UGM Boston Oct. 4th 23

Chemaxon

Generative AI Begins to Dominate the AI Conversation in Early Drug Discovery

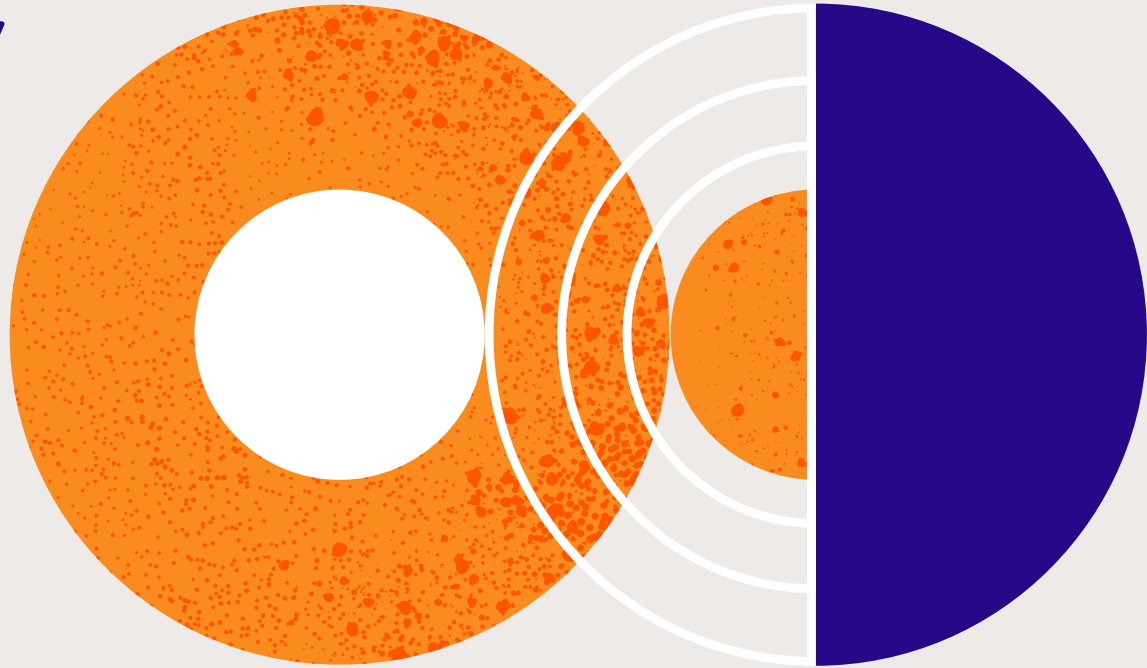
Joe Michel

Director of Informatics,
Cytokinetics

Generative AI Begins to Dominate the AI Conversation in Early Drug Discovery

UGM Boston Oct. 4th 23
Chemaxon

Joe Michel,
Head of Informatics,
Cytokinetics Inc.
South San Francisco, CA



New Focus of AI: Generative AI

How Generative AI shifts our conversation from predictors to generating medicines

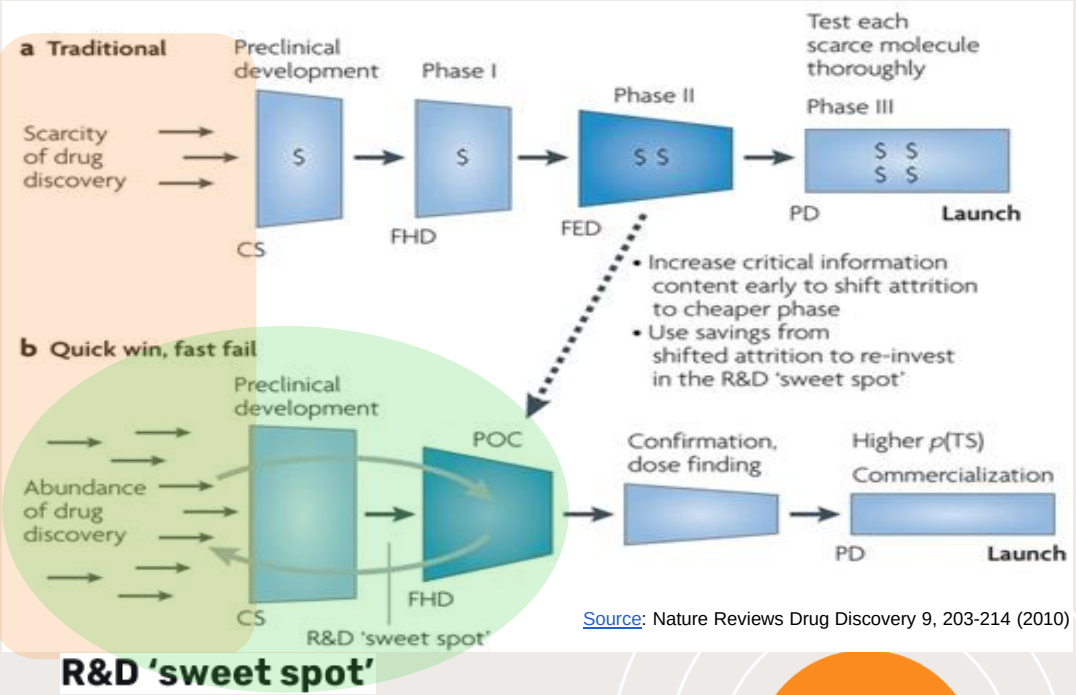
Agenda:

- Need for advancements in Research Drug Discovery
- Generative AI has made impact
- Recent Headlines - Generative AI (Gen AI) continues contributes
- What has changed? What is Generative AI / Foundation models?
- Generative AI Framework
- Generative AI Examples in Drug Discovery Ligand Discovery & VS

New Focus of AI: Generative AI

What is the need? Cost, Time, and Quality…

Quick win, fast fail drug development paradigm



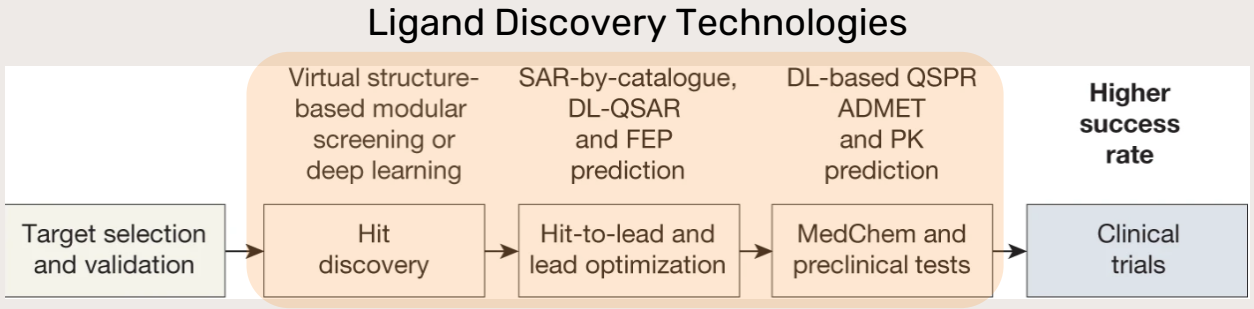
Source: Nature Reviews Drug Discovery 9, 203-214 (2010)

R&D 'sweet spot'

New Focus of AI: Generative AI

One of the areas AI technology could make an impact

- Computer-aided drug discovery (CADD) been around for decades, as well as Ligand Discovery Technologies.
- How can Ligand Discovery Technologies improve “quick win, fast fail”?

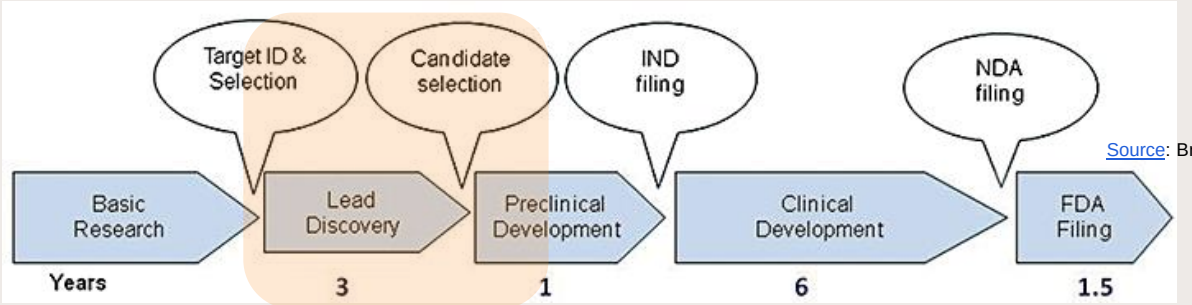


Source: Nature 616, 673-685 (2023) , ISSN 1476-4687

New Focus of AI: Generative AI

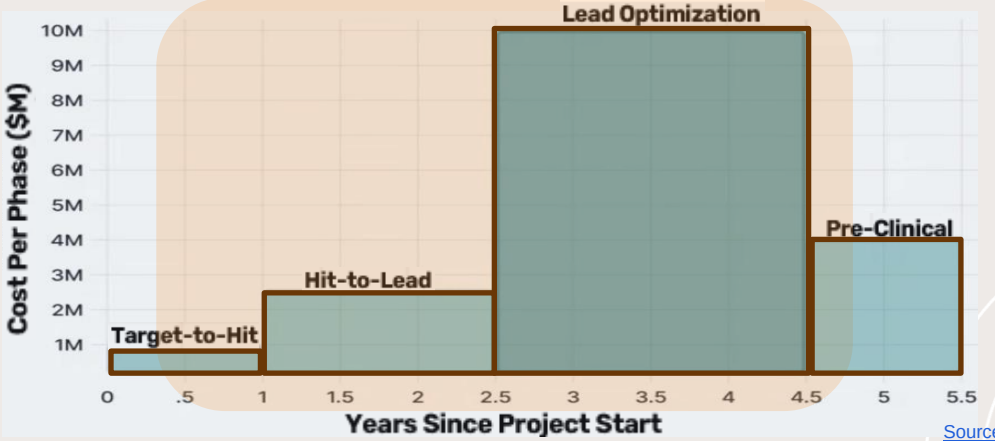
One of the areas AI technology could make an impact

Typical Years of Drug Discovery Phases:



Source: Br J Pharmacol 2011; 162(6): 1239-1249

Within Discovery Phases, Typically, Lead Optimization Phase is the most costly:



Source: Nature Reviews Drug Discovery 9, 203-214 (2010)

Generative AI has made “a breakthrough”

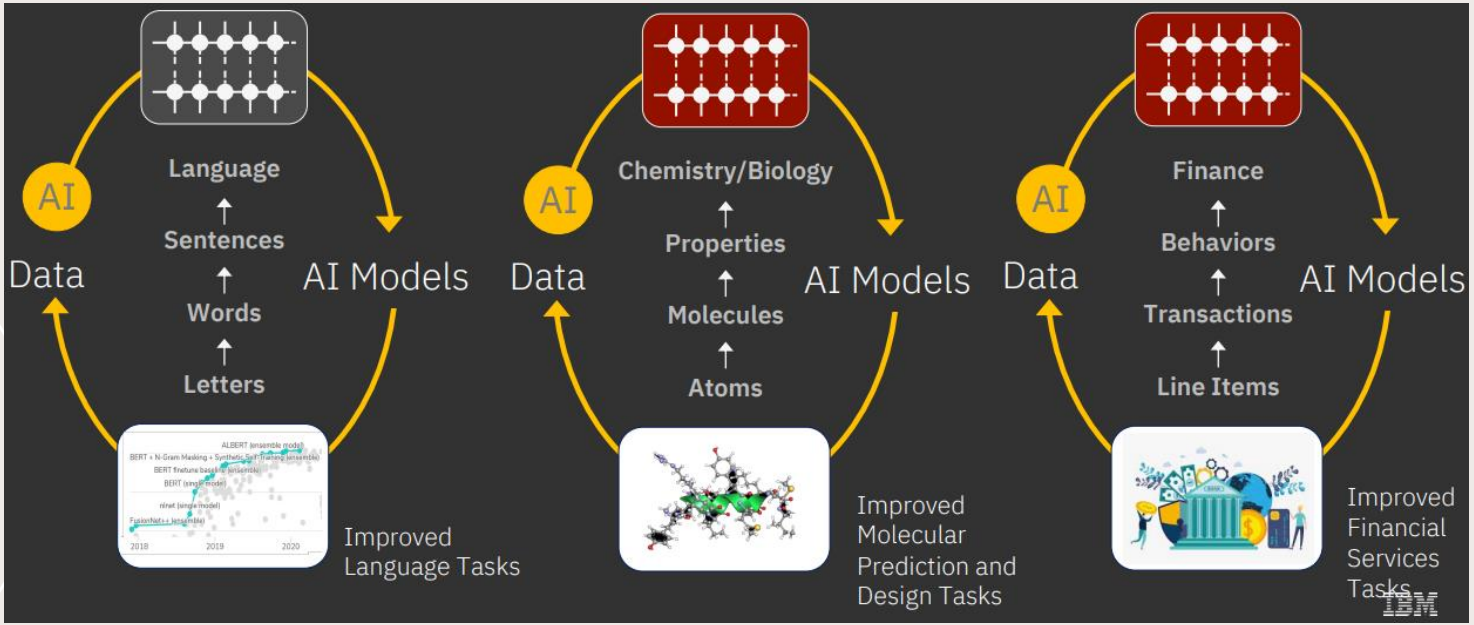
It’s here, it’s real and starting to impact Drug Discovery

- In 2023, ChatGPT (a chatbot developed by Microsoft's OpenAI) was "born“
- DeepMind (AlphaFold) CEO, Demis Hassabis, spoke on this topic, **AlphaFold IS A major break-through** Understanding and controlling protein folding is arguably the most important challenge in structural biology (Martnez, 2014; Jankovic and Polovic, 2017)

- Nvidia’s CEO Jensen Huang stated, “**We are in the 'iPhone moment' of AI**“. "2023 is expected to be the most exciting year in the field of AI" (source: keynote speech 2023)

Generative AI continues to “breakthrough”

The same AI breakthroughs happening in language are Impacting other scientific and enterprise applications



Source: The Promise of Foundation Models and Generative AI (2023), Payel Das, Principal Research & Master Investor

Generative AI impact in recent years

2016

AlphaGo / Go

DeepMind AlphaGo beats world champion (a)

2020

Exscientia

1st AI drug enter Clinical Trials (1)

2021

AlphaBet / Google

DeepMind AlphaFold predicts protein structures(a)

2022

Insilico Med.

AI drug from AI-discovered novel target enters Clinical Trials (2)

2022

Schrodinger

AI drug enters Clinical Trials from target-to-clinic under 1 year (3)

Nov 2022

OpenAI

ChatGPT launches (b)

Jan 2023

Absci Corp

AI Ab drug designed for HER2 (4)

Feb 2023

Insilico Med.

AI drug receives Orphan Drug Designation (5)

June 2023

Insilico Med.

1st AI drug enters Phase II Clinical Trials (7)

* 30 months from Discovery to Phase I

Sept 2023

Insilico Med.

Exelixis pharmaceutical company purchases Phase I AI drug for $80 Million (8)

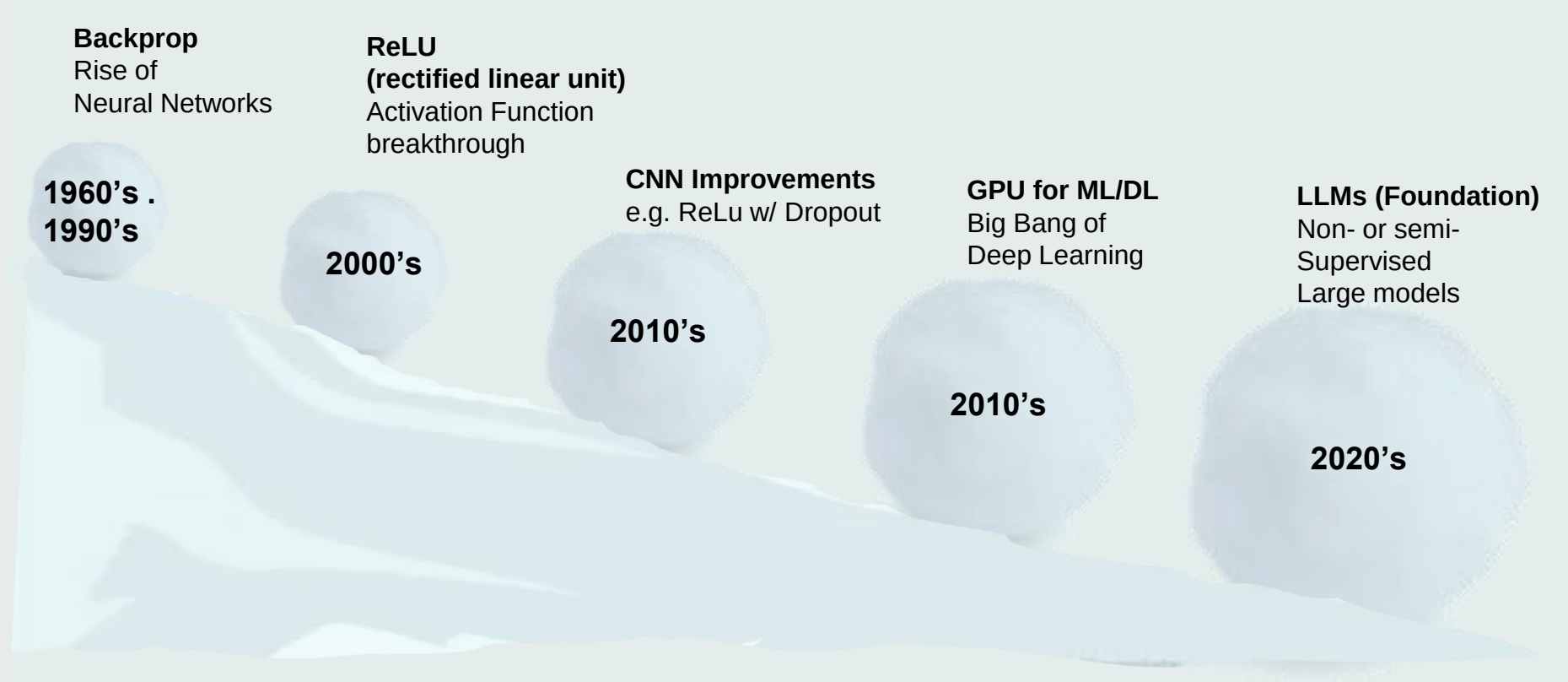
Gen AI – Recent Headlines

AI Year in Review (2023)

Jan 2023: Absci - deploys zero-shot generative AI in antibody design
March 2023: MIT reveals DiffDock, which could support faster, safer drug development
March 2023: Nvidia launches BioNeMo Cloud to boost drug discovery with generative AI
June 2023: Insilico Medicine's AI drug enters phase 2 study
June 2023: Sanofi reveals plan to put AI at the center of its operations
July 2023: AI-aided drug ulotaront fails phase 3 studies
July 2023: Nvidia invests $50 million in Recursion to boost AI-driven drug discovery
July 2023: AI startup Causaly raises $60 million to expedite drug development
August 2023: Pharos iBio develops anticancer drug with AI platform
August 2023: Generative AI tool boasts 79% accuracy in predicting clinical trial outcomes
August 2023: UCF researchers reveal AI-assisted drug screening technology
August 2023: Recursion uses AI to bridge protein/chemical universe, predicted targets for 36 bill. Compounds Recursion is using Cyclica's MatchMaker technology, NVIDIA DGX Cloud supercomputing and DeepMind's AlphaFold2 database to screen Enamine REAL Space

Source: A year in review: AI's evolving role in drug discovery and development in 2023 (Aug 4, 2023)

AI (DL) Breakthroughs over the decades



What's required for Gen AI?

LLMs or Foundation models is a fundamental AI "breakthrough"

- **Pre-trained** on unlabeled **large** datasets of different modalities (e.g. molecules, proteins)
- Leverage **self-supervised learning**
- Learn **generalizable & adaptable data representations** which can be effective used in multiple downstream tasks (e.g. molecule generation)

e.g.

GPT-3 (ChatGPT) - 175 billion parameters
IBM MolFormer-XL – 1.1 billion molecules

Source: Nature Machine Intelligence 4, 1256-1264

~ Training DataSet Sizes

Traditional ML/DL		Foundational DL
100's to 10,000's (thousands)	> 100x →	1,000,000's (billions)

Source: Definition of Foundation models AWS

More input records, more potential input parameters…

What's definition of Generative AI?

It makes something, not just a predictor…

Generally Speaking:

- Generative AI is artificial intelligence capable of generating text, images, other media, using generative models. It produces beyond just a mathematical predictor value.

For Ligand Discovery Technologies:

- Generative AI is trained to understand molecules and can generate new molecules randomly or can be controlled by specific criteria of interest (**molecule generator**).
    - Represented Molecular Space – (VAEs or GANs) Gen AI that can produce a representation of molecules based on an input molecule or information (approx. Gaussian distribution)
    - Reinforcement Learning - This control mechanism usual requires additional AI models separate from the molecule generator

Source: What is Generative AI and How does it work?

Another “breakthrough” in DD?

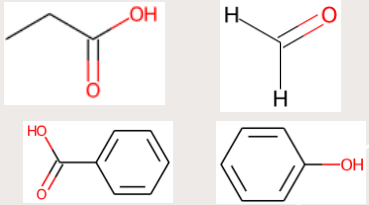
Producing beyond predictions or classifications

DALL-E (GPT-3)

Text Description:
Chemist synthesizing molecule
in Biotech lab
in Boston Massachusetts

Molecule Gen AI

Text Description:
SMILE String
[CH3][CH2]C(O)=O
&
Potentially other ligand info



Example Tools:
Application: Chemistry42 by Insilico Med
Software Tool Kit: IBM GT4SD by IBM (open source release March 2023)

Nuts & Bolts - 3 popular Gen AI model types

- Breakthrough – Foundation models -> Ability to leverage different learning approaches, including unsupervised or semi-supervised learning
- 3 Main Deep Learning Models used for Generative AI:
  - Denoising **Diffusion** Probabilistic **models** (DDPMs) – (also considered Foundation models)

    Two-step process during training. The two steps are forward diffusion and reverse diffusion
  - Variational autoencoders (VAEs)

    Collaboration of two neural networks (encoder and decoder), with a subsampling middle layer
  - Generative adversarial networks (GANs)

    Pitting two neural networks against each other: a generator that generates new examples and a discriminator

Source: What is Generative AI and How does it work?

How can Gen AI help discover Ligands

Guided molecule generation is required

Power comes in combination of Deep Learning Techniques (supervised & unsupervised):

- BASIS: VAEs & GANs capable of generating molecule set based on representative input (e.g. IBM)
- Predictors – use or train needed predictors (bread and butter)
- Additionally apply Reinforcement Learning (RL) – Can be combined with VAE or GAN to further optimize generative molecular structure & improve chances of success

Source: What is Generative AI and How does it work?
Source: Saturn Cloud Generative AI explanation

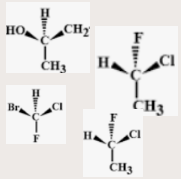
Gen AI Framework

Several model types can help strengthen your system

Do it again /iterate [generations]

Molecule Starter (optional)

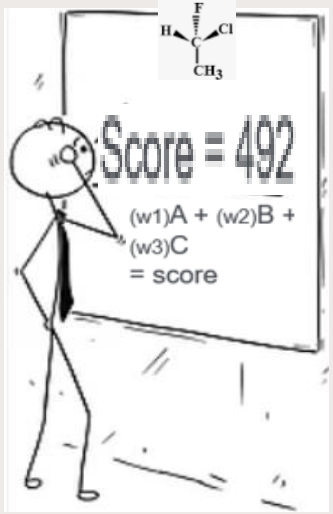
Molecule Generator (Gen AI)



Represented Molecular Space

GAN or VAE

Reinforcement Learning



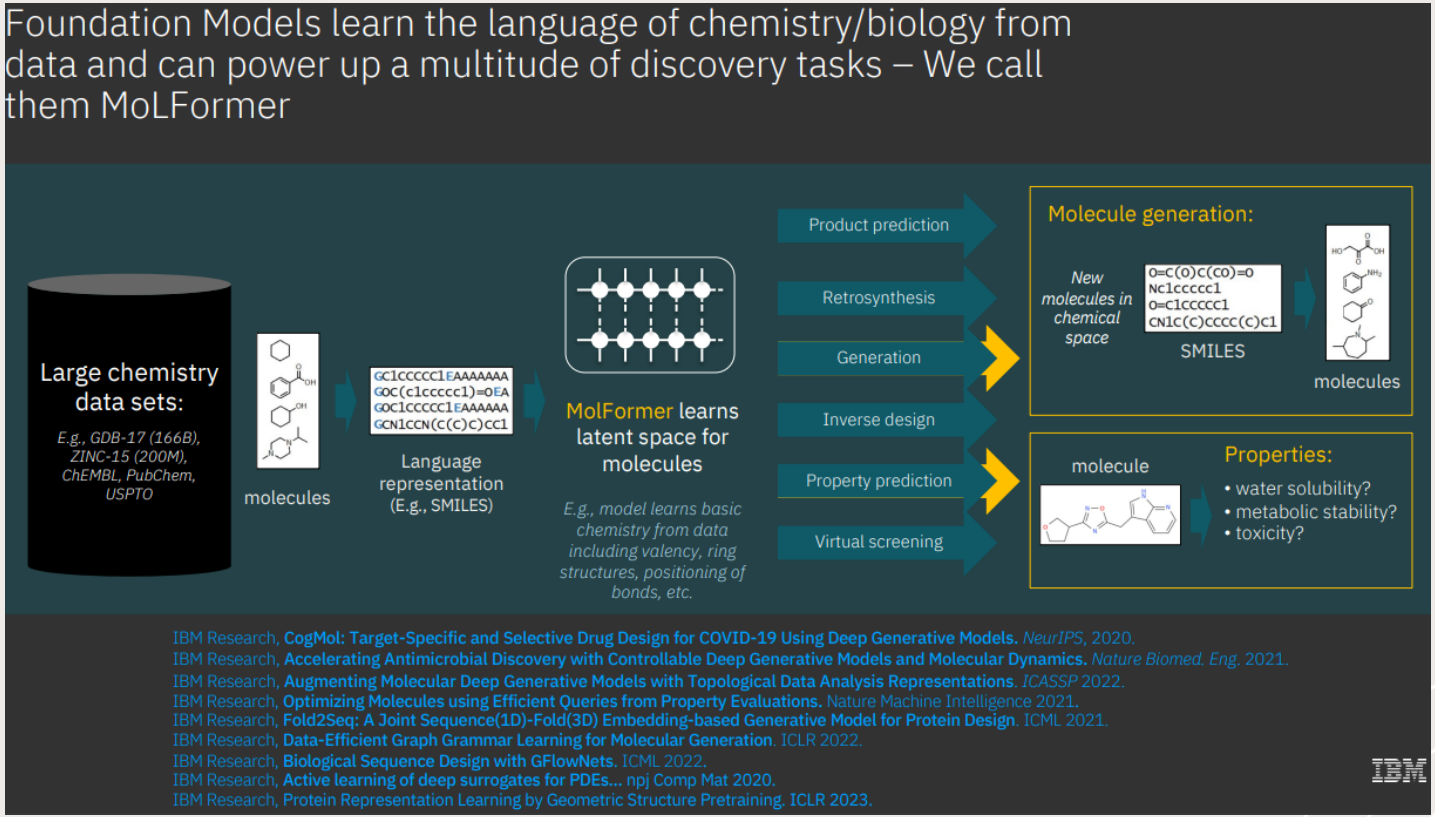
Rewards & Penalties

Goal: generate structures with high predicted activity or low predicted toxicity

Predictors

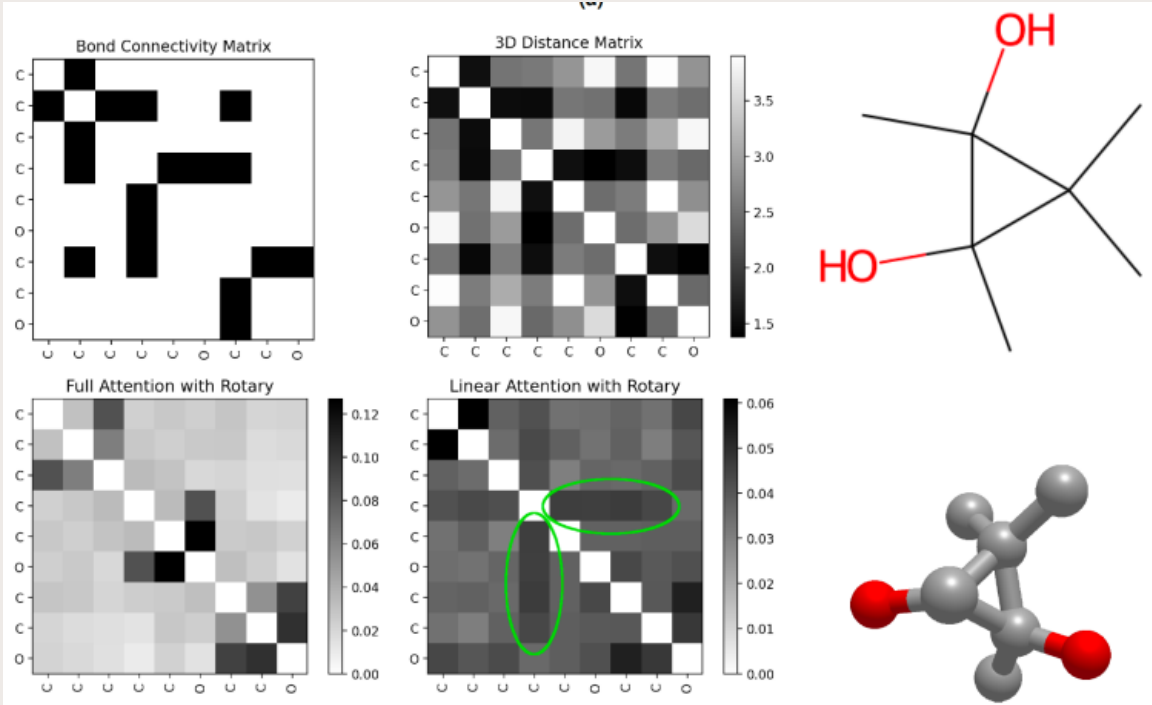
- Potency & Selectivity Predictors
- Property Predictors (PhysChem, ADMET, Tox …)
- Predictive Retrosynthesis

IBM Generative AI / Foundation AI



Source: The Promise of Foundation Models and Generative AI (2023), Payel Das, Principal Research & Master Investor

IBM Molformer captures structural info

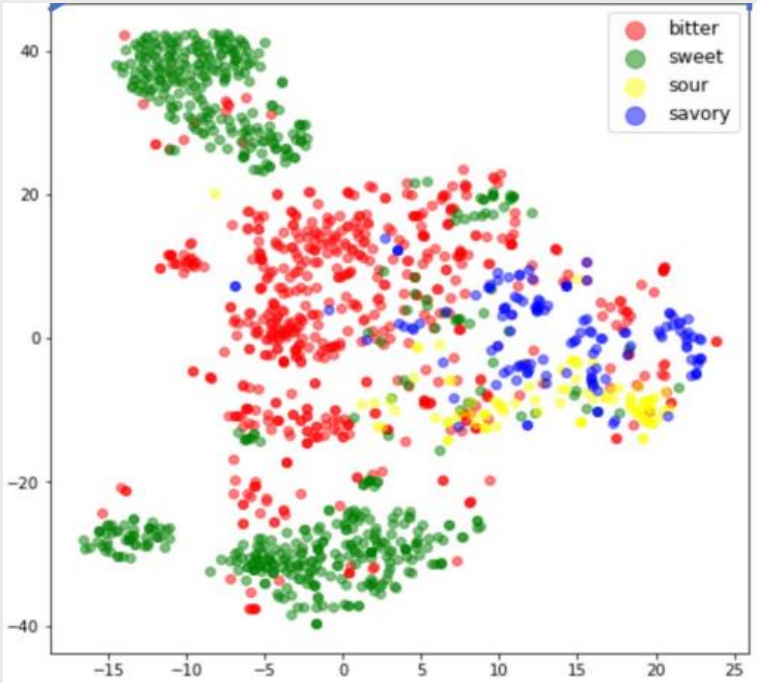


Source: The Promise of Foundation Models and Generative AI (2023), Payel Das, Principal Research & Master Investor

Molformer Learns molecular taste w/o labels (unsupervised)

Visualization of unsupervised MolFormer Embeddings in t-SNE space and separation of flavor molecules in that space.
source

t-SNE (t-distributed Stochastic Neighbor Embedding) is an unsupervised non-linear dimensionality reduction technique for data exploration and visualizing high-dimensional data



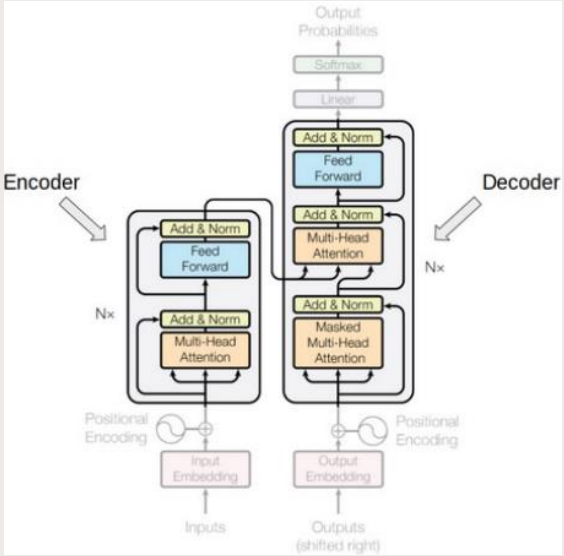
[IBM Research, **Cloud-Based Real-Time Molecular Screening Platform with MolFormer** , ECML PKDD (2022). https://2022.ecmlpkdd.org/wp-content/uploads/2022/09/sub_1455.pdf

Source: The Promise of Foundation Models and Generative AI (2023), Payel Das, Principal Research & Master Investor

MolFormer: Foundational transformer

MoLFormer-XL – a specific example from MoLFormer family

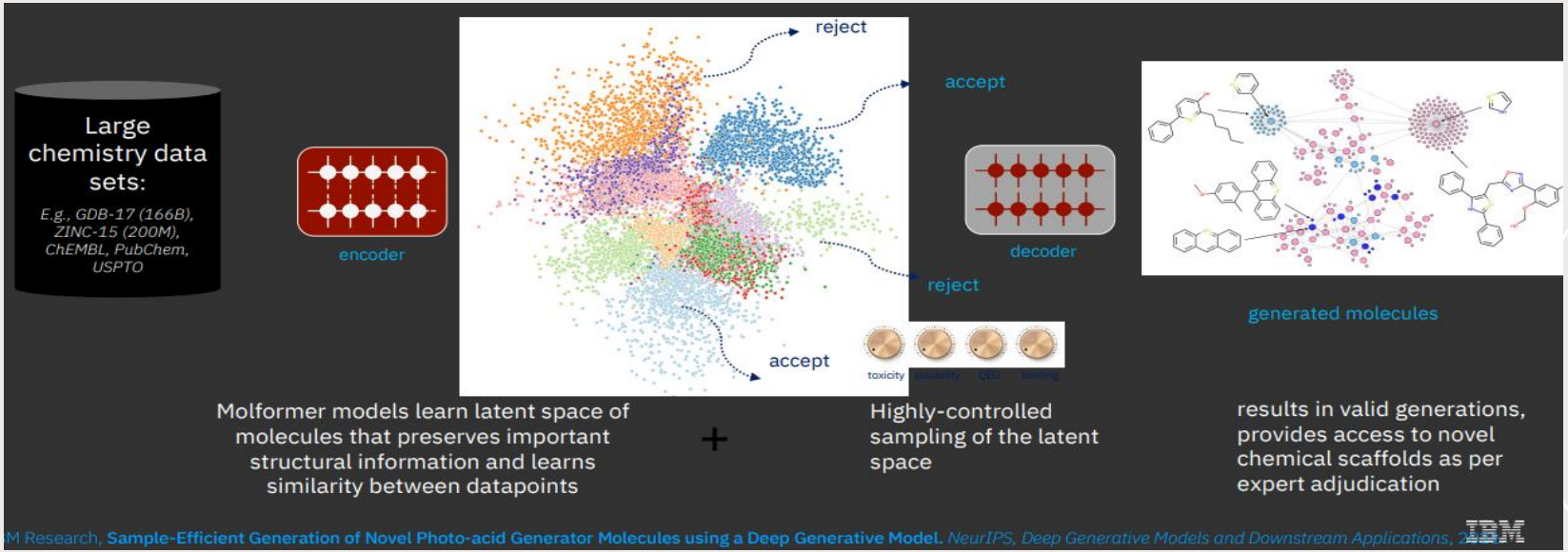
- Trained on up to over **a billion** molecular text strings (SMILES), with relatively limited hardware resources (16 V100 GPUs).
- Scalable and fast to train linear time attention transformers as encoders and decoders
- Relative position embeddings facilitate learning on SMILES
- State-of-the-art, universal chemical language model for wide ranges of **70+ molecular property prediction**
- Shows **emergent behavior**, such as geometry, taste, etc.



Source: The Promise of Foundation Models and Generative AI (2023), Payel Das, Principal Research & Master Investor

Controllable Generation of Novel Molecules

Large-scale unsupervised pretraining, novel sampling, and optimization methods enable controllable generation of novel molecules w/ desired properties



Source: The Promise of Foundation Models and Generative AI (2023), Payel Das, Principal Research & Master Investor

Comparison of MolFormer w/ others

Comparison with existing baselines on classification and regression benchmarks

Dataset	BBBP	Tox21	ClinTox	HIV	BACE	SIDER
Tasks	1	12	2	1	1	27
RF	71.4	76.9	71.3	78.1	**86.7**	**68.4**
SVM	72.9	**81.8**	66.9	**79.2**	86.2	68.2
MGCN [56]	**85.0**	70.7	63.4	73.8	73.4	55.2
D-MPNN [57]	71.2	68.9	**90.5**	75.0	85.3	63.2
Hu, et al. [58]	70.8	78.7	78.9	80.2	85.9	65.2
N-Gram [44]	91.2	76.9	85.5	**83.0**	87.6	63.2
MolCLR [24]	73.6	79.8	93.2	80.6	**89.0**	68.0
MoLFormer-XL	**93.7**	**84.7**	**94.8**	82.2	88.21	**69.0**

Source: The Promise of Foundation Models and Generative AI (2023), Payel Das, Principal Research & Master Investor

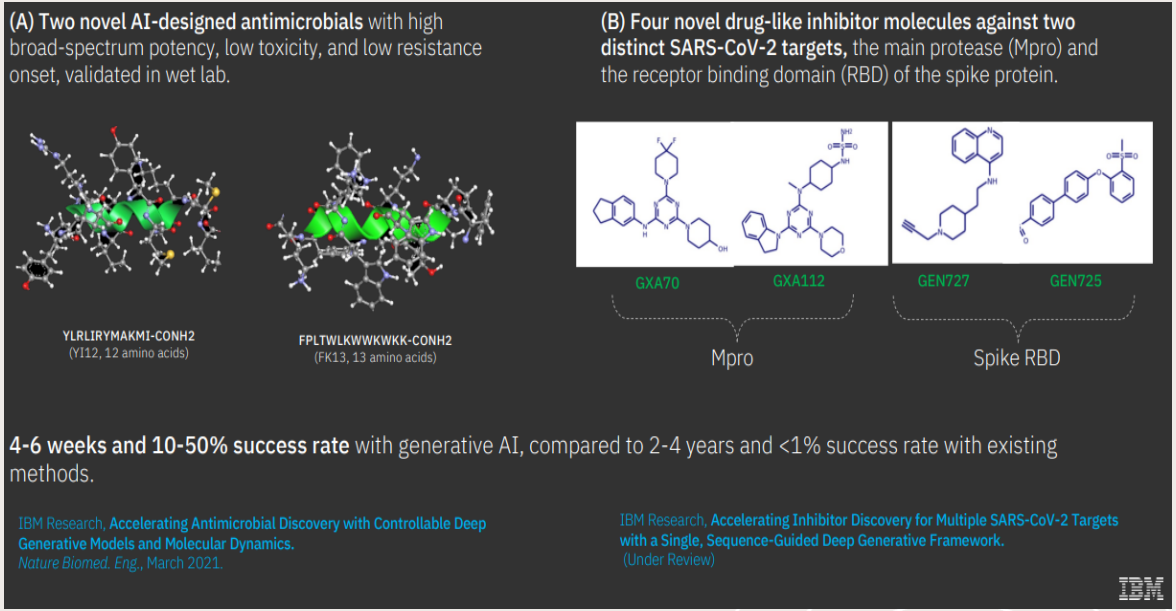
IBM Utilizes LLMs & Gen AI

SARS-CoV-2 potential molecules discovered

IBM research created a deep generative framework, CogMol, to design small molecule inhibitors for two different targets —the spike protein receptor binding domain (RBD) and the main protease from SARS-CoV-2

IBM synthesized 4 potential inhibitors for Mpro

2 of them had inhibitory activity (43 and 34.2 μM)



Source: Four novel SARS-CoV-2 molecules

IBM Utilizes LLMs & Gen AI

SARS-CoV-2 potential molecules discovered

VAE (encoder / decoder) - molecules

Pre-trained embeddings of protein

UC San Diego
SKAGGS SCHOOL OF PHARMACY AND PHARMACEUTICAL SCIENCES

BindingDB database

Gen AI
Used Conditional Latent Space Sampling (CLaSS) for attribute-controlled

IBM RXN platform

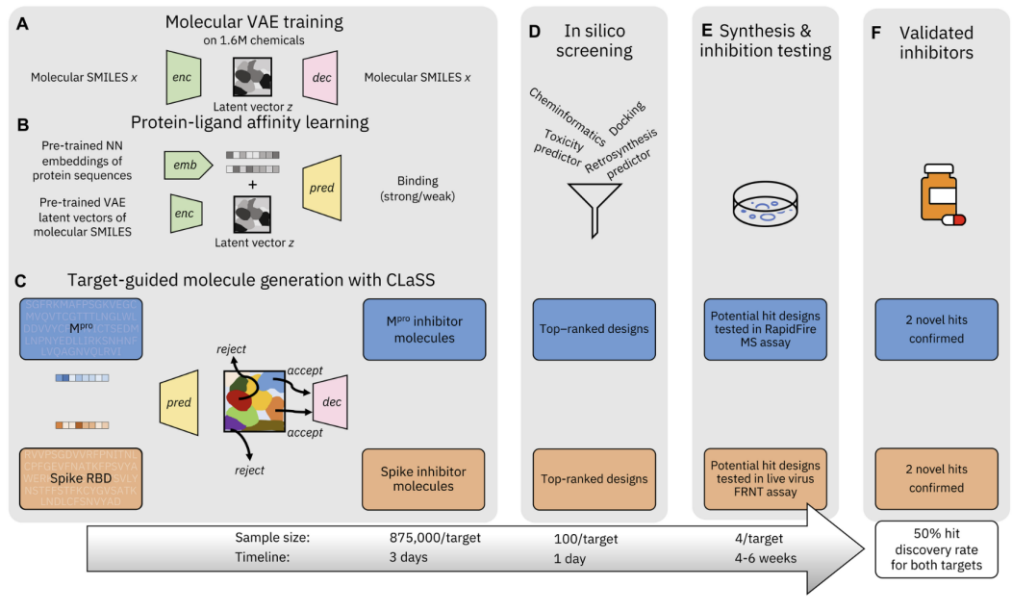


Fig. 1. Overview of our inhibitor discovery workflow driven by CogMol, a sequence-guided deep generative

Source: Target Guided

the CogMol Generative Framework relies on

- a chemical VAE,
- a protein sequence encoder,
- and a set of molecular property predictors,

all of which are pretrained on large amount of broad data—i.e., chemical SMILES, protein sequences, and available protein-ligand binding affinities.

Cytokinetics begins utilization of GT4SD

In-House Virtual Screen

Enamine Real Lead-Like Compounds
~3.5 Billion

Reduced ~1,000 x

Pharmacophore Matching
~45 Million

Reduced ~1,000 x

GT4SD Foundation AI
19,341

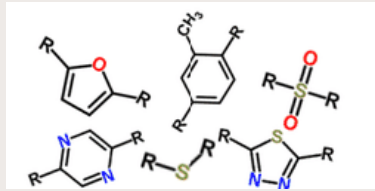
Reduced >10 x

Filter and Cluster
1,164
(3 clusters)

Reduced >5 x

Docked
200 Compounds
(3 clusters)

Ordered



RESULTS:

**** 1 VS cluster core matched top hit of traditional HTS screen!

VS & traditional HTS conducted independently

Acknowledgements

Cytokinetics Informatics Team
Marc Garard
Scott Rowland

Cytokinetics Research Management
Bradley Morgan, SVP, Research and Non-Clinical Development
Anne Murphy, VP, Biology
Ajay Chawla, VP, Translational Sciences

UGM Boston Oct. 4th 23

Chemaxon

Thank you

Joe Michel

jmichel@cytokinetics.com