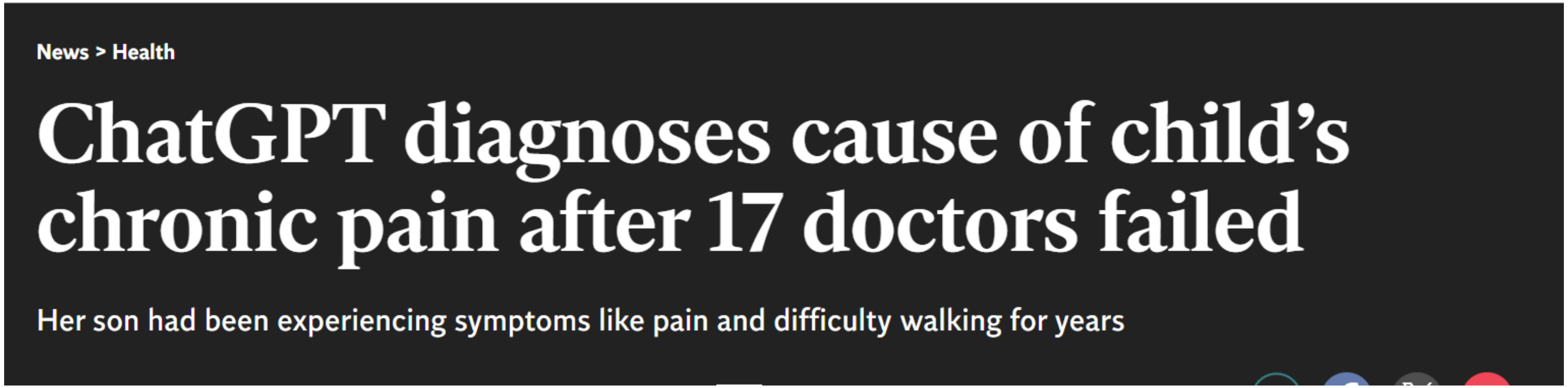
# Knowledge Discovery and Extraction in the Biomedical Domain

## ACISN’24

– PRESENTED BY DR AID AICHA

https://sites.google.com/univ-bouira.dz/acisn24

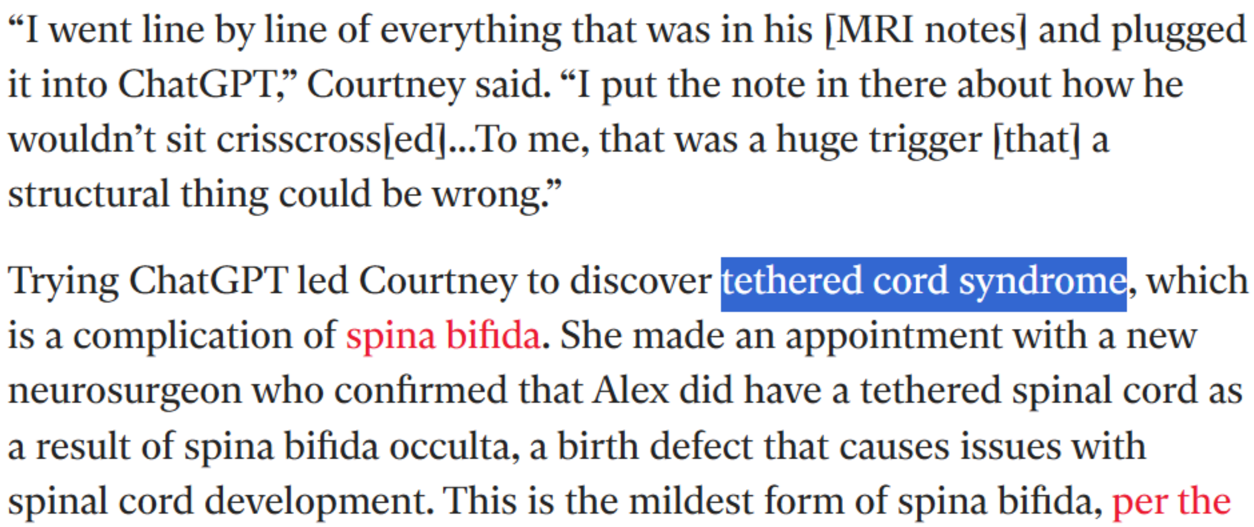
# Opening Story - Public



News > Health

## ChatGPT diagnoses cause of child’s chronic pain after 17 doctors failed

Her son had been experiencing symptoms like pain and difficulty walking for years

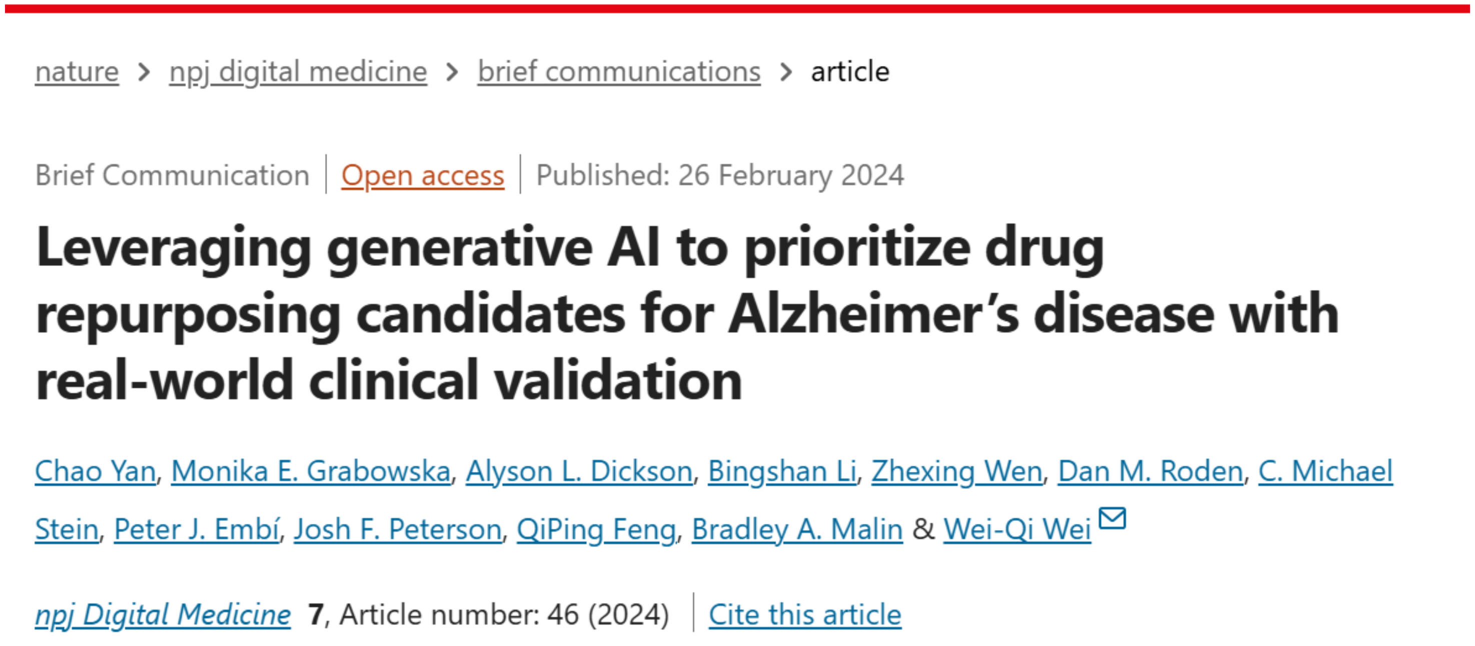


“I went line by line of everything that was in his [MRI notes] and plugged it into ChatGPT,” Courtney said. “I put the note in there about how he wouldn’t sit crisscross[ed]...To me, that was a huge trigger [that] a structural thing could be wrong.”

Trying ChatGPT led Courtney to discover tethered cord syndrome, which is a complication of spina bifida. She made an appointment with a new neurosurgeon who confirmed that Alex did have a tethered spinal cord as a result of spina bifida occulta, a birth defect that causes issues with spinal cord development. This is the mildest form of spina bifida, per the

https://www.independent.co.uk/news/health/chatgpt-diagnosis-spina-bifida-mother-son-b2410361.html

# Opening Story - Research



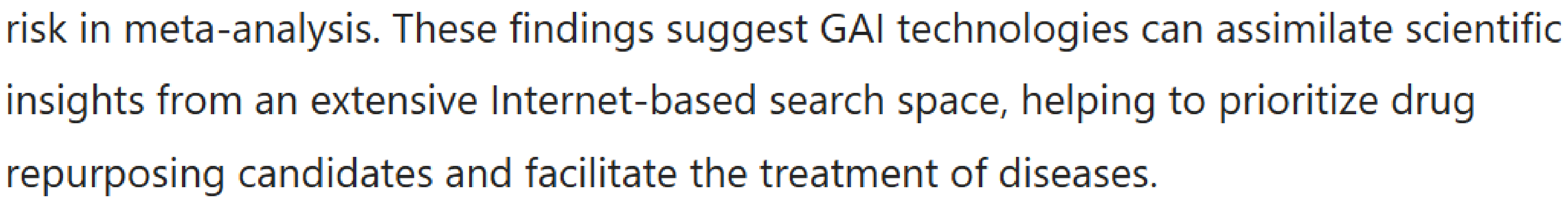
nature › npj digital medicine › brief communications › article

Brief Communication | Open access | Published: 26 February 2024

**Leveraging generative AI to prioritize drug repurposing candidates for Alzheimer's disease with real-world clinical validation**

Chao Yan, Monika E. Grabowska, Alyson L. Dickson, Bingshan Li, Zhexing Wen, Dan M. Roden, C. Michael Stein, Peter J. Embí, Josh F. Peterson, QiPing Feng, Bradley A. Malin & Wei-Qi Wei ✉

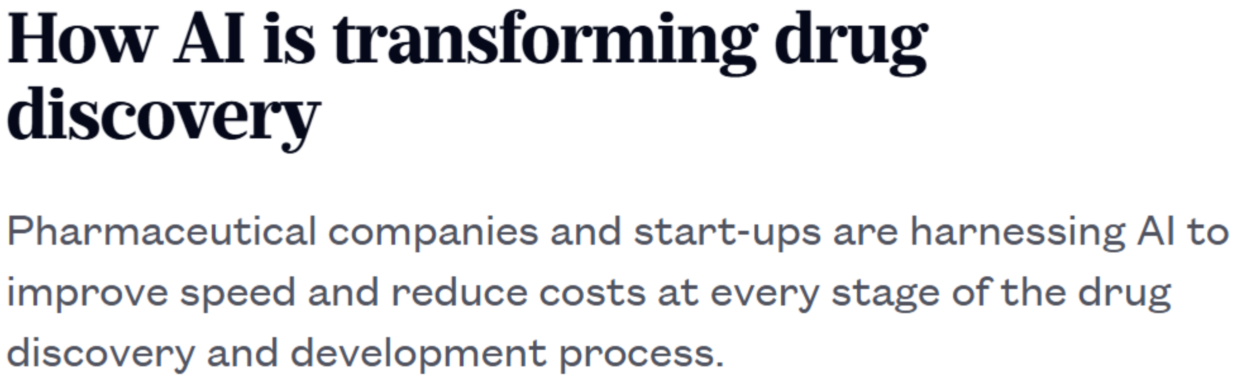
*npj Digital Medicine* **7**, Article number: 46 (2024) | Cite this article



risk in meta-analysis. These findings suggest GAI technologies can assimilate scientific insights from an extensive Internet-based search space, helping to prioritize drug repurposing candidates and facilitate the treatment of diseases.

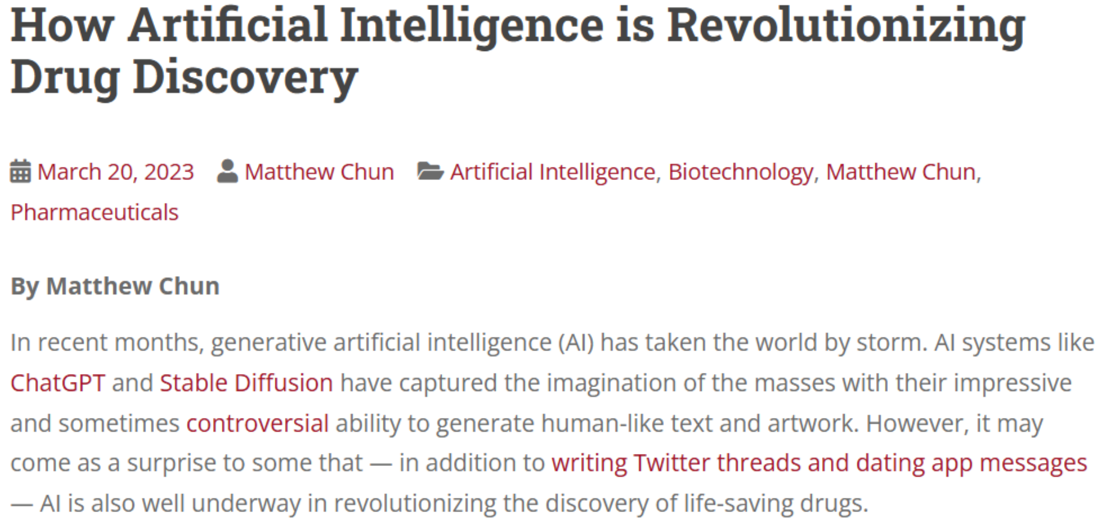
https://www.nature.com/articles/s41746-024-01038-3

# Opening Story - Business



## How AI is transforming drug discovery

Pharmaceutical companies and start-ups are harnessing AI to improve speed and reduce costs at every stage of the drug discovery and development process.



## How Artificial Intelligence is Revolutionizing Drug Discovery

March 20, 2023 Matthew Chun Artificial Intelligence, Biotechnology, Matthew Chun, Pharmaceuticals

**By Matthew Chun**

In recent months, generative artificial intelligence (AI) has taken the world by storm. AI systems like ChatGPT and Stable Diffusion have captured the imagination of the masses with their impressive and sometimes controversial ability to generate human-like text and artwork. However, it may come as a surprise to some that — in addition to writing Twitter threads and dating app messages — AI is also well underway in revolutionizing the discovery of life-saving drugs.

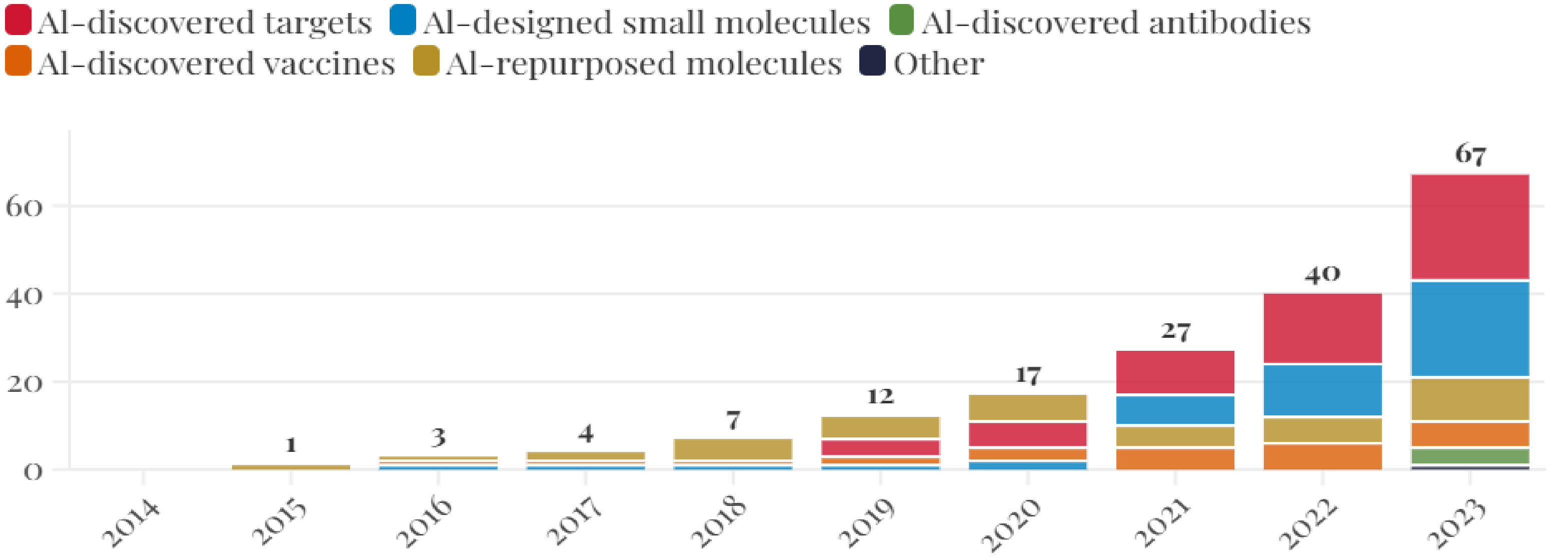
https://blog.petrieflom.law.harvard.edu/2023/03/20/how-artificial-intelligence-is-revolutionizing-drug-discovery/

https://pharmaceutical-journal.com/article/feature/how-ai-is-transforming-drug-discovery

# Opening Story - Business

## Figure: Growth in number of AI-discovered drugs in clinical trials

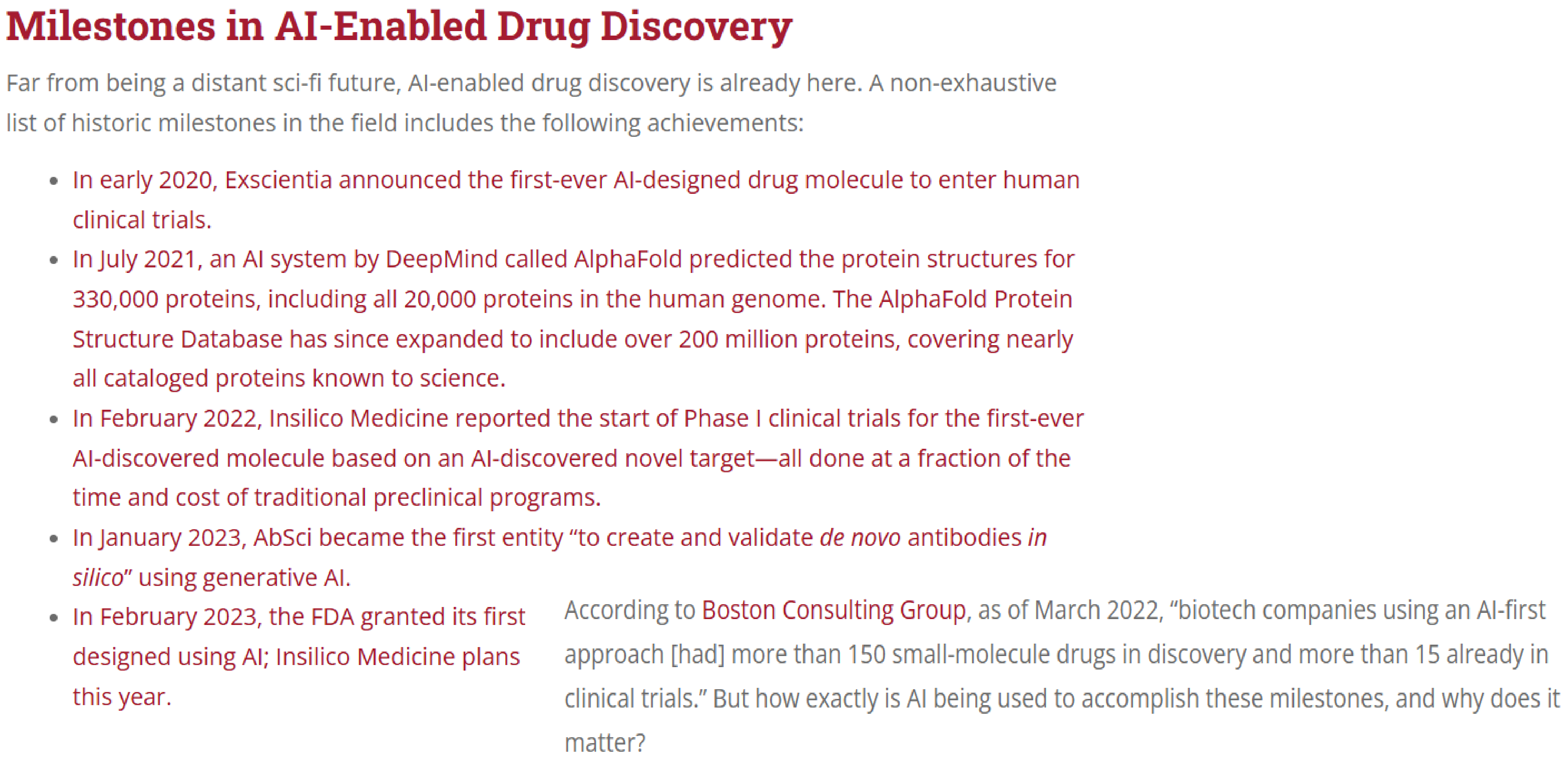
In the past decade, the number of AI-discovered drugs has increased exponentially, with 46 reaching phase II and III clinical trials in 2023. AI-repurposed drugs dominated before 2020 but in recent years there has been an increasingly diverse range of modes of discovery.



Source: Drug Discovery Today 2024;29 (6):104009 •

https://pharmaceutical-journal.com/article/feature/how-ai-is-transforming-drug-discovery

# Opening Story - Business



## Milestones in AI-Enabled Drug Discovery

Far from being a distant sci-fi future, AI-enabled drug discovery is already here. A non-exhaustive list of historic milestones in the field includes the following achievements:

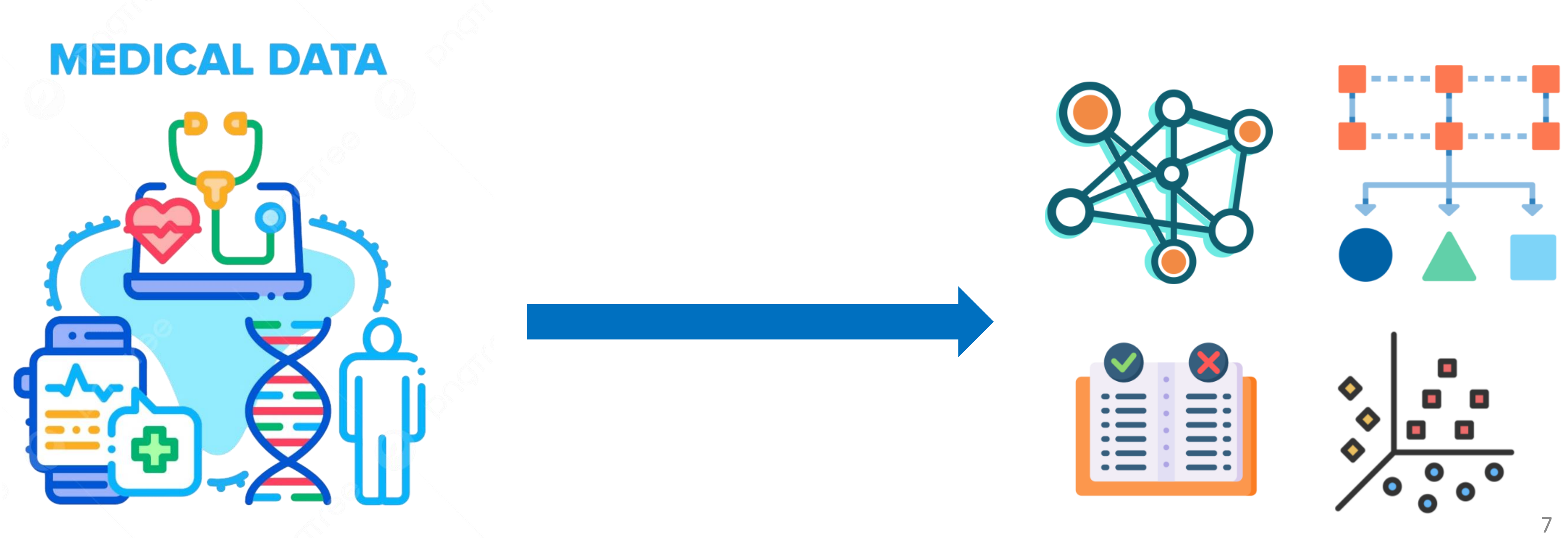
- In early 2020, Exscientia announced the first-ever AI-designed drug molecule to enter human clinical trials.
- In July 2021, an AI system by DeepMind called AlphaFold predicted the protein structures for 330,000 proteins, including all 20,000 proteins in the human genome. The AlphaFold Protein Structure Database has since expanded to include over 200 million proteins, covering nearly all cataloged proteins known to science.
- In February 2022, Insilico Medicine reported the start of Phase I clinical trials for the first-ever AI-discovered molecule based on an AI-discovered novel target—all done at a fraction of the time and cost of traditional preclinical programs.
- In January 2023, AbSci became the first entity "to create and validate *de novo* antibodies *in silico*" using generative AI.
- In February 2023, the FDA granted its first designed using AI; Insilico Medicine plans this year.

According to Boston Consulting Group, as of March 2022, "biotech companies using an AI-first approach [had] more than 150 small-molecule drugs in discovery and more than 15 already in clinical trials." But how exactly is AI being used to accomplish these milestones, and why does it matter?

# Biomedical Knowledge Discovery

- Extracting meaningful patterns, relationships, and insights from vast and complex biomedical data sources. Transforming **unstructured** biomedical data into **structured**, **actionable** knowledge.

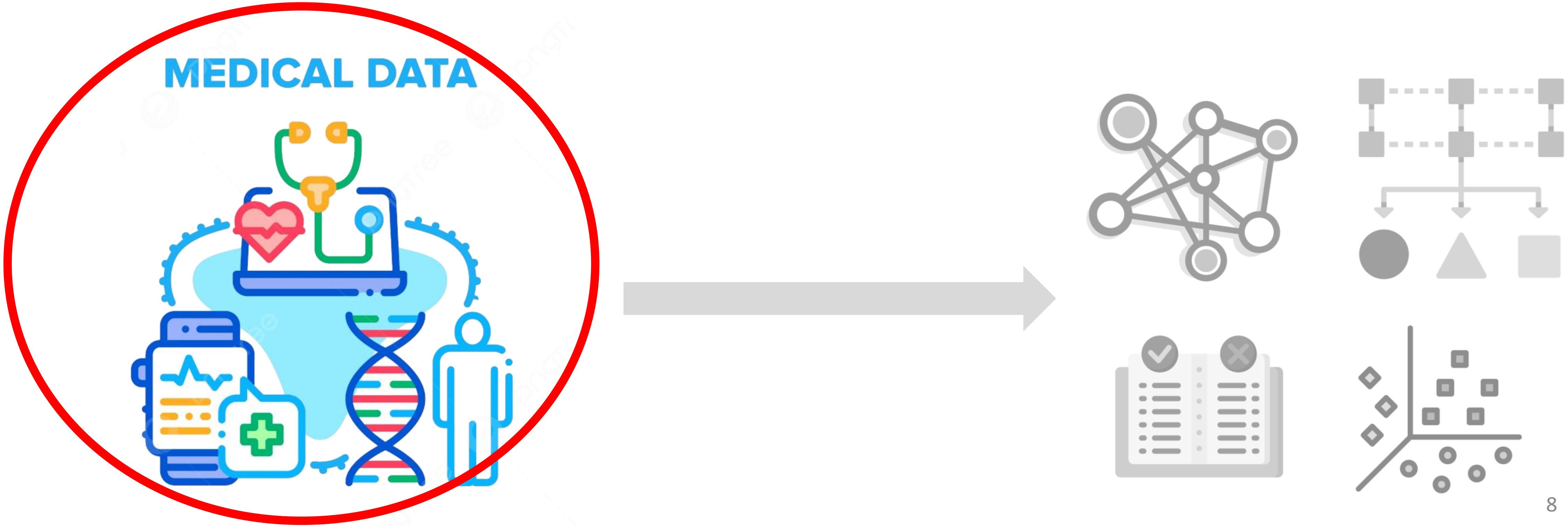
- Converting raw biomedical information into actionable knowledge that can support various applications within healthcare and life sciences.



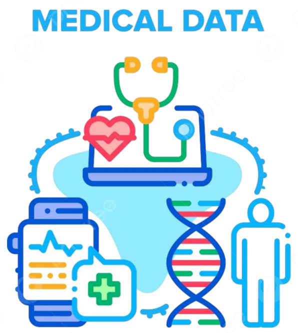
# Biomedical Knowledge Discovery

- Extracting meaningful patterns, relationships, and insights from vast and complex biomedical data sources. Transforming **unstructured** biomedical data into **structured**, **actionable** knowledge.

- Converting raw biomedical information into actionable knowledge that can support various applications within healthcare and life sciences.



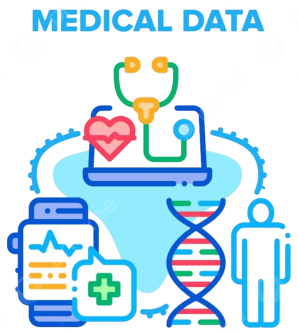
# Biomedical Data Sources



## Literature and Textual Databases

- **PubMed/MEDLINE**: A comprehensive database of biomedical and life sciences literature, providing access to research papers, clinical studies, and review articles.
- **ClinicalTrials.gov**: A registry and database of clinical trials, offering details on study design, interventions, and results.
- **PubChem**: Provides information on the biological activities of small molecules, including compound structures, properties, and biological assay results.

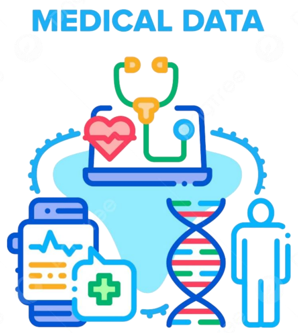
# Biomedical Data Sources



## ➢ Clinical and Patient Data

- **Electronic Health Records (EHRs)**: Patient records maintained by hospitals and clinics, containing medical histories, diagnoses, treatment plans, and lab results.
- **OMOP Common Data Model**: A standardized data model for organizing and analyzing observational health data from multiple sources like EHRs and insurance claims.
- **The MIMIC-III Database**: Contains de-identified health data of ICU patients, widely used in clinical research for analyzing treatment outcomes and disease progression.

# Biomedical Data Sources



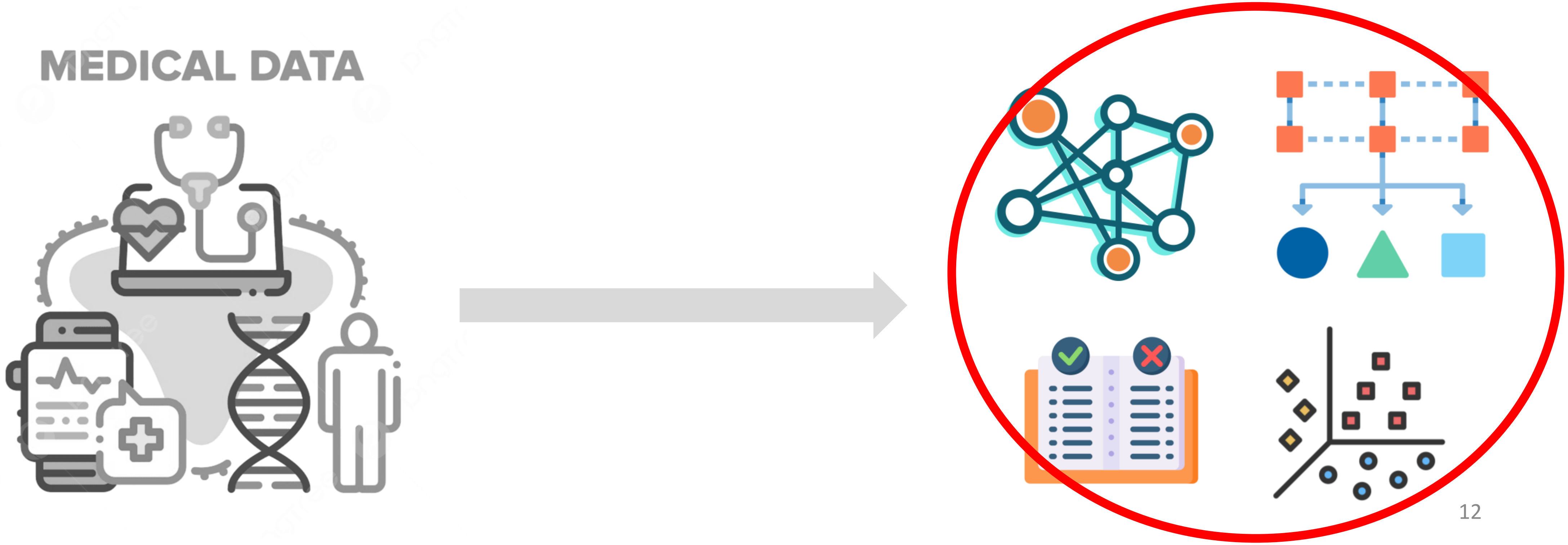
## ➢ Pharmacological and Drug Databases & Genomic and Molecular Databases

- **DrugBank**: A resource combining detailed drug information with comprehensive drug target data, including mechanisms, interactions, and side effects.
- **ChEMBL**: A database of bioactive molecules with drug-like properties, used for bioactivity, drug discovery, and pharmacology research.
- **UniProt**: A comprehensive protein sequence and functional information database, offering details on protein structures, functions, and interactions.
- **GenBank**: A repository of nucleotide sequences and supporting bibliographic and biological annotation, managed by NCBI.

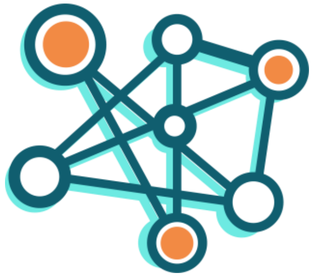
# Biomedical Knowledge Discovery

- Extracting meaningful patterns, relationships, and insights from vast and complex biomedical data sources. Transforming **unstructured** biomedical data into **structured**, **actionable** knowledge.

- Converting raw biomedical information into actionable knowledge that can support various applications within healthcare and life sciences.



# Biomedical Actionable Knowledge



- **Patterns and Trends**: Recognizable patterns in the data, such as correlations, associations, and time-based trends, which can reveal insights like disease prevalence over time, patient responses to treatments, or relationships between different genes and diseases.

- **Rules and Associations**: Formalized rules or associations derived from the data, such as "if-then" statements or statistical correlations. For instance, a rule might indicate that patients with a certain gene variant are more likely to respond to a specific drug.

- **Relationships and Links**: Connections between different entities, such as relationships between diseases and genes, drugs and side effects, or treatments and outcomes. These relationships are particularly useful in building **knowledge graphs** or relational databases.

# Biomedical Actionable Knowledge

- **Clusters and Groupings**: Groups or clusters of data that reveal similarities among entities (like patients, symptoms, or treatments) based on their characteristics. This is often used to classify patients into subtypes or categorize diseases with similar features.

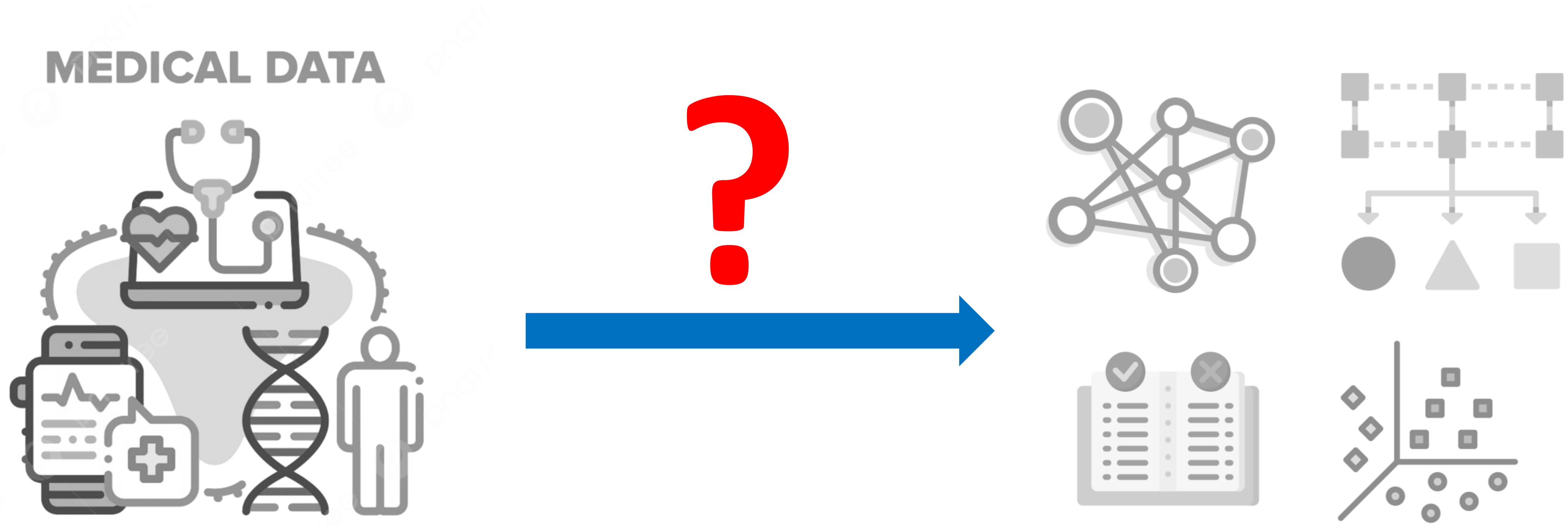
- **Predictions**: Models or insights that can predict outcomes, such as disease progression, treatment responses, or risk factors for certain conditions. These predictive outputs help in making informed clinical or research-based decisions.

- **Summaries and Reports**: Condensed summaries of key insights, often in the form of visualizations, charts, or textual reports, that highlight the main findings. These can support decision-makers by providing a clear overview of the discovered knowledge.

# Biomedical Knowledge Discovery

- The need for **tools/techniques** to transform data into actionable insights.

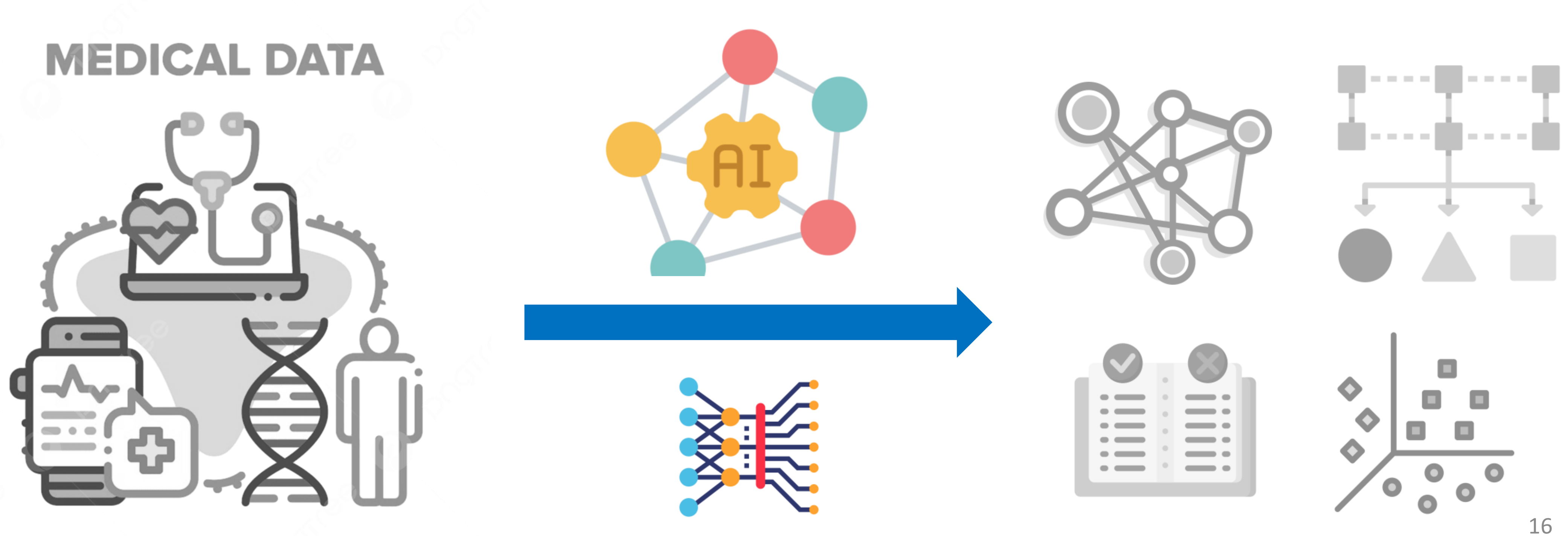
- Goal : Accelerates literature review, personalizes treatment options, supports early diagnosis, improves risk prediction and patient monitoring, drug development, etc.



# Biomedical Knowledge Discovery

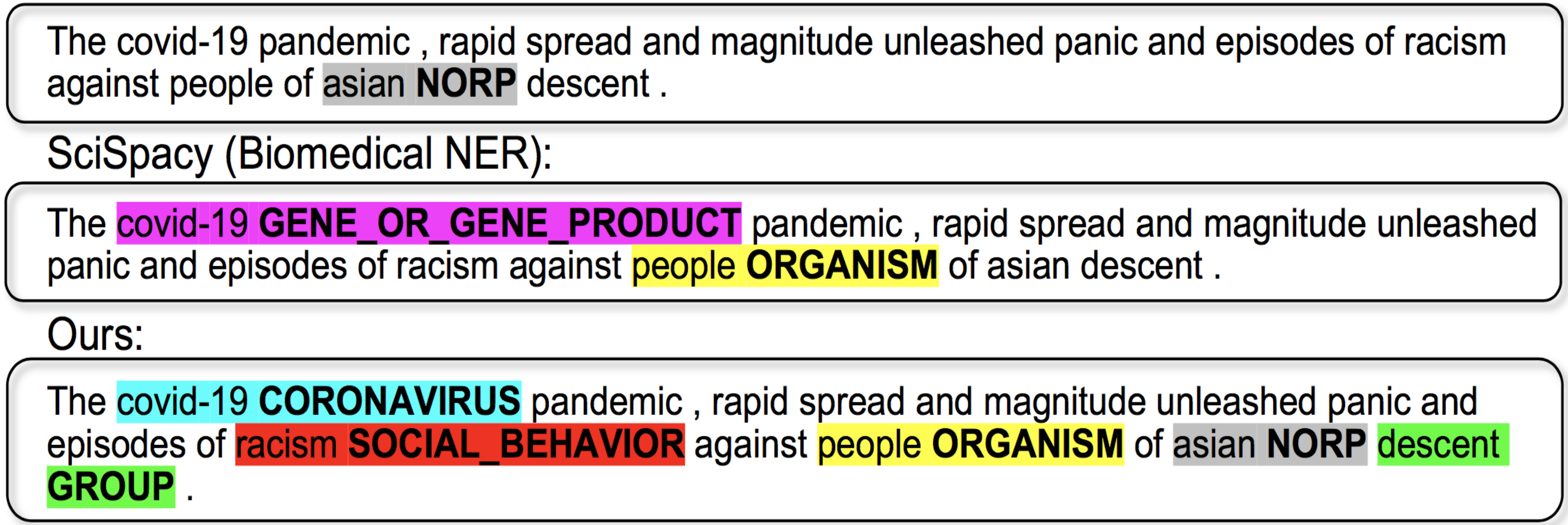
- The need for **tools/techniques** to transform data into actionable insights.

- Goal : Accelerates literature review, personalizes treatment options, supports early diagnosis, improves risk prediction and patient monitoring, drug development, etc.



# Text Mining and NLP in the Biomedical Domain

- Text Mining and NLP Taks: **Named Entity Recognition (BioNER)**



The covid-19 pandemic , rapid spread and magnitude unleashed panic and episodes of racism against people of asian **NORP** descent .

SciSpacy (Biomedical NER):

The covid-19 **GENE_OR_GENE_PRODUCT** pandemic , rapid spread and magnitude unleashed panic and episodes of racism against people **ORGANISM** of asian descent .

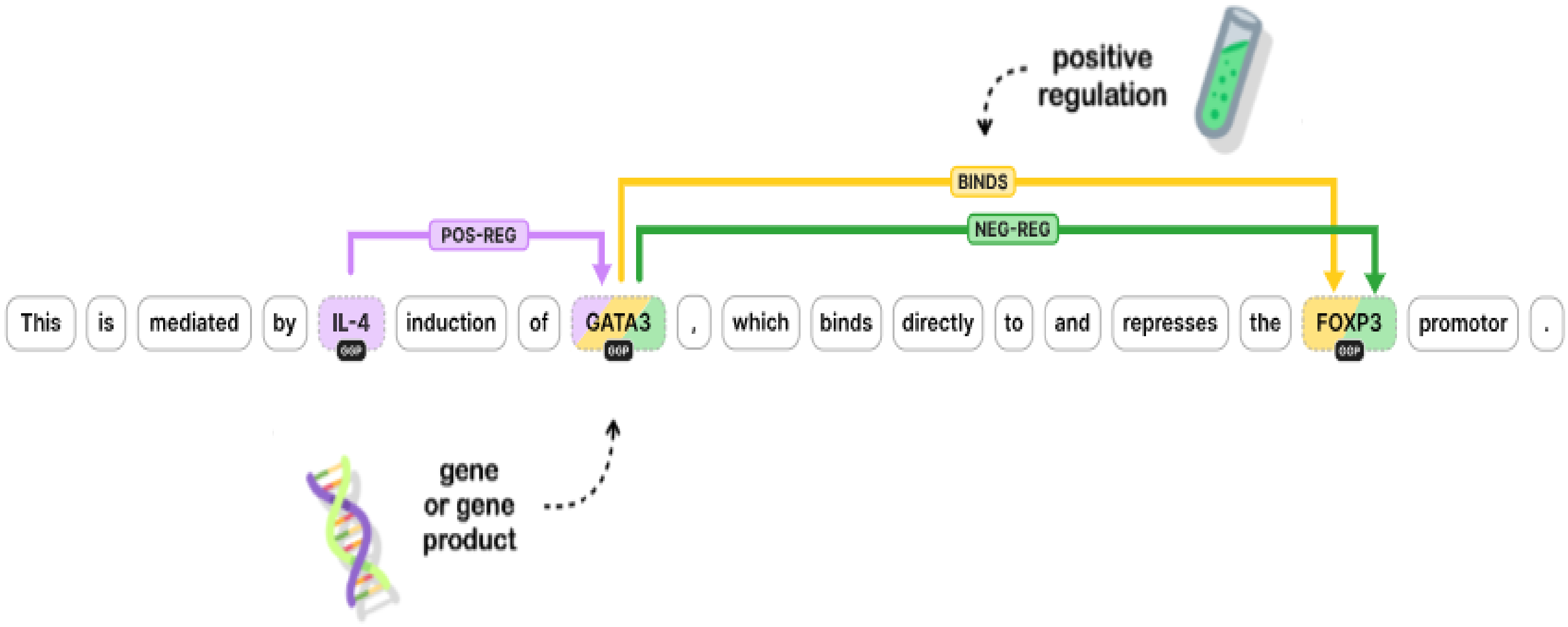
Ours:

The covid-19 **CORONAVIRUS** pandemic , rapid spread and magnitude unleashed panic and episodes of racism **SOCIAL_BEHAVIOR** against people **ORGANISM** of asian **NORP** descent **GROUP** .

1. Genes
2. Proteins
3. Diseases
4. Drugs
5. Chemical compounds
6. Cells
7. Biomarkers
8. Symptoms

# Text Mining and NLP in the Biomedical Domain

- Text Mining and NLP **Taks** : **Relation Extraction (BioRE)**

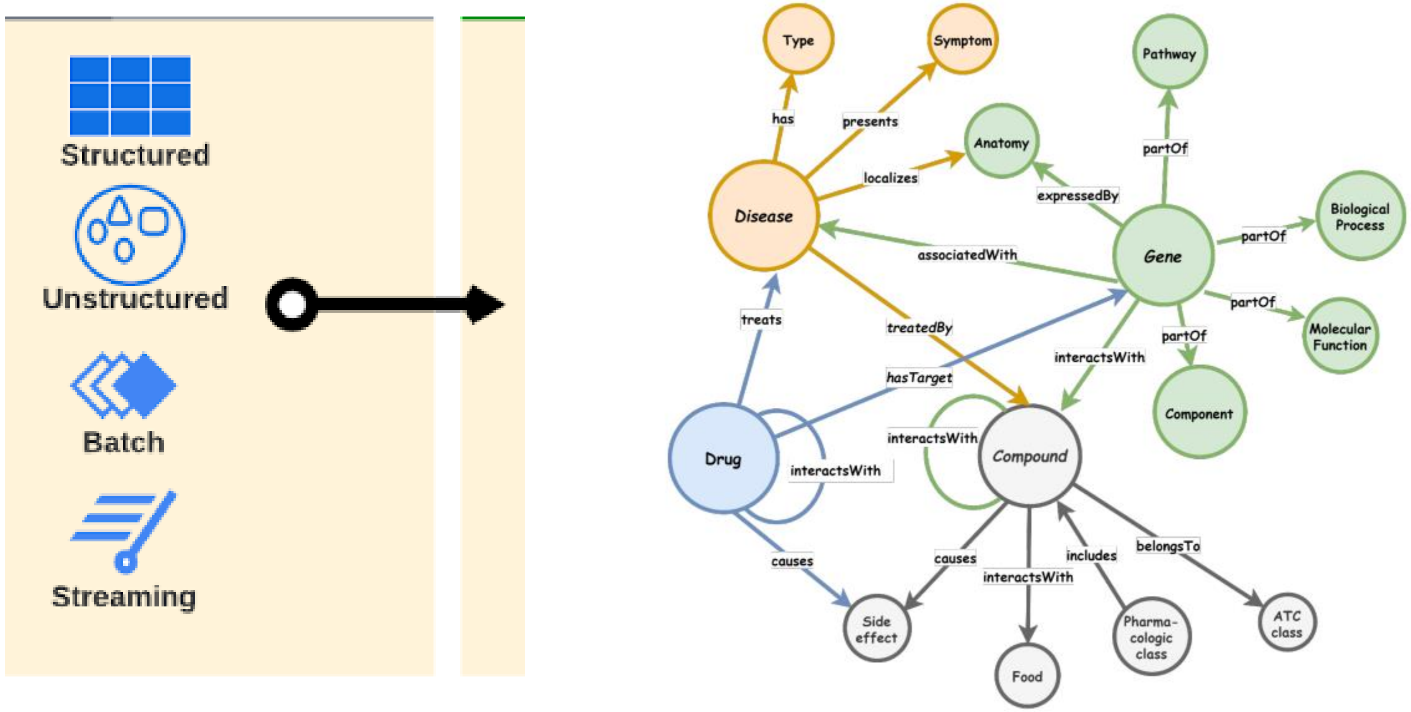


- Disease-Drug Interaction
- Gene-Disease Association
- Drug-Protein Interaction
- Chemical-Disease Relationship
- Gene-Gene Interaction
- Chemical-Protein Interaction
- Drug-Drug Interaction
- Disease-Treatment Effect
- Drug-Side Effect
- Drug-Target Binding

# Text Mining and NLP in the Biomedical Domain

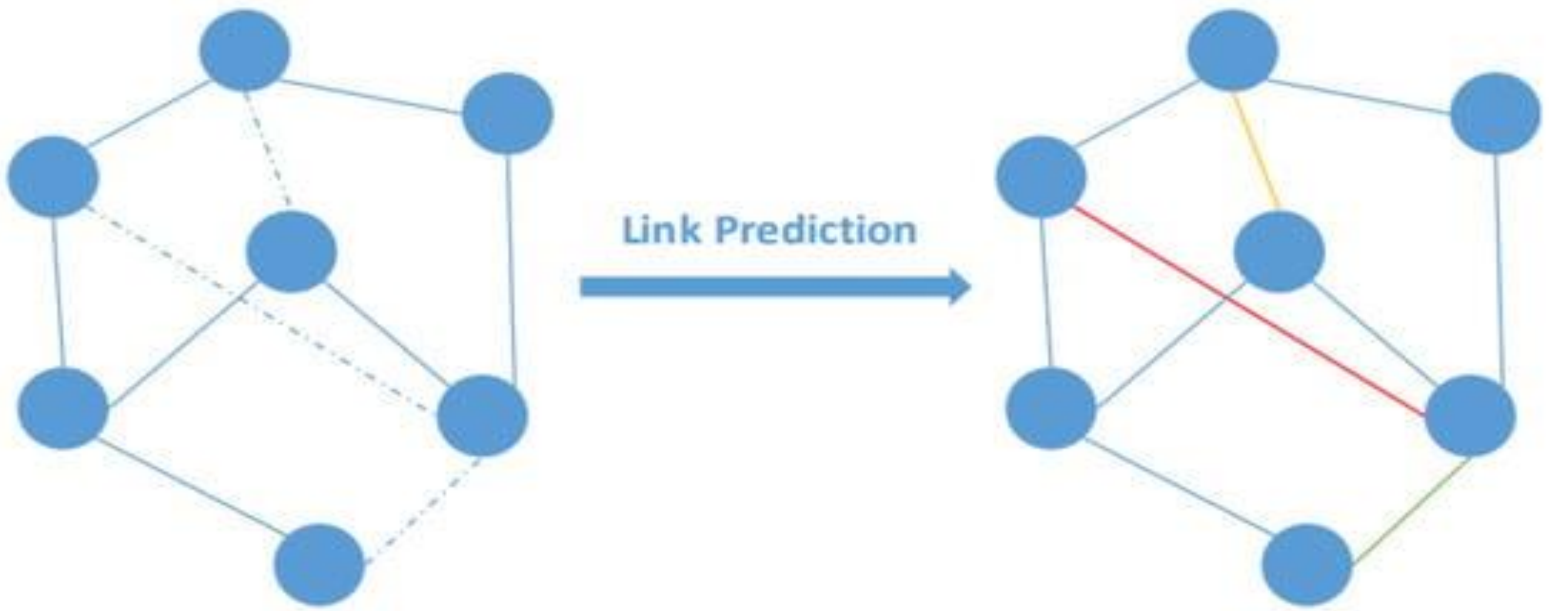
- Text Mining and NLP **Taks** : **Knowledge Graph Construction**

- Capturing relationships and entities (such as genes, diseases, proteins, drugs, and their interactions) from unstructured data sources and integrating them into a unified knowledge graph.



# Text Mining and NLP in the Biomedical Domain

- Text Mining and NLP **Taks** : **Link Prediction (BioLP)**
- Predicting missing relationships (or links) between entities in a graph, based on the existing graph structure. Discovering novel or unobserved relationships between biomedical entities.

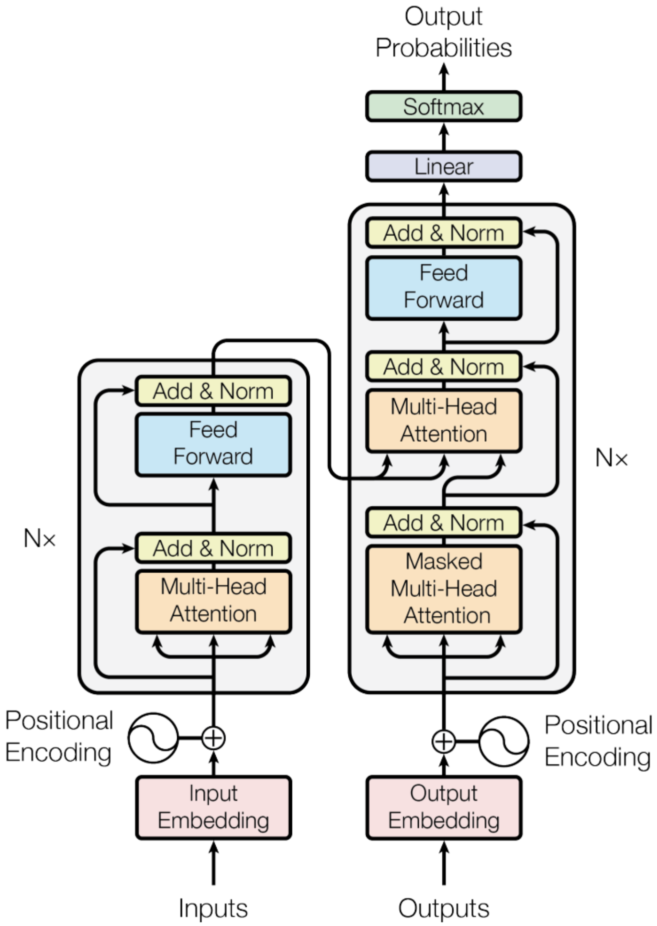


# NLP Methods and Models

➢ **Transformer-Based Models for Biomedical NLP**

BERT

Encoder



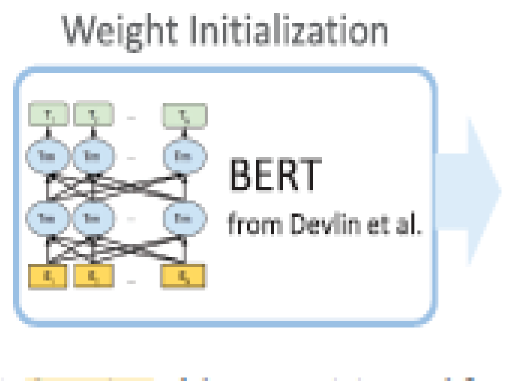
GPT

Decoder

# NLP Methods and Models

## ➢ Transformer-Based Models for Biomedical NLP

- Pretraining BERT for Biomedical Text: BioBERT, SciBERT, PubMedBERT, BioM-BERT, ClinicalBERT.

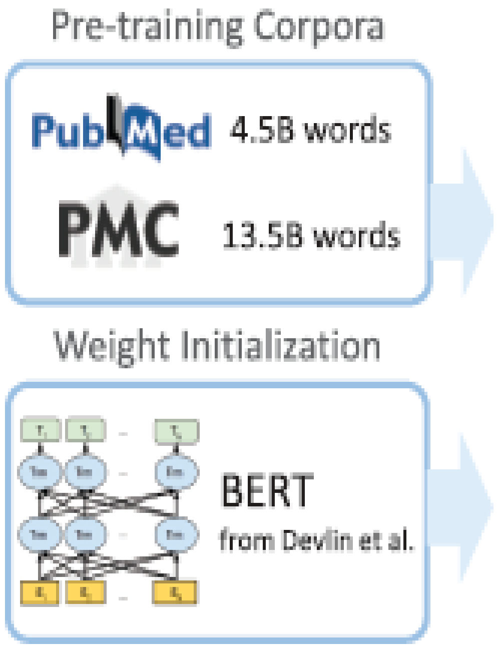


https://arxiv.org/abs/1901.08746

# NLP Methods and Models

- **Transformer-Based Models for Biomedical NLP**

- Pretraining BERT for Biomedical Text: BioBERT, SciBERT, PubMedBERT, BioM-BERT, ClinicalBERT.

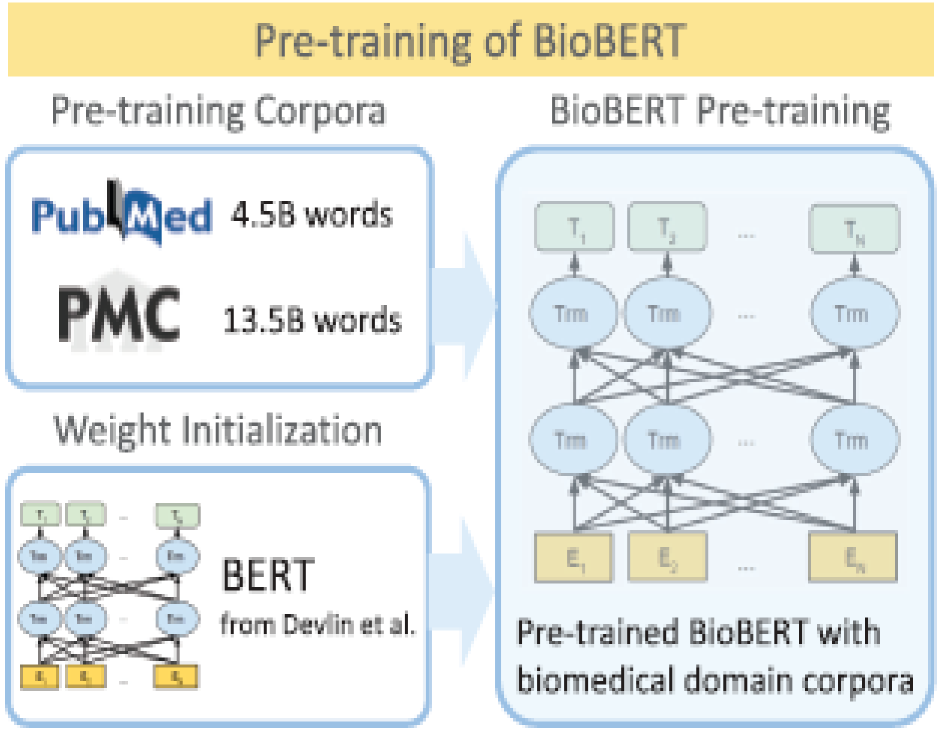


https://arxiv.org/abs/1901.08746

# NLP Methods and Models

## ➢ Transformer-Based Models for Biomedical NLP

- Pretraining BERT for Biomedical Text: BioBERT, SciBERT, PubMedBERT, BioM-BERT, ClinicalBERT.

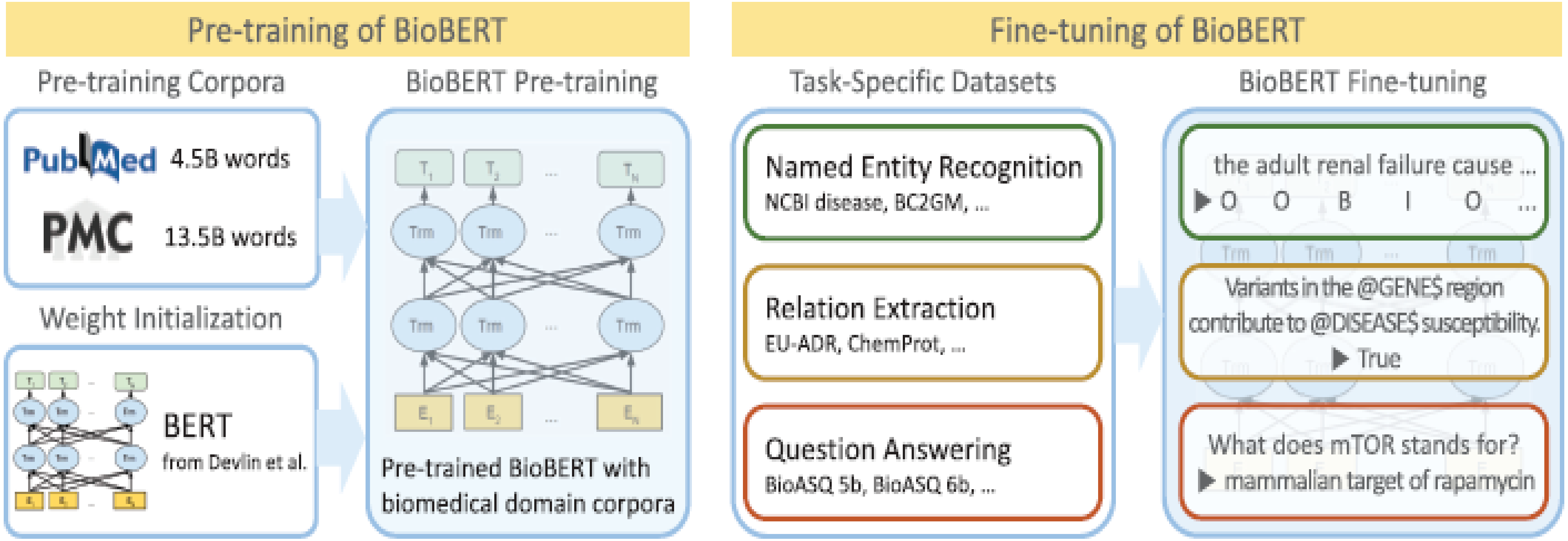


https://arxiv.org/abs/1901.08746

# NLP Methods and Models

➢ **Transfer Learning for Biomedical NLP**

- Fine-tuning for specific biomedical tasks.

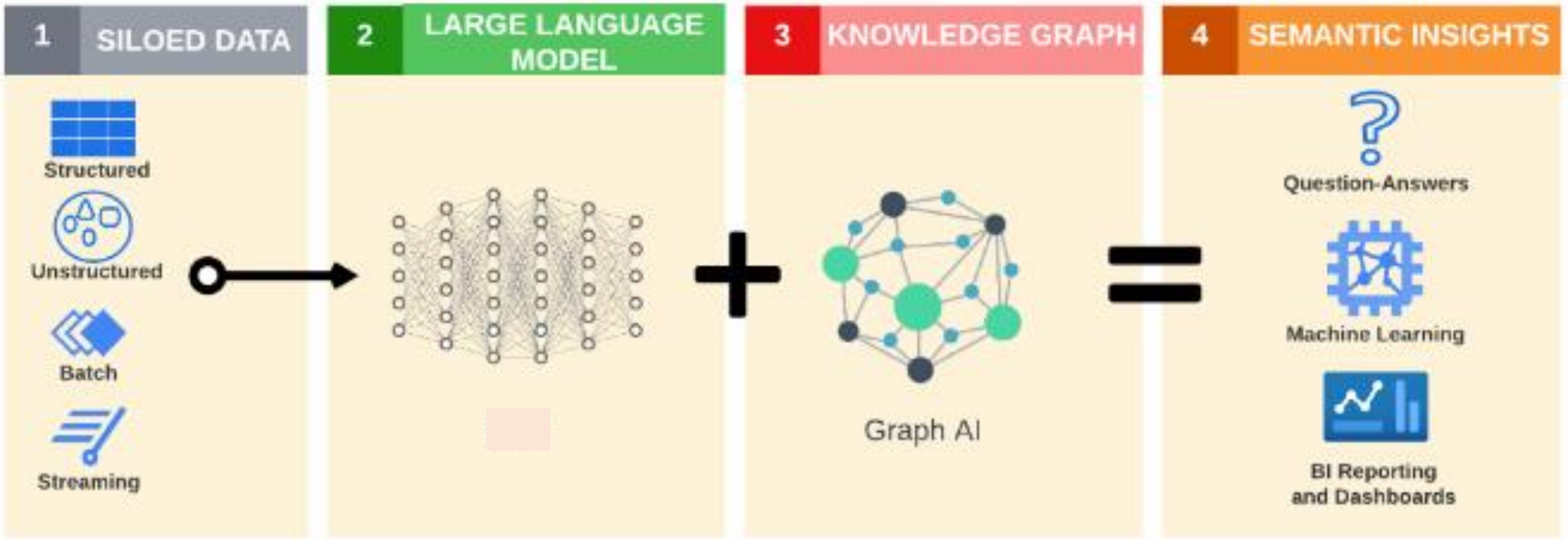


https://arxiv.org/abs/1901.08746

# NLP Methods and Models

- **Knowledge Graph Embeddings and Domain-Specific Vocabularies**

- Enriching models with additional domain-specific knowledge (Entity & Relation Enrichment).



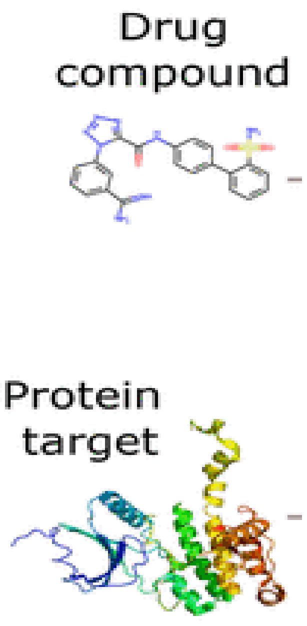
# Case Studies / Applications

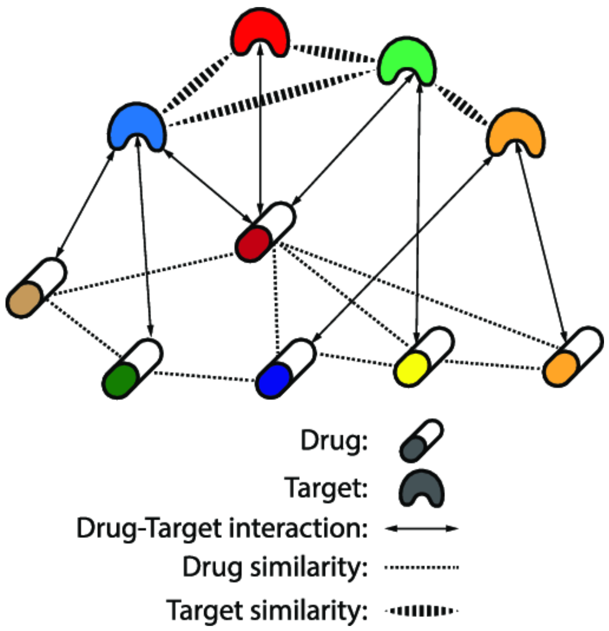
- **Drug Development** : Drug Discovery, Drug Target Identification, Drug Repurposing, etc.
- **Enhances Safety and Efficacy Analysis :** By mining adverse event reports, clinical trial data, and patient feedback, knowledge discovery can detect safety signals and efficacy trends, informing decisions about whether to pursue or halt drug development.
- **Healthcare Research:** Accelerates Literature Review, Identifies Emerging Trends and Hypotheses, Facilitates Genomic and Proteomic Insights.
- **Disease Prediction**: Risk assessment and early detection from patient data.
- **Precision Medicine.**
- **Clinical Decision-Making:** Personalizes Treatment Options, Supports Early Diagnosis, Improves Risk Prediction and Patient Monitoring.

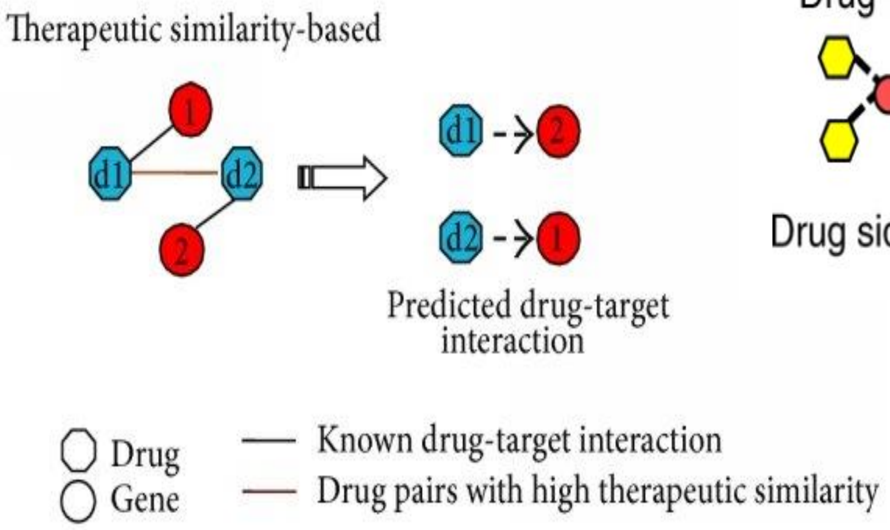
# Case Study / Application

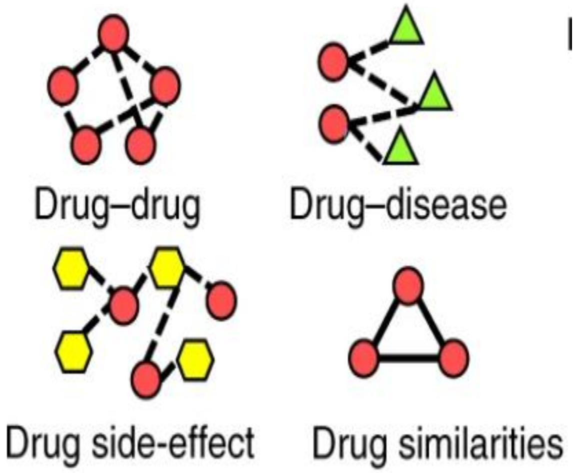
- **Drug Development** : **Drug Target Identification**, Drug Repurposing, De Novo Drug Design, etc.

- Identifying whether a particular drug interacts with a specific target - biological molecules (e.g., proteins, genes, enzymes, or RNA).

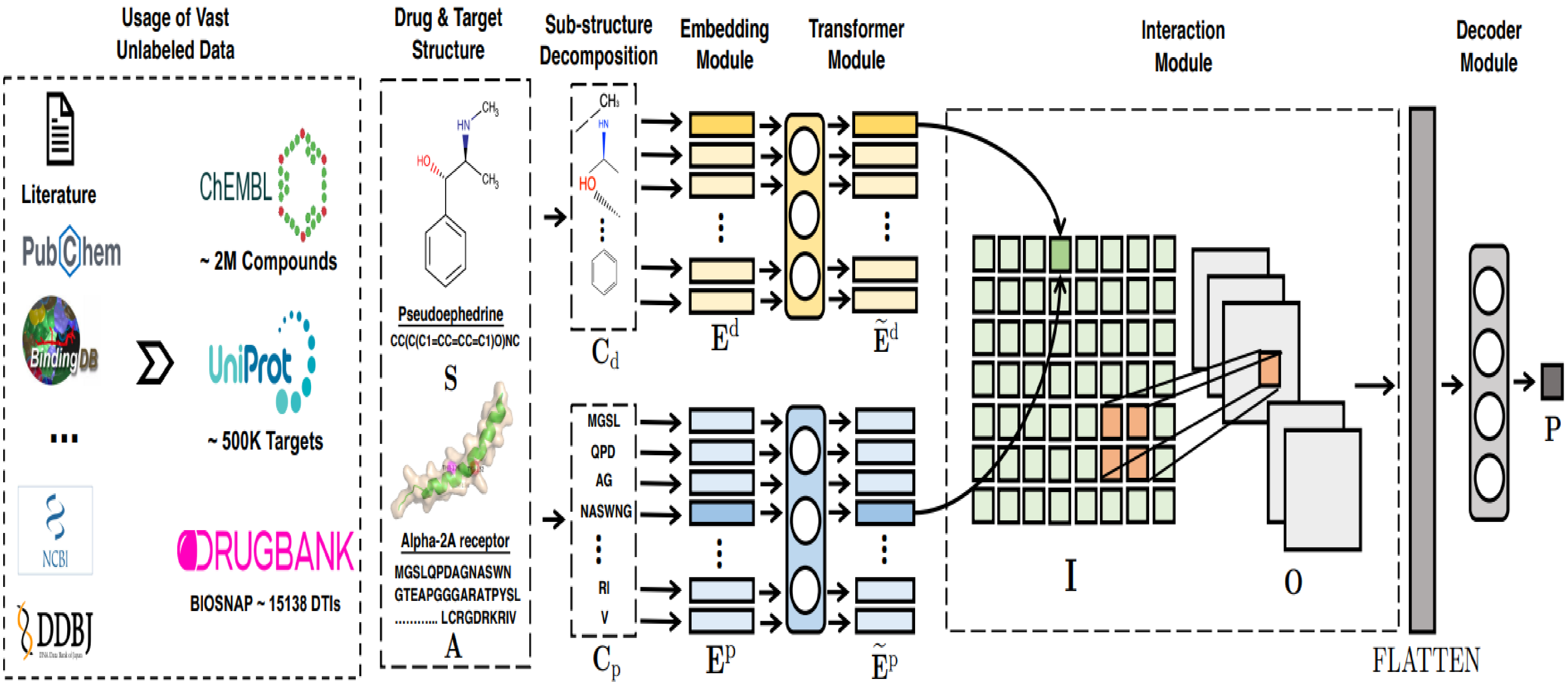




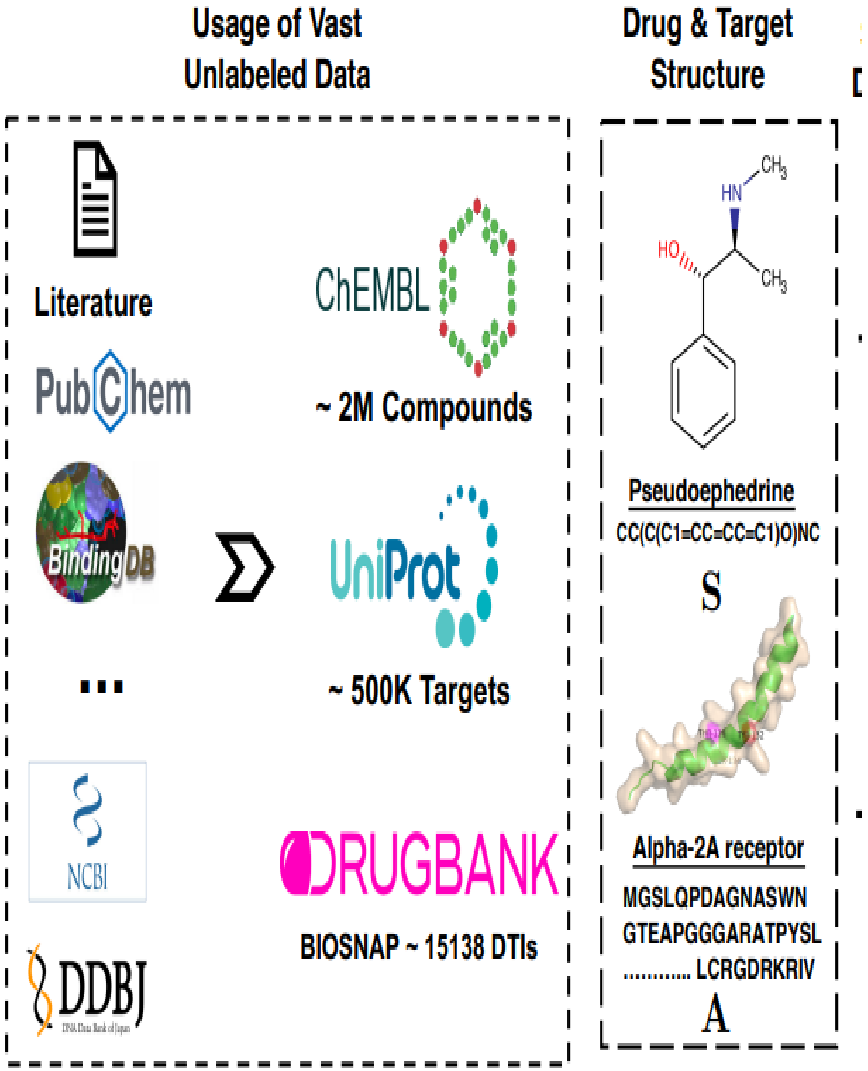

Therapeutic similarity-based
1
d1
d2
2
d1
2
d2
1
Predicted drug-target interaction
Drug
Gene
Known drug-target interaction
Drug pairs with high therapeutic similarity



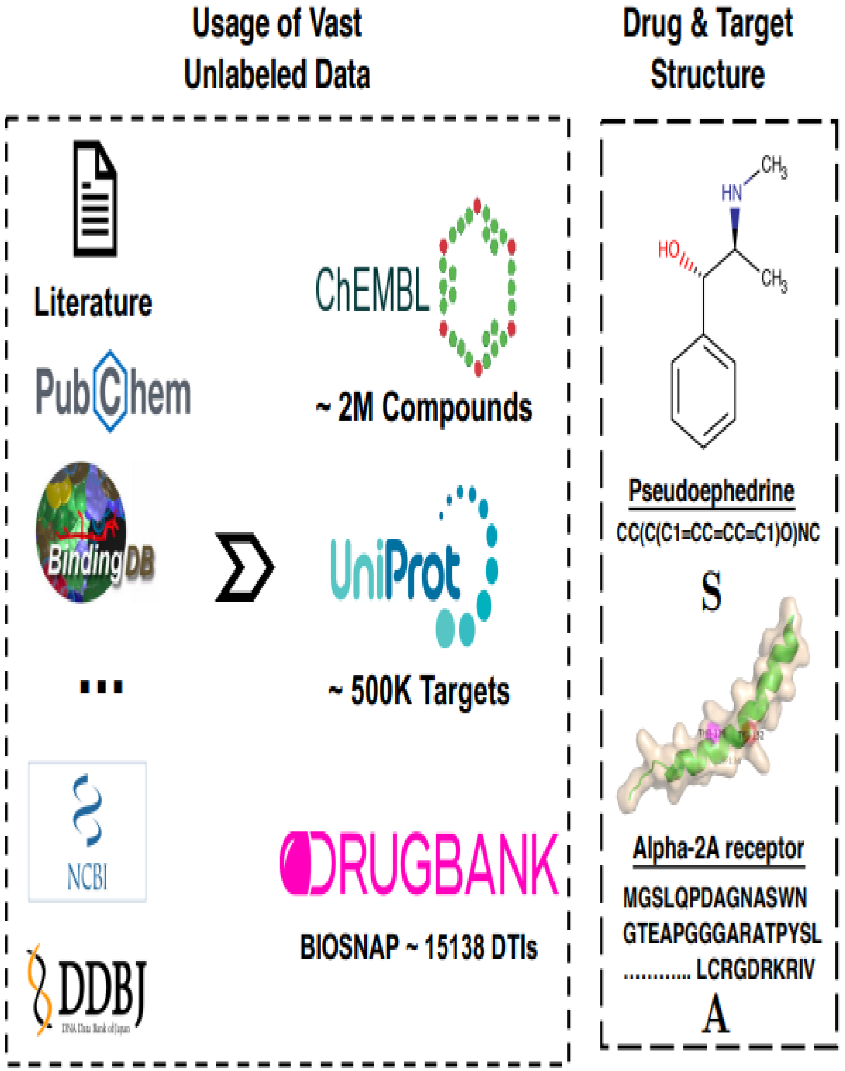
- **MolTrans**: Molecular Interaction Transformer for Drug Target Interaction Prediction https://arxiv.org/pdf/2004.11424

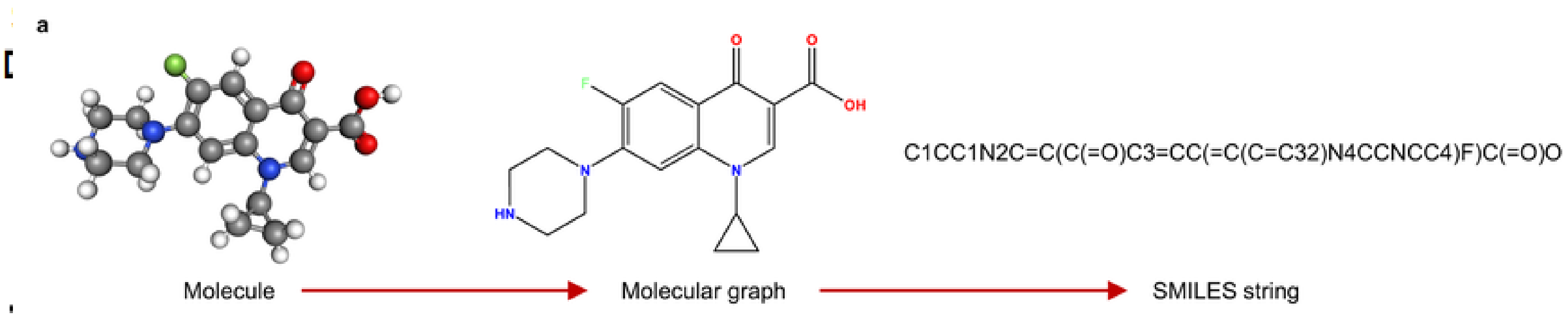


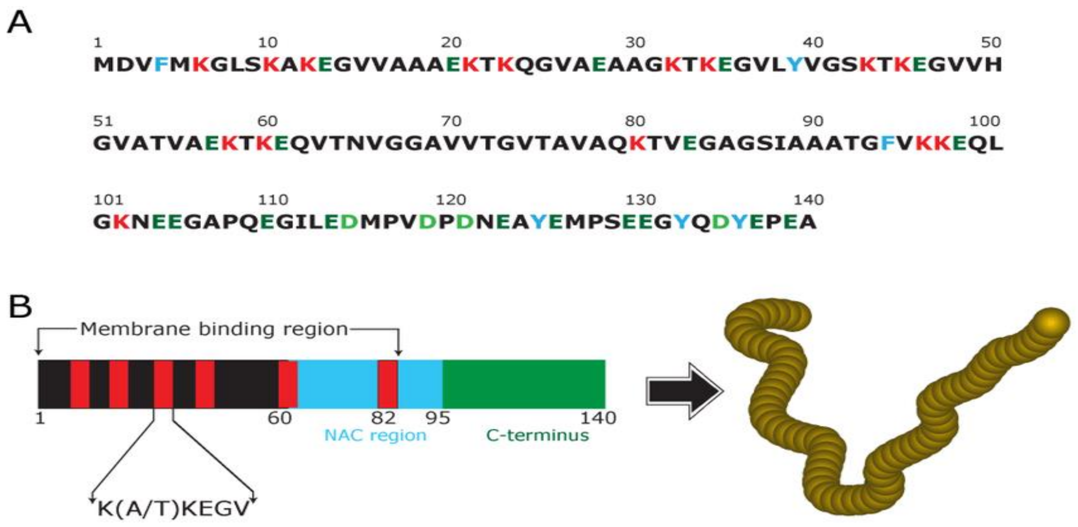
- **MolTrans**: Molecular Interaction Transformer for Drug Target Interaction Prediction https://arxiv.org/pdf/2004.11424



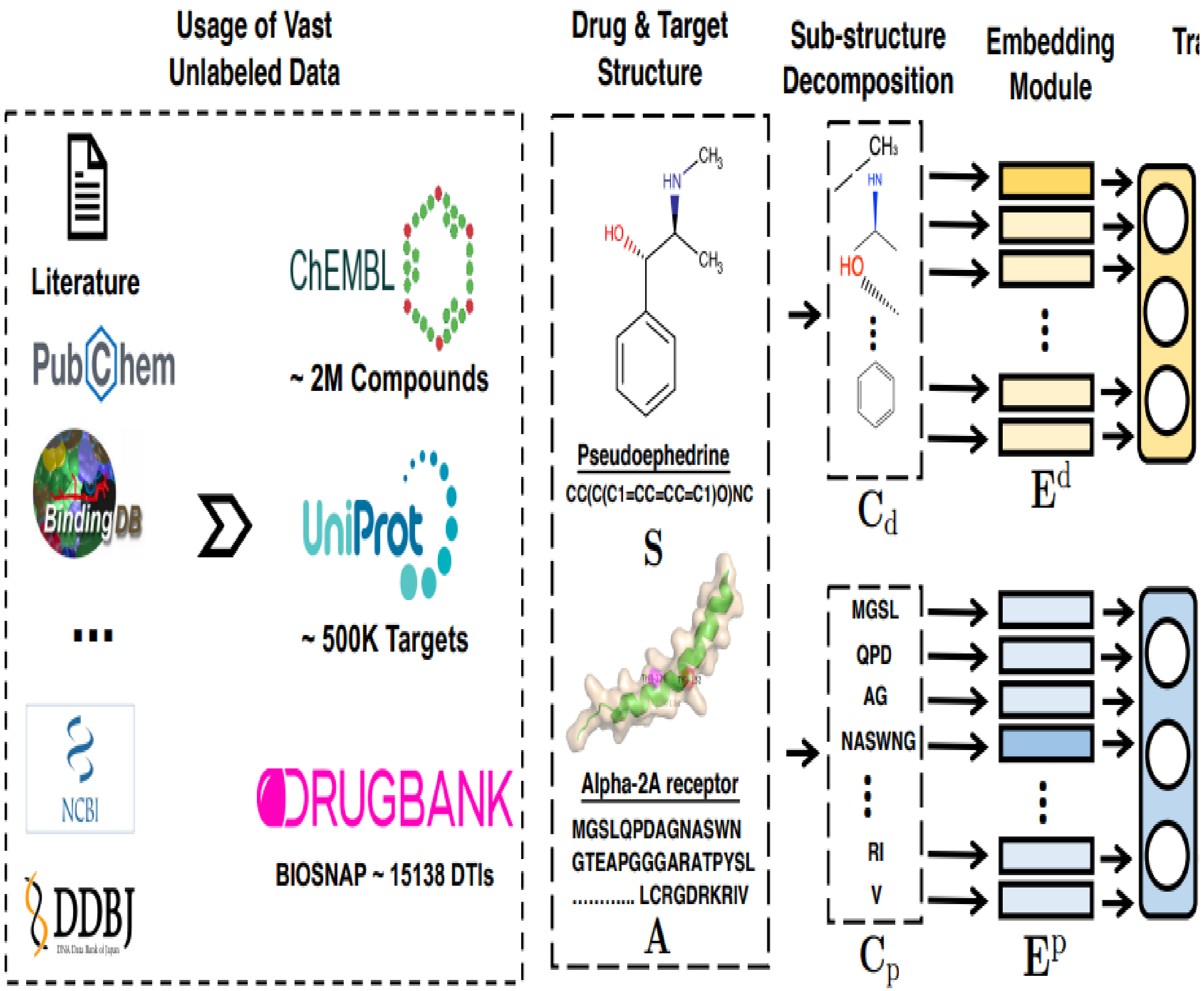
- **MolTrans**: Molecular Interaction Transformer for Drug Target Interaction Prediction https://arxiv.org/pdf/2004.11424



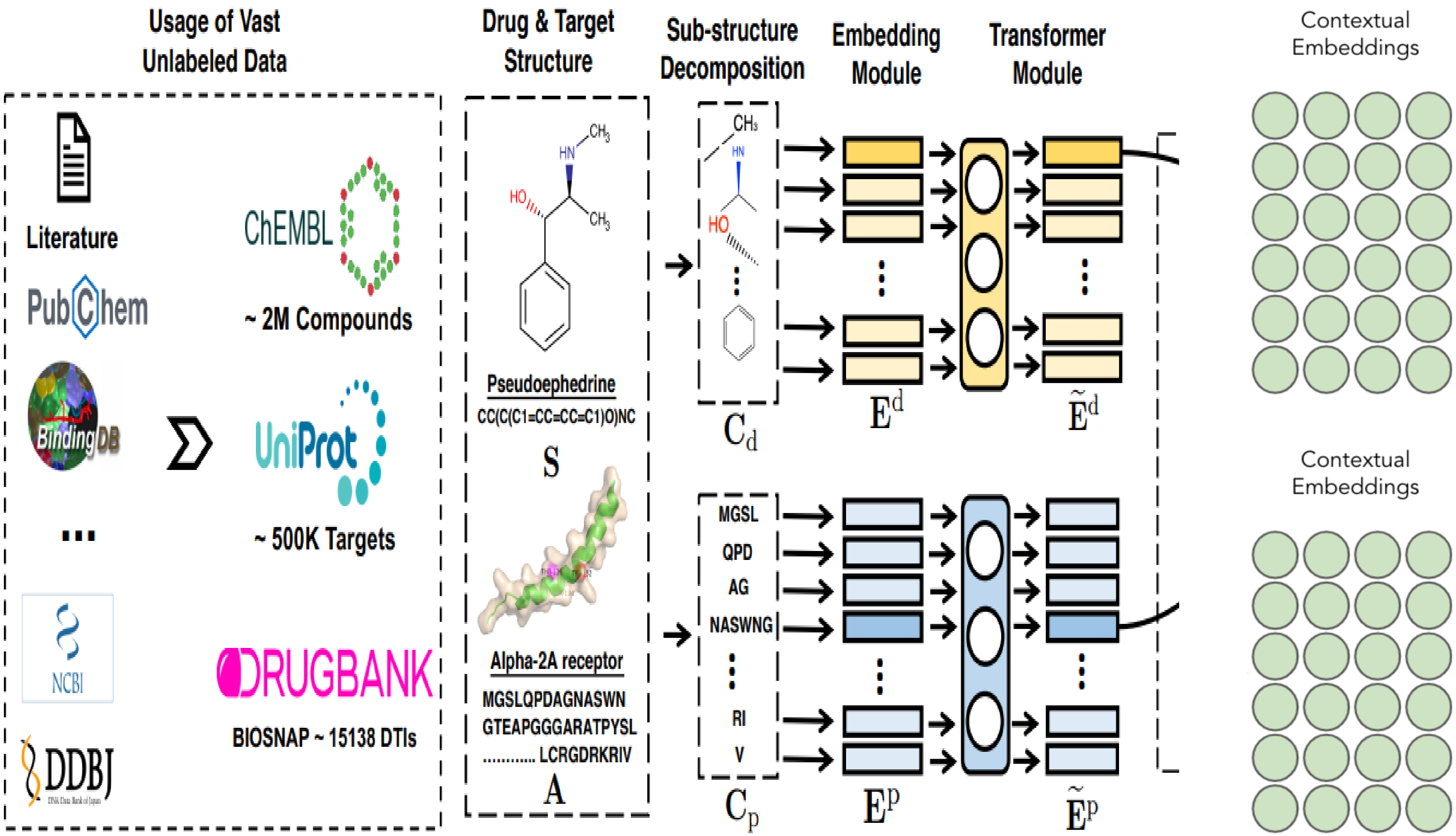



A
1
10
20
30
40
50
MDVFMKGLSKAKEGVVAAAEKTKQGVAEAAGKTKEGVLYVGSKTKEGVVH
51
60
70
80
90
100
GVATVAEKTKEQVTNVGGAVVTGVTAVAQKTVEGAGSIAAATGFVKKEQL
101
110
120
130
140
GKNEEGAPQEGILEDMPVDPDNEAYEMPSEEGYQDYEPEA
B
Membrane binding region
1
60
82
95
140
NAC region
C-terminus
K(A/T)KEGV

- **MolTrans**: Molecular Interaction Transformer for Drug Target Interaction Prediction https://arxiv.org/pdf/2004.11424



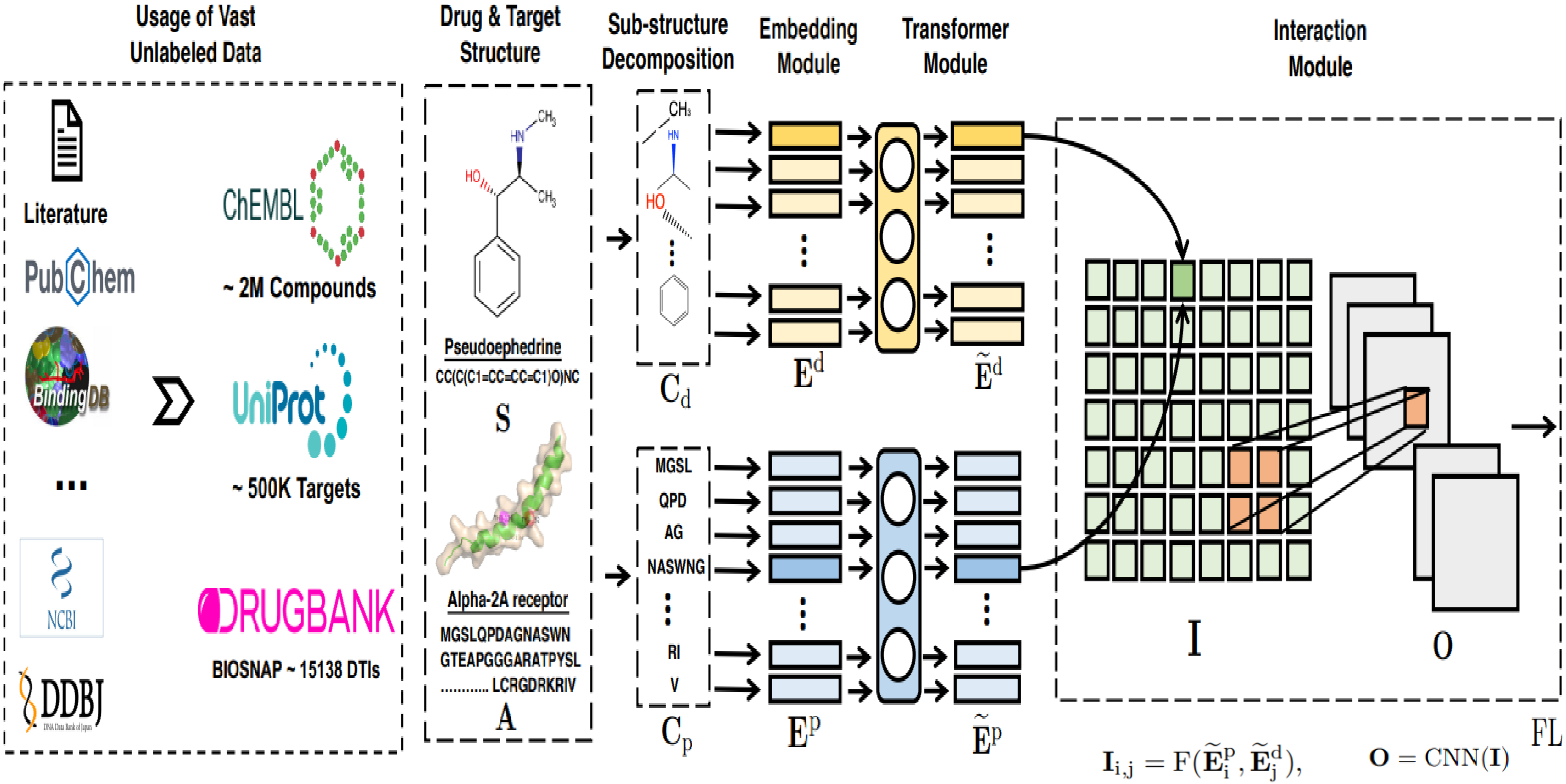
- **MolTrans**: Molecular Interaction Transformer for Drug Target Interaction Prediction https://arxiv.org/pdf/2004.11424



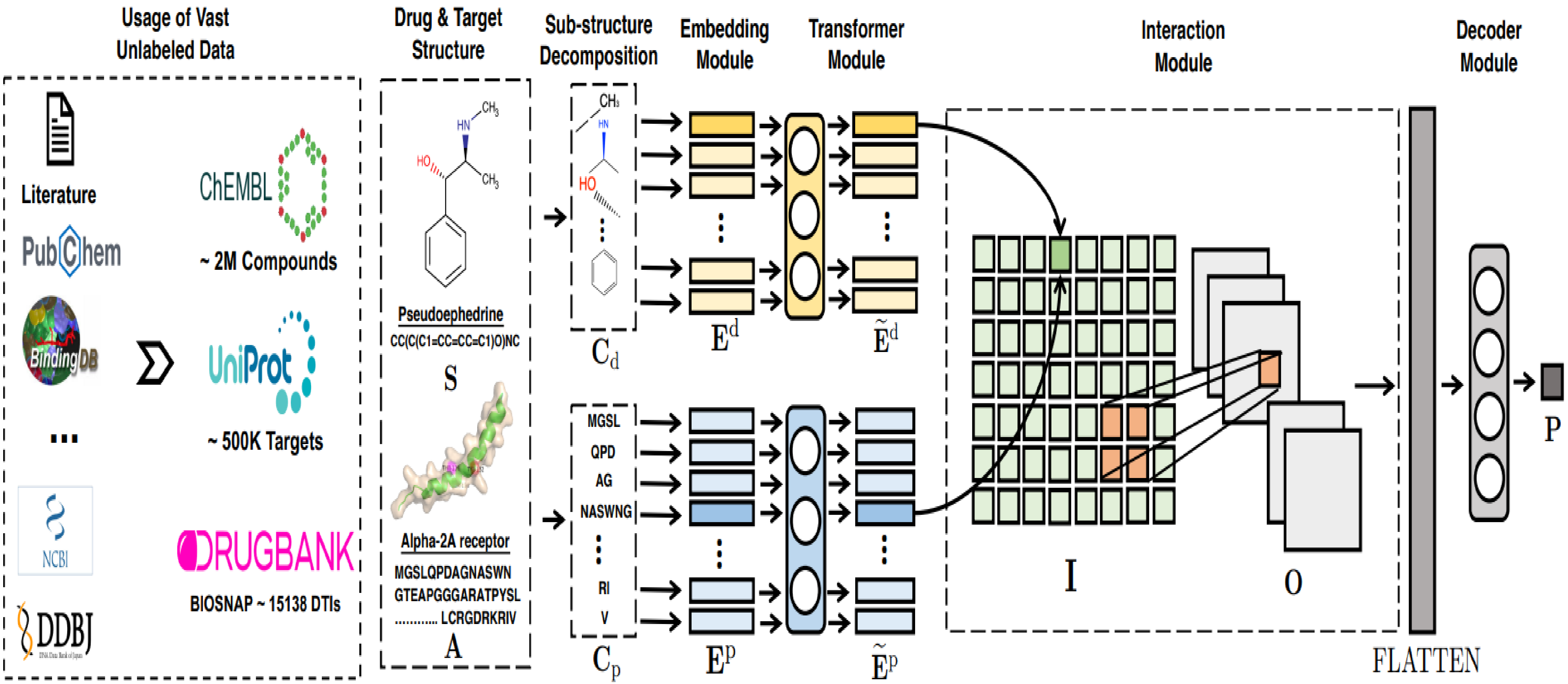
Contextual Embeddings

Contextual Embeddings

- **MolTrans**: Molecular Interaction Transformer for Drug Target Interaction Prediction https://arxiv.org/pdf/2004.11424



- **MolTrans**: Molecular Interaction Transformer for Drug Target Interaction Prediction https://arxiv.org/pdf/2004.11424



# Current Trends and Future Directions

- Knowledge discovery and extraction are transforming biomedicine.

- **Knowledge Graphs with AI**: Integration of graphs with AI models for complex queries. This combination enhances reasoning over biomedical data, enabling better insights. Ex: Integration of DrugBank and UniProt with text data to enhance prediction accuracy in drug discovery.

- **Multimodal Data**: Merging text, EHRs, images, genomics, and other data types for richer insights. Enables a more comprehensive understanding of patient conditions and biological mechanisms.

- **Explainable AI (XAI)** in Biomedical NLP: Shift toward making AI models interpretable for better adoption in clinical settings and to increase trust in critical applications. Ex: Explaining why a model predicts a drug-disease interaction or prioritizes a gene for study.

- **Ethical and Practical Challenges**: Addressing multilingual data, ethical considerations of data privacy, especially in handling sensitive patient data. Challenge of scaling these models and maintaining data quality across diverse sources.

# Thank You

HAPPY CODING

DO DIGITAL. STAY HUMAN.