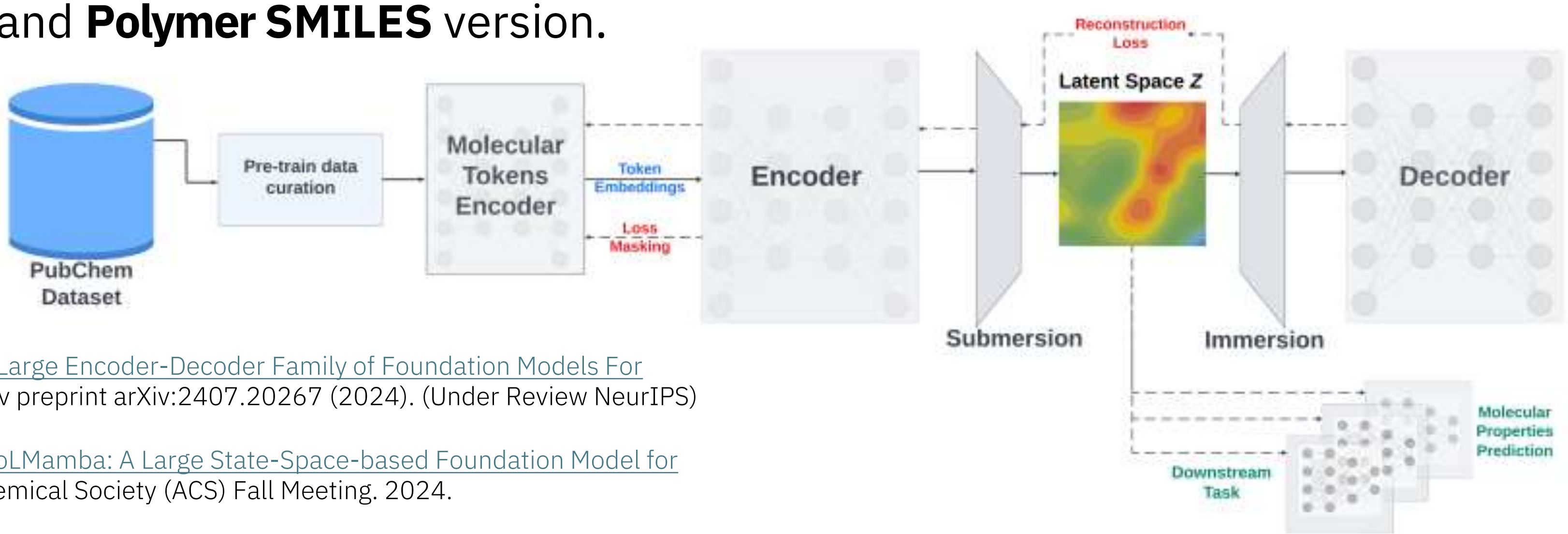
Downstream tasks

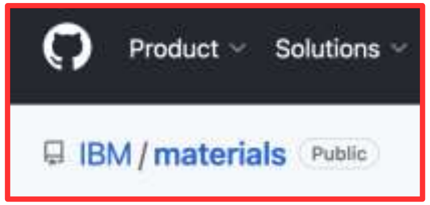


(1) SMILES Model

Transformers-based encoder-decoder model pre-trained on 91 M samples curated from PubChem (4 billion molecular tokens). We also include a **Mamba-based** variant for faster inference, and **Polymer SMILES** version.

Eduardo Almeida Soares





Product Solutions

IBM / materials Public

Open-released in July '24

Eduardo Soares, et al. "A Large Encoder-Decoder Family of Foundation Models For Chemical Language." arXiv preprint arXiv:2407.20267 (2024). (Under Review NeurIPS)

Eduardo Soares, et al. "MoLMamba: A Large State-Space-based Foundation Model for Chemistry." American Chemical Society (ACS) Fall Meeting. 2024.

Variety of models

Model	Num Parameters	Architecture	Input	Open- Source
Base	289M	Transformers	SMILES	Yes
Large	738M	Transformers	SMILES	No
XL	2.5B	Transformers	SMILES	No
MoE	8 X 289M	Transformers	SMILES	Yes
SSM	336M	SSM - Mamba	SMILES	In process
PSMILES	289M	Transformers	Polymers	No

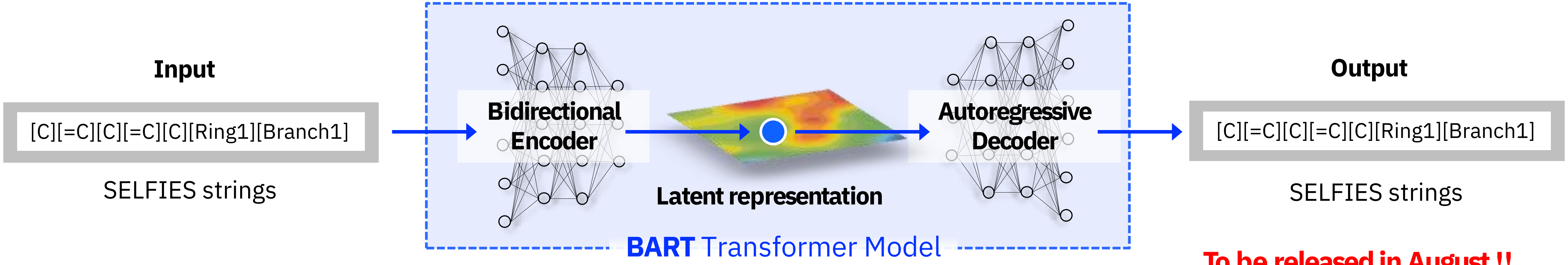
Key Performance example (MoleculeNet and QM9)

Methods	BBBP	HIV	BACE	ClinTox	SIDER	TOX21	QM9
MolFM	72.9	78.8	78.8	79.7	64.2	77.2	-
MoLFormer	90.9	**80.5**	86.3	91.2	65.5	80.4	1.59
IBM.materials.smi-TED289M (Frozen)	91.4	**80.5**	85.6	93.5	**66.01**	81.5	7.49
IBM.materials.smi-TED289M (Fine-Tuned)	**92.2**	76.8	**88.2**	**94.3**	65.6	**81.9**	**1.32**

(2) SELFIES Model

Indra Priyadarsini

BART model is trained with SELFIES strings in a self-supervised manner with/without masking tokens.
1 billion training samples extracted from ZINC22 and PubChem are used.



To be released in August !!

Variety of models

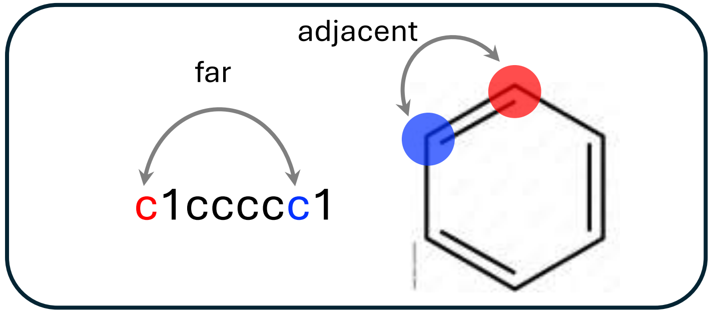
Model	Num Parameters	Dataset	Tokens	Samples Trained
Mini	2.2 M	ZINC 22	173	8 B
Base	354M	ZINC 22	173	1B
Base-mix	354M	ZINC + PubChem	3160	1B
Large	1 B	ZINC + PubChem	3160	100 M

Key Performance example

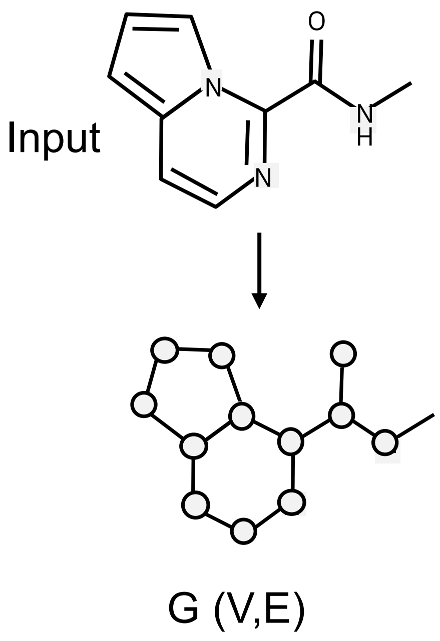
Model	BBBP	HIV	BACE	ClinTox	Sider	Tox21
ChemBERTa	64.3	62.2	79.9	73.3	-	72.8
MolFormer-ZINC	89.9	78.4	87.7	82.2	66.8	83.2
MolFormer-XL	93.7	82.2	88.2	94.8	69	**84.7**
SELFormer	90.2	68.1	83.2	-	**74.5**	65.3
SELF-BART (mini)	92.6	74.2	87	88.3	62.4	75.1
SELF-BART (base)	**95.2**	**83**	**88.8**	**96.9**	65	76.5

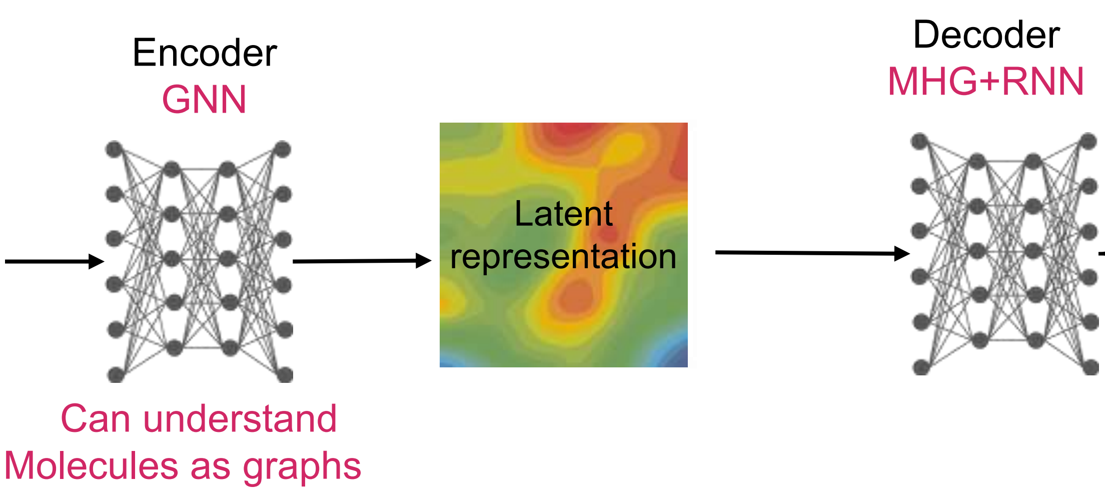
Indra Priyadarsini et al., "A transformer based large-scale molecular representation model." Materials Research Society (MRS) Fall Meeting. 2023.

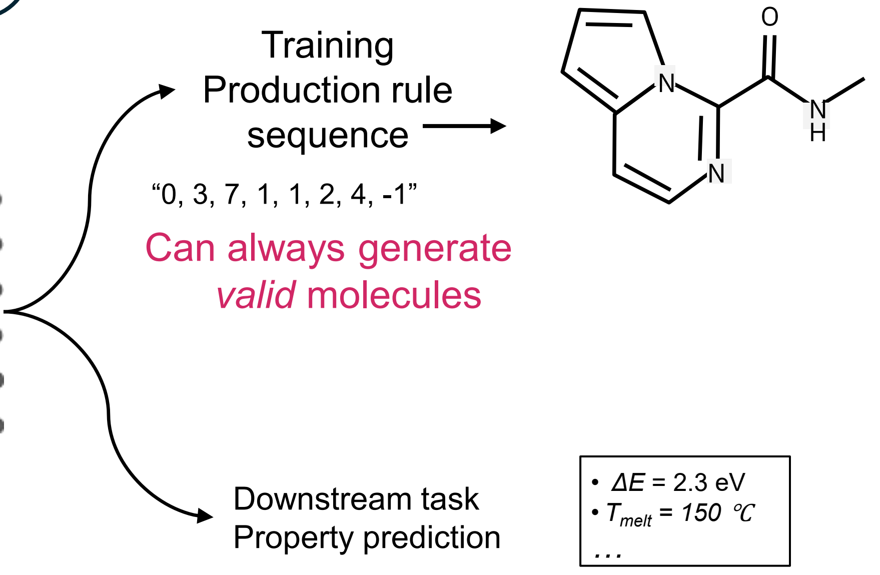
(3) Molecular Graph



Akihiro Kishimoto







Example of MHG-GNN performance evaluation (R2 score ↑)

	Polymer			Photoresist		Chromophore
Method	Density	Bulk Modulus	Refractive Index	Homo	Lumo	Λ_{max} on NIR
MHG-GNN	**0.578**	**0.516**	**0.865**	**0.896**	**0.845**	**0.845**
ECFP6	0.523	0.482	0.823	0.791	0.782	0.708
Modred	0.567	0.505	0.859	0.894	0.830	0.842

Kishimoto, Akihiro, et al. "Autoencoder based on Graph and Recurrent Neural Networks and Application to Property Prediction." *Materials Research Society (MRS) Fall Meeting*. 2023.

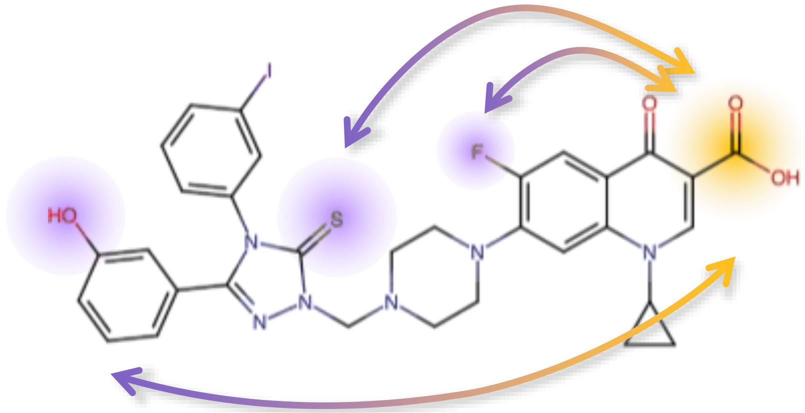
Kishimoto, Akihiro, et al. "MHG-GNN: Combination of Molecular Hypergraph Grammar with Graph Neural Network." *AI4Mat Workshop @ NeurIPS* (2023).

NeurIPS '23 AI4Mat

(4) Topological distance descriptor

Lisa Hamada

Algorithm based model for extracting topological distance between substructures pairs considering intramolecular interaction.



- ■ Can target any substructure without size limitations
- ■ Can handle distances of more than five bonds (a limitation in GNN)
- ■ Consider multiple exception handling scenarios

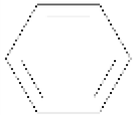
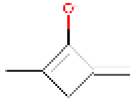
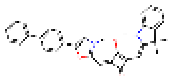
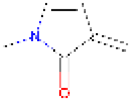
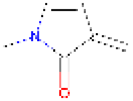
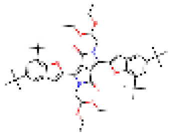
L. Hamada, *et al*. "Molecular Descriptors Accounting for Intramolecular Interactions and Application to Chemical Property Prediction." *American Chemical Society (ACS) Fall Meeting*. 2022.

Key Performance example (R^2 score)

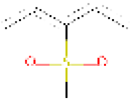
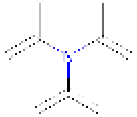
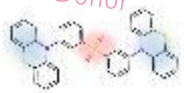
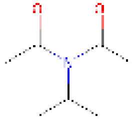
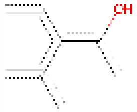
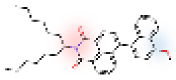
Dataset	Cmp. $(CHCl_3)$		Cmp. (CH_3OH)	
Structure	Large		Large - Medium	
property	$\lambda_{abs.}^{max}$	$\Delta\lambda$	$\lambda_{abs.}^{max}$	$\Delta\lambda$
MolFormer	0.83	0.56	0.87	0.45
MolCLR	0.73	0.43	0.78	0.40
Mordred	0.84	0.51	0.86	0.46
Atom Pair	0.84	0.58	0.84	0.50
Our method	**0.94**	**0.77**	**0.93**	**0.72**

High explainability of Chemical Insight

Cmp. (CH_3OH) $\lambda_{abs.}^{max}$

Rank	Sub1	Sub2	Rep. Mol.
1st			 Squarium
3rd			 DP P

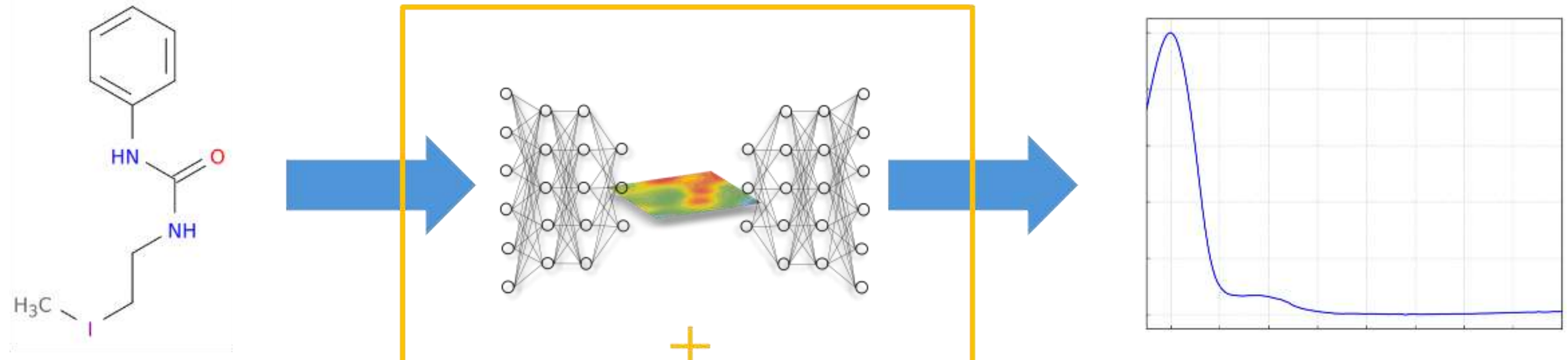
Cmp. (CH_3OH) $\Delta\lambda$

Rank	Sub1	Sub2	Rep. Mol.
2nd			Donor  Acceptor
4th			

(5) Spectrum

Hajime Shinohara

UV spectrum is limited data availability due to the experimental setting.
By implementing domain-specific knowledge into the model, the prediction performance has been increased.



Input
SMILES strings

Domain-specific knowledge

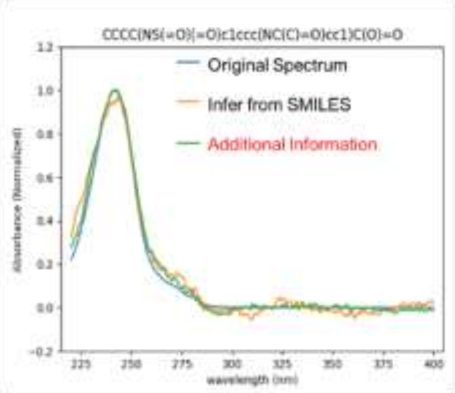
Output
UV spectra

Shinohara, Hajime, et al. "Pre-Treatment Methods for Machine Learning in Finer UV Spectrum Inference." *Materials Research Society (MRS) Fall Meeting*. 2023.

Domain-specific knowledge implementation of UV spectrum from organic molecules

Ex)
Peak position addition
Curvature limitation
Curriculum learning method

Key results example



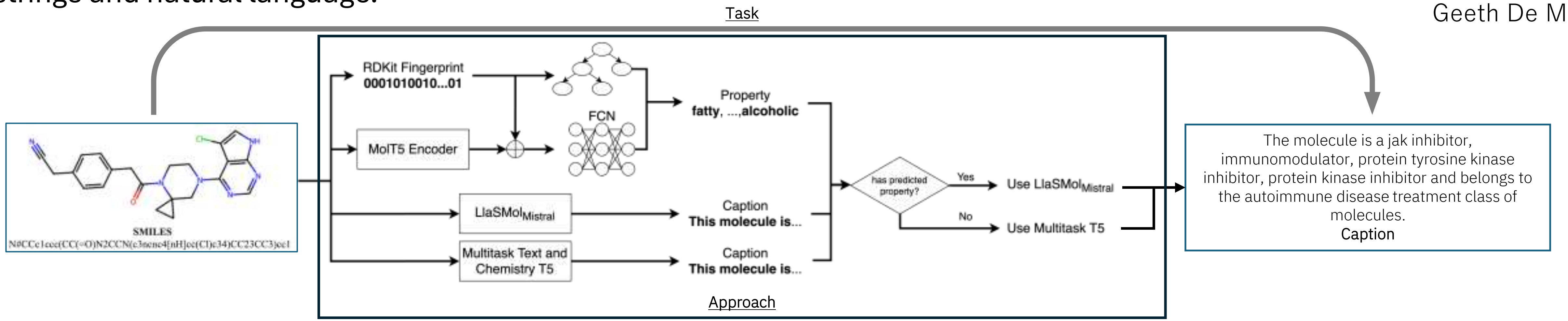
By implementing domain-specific knowledge of UV spectrum from organic molecules, the prediction performance of the spectrum has increased in various models even with small dataset (~3k)

Model	FCN	LSTM	GRU	RNN
Baseline	0.1513	0.1712	0.1597	0.1733
Our model	**0.1377**	**0.1451**	**0.1445**	**0.1716**

(6) Text description - Molecule Captioning from SMILES

“Translating” between molecules encoded in SMILES strings and natural language.

Geeth De Mel



The molecule is a jak inhibitor, immunomodulator, protein tyrosine kinase inhibitor, protein kinase inhibitor and belongs to the autoimmune disease treatment class of molecules.

Caption

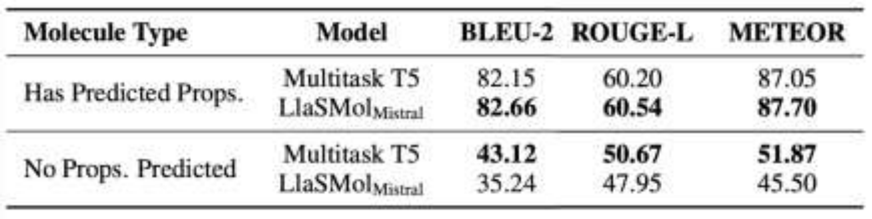
Data Sources

- PubChem
- Chemical Function (CheF)
- ChemFOnt

Models

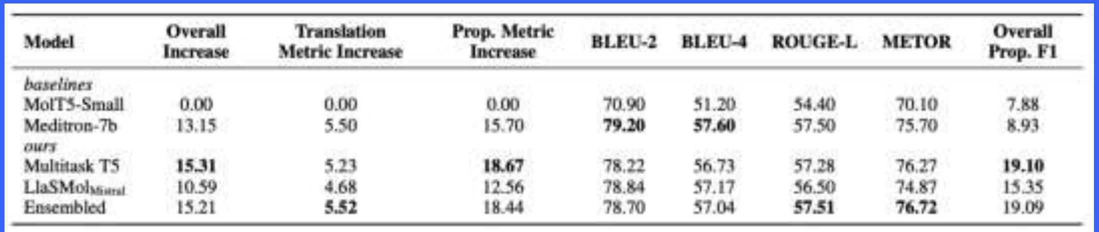
- MolT5 Small/Base/Large - 60M ~ 770M
- Multitask text and Chemistry Small/Base augm - 60M ~ 223M
- Meditron - 7B
- Mistral - 7B
- XGBoost-based molecule property predictor

Results



Molecule Type	Model	BLEU-2	ROUGE-L	METEOR
Has Predicted Props.	Multitask T5	82.15	60.20	87.05
	LlaSMol$_{Mistral}$	**82.66**	**60.54**	**87.70**
No Props. Predicted	Multitask T5	**43.12**	**50.67**	**51.87**
	LlaSMol$_{Mistral}$	35.24	47.95	45.50

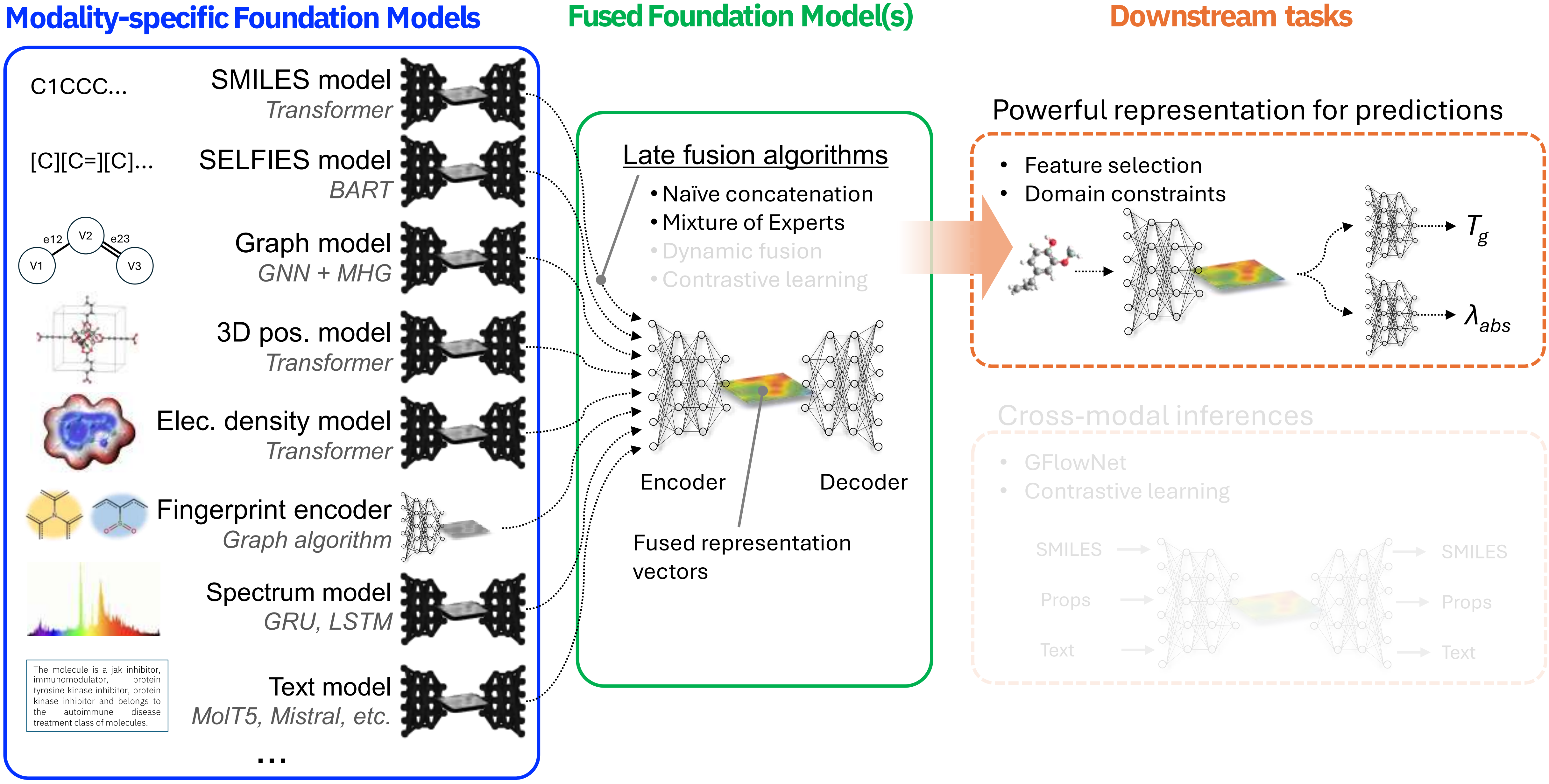
Translation metrics by molecular type on dev. set



Model	Overall Increase	Translation Metric Increase	Prop. Metric Increase	BLEU-2	BLEU-4	ROUGE-L	METOR	Overall Prop. F1
baselines								
MolT5-Small	0.00	0.00	0.00	70.90	51.20	54.40	70.10	7.88
Meditron-7b	13.15	5.50	15.70	**79.20**	**57.60**	57.50	75.70	8.93
ours								
Multitask T5	**15.31**	5.23	**18.67**	78.22	56.73	57.28	76.27	**19.10**
LlaSMol$_{Mistral}$	10.59	4.68	12.56	78.84	57.17	56.50	74.87	15.35
Ensembled	15.21	**5.52**	18.44	78.70	57.04	**57.51**	**76.72**	19.09

Overall increase from MolT5-Small baseline and translation metrics results on dev. set

Overview of Model Architecture



Modality-specific Foundation Models

C1CCC...
SMILES model
Transformer

[C][C=][C]...
SELFIES model
BART

Graph model
GNN + MHG

3D pos. model
Transformer

Elec. density model
Transformer

Fingerprint encoder
Graph algorithm

Spectrum model
GRU, LSTM

The molecule is a jak inhibitor, immunomodulator, protein tyrosine kinase inhibitor, protein kinase inhibitor and belongs to the autoimmune disease treatment class of molecules.

Text model
MolT5, Mistral, etc.

...

Fused Foundation Model(s)

Late fusion algorithms

- Naïve concatenation
- Mixture of Experts
- Dynamic fusion
- Contrastive learning

Encoder

Decoder

Fused representation vectors

Downstream tasks

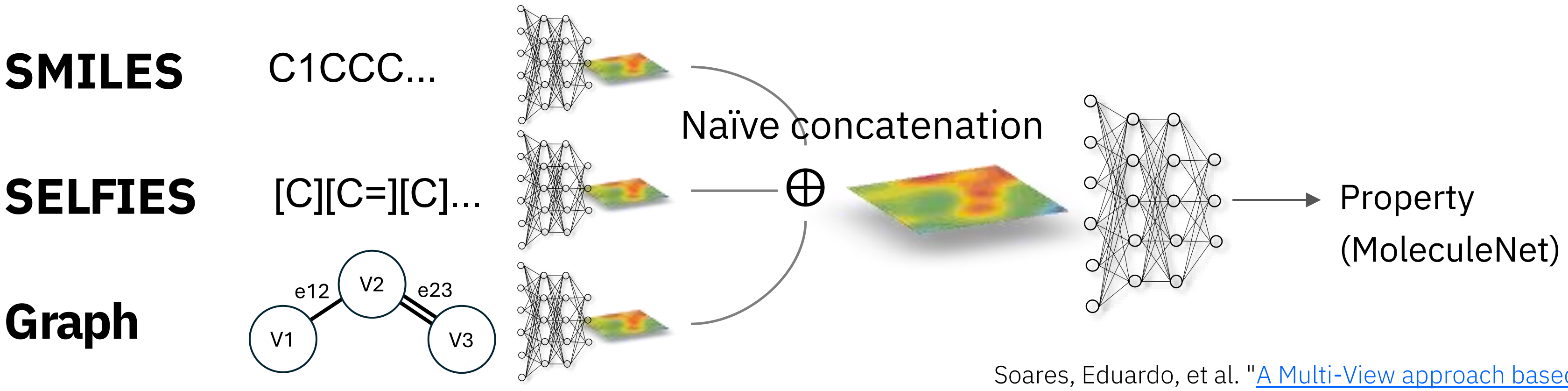
Powerful representation for predictions

- Feature selection
- Domain constraints

Cross-modal inferences

- GFlowNet
- Contrastive learning

Multi-modal feature representations improve downstream prediction accuracy



SMILES C1CCC...

SELFIES [C][C=][C]...

Graph

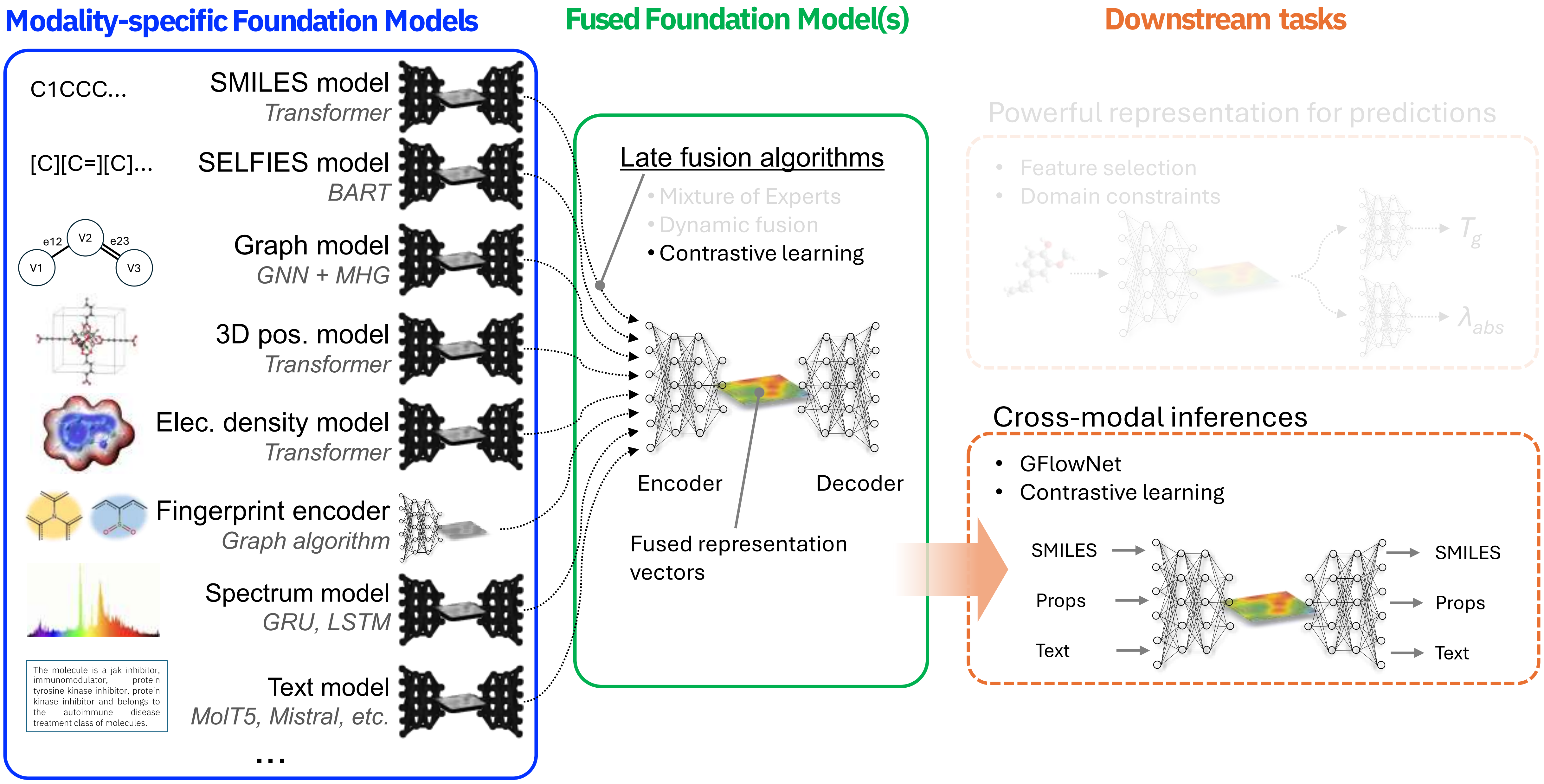
Naïve concatenation ⊕

Property (MoleculeNet)

Soares, Eduardo, et al. "A Multi-View approach based on Graphs and Chemical Language Foundation Model for Molecular Properties Prediction." AAAI 2024.

		BBBP	HIV	BACE	ClinTox	SIDER	Tox21
Single modality	SELFIES (BART)	95.22	83.04	88.77	96.86	64.95	76.53
	Graph (MHG)	93.54	82.89	89.53	87.46	66.93	79.32
	SMILES (MolFormer-XL)	93.70	82.20	88.21	94.80	**69.00**	**84.70**
Multi-modality	SELFIES + Graph	95.35	83.93	88.66	92.31	66.02	78.81
	SELFIES + SMILES	96.42	83.99	89.80	99.82	64.41	80.48
	SMILES + Graph	**96.60**	84.93	**90.48**	99.59	65.20	78.15
	SELFIES + SMILES + Graph	96.06	**85.28**	90.00	**99.94**	65.65	79.19

Overview of Model Architecture



Modality-specific Foundation Models

C1CCC... SMILES model *Transformer*

[C][C=][C]... SELFIES model *BART*

Graph model *GNN + MHG*

3D pos. model *Transformer*

Elec. density model *Transformer*

Fingerprint encoder *Graph algorithm*

Spectrum model *GRU, LSTM*

Text model *MolT5, Mistral, etc.*

The molecule is a jak inhibitor, immunomodulator, protein tyrosine kinase inhibitor, protein kinase inhibitor and belongs to the autoimmune disease treatment class of molecules.

...

Fused Foundation Model(s)

Late fusion algorithms

- Mixture of Experts
- Dynamic fusion
- Contrastive learning

Fused representation vectors

Downstream tasks

Powerful representation for predictions

- Feature selection
- Domain constraints

Cross-modal inferences

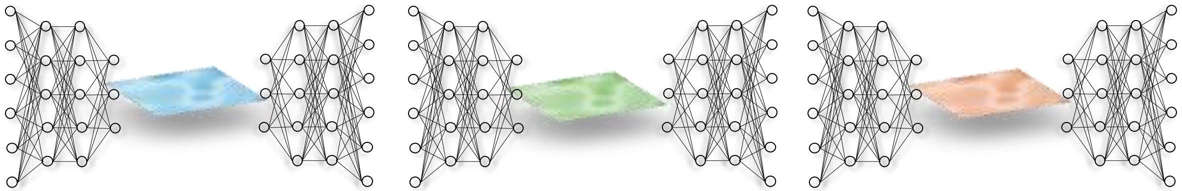
- GFlowNet
- Contrastive learning

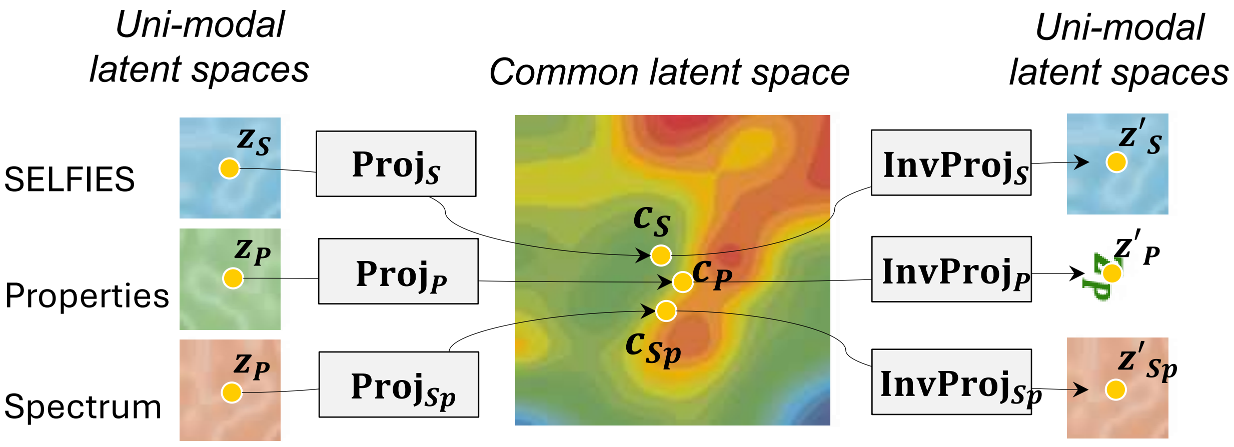
Cross-Modal Inferences

Projectors/inverse-projectors are trained to align different modal representations on a common latent space so that modality-to-modality (cross-modal) inferences are achieved.

Seiji Takeda

SELFIES (BART) | DFT Properties (FC VAE) | UV Spectrum (GRU VAE)

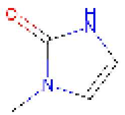




Uni-modal latent spaces | *Common latent space* | *Uni-modal latent spaces*

$$\min_{\mathbf{c}_i,\mathbf{c}_j} \sum_{i,j\in\{S,P,Sp\},i\neq j} \|\mathbf{c}_i - \mathbf{c}_j\|^2$$

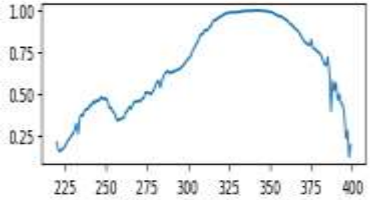
Example of cross-modal inferences



[C][N][C][=C][NH1][C][Ring1][Branch1][=O]

SELFIES

Dipole moment	3.4
Isotropic coefficient	58
HOMO energy	-0.3
LUMO energy	0
Energy gap	0.3
Spatial extention	730
…	…

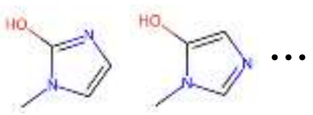


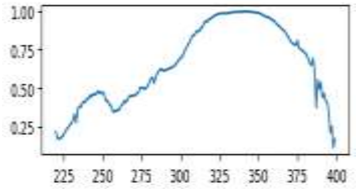
DFT Properties

UV Spectrum

Dipole moment	3.5
Isotropic coefficient	60
HOMO energy	-0.3
LUMO energy	0
Energy gap	0.3
Spatial extention	730
…	…

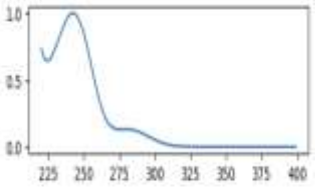
DFT Properties



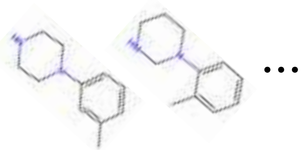


SELFIES

UV Spectrum



UV Spectrum



Dipole moment	2.6
Isotropic coefficient	68
HOMO energy	-0.2
LUMO energy	0.1
Energy gap	0.3
Spatial extention	912
…	…

SELFIES

DFT Properties

Battery Materials Discovery Empowered by AI and Foundation Models

Young-Hye Na

MOLECULE
MIXTURE
DEVICE

Battery Materials

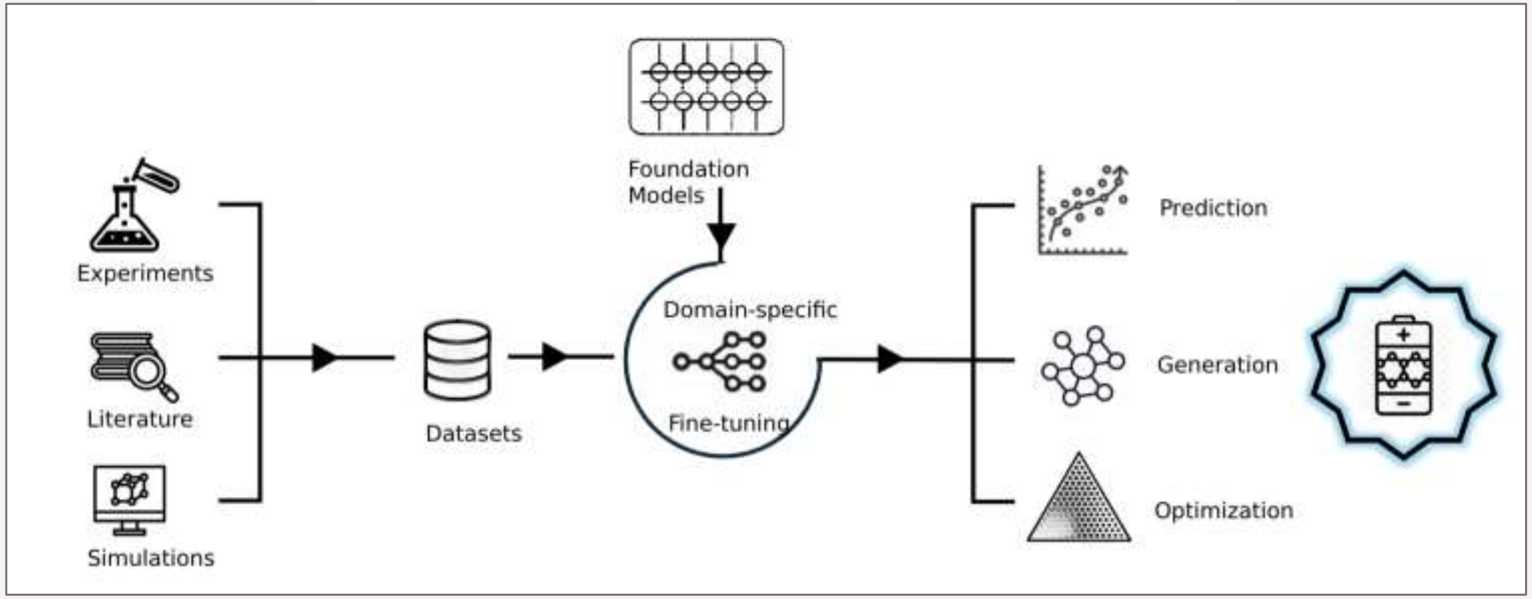
- Large design space and multivariable relationships
- Complexities at multiple-length scale

✓ **Foundation Models (FM) are fine-tuned for domain specific tasks** using labeled datasets derived from literature, simulations, or lab-experimentation

✓ **Our fine-tuned FM models effectively map material structures, compositions, and device performance,** predicting and optimizing the properties and performance of complex mixed materials (e.g., electrolyte formulations)

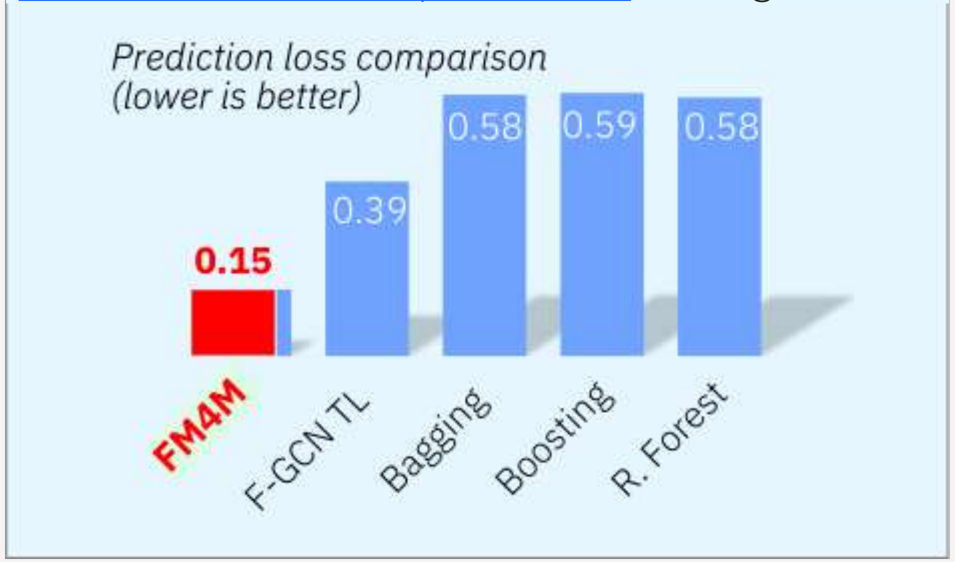
✓ **These models, along with our customized AI-workflows,** are currently being used to discover new electrolytes for our industrial partners

AI-assisted Electrolyte Discovery Workflow



Prediction of Battery Performance (LCE)

Indrapriyadarsini et.al, "Improving Performance Prediction of Electrolyte Formulations with Transformer-based Molecular Representation Model." ML4LMS @ *ICML*. 2024.



IBM **Research/ Sustainable Battery Materials Discovery/ Young-Hye Na:** yna@us.ibm.com

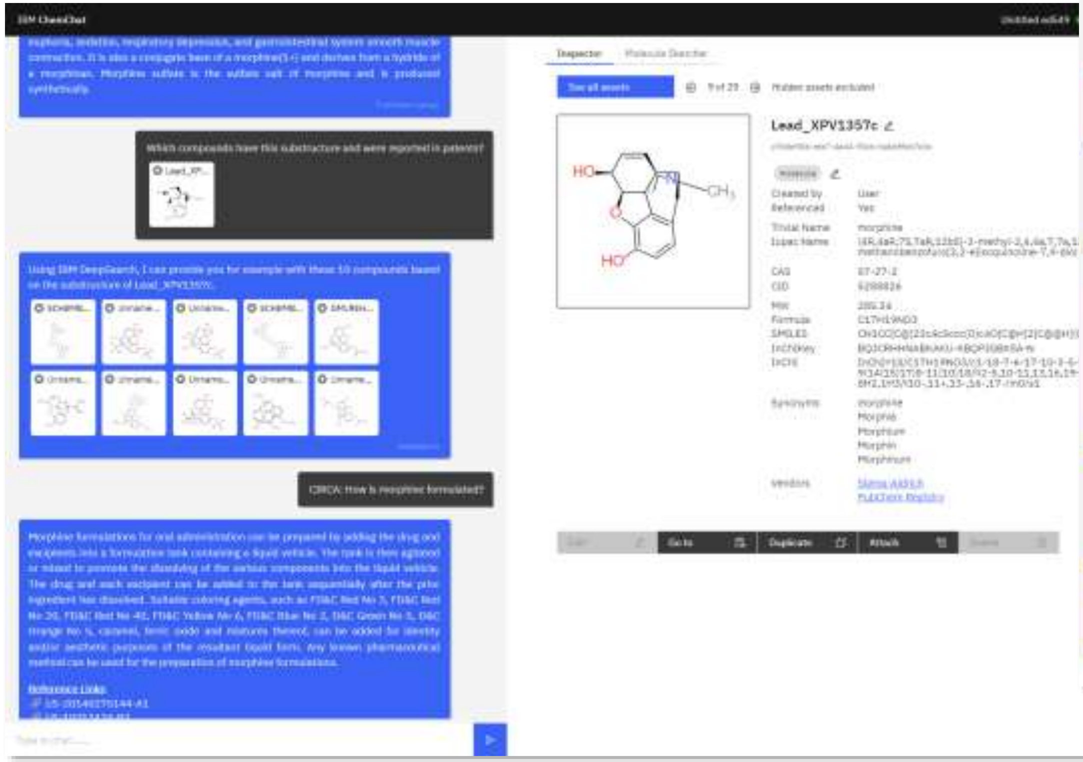
ChemChat Material Science and Data Visualization Assistant

Tim Erdmann

Molecules: Identification & description, Vendor check, Property calculations & predictions by cheminformatics, conventional AI and FM models, PFAS determination after EPA & ECHA, Retrosynthesis prediction, Generation by target properties, Locating molecules and substructure matches in patents

Documents: Querying 30M material and chemistry-focused patents, Querying user-defined collection of pdfs

Visualizations: 11 chart types incl. request for aggregations and interactive elements



euphoria, sedation, respiratory depression, and gastrointestinal system smooth muscle contraction. It is also a conjugate base of a morphine(1+) and derives from a hydride of a morphinan. Morphine sulfate is the sulfate salt of morphine and is produced synthetically.

Which compounds have this substructure and were reported in patents?

Using IBM DeepSearch, I can provide you for example with these 10 compounds based on the substructure of Lead_XPV1357c.

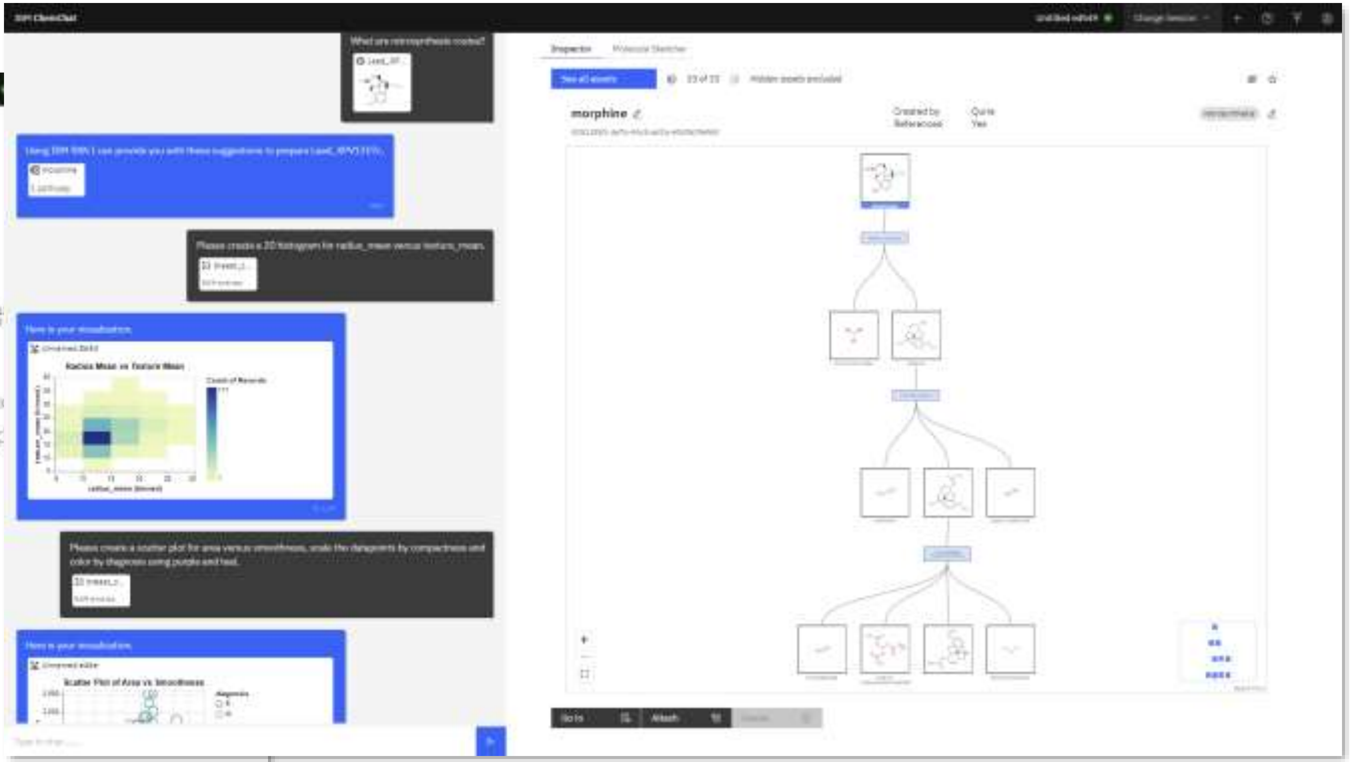
CIRCA: How is morphine formulated?

Morphine formulations for oral administration can be prepared by adding the drug and excipients into a formulation tank containing a liquid vehicle. The tank is then agitated or mixed to promote the dissolving of the various components into the liquid vehicle. The drug and each excipient can be added to the tank sequentially after the prior ingredient has dissolved. Suitable coloring agents, such as FD&C Red No 3, FD&C Red No 20, FD&C Red No 40, FD&C Yellow No 6, FD&C Blue No 2, D&C Green No 5, D&C Orange No 5, caramel, ferric oxide and mixtures thereof, can be added for identity and/or aesthetic purposes of the resultant liquid form. Any known pharmaceutical method can be used for the preparation of morphine formulations.

Reference Links
US-20140275144-A1

Lead_XPV1357c

Created by: User
Referenced: Yes
Trivial Name: morphine
Iupac Name: (4R,4aR,7S,7aR,12bS)-3-methyl-2,4,4a,7,7a,1
methanobenzofuro[3,2-e]isoquinoline-7,9-diol
CAS: 57-27-2
CID: 5288826
MW: 285.34
Formula: C17H19NO3
SMILES: CN1CC[C@]23c4c5ccc(O)c4O[C@H]2[C@@H](
InChIKey: BQJCRHHNABKAKU-KBQPJGBKSA-N
InChI: InChI=1S/C17H19NO3/c1-18-7-6-17-10-3-5-
9(14)15(17)8-11(10)18/h2-5,10-11,13,16,19-
6H2,1H3/t10-,11+,13-,16-,17-/m0/s1
Synonyms: morphine
Morphia
Morphium
Morphin
Morphinum
Vendors: Sigma Aldrich
PubChem Registry



What are retrosynthesis routes?

Using IBM RXN I can provide you with these suggestions to prepare Lead_XPV1357c.

Please create a 2D histogram for radius_mean versus texture_mean.

Here is your visualization.

Please create a scatter plot for area versus smoothness, scale the datapoints by compactness and color by diagnosis using purple and teal.

Here is your visualization.

The AI Alliance & Open Innovation

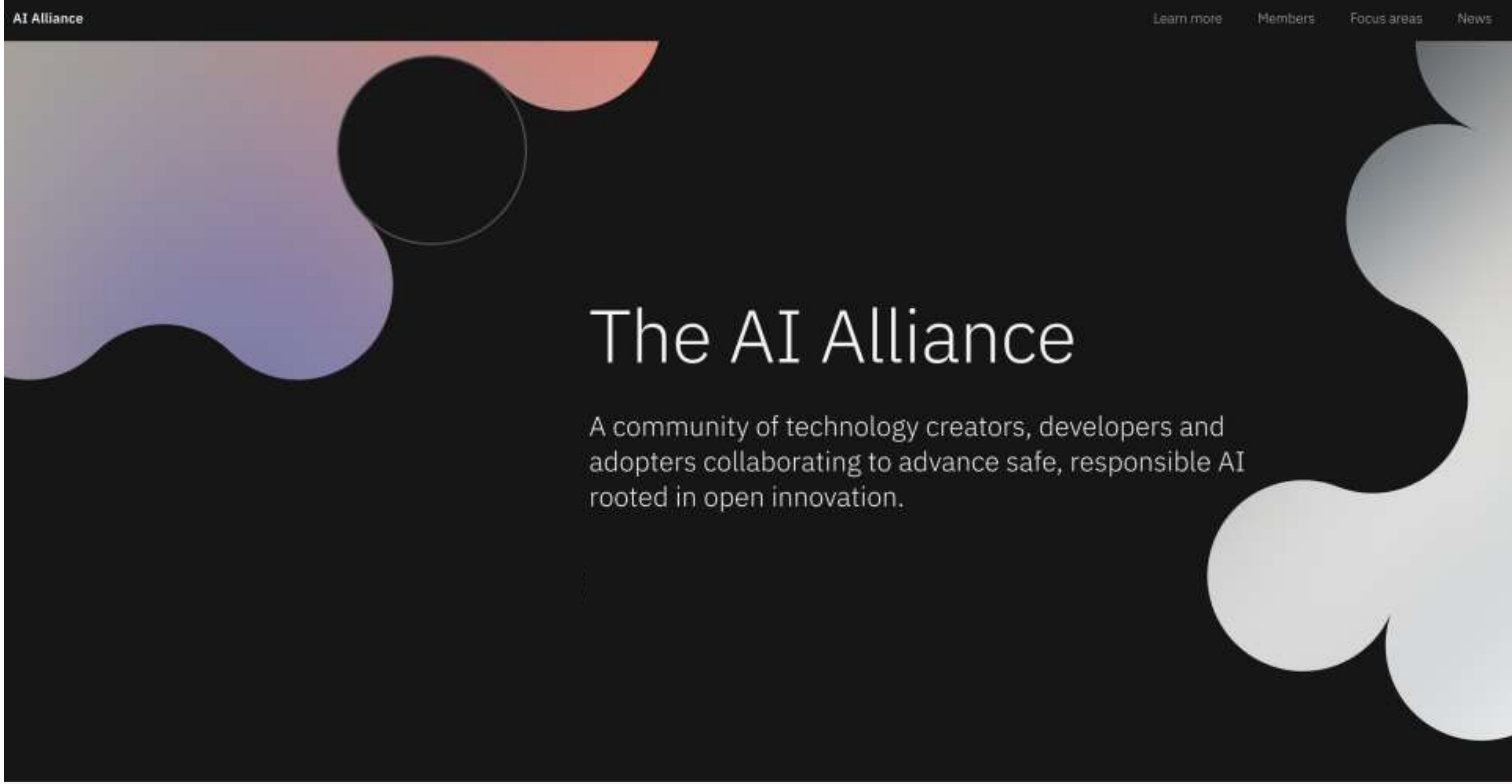
Build and support open technologies across software, models and tools.

Enable developers and scientists to understand, experiment, and adopt open technologies.

Advocate for open innovation with organizational and societal leaders, policy and regulatory bodies, and the public.

https://thealliance.ai/

IBM **Research**



The AI Alliance

A community of technology creators, developers and adopters collaborating to advance safe, responsible AI rooted in open innovation.

WG4M – Working group for Materials and Chemistry

- Community development of open-source code and models
  - Defining target domains/use-cases (PFAS, polymer etc)
  - Rich model family of modalities and architectures
- Open-source of base models at https://github.com/IBM/materials (one model released, more planned)
- Sharing: model architecture and contact points for members, and workstream-based engagement approach
- 20+ materials company and worldwide researchers involved

AI Alliance Working Group for Materials & Chemistry (WG4M)

(1) Kick-off Technical Workshop

(16th May 2024, Tokyo)

Keynote Talks & Lightning Pitches

Group discussion & Readout

Functions do you expect on
AI x Materials in this working Group
リソース
What do you expect on
AI x Material working group?
Possible Actions
- Survey of public data
- Survey of available models
- Survey of material classes & applications
- workshops & meetings

(2) Social Gathering @ ICML AI4Science

(25th July 2024, Vienna)

ICML
International Conference On Machine Learning
AI Alliance
Working Group for Materials and Chemistry
Our SMILES Foundation Model
github.com/IBM/materials
Join us!
think
Rezeption

AI Alliance Working Group for Materials & Chemistry (WG4M)

Community Development

- 20+ materials companies
- Bi-weekly calls with material companies
- Seeking co-chairs, topic leads

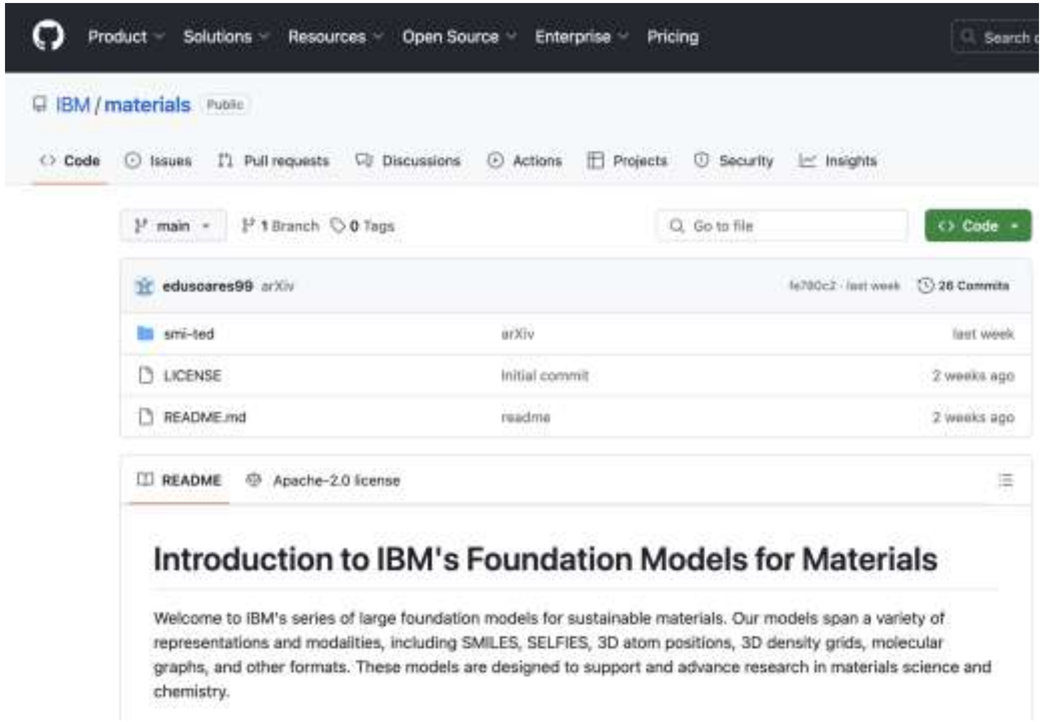
JSR
W&B
TEL
TOKYO ELECTRON
Sony AI
AsahiKASEI
SUMITOMO BAKELITE CO., LTD
RICOH
SUMITOMO ELECTRIC
SMIC
MITSUBISHI CHEMICAL GROUP
Panasonic
tok
Canon
HITACHI
UTokyo
Denka
The Institute of Statistical Mathematics
NAGASE
RESONAC
AGC
HONDA

Model development

Open-source of base models at github.com/IBM/materials

- SMILES Transformer model was released on 25 July
- Other modalities models (SELFIES, Graph, etc.) will soon be released!

https://thealliance.ai/



edusoares99 arXiv fe780c2 · last week 26 Commits

smi-ted	arXiv	last week
LICENSE	Initial commit	2 weeks ago
README.md	readme	2 weeks ago

Introduction to IBM's Foundation Models for Materials

Welcome to IBM's series of large foundation models for sustainable materials. Our models span a variety of representations and modalities, including SMILES, SELFIES, 3D atom positions, 3D density grids, molecular graphs, and other formats. These models are designed to support and advance research in materials science and chemistry.

Thank you

https://www.linkedin.com/in/indra-ipd/

Backup

Working Group for Materials (WG4M)

Vision

To build an open development community for Materials Informatics (MI) that transcends national and organizational boundaries, and to accelerate the speed of global materials development by creating a suite of MI tools and models that are accessible to everyone.

1. Open development and consumption of MI AI tools and models
2. Data sharing (to the extent possible)
3. Creation of benchmarks
4. Provision of platforms for information exchange and discussion
5. Development of talent in Materials x AI