NLP SUMMIT HEALTHCARE

Natural Language Processing for Drug Discovery: The State of Practices, Opportunities, and Challenges

Suneel Kumar BVS, Ph.D.,

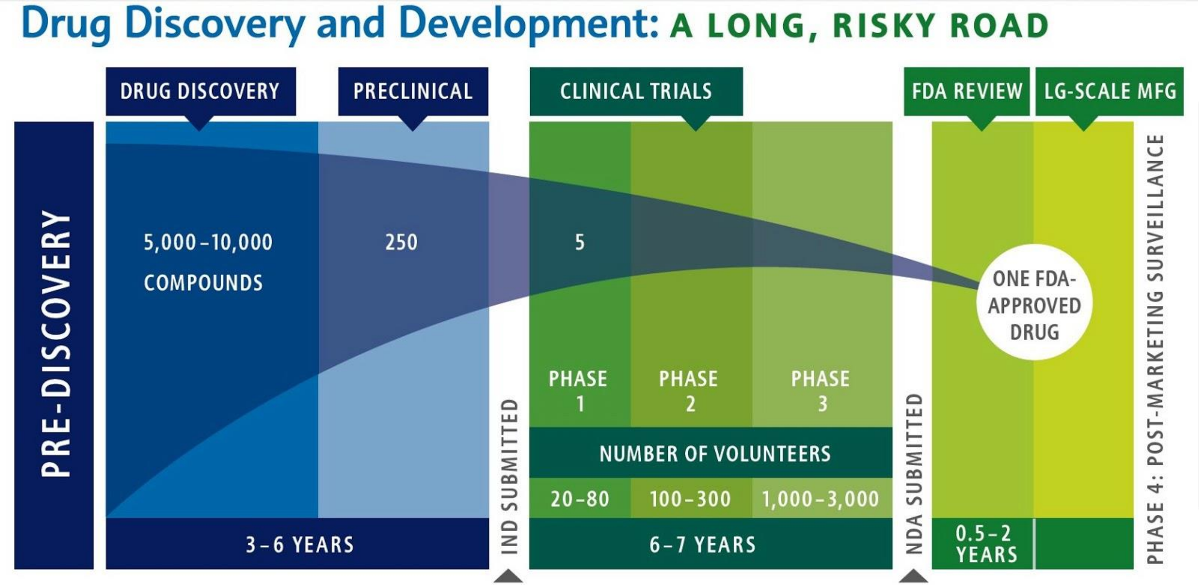
Director of Drug Design & AI
Molecular Forecaster

NLP for DD: The State of Practices, Opportunities, and Challenges

Hit identification is a crucial step in the drug discovery process, where potential drug candidates are identified from a large chemical space. There are several methods for hit identification, including structure and ligand based-virtual screening, fragment-based drug design, and AI-driven drug design approaches.

In the context of de novo drug design, both generative methods and NLP are used to extract the information about existing data points, generate and optimize molecular structures based on desired properties. These algorithms can be trained on a variety of data sources, including molecular data, its biological profile, leading to a more informed hit generation process. In this talk, I will share an overview of these methods, practices, limitations, and case studies.

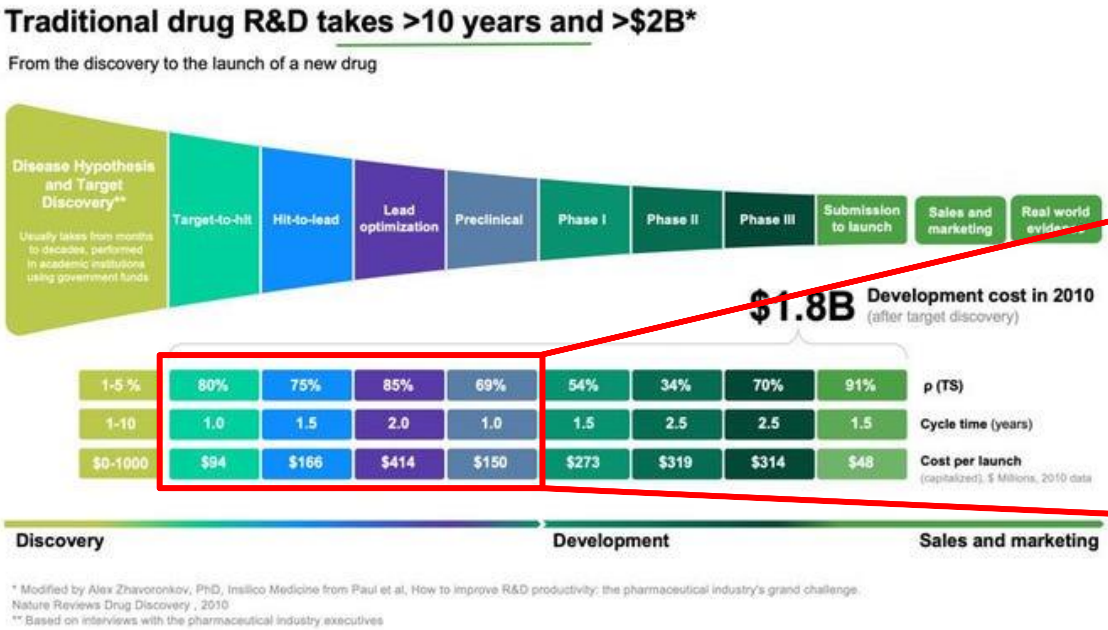
Drug Discovery - Workflow



Average time/cost for designing one drug = 10 years + $2.6B

Figure source: Pharmaceutical Research and Manufacturers of America

Drug Discovery - Workflow



Traditional drug R&D takes >10 years and >$2B*

From the discovery to the launch of a new drug

	Disease Hypothesis and Target Discovery**	Target-to-hit	Hit-to-lead	Lead optimization	Preclinical	Phase I	Phase II	Phase III	Submission to launch
p (TS)	1-5 %	80%	75%	85%	69%	54%	34%	70%	91%
Cycle time (years)	1-10	1.0	1.5	2.0	1.0	1.5	2.5	2.5	1.5
Cost per launch (capitalized), $ Millions, 2010 data	$0-1000	$94	$166	$414	$150	$273	$319	$314	$48

$1.8B Development cost in 2010 (after target discovery)

Discovery | Development | Sales and marketing

* Modified by Alex Zhavoronkov, PhD, Insilico Medicine from Paul et al, How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature Reviews Drug Discovery , 2010

** Based on interviews with the pharmaceutical industry executives

Hit to Clinical Candidate:

Cost Involved : 824$ Millions

Time : 5.5 Years

Failure Ratio (range) : 69-85%

NLP SUMMIT

Figure source: Pharmaceutical Research and Manufacturers of America

Insilico Medicine –

AI generated Lead Candidate

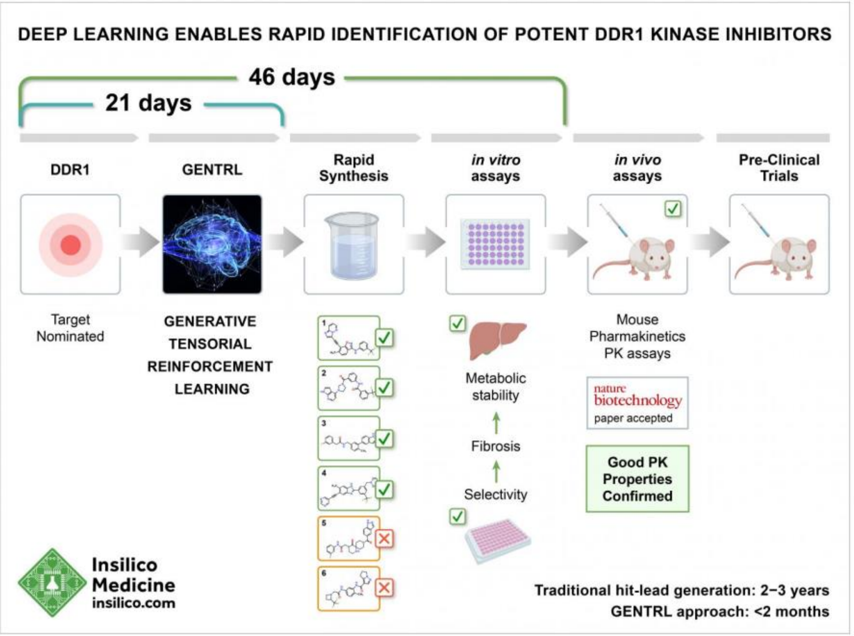
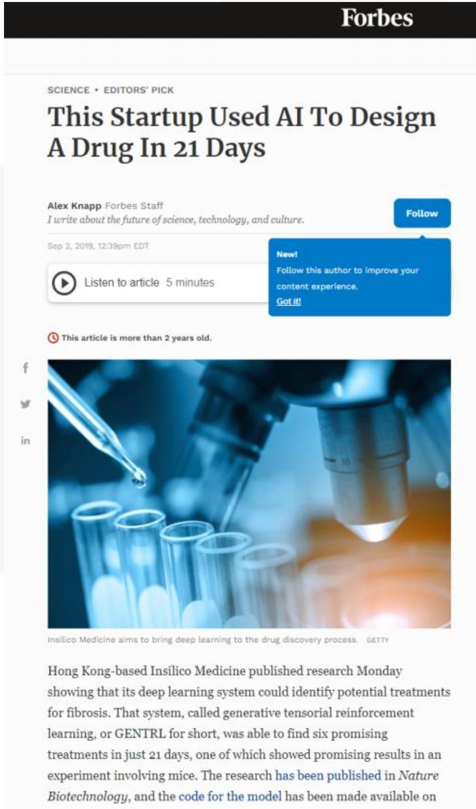

DEEP LEARNING ENABLES RAPID IDENTIFICATION OF POTENT DDR1 KINASE INHIBITORS
46 days
21 days
DDR1
GENTRL
Rapid Synthesis
in vitro assays
in vivo assays
Pre-Clinical Trials
Target Nominated
GENERATIVE TENSORIAL REINFORCEMENT LEARNING
1
2
3
4
5
6
Metabolic stability
Fibrosis
Selectivity
Mouse Pharmakinetics PK assays
nature biotechnology
paper accepted
Good PK Properties Confirmed
Insilico Medicine
insilico.com
Traditional hit-lead generation: 2–3 years
GENTRL approach: <2 months

Figure source: Insilico Medicine



SCIENCE • EDITORS' PICK

This Startup Used AI To Design A Drug In 21 Days

Alex Knapp Forbes Staff
I write about the future of science, technology, and culture.

Sep 2, 2019, 12:39pm EDT

Listen to article 5 minutes

New!
Follow this author to improve your content experience.
Got it!

This article is more than 2 years old.

Insilico Medicine aims to bring deep learning to the drug discovery process. GETTY

Hong Kong-based Insilico Medicine published research Monday showing that its deep learning system could identify potential treatments for fibrosis. That system, called generative tensorial reinforcement learning, or GENTRL for short, was able to find six promising treatments in just 21 days, one of which showed promising results in an experiment involving mice. The research has been published in *Nature Biotechnology*, and the code for the model has been made available on

Insilico Medicine –

AI generated Lead Candidate

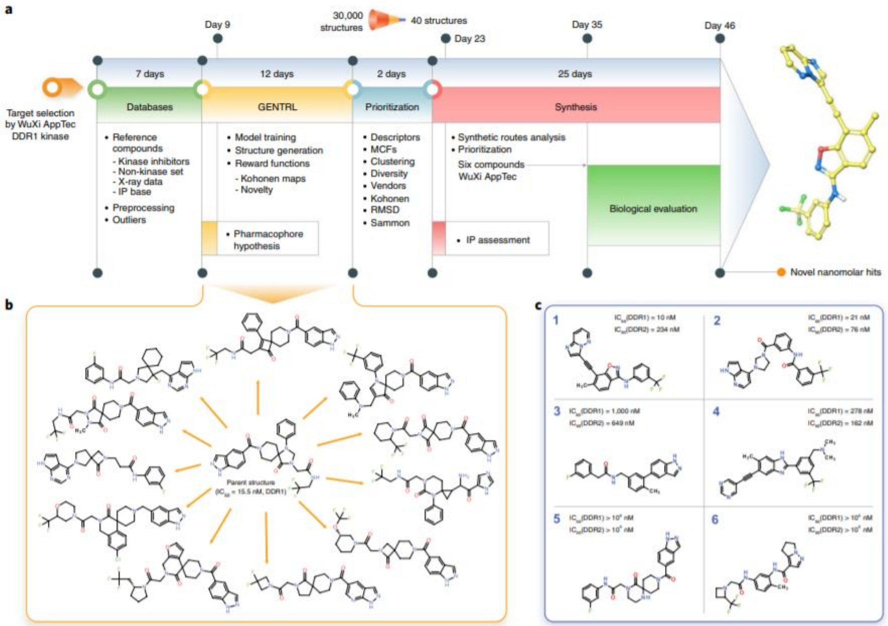
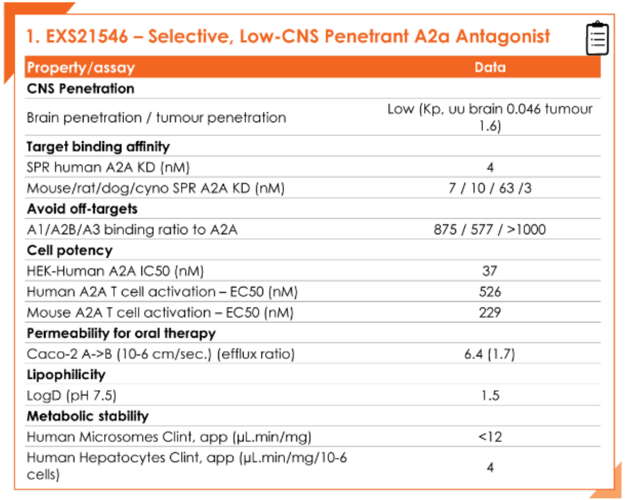

a
Target selection by WuXi AppTec DDR1 kinase
Day 9
30,000 structures
40 structures
Day 23
Day 35
Day 46
7 days
12 days
2 days
25 days
Databases
GENTRL
Prioritization
Synthesis
• Reference compounds
- Kinase inhibitors
- Non-kinase set
- X-ray data
- IP base
• Preprocessing
• Outliers
• Model training
• Structure generation
• Reward functions
- Kohonen maps
- Novelty
• Pharmacophore hypothesis
• Descriptors
• MCFs
• Clustering
• Diversity
• Vendors
• Kohonen
• RMSD
• Sammon
• Synthetic routes analysis
• Prioritization
Six compounds WuXi AppTec
• IP assessment
Biological evaluation
Novel nanomolar hits
b
Parent structure (IC_{50} = 15.5 nM, DDR1)
c
1 IC_{50}(DDR1) = 10 nM IC_{50}(DDR2) = 234 nM
2 IC_{50}(DDR1) = 21 nM IC_{50}(DDR2) = 76 nM
3 IC_{50}(DDR1) = 1,000 nM IC_{50}(DDR2) = 649 nM
4 IC_{50}(DDR1) = 278 nM IC_{50}(DDR2) = 162 nM
5 IC_{50}(DDR1) > 10^4 nM IC_{50}(DDR2) > 10^4 nM
6 IC_{50}(DDR1) > 10^4 nM IC_{50}(DDR2) > 10^4 nM

Fig. GENTRL model design, workflow, and nanomolar hits. a, The general workflow and timeline for the design of lead candidates against human DDR1 kinase.

Figure source: Insilico Medicine

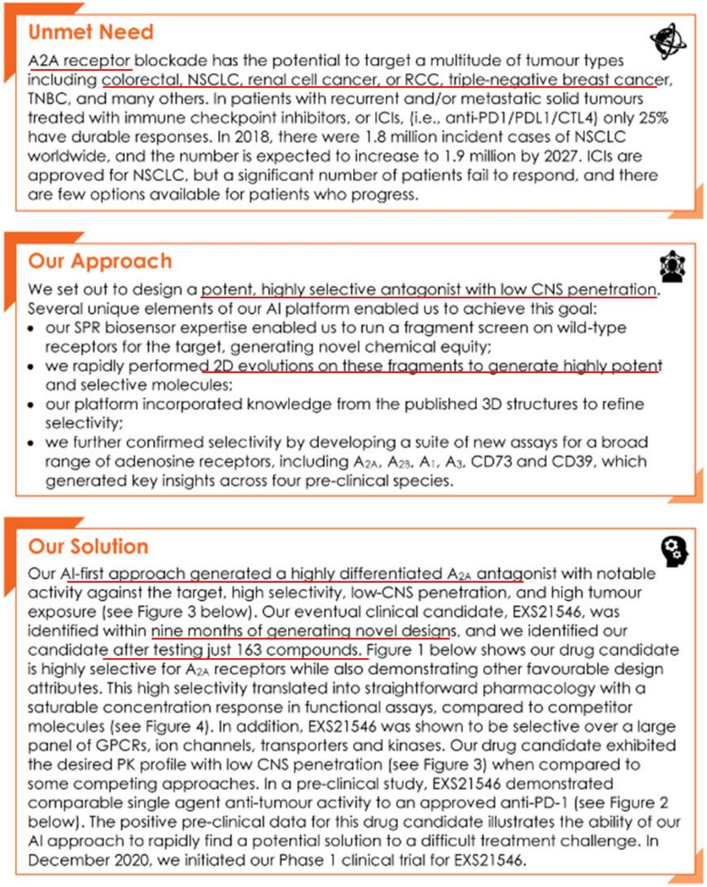
Exscientia

– AI Clinical Candidate



1. EXS21546 – Selective, Low-CNS Penetrant A2a Antagonist

Property/assay	Data
CNS Penetration	
Brain penetration / tumour penetration	Low (Kp, uu brain 0.046 tumour 1.6)
Target binding affinity	
SPR human A2A KD (nM)	4
Mouse/rat/dog/cyno SPR A2A KD (nM)	7 / 10 / 63 /3
Avoid off-targets	
A1/A2B/A3 binding ratio to A2A	875 / 577 / >1000
Cell potency	
HEK-Human A2A IC50 (nM)	37
Human A2A T cell activation – EC50 (nM)	526
Mouse A2A T cell activation – EC50 (nM)	229
Permeability for oral therapy	
Caco-2 A->B (10-6 cm/sec.) (efflux ratio)	6.4 (1.7)
Lipophilicity	
LogD (pH 7.5)	1.5
Metabolic stability	
Human Microsomes Clint, app (µL.min/mg)	<12
Human Hepatocytes Clint, app (µL.min/mg/10-6 cells)	4



Unmet Need

A2A receptor blockade has the potential to target a multitude of tumour types including colorectal, NSCLC, renal cell cancer, or RCC, triple-negative breast cancer, TNBC, and many others. In patients with recurrent and/or metastatic solid tumours treated with immune checkpoint inhibitors, or ICIs, (i.e., anti-PD1/PDL1/CTL4) only 25% have durable responses. In 2018, there were 1.8 million incident cases of NSCLC worldwide, and the number is expected to increase to 1.9 million by 2027. ICIs are approved for NSCLC, but a significant number of patients fail to respond, and there are few options available for patients who progress.

Our Approach

We set out to design a potent, highly selective antagonist with low CNS penetration. Several unique elements of our AI platform enabled us to achieve this goal:

- our SPR biosensor expertise enabled us to run a fragment screen on wild-type receptors for the target, generating novel chemical equity;
- we rapidly performed 2D evolutions on these fragments to generate highly potent and selective molecules;
- our platform incorporated knowledge from the published 3D structures to refine selectivity;
- we further confirmed selectivity by developing a suite of new assays for a broad range of adenosine receptors, including A_{2A}, A_{2B}, A_1, A_3, CD73 and CD39, which generated key insights across four pre-clinical species.

Our Solution

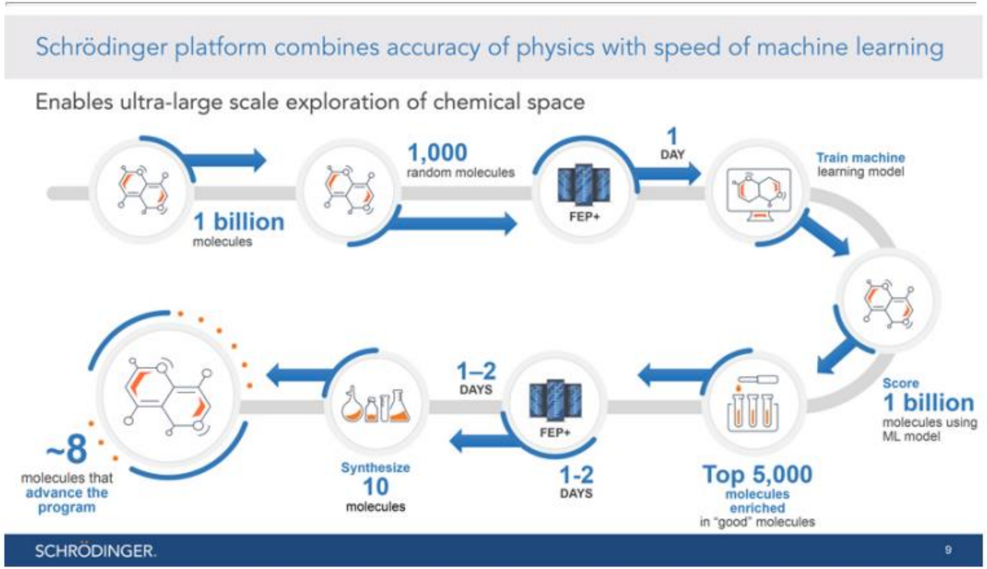
Our AI-first approach generated a highly differentiated A_{2A} antagonist with notable activity against the target, high selectivity, low-CNS penetration, and high tumour exposure (see Figure 3 below). Our eventual clinical candidate, EXS21546, was identified within nine months of generating novel designs, and we identified our candidate after testing just 163 compounds. Figure 1 below shows our drug candidate is highly selective for A_{2A} receptors while also demonstrating other favourable design attributes. This high selectivity translated into straightforward pharmacology with a saturable concentration response in functional assays, compared to competitor molecules (see Figure 4). In addition, EXS21546 was shown to be selective over a large panel of GPCRs, ion channels, transporters and kinases. Our drug candidate exhibited the desired PK profile with low CNS penetration (see Figure 3) when compared to some competing approaches. In a pre-clinical study, EXS21546 demonstrated comparable single agent anti-tumour activity to an approved anti-PD-1 (see Figure 2 below). The positive pre-clinical data for this drug candidate illustrates the ability of our AI approach to rapidly find a potential solution to a difficult treatment challenge. In December 2020, we initiated our Phase 1 clinical trial for EXS21546.

NLP SUMMIT

Figure source: Exscientia reports

Schrodinger

– AI/ML Clinical Candidate



Schrödinger platform combines accuracy of physics with speed of machine learning

Enables ultra-large scale exploration of chemical space

1 billion molecules

1,000 random molecules

FEP+

1 DAY

Train machine learning model

Score 1 billion molecules using ML model

Top 5,000 molecules enriched in "good" molecules

1-2 DAYS

FEP+

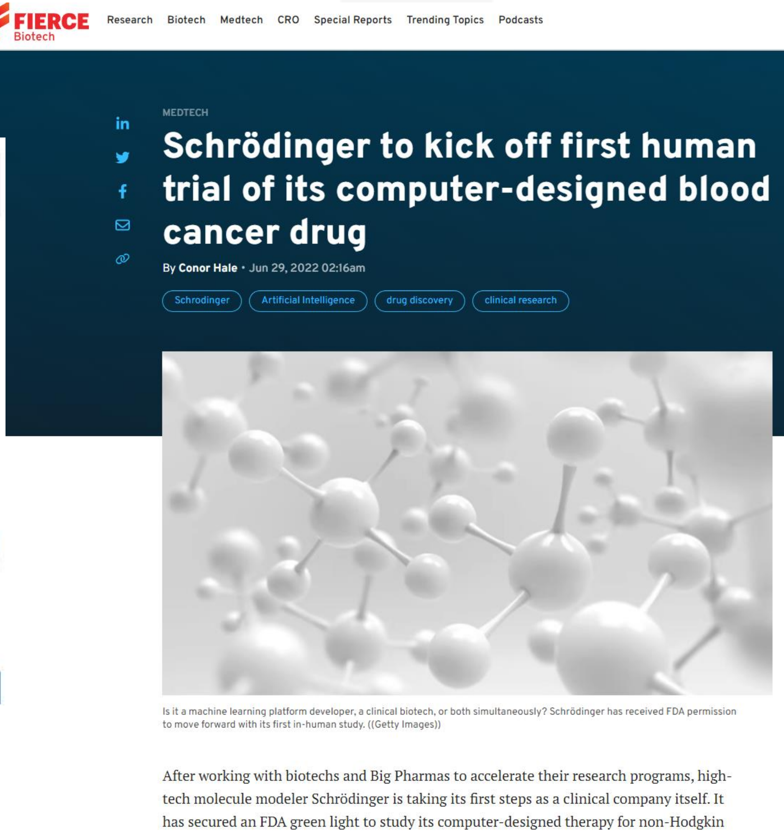
1–2 DAYS

Synthesize 10 molecules

~8 molecules that advance the program

SCHRÖDINGER.

9



FIERCE Biotech

Research Biotech Medtech CRO Special Reports Trending Topics Podcasts

MEDTECH

Schrödinger to kick off first human trial of its computer-designed blood cancer drug

By Conor Hale • Jun 29, 2022 02:16am

Schrodinger | Artificial Intelligence | drug discovery | clinical research

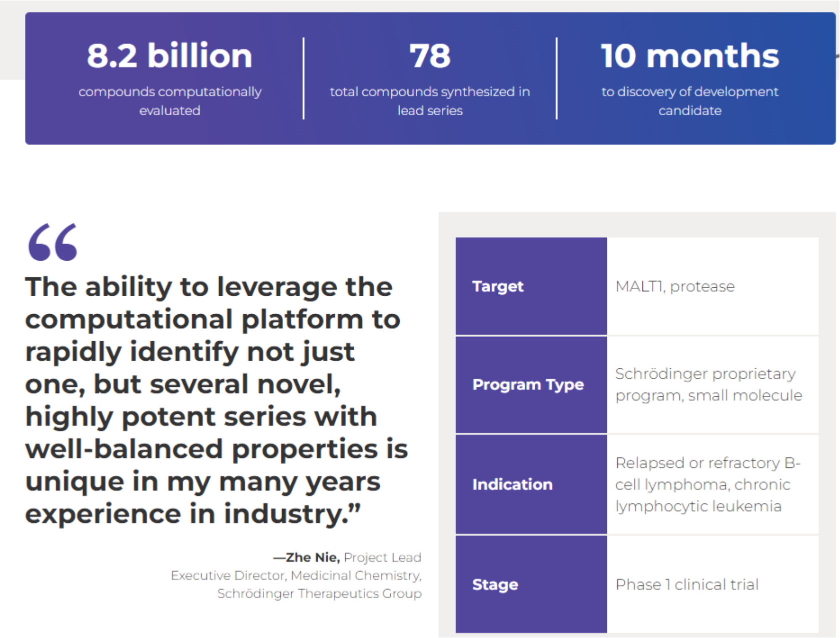
Is it a machine learning platform developer, a clinical biotech, or both simultaneously? Schrödinger has received FDA permission to move forward with its first in-human study. ((Getty Images))

After working with biotechs and Big Pharmas to accelerate their research programs, high-tech molecule modeler Schrödinger is taking its first steps as a clinical company itself. It has secured an FDA green light to study its computer-designed therapy for non-Hodgkin

Figure source: Schrodinger.com

Schrodinger

– AI/ML Clinical Candidate



8.2 billion
compounds computationally evaluated

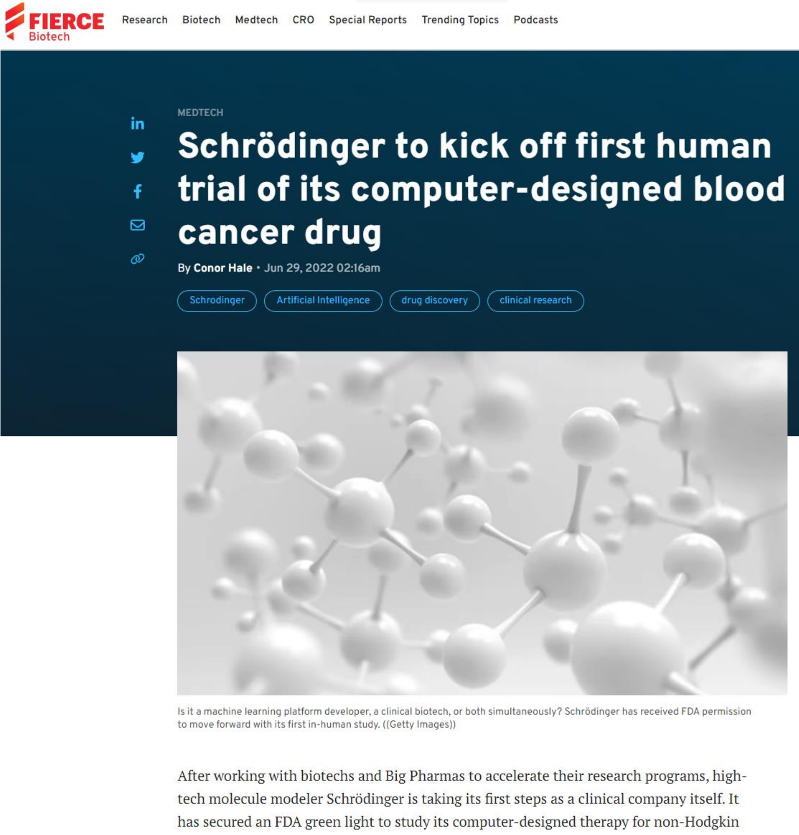
78
total compounds synthesized in lead series

10 months
to discovery of development candidate

> “The ability to leverage the computational platform to rapidly identify not just one, but several novel, highly potent series with well-balanced properties is unique in my many years experience in industry.”

—**Zhe Nie,** Project Lead
Executive Director, Medicinal Chemistry,
Schrödinger Therapeutics Group

Target	MALT1, protease
Program Type	Schrödinger proprietary program, small molecule
Indication	Relapsed or refractory B-cell lymphoma, chronic lymphocytic leukemia
Stage	Phase 1 clinical trial



MEDTECH

Schrödinger to kick off first human trial of its computer-designed blood cancer drug

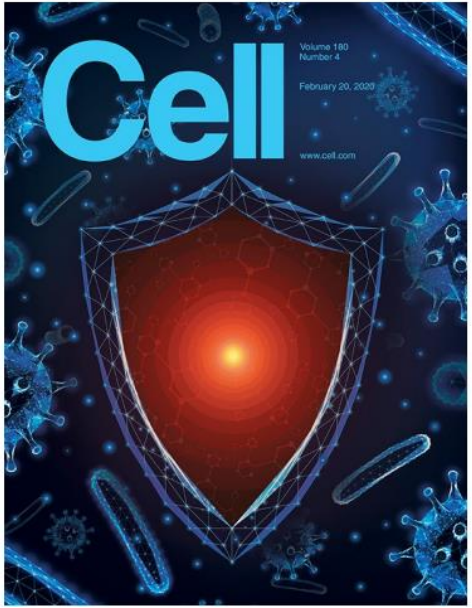
By **Conor Hale** • Jun 29, 2022 02:16am

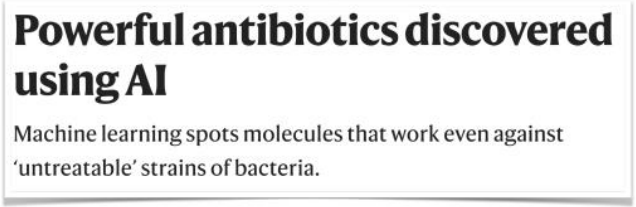
Is it a machine learning platform developer, a clinical biotech, or both simultaneously? Schrödinger has received FDA permission to move forward with its first in-human study. ((Getty Images))

After working with biotechs and Big Pharmas to accelerate their research programs, high-tech molecule modeler Schrödinger is taking its first steps as a clinical company itself. It has secured an FDA green light to study its computer-designed therapy for non-Hodgkin

Figure source: Schrodinger.com

And Many More Success Stories

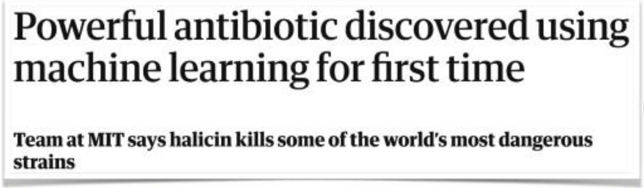




Powerful antibiotics discovered using AI

Machine learning spots molecules that work even against 'untreatable' strains of bacteria.

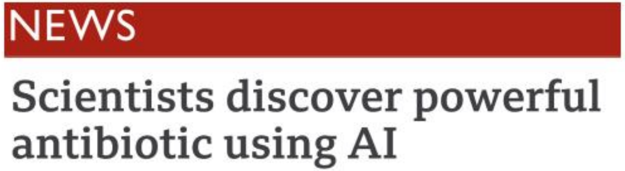
nature



Powerful antibiotic discovered using machine learning for first time

Team at MIT says halicin kills some of the world's most dangerous strains

The Guardian



NEWS

Scientists discover powerful antibiotic using AI

BBC

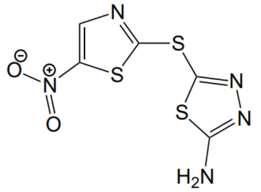
NLP SUMMIT

Figure Source: Stokes, Yang, Swanson, Jin et al, Cell 2020

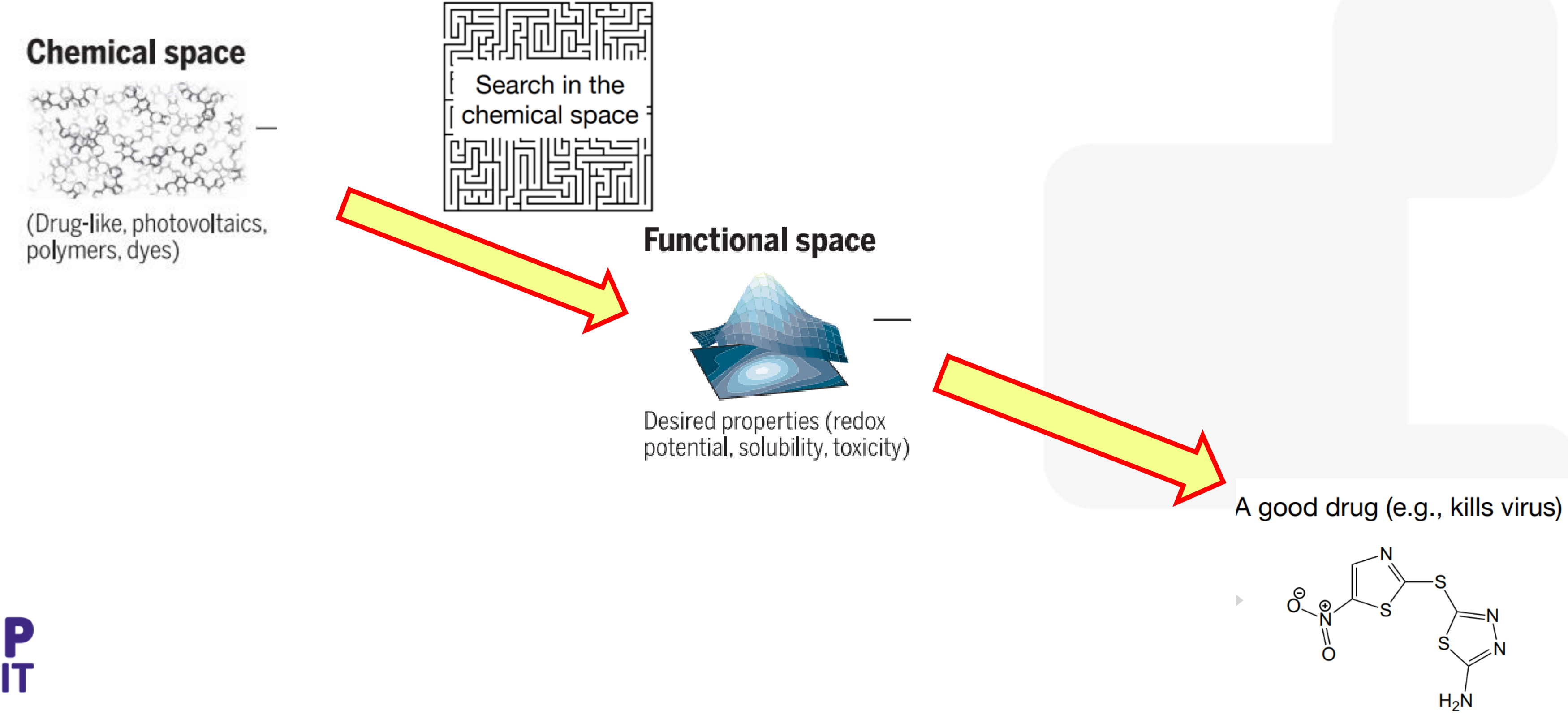
Drug Discovery – Complex, Challenging Process

Biological target: macromolecule (such as., protein, RNA) involved in the disease pathway
Compound/Hit/Lead/Candidate : Binds and interact with the target
Toxicity : isn't harmful to organism
Selectivity : binding specifically to desired biological target
and many more such as., **Solubility, Caco2, ADME, t1/2, and PK/PD**

A good drug (e.g., kills virus)



Title



Chemical space

(Drug-like, photovoltaics, polymers, dyes)

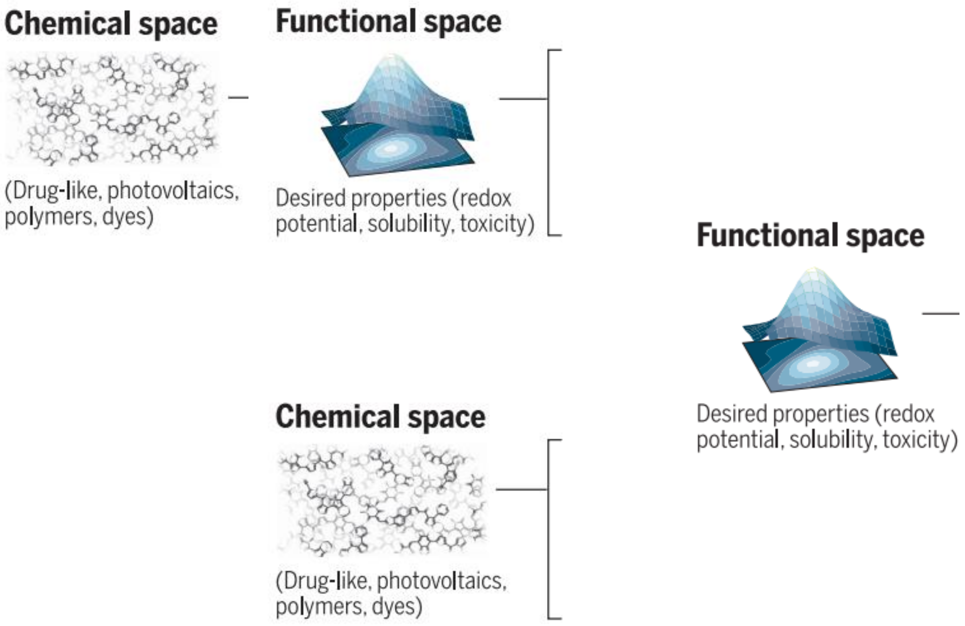
Search in the chemical space

Functional space

Desired properties (redox potential, solubility, toxicity)

A good drug (e.g., kills virus)

Drug Discovery – Complex, Challenging Process



Chemical space

(Drug-like, photovoltaics, polymers, dyes)

Functional space

Desired properties (redox potential, solubility, toxicity)

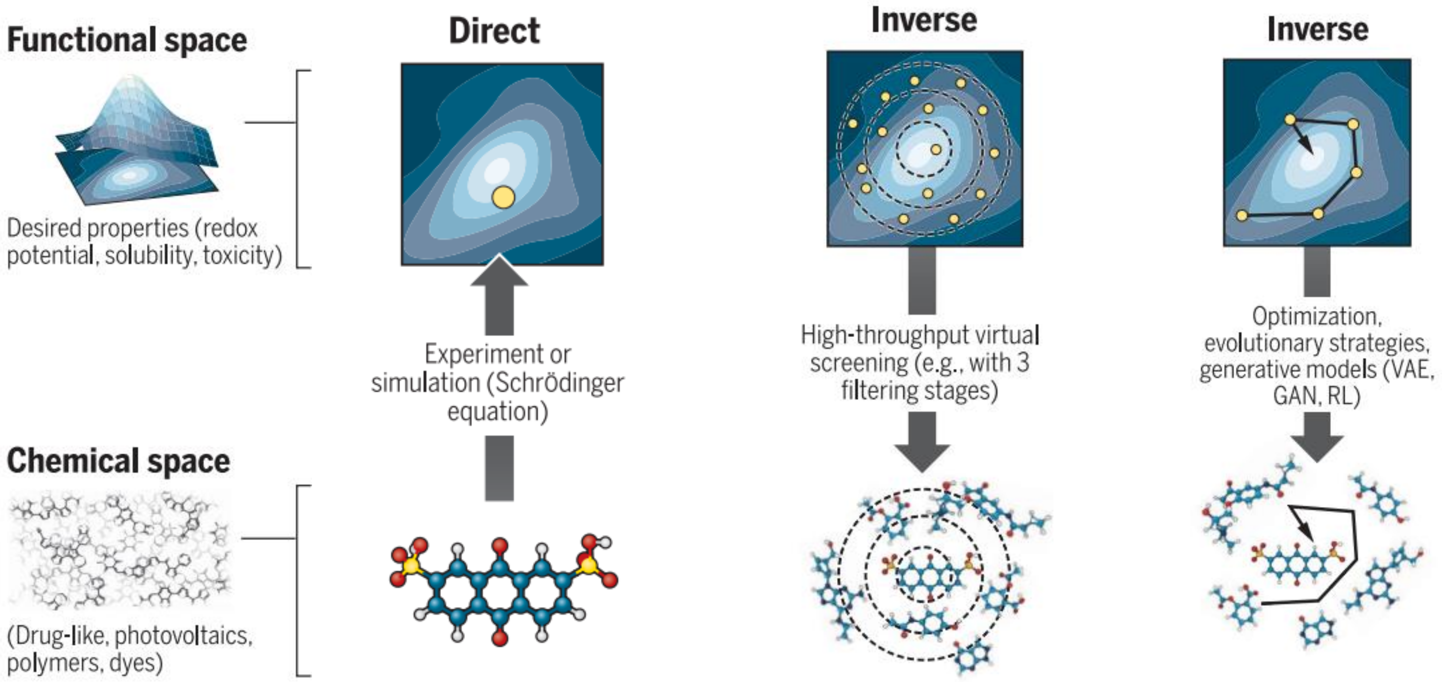
Chemical space

(Drug-like, photovoltaics, polymers, dyes)

Functional space

Desired properties (redox potential, solubility, toxicity)

Drug Designing – Strategies


Functional space
Desired properties (redox potential, solubility, toxicity)
Chemical space
(Drug-like, photovoltaics, polymers, dyes)
Direct
Experiment or simulation (Schrödinger equation)
Inverse
High-throughput virtual screening (e.g., with 3 filtering stages)
Inverse
Optimization, evolutionary strategies, generative models (VAE, GAN, RL)

Molecular Representations

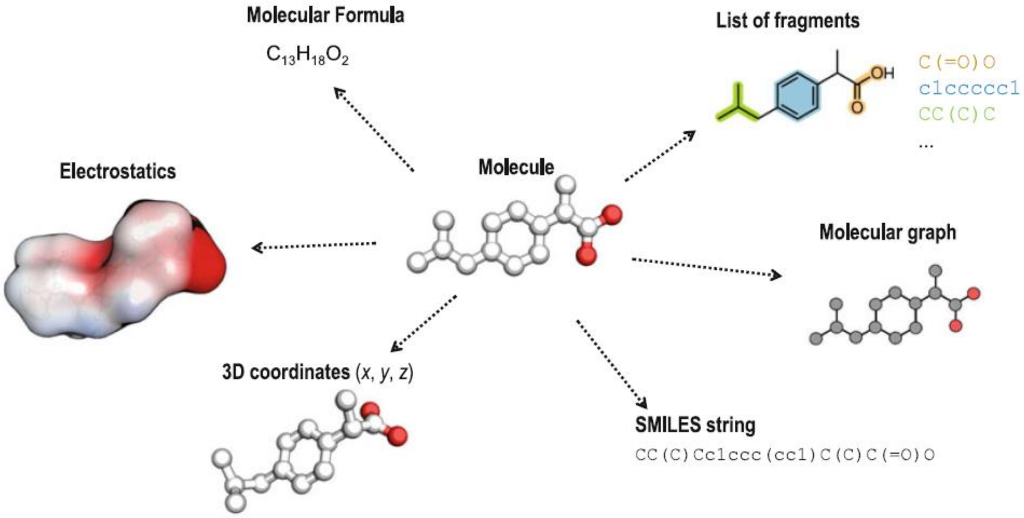
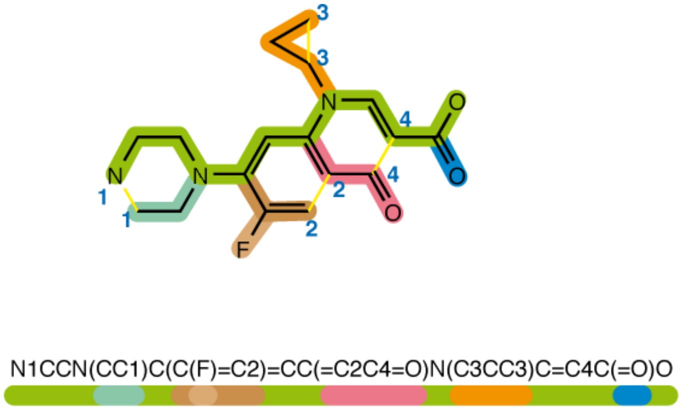


Figure source: Exscientia reports

Molecular Representations

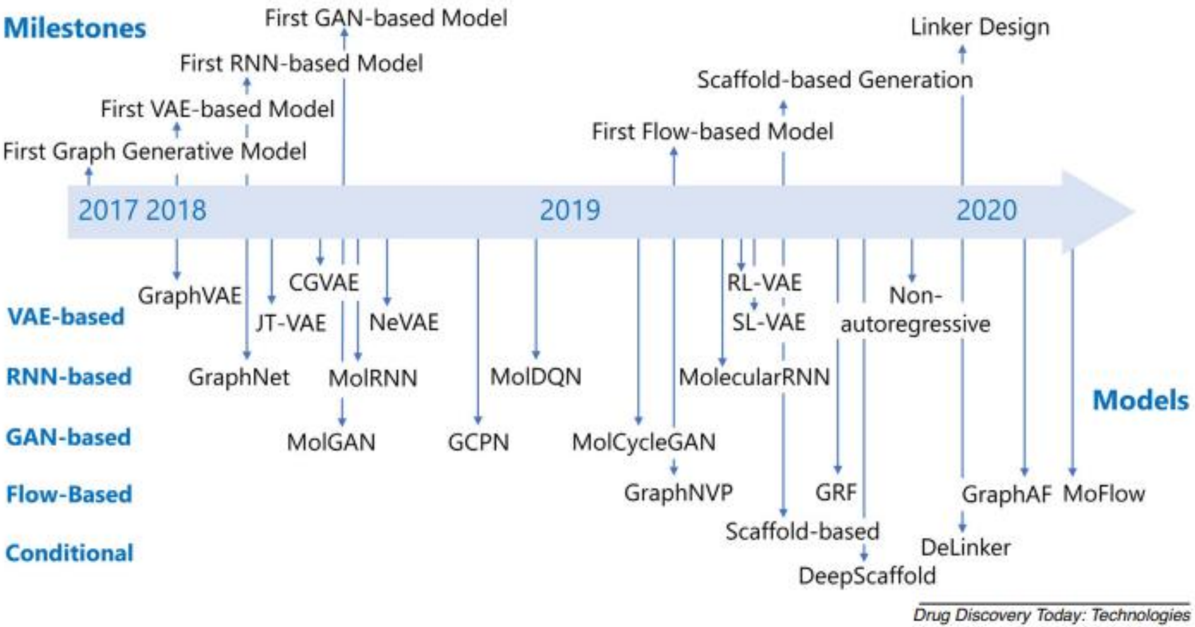


SMILES string for ciprofloxacin

- SMILES (Simplified molecular-input line-entry system)
- string representation

Figure source: Wikipedia

Generative AI methods in DD



Drug Discovery Today: Technologies

NLP SUMMIT

Generative AI methods in DD

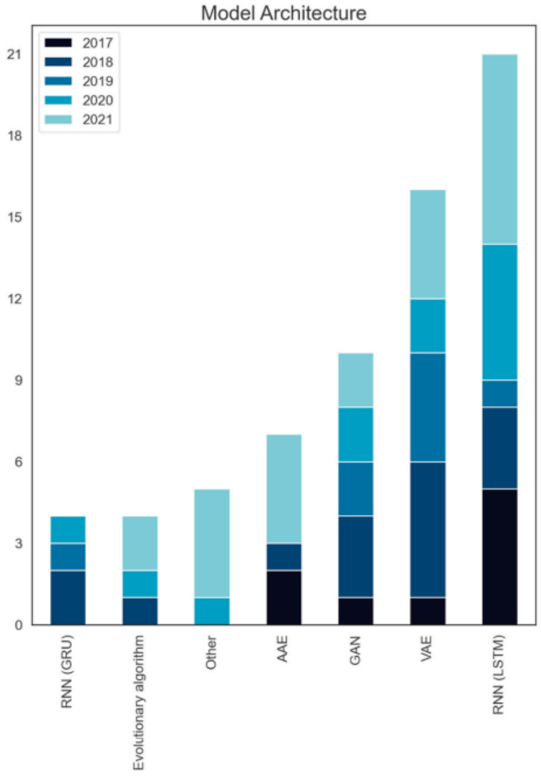
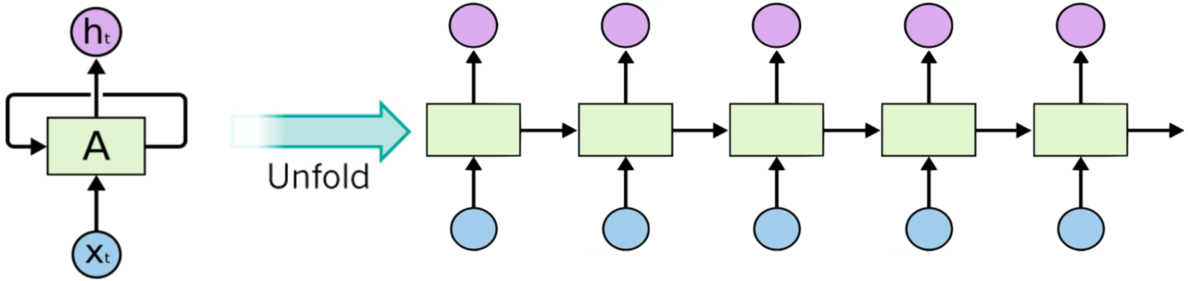
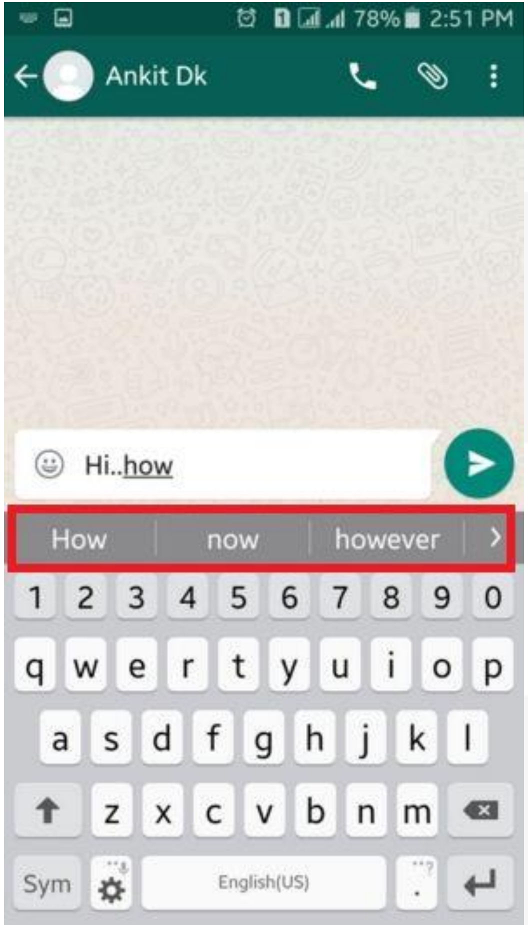


Fig. The relative frequencies of the various AI/ML model architectures observed in the review and their chronological progression.

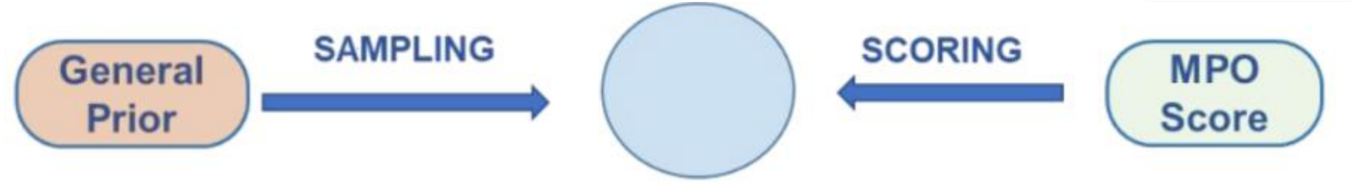
Title


h
t
A
X
t
Unfold


78% 2:51 PM
Ankit Dk
How
now
however
1 2 3 4 5 6 7 8 9 0
q w e r t y u i o p
a s d f g h j k l
z x c v b n m
Sym
English(US)

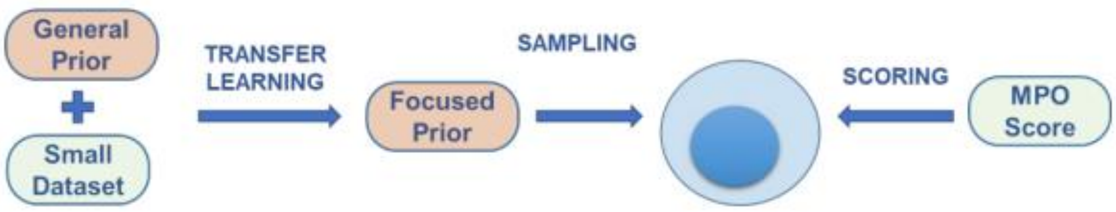
Hi..how

REINVENT – WORKFLOWS for denovo design



Source : J. Chem. Inf. Model. 2020, 60, 12, 5918–5922

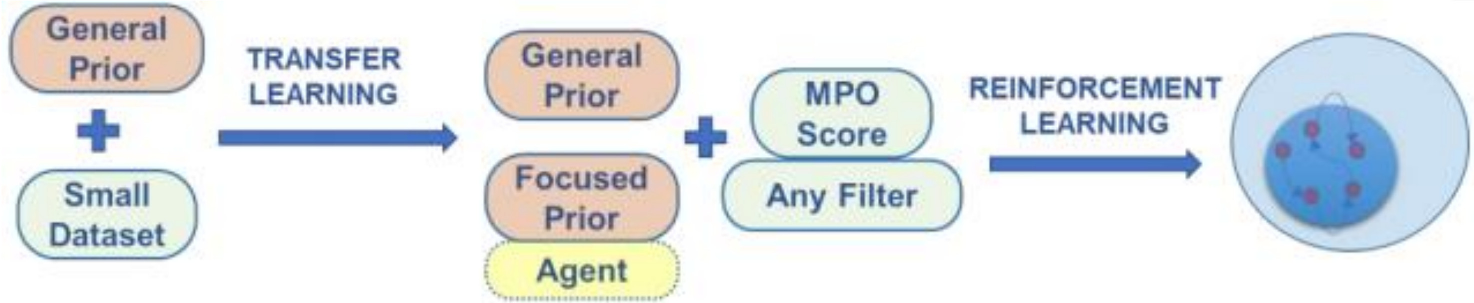
REINVENT – WORKFLOWS for denovo design



- The generative model is subjected to transfer learning with a smaller set of compounds that are relevant to the project of interest
- This will bias the resulting model to produce project specific compounds with much higher probability than any random compounds
- Therefore much smaller dataset can suffice to find good hits when using the scoring function

Source : J. Chem. Inf. Model. 2020, 60, 12, 5918–5922

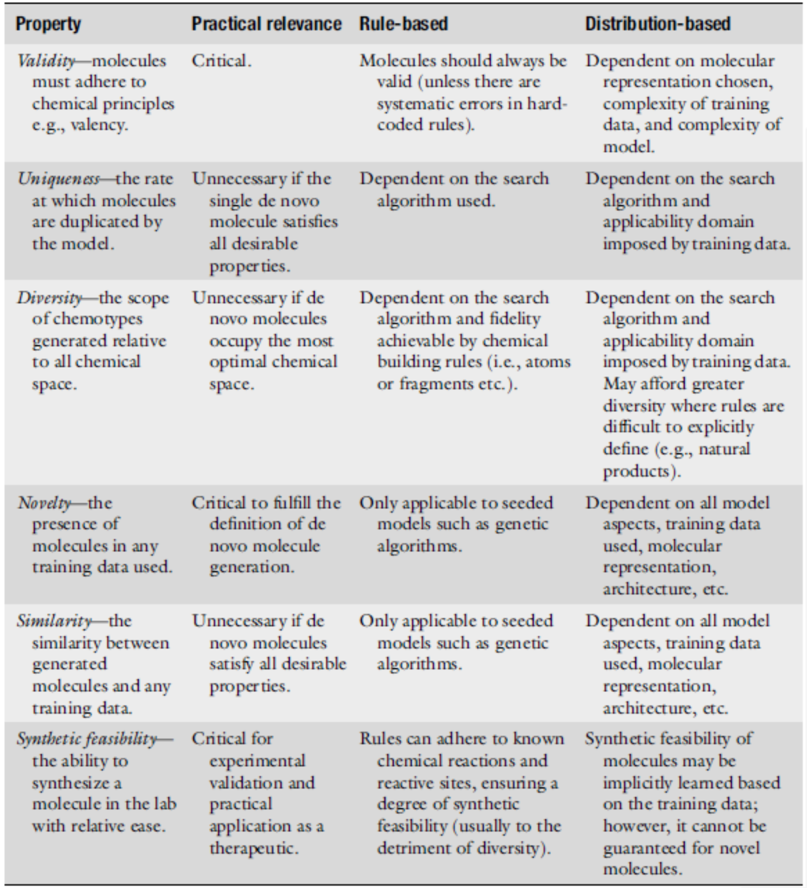
REINVENT – WORKFLOWS for denovo design



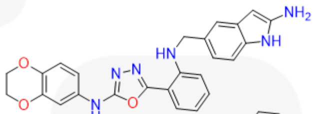
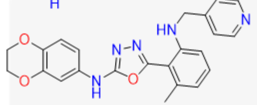
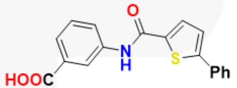
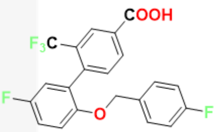
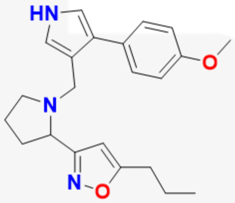
- Transfer learning is conducted with a set of compounds that have the desired properties
- For example, if we aim to maximize a predictive model among the other components we would use all the compounds that are considered as active by this model; If we aim towards certain subseries of compounds we would only use those that share the specific features for transfer learning
- After concluding with transfer learning the resulting agent is **“focused”** on the specific set
- We can conduct TL for multiple epochs & observe the stats from each epoch thus deciding which agent is focused enough

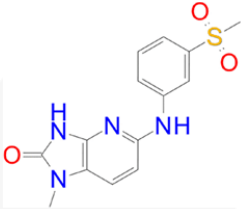
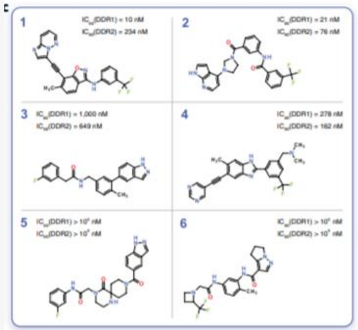
Source : J. Chem. Inf. Model. 2020, 60, 12, 5918–5922

Benchmarking?



Property	Practical relevance	Rule-based	Distribution-based
Validity—molecules must adhere to chemical principles e.g., valency.	Critical.	Molecules should always be valid (unless there are systematic errors in hard-coded rules).	Dependent on molecular representation chosen, complexity of training data, and complexity of model.
Uniqueness—the rate at which molecules are duplicated by the model.	Unnecessary if the single de novo molecule satisfies all desirable properties.	Dependent on the search algorithm used.	Dependent on the search algorithm and applicability domain imposed by training data.
Diversity—the scope of chemotypes generated relative to all chemical space.	Unnecessary if de novo molecules occupy the most optimal chemical space.	Dependent on the search algorithm and fidelity achievable by chemical building rules (i.e., atoms or fragments etc.).	Dependent on the search algorithm and applicability domain imposed by training data. May afford greater diversity where rules are difficult to explicitly define (e.g., natural products).
Novelty—the presence of molecules in any training data used.	Critical to fulfill the definition of de novo molecule generation.	Only applicable to seeded models such as genetic algorithms.	Dependent on all model aspects, training data used, molecular representation, architecture, etc.
Similarity—the similarity between generated molecules and any training data.	Unnecessary if de novo molecules satisfy all desirable properties.	Only applicable to seeded models such as genetic algorithms.	Dependent on all model aspects, training data used, molecular representation, architecture, etc.
Synthetic feasibility—the ability to synthesize a molecule in the lab with relative ease.	Critical for experimental validation and practical application as a therapeutic.	Rules can adhere to known chemical reactions and reactive sites, ensuring a degree of synthetic feasibility (usually to the detriment of diversity).	Synthetic feasibility of molecules may be implicitly learned based on the training data; however, it cannot be guaranteed for novel molecules.

Target	Training	Approach	Cpds Generated	Filters	Screening?	Active Hits
2017 VEGFR2 University of Oxford	VEGFR2 actives (25,000) from Binding DB	char-RNN	10,000 Compounds	• Solubility • Synthesizability • Similarity with VEGFR2 actives • Docking Score Cutoff	5 Compounds Tested 2 Compounds Active @ (1-1000 nM)	 
2018 PPARs & RXRs agonists ETH Zurich	25 RXR/PPAR fatty acid Active Mimetics	LSTM	1000 Compounds (–COOH)	• SPIDER – target predic tion • BB's availability • Similarity with known actives • Docking Score Cut-off	5 Compounds Tested 4 Compounds Active @ (0.06 - 14 uM)	 
2018 Kinase Panel Insilico Medicine	30,000 Kinas e inhibitors fr om Thomson Integrity DB	ATNC	30,000 Compounds	• MedChem Filters • Clustering • Reward function for di versity • Similarity with known actives (> 0.7)	50 cpds Tested 7 cpds Active @ (100% inh 10 uM)	

Target	Training	Approach	Cpds Generated	Filters	Screening?	Active Hits
2018 JAK2/JAK3 **Insilico Medicine**	Known actives	SSAAE	30,000 Compounds	• Docking Score • Off target effects • Property Analysis • Molecular dynamics • MedChem Feedback	1 Compound Tested 1 Compound Active @ (6 M)	
2018 DDR1 **Insilico Medicine**	17,000 Kinase inhibitors + DDR1 active cpds	GENTRL	30,000 Compounds	• MedChem Filters • Clustering • Pharmacophore • Synthesizability • MedChem filters	6 cpds Tested 4 cpds Active @ nM range	

Limitations and future scope

- Generation in Low-data Regime
- Lack of Unified Evaluation Protocols
- Lack of Large-scale Study and Benchmark
- Out-of-distribution Generation
- Unrealistic Problem Formulation
- Ideal assumptions/expensive Oracle Calls
- Lack of Interpretability

NLP SUMMIT
HEALTHCARE

Presented by

John Snow LABS

www.nlpsummit.org