

Microsoft Research

**Summit 2022**

# **Generative Models and Transformers for Chemistry**

Morris Sharp

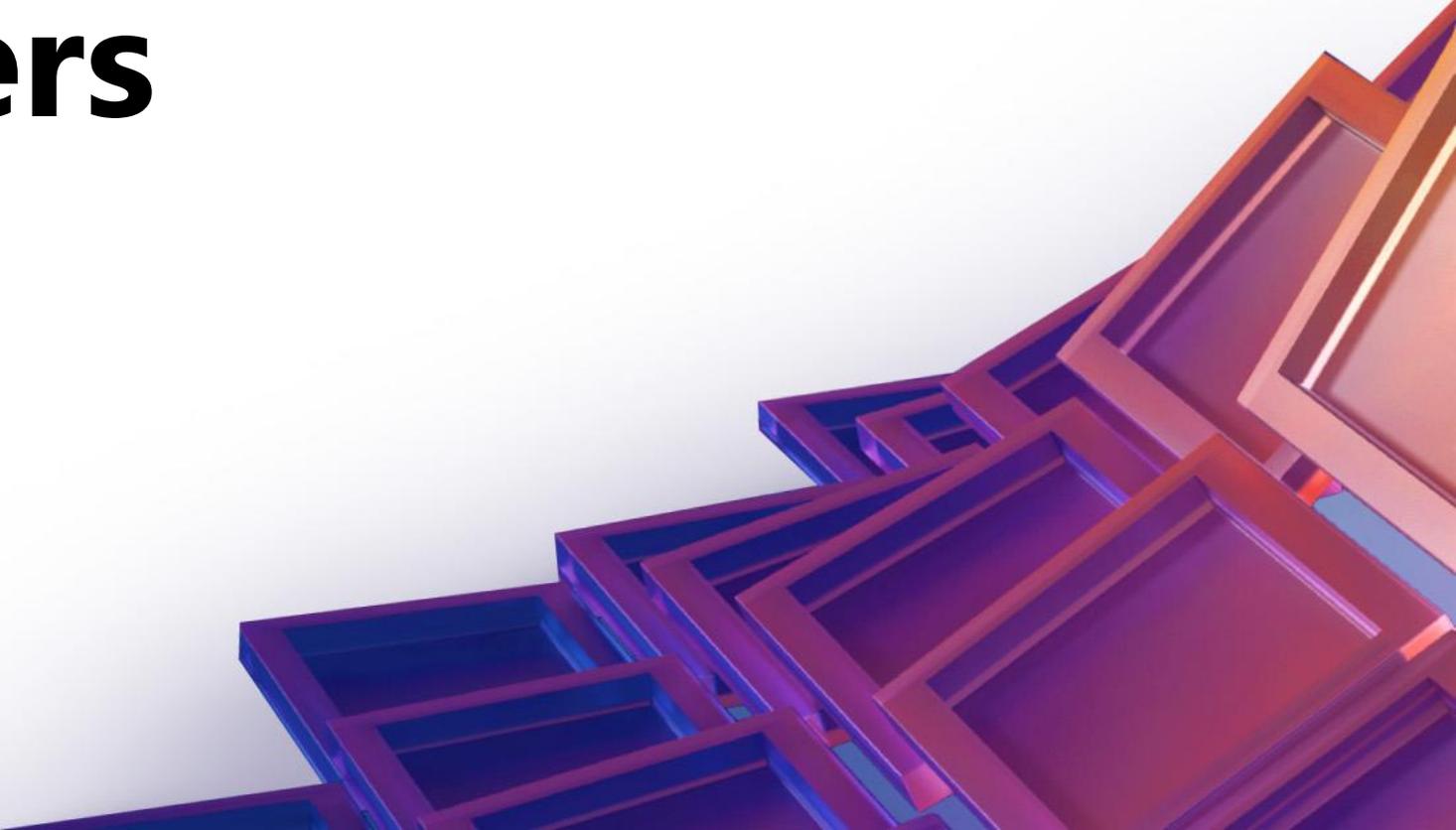
Microsoft Research

# 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

```
ts sentiments.ts  write_sql.go  parse_expenses.py  addresses.rb

1 import datetime
2
3 def parse_expenses(expenses_string):
4     """Parse the list of expenses and return the list of triples (date, value, currency).
5     Ignore lines starting with #.
6     Parse the date using datetime.
7     Example expenses_string:
8         2016-01-02 -34.01 USD
9         2016-01-03 2.59 DKK
10        2016-01-03 -2.72 EUR
11    """
12    expenses = []
13    for line in expenses_string.splitlines():
14        if line.startswith("#"):
15            continue
16        date, value, currency = line.split(" ")
17        expenses.append((datetime.datetime.strptime(date, "%Y-%m-%d"),
18                        float(value),
19                        currency))
20    return expenses
```

 Copilot

# Fine-tuning a transformer is relatively quick

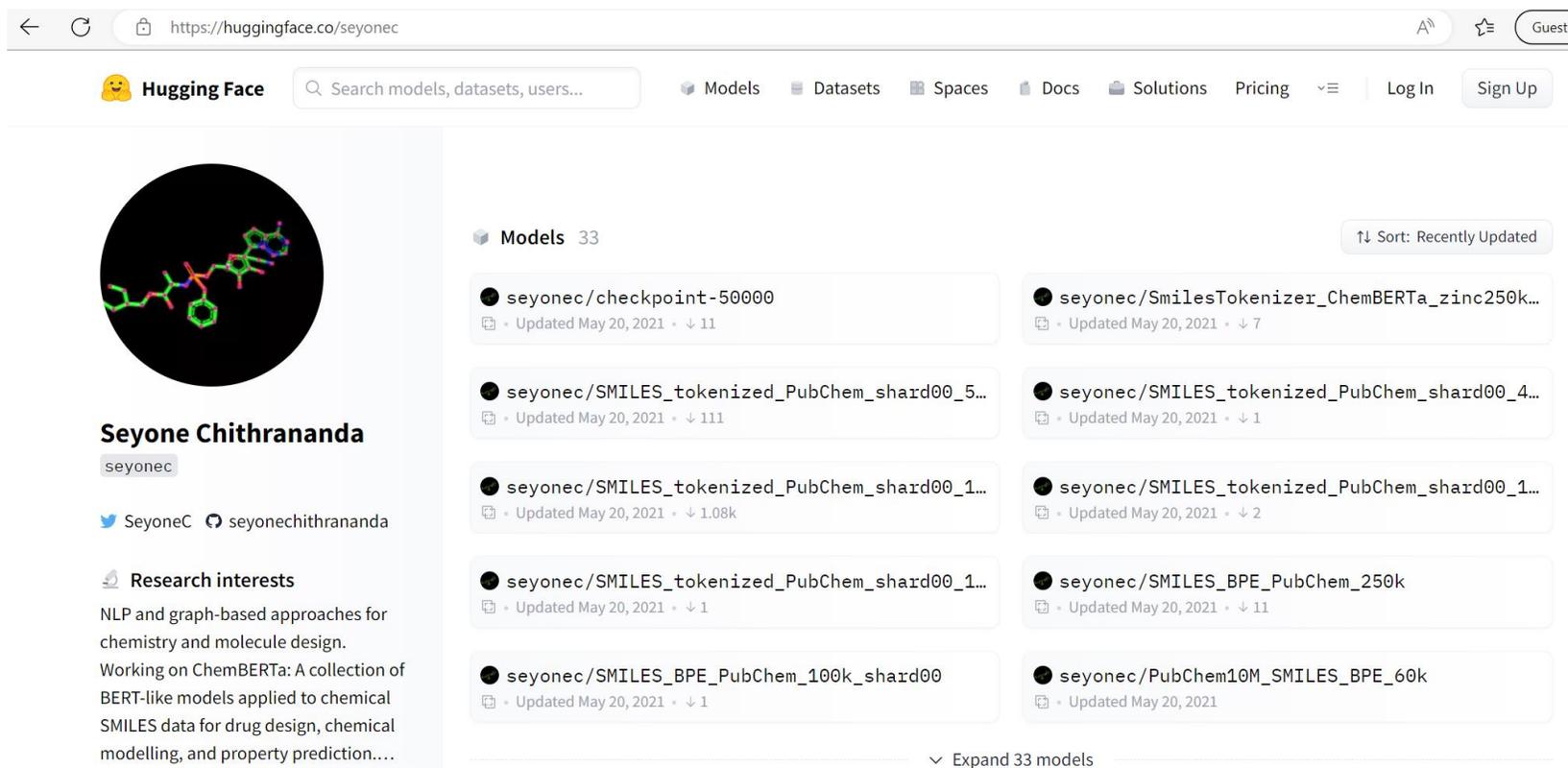
## 1. Pre-training

- 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



Browser address bar: <https://huggingface.co/seyonec>

Hugging Face Search models, datasets, users... Models Datasets Spaces Docs Solutions Pricing Log In Sign Up

**Seyone Chithrananda**  
seyonec  
SeyoneC seyonechithrananda

**Research interests**  
NLP and graph-based approaches for chemistry and molecule design.  
Working on ChemBERTa: A collection of BERT-like models applied to chemical SMILES data for drug design, chemical modelling, and property prediction...

**Models 33** Sort: Recently Updated

- seyonec/checkpoint-50000  
Updated May 20, 2021 · ↓ 11
- seyonec/SmilesTokenizer\_ChemBERTa\_zinc250k...  
Updated May 20, 2021 · ↓ 7
- seyonec/SMILES\_tokenized\_PubChem\_shard00\_5...  
Updated May 20, 2021 · ↓ 111
- seyonec/SMILES\_tokenized\_PubChem\_shard00\_4...  
Updated May 20, 2021 · ↓ 1
- seyonec/SMILES\_tokenized\_PubChem\_shard00\_1...  
Updated May 20, 2021 · ↓ 1.08k
- seyonec/SMILES\_tokenized\_PubChem\_shard00\_1...  
Updated May 20, 2021 · ↓ 2
- seyonec/SMILES\_tokenized\_PubChem\_shard00\_1...  
Updated May 20, 2021 · ↓ 1
- seyonec/SMILES\_BPE\_PubChem\_250k  
Updated May 20, 2021 · ↓ 11
- seyonec/SMILES\_BPE\_PubChem\_100k\_shard00  
Updated May 20, 2021 · ↓ 1
- seyonec/PubChem10M\_SMILES\_BPE\_60k  
Updated May 20, 2021

Expand 33 models

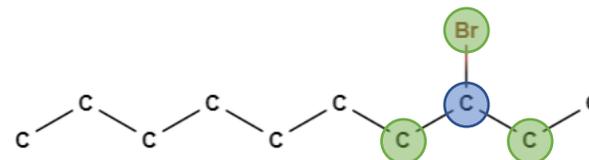
# Understanding context is key for transformers

## 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)

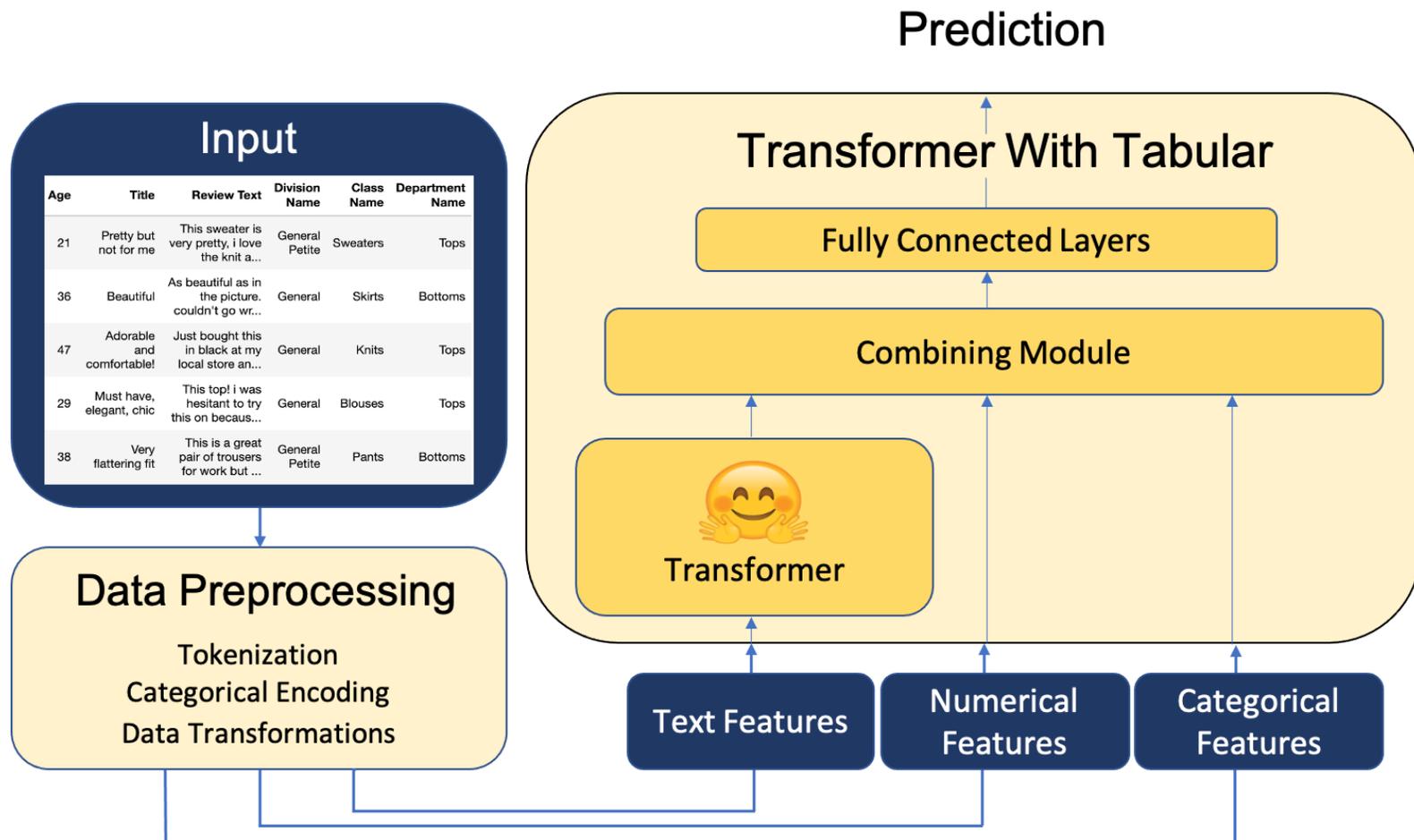


Open Catalyst challenge (2021)

Graphormer	TransMol	ChemFormer
Molecular Transformer	Molecular Attention Transformer	ChemBERTa

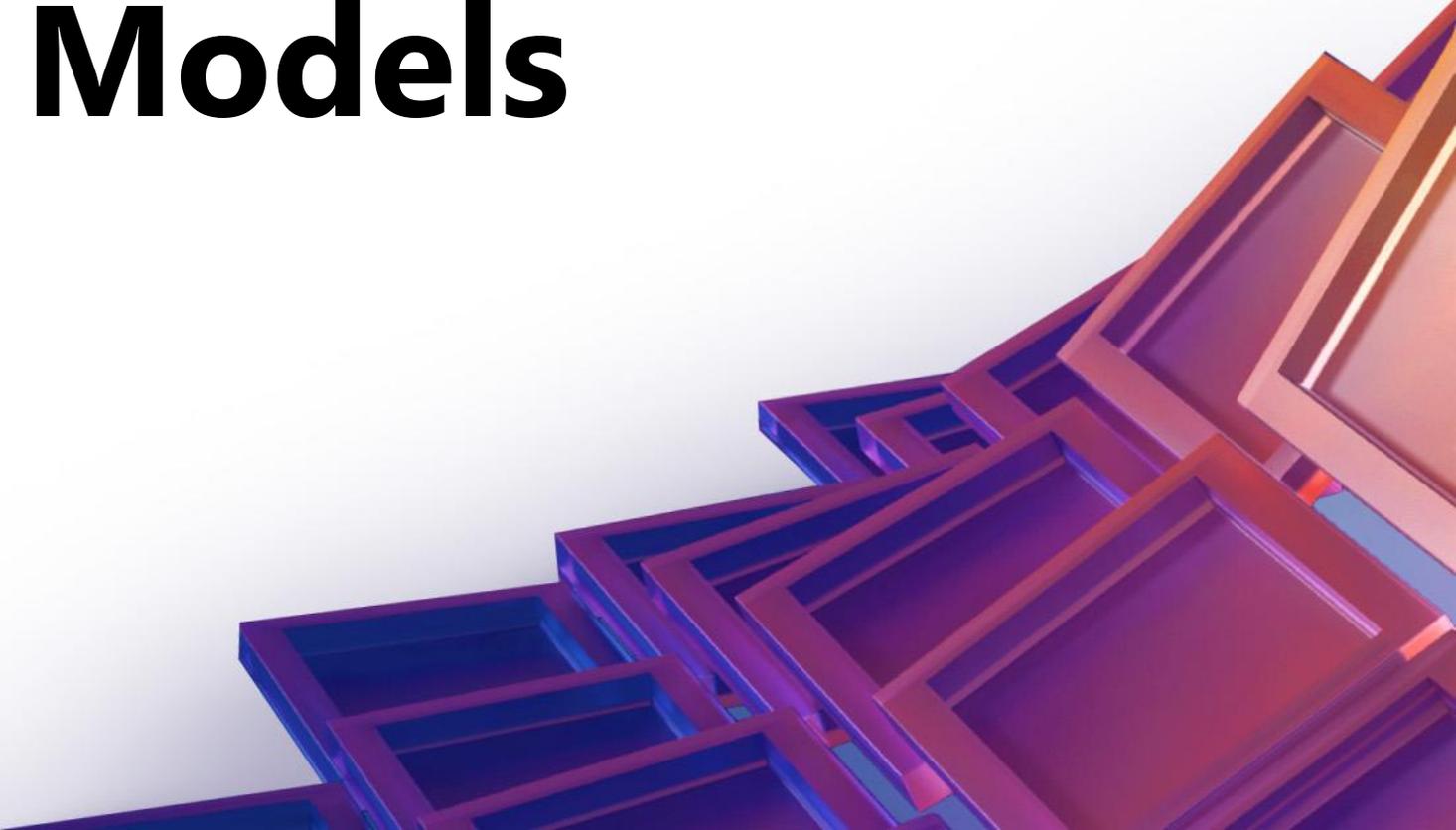
# Chemistry transformers can be combined with numerical data

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



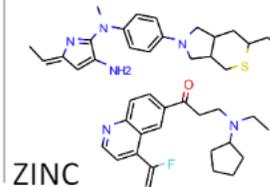
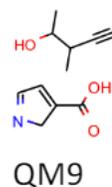
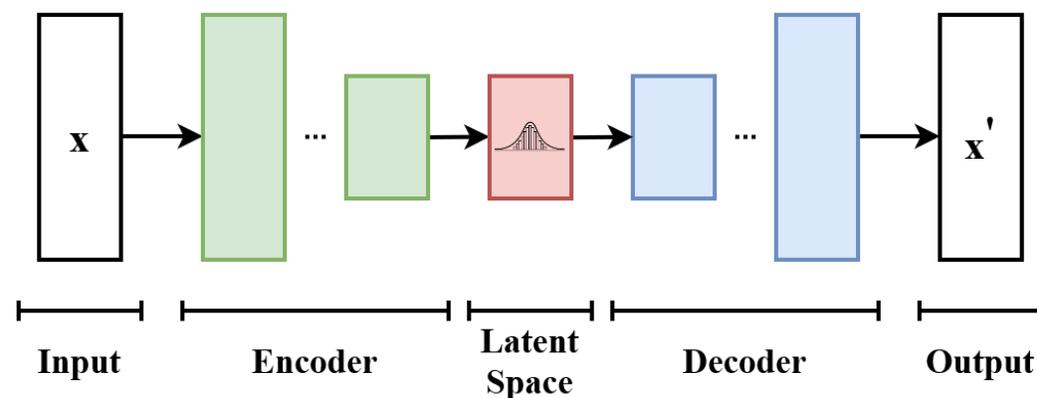


# Generative Models



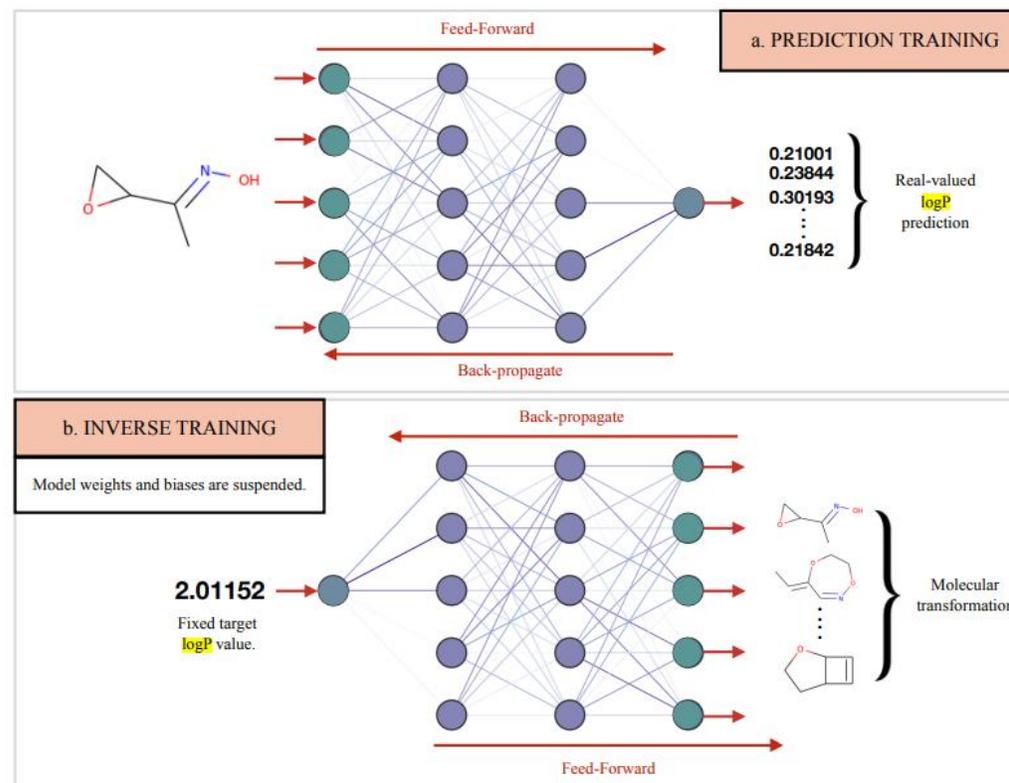
# How do we generate new molecules?

- 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**