Microsoft Research
Summit 2022

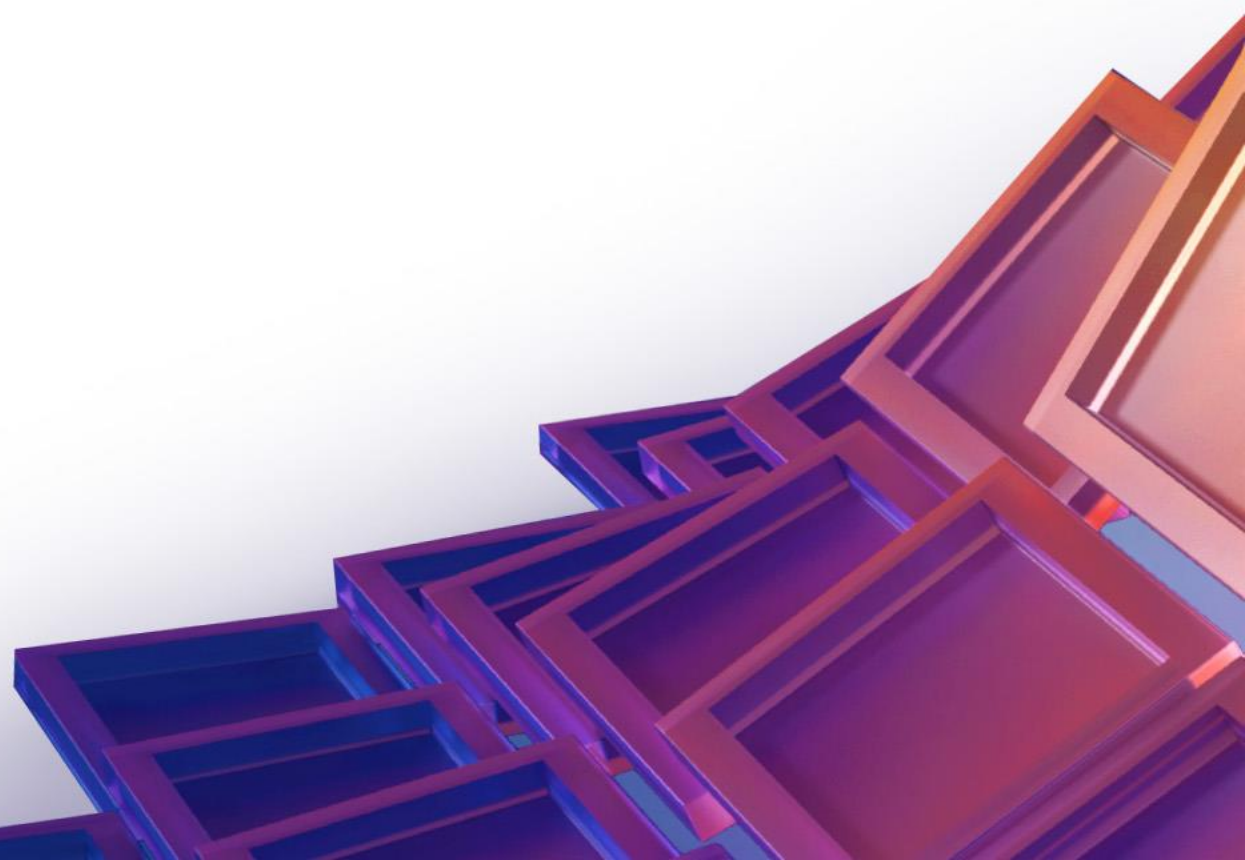
Generative Models and Transformers for Chemistry

Morris Sharp
Microsoft Research

Advances in AI applied to chemistry problems

- New AI models can capture large amounts of data
- Latest research applied these models to chemistry problems
- Goal: Optimize time spent in the lab

Transformers



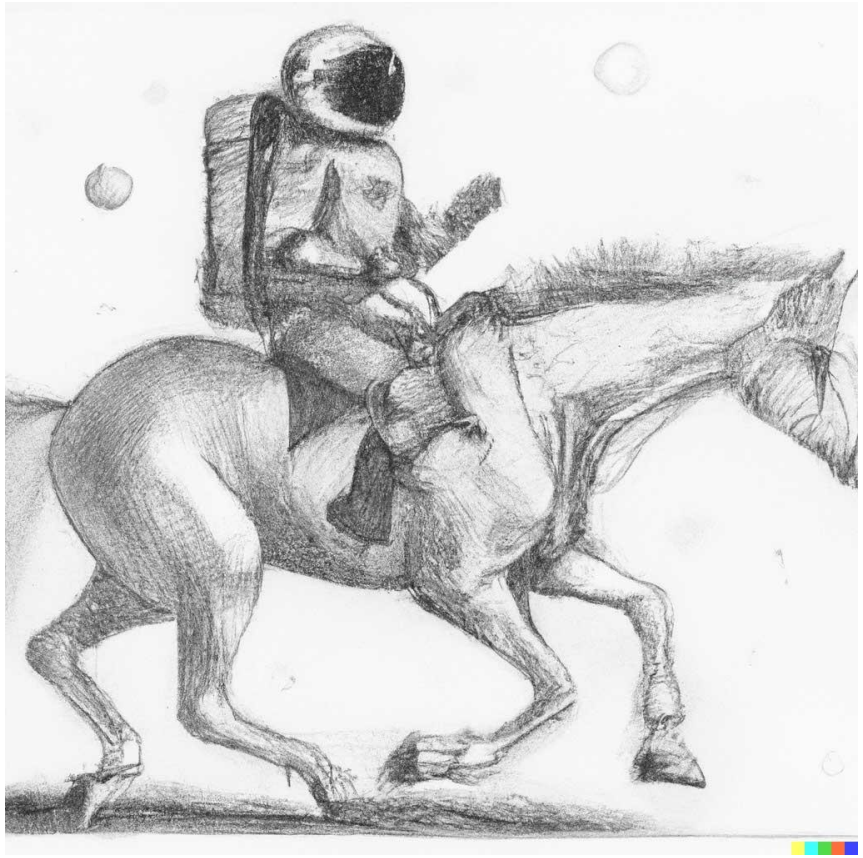
The Transformer Revolution

- Neural network that can learn the context (“attention”) for each part of a given input
- Context is important for language tasks such as translation and document summarization.
- Can accommodate and require huge amounts of data
- Led to the creation of a number of transformer models:
  - GPT
  - BERT
  - Turing

Transformers in action

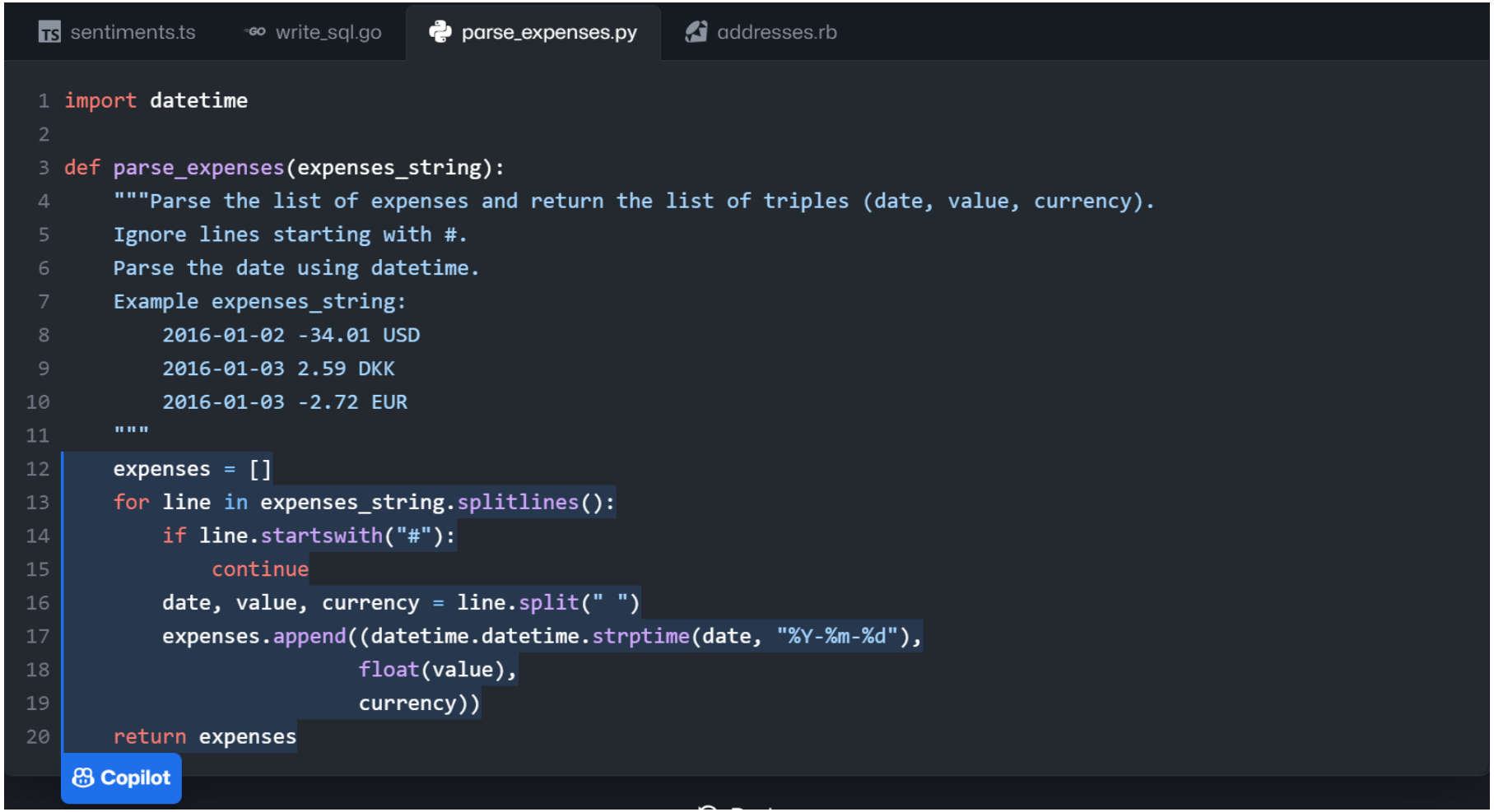
DALL·E 2

"An astronaut riding a horse as a pencil drawing"



Transformers in action

Github Copilot



Fine-tuning a transformer is relatively quick

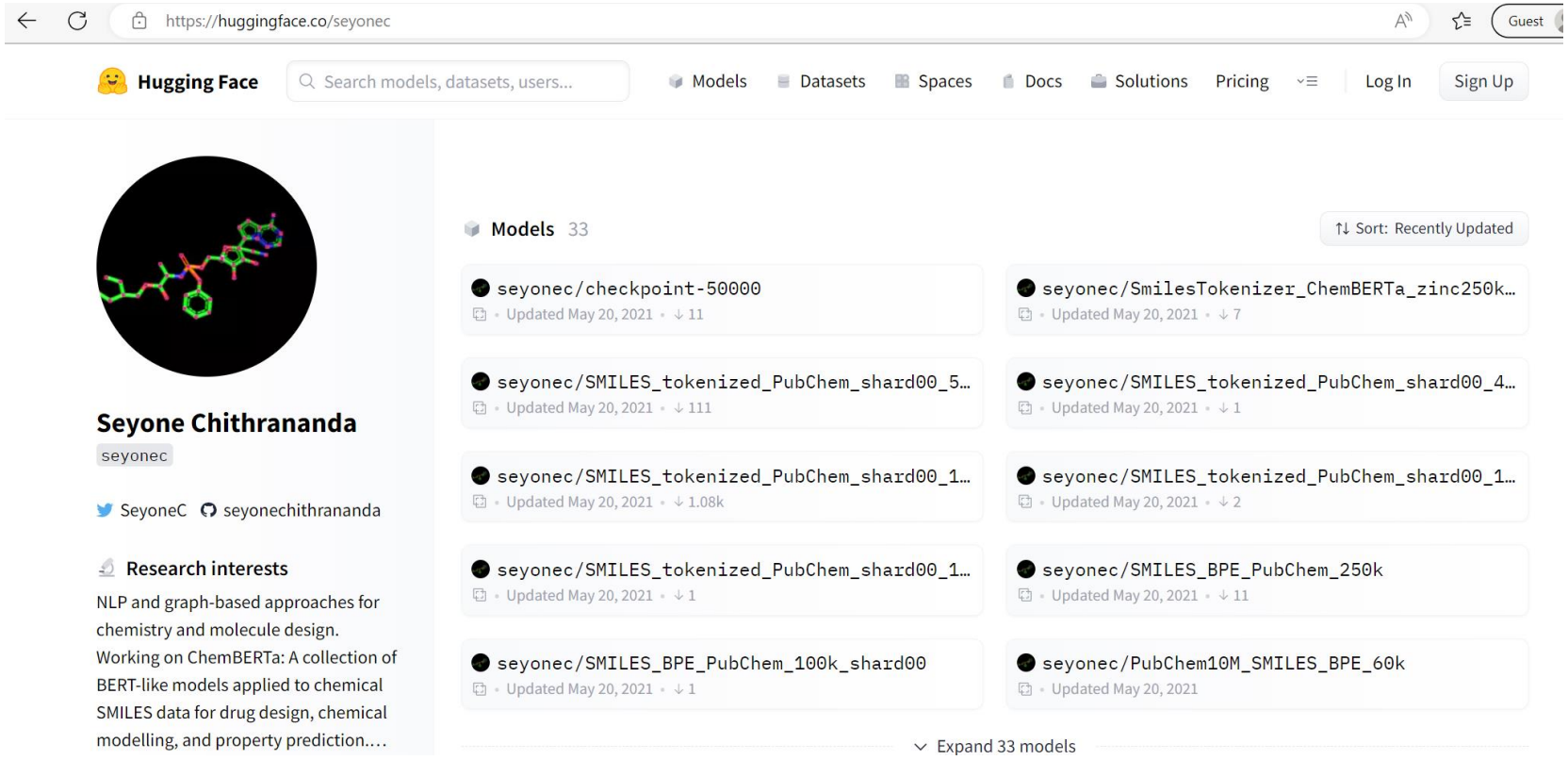
1. Pre-training

- Train model on a general task
  - Predict missing atom in sequence
  - Predict chemical properties
- Needs a lot of data
- Expensive and slow
- Can be re-used

2. Fine-tuning

- Fine tune model for specific task
  - Predict homo-lumo gap
  - Predict toxicity
- Only needs small dataset
- Cheap and fast

Download a pre-trained model from the web

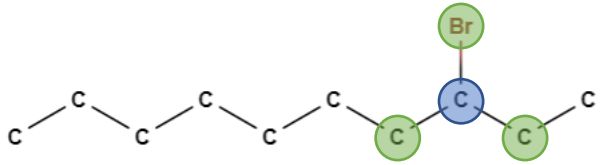


Understanding context is key for transformers

Language

I went to the grocery store

Chemistry



Transformer models excel at many tasks

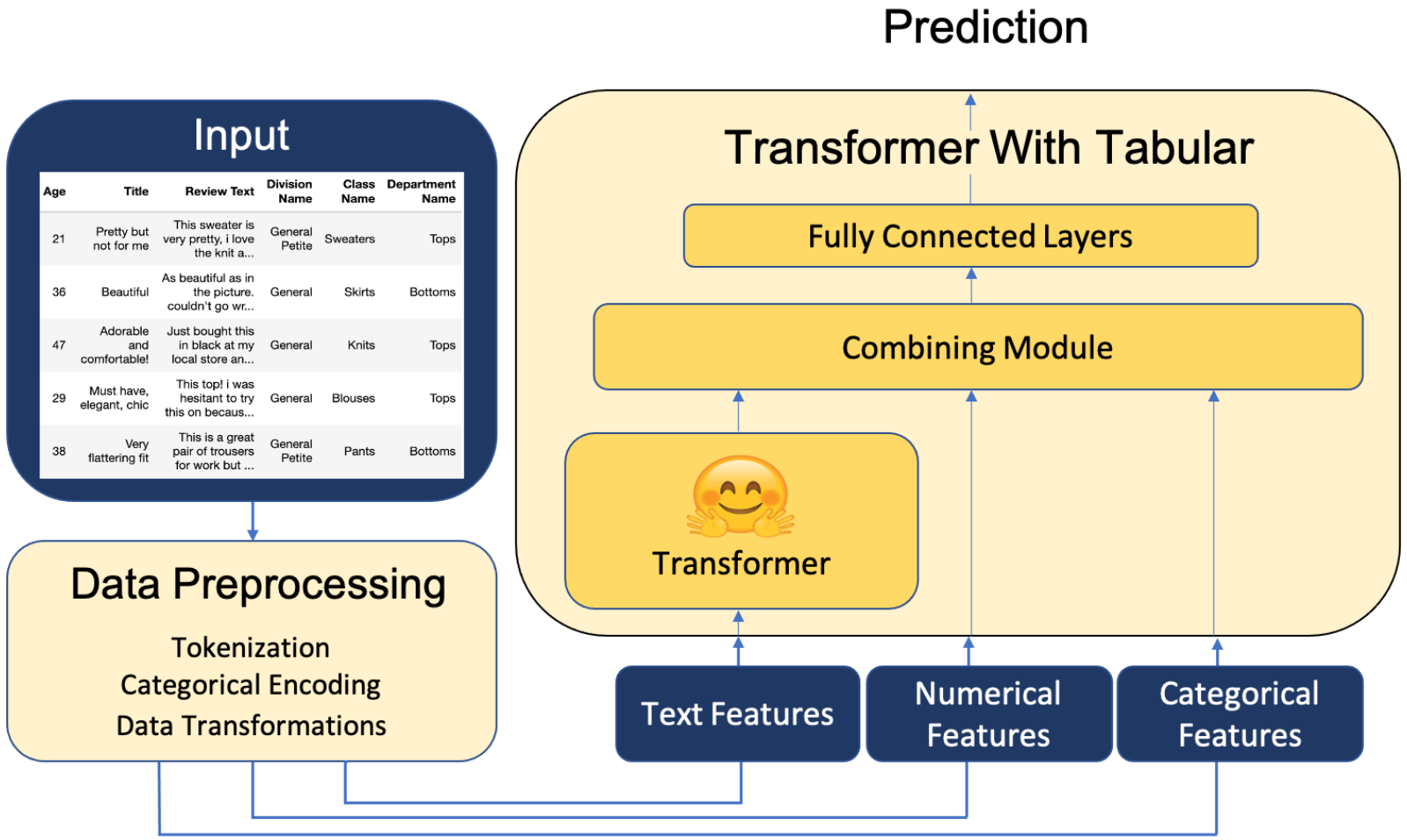
Graphormer

- Quantum prediction track of Open Graph Benchmark Large-Scale Challenge (2021)
- Open Catalyst challenge (2021)

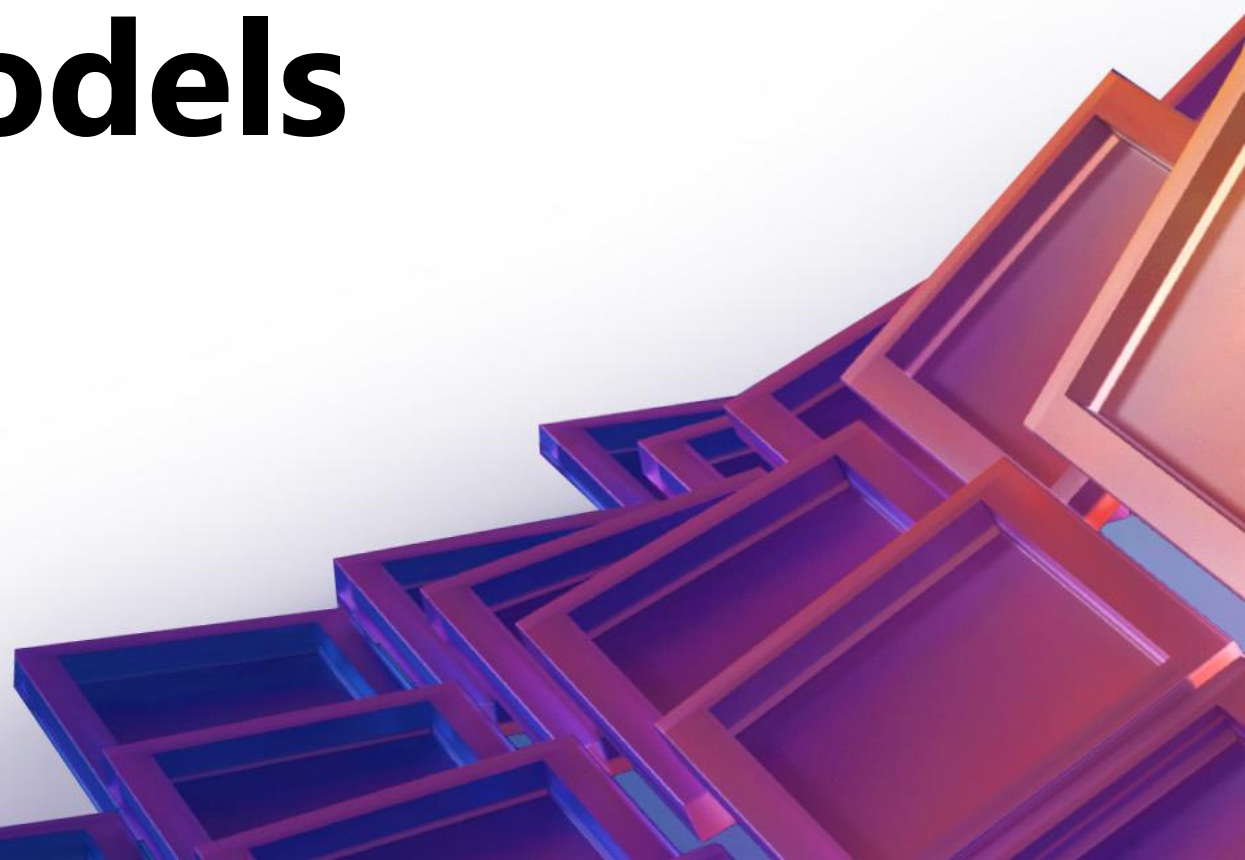
Graphormer	TransMol	ChemFormer
Molecular Transformer	Molecular Attention Transformer	ChemBERTa

Chemistry transformers can be combined with numerical data

- Often have experimental conditions that affect predictions
- Combine pre-trained transformers with numerical and categorical features

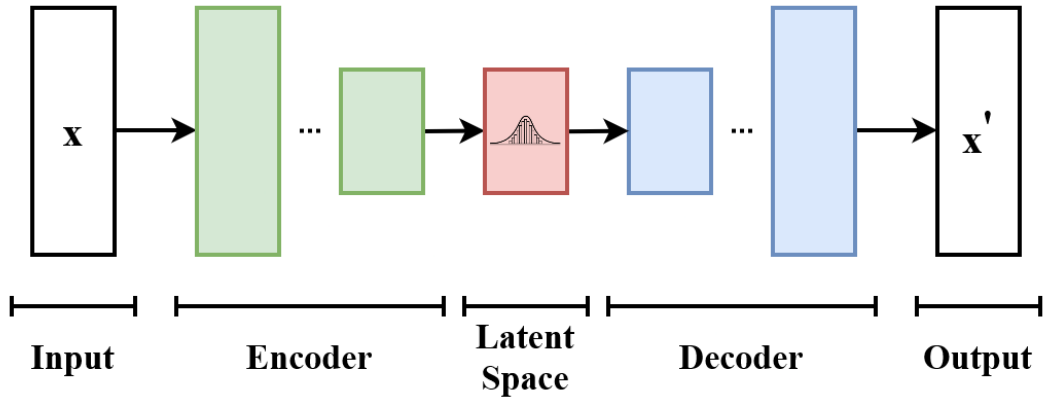


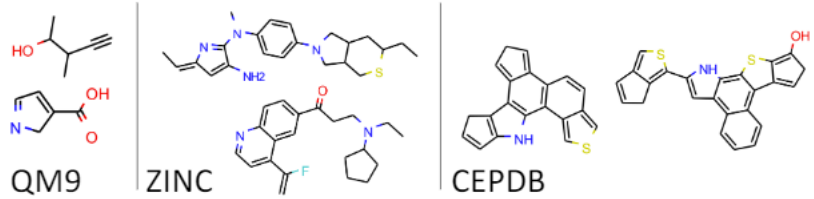
Generative Models



How do we generate new molecules?

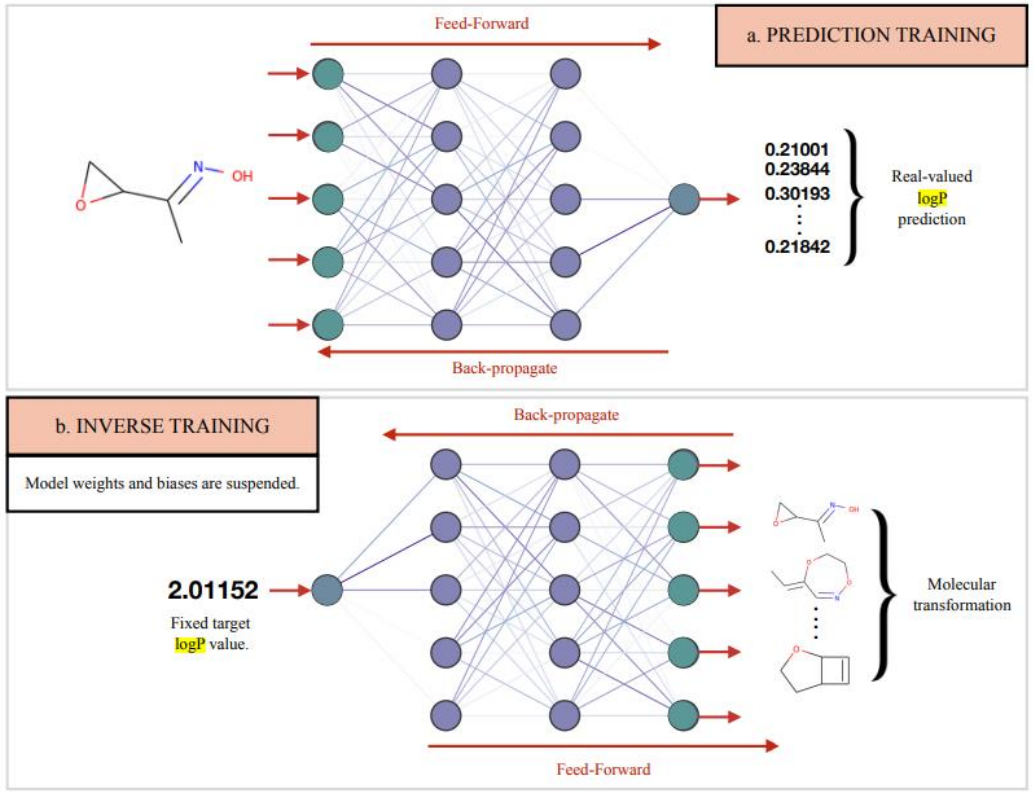
- Variational Autoencoders useful for generating similar data to inputs
- For molecular design, need to generate chemically valid structures
- Constrained Graph Variational Autoencoders for Molecule Design





Generating new molecules with a specific property

- PASITHEA: Inverse ML for molecular design
- Generate molecules based on a target property value



Microsoft Research
Summit 2022

Thank you