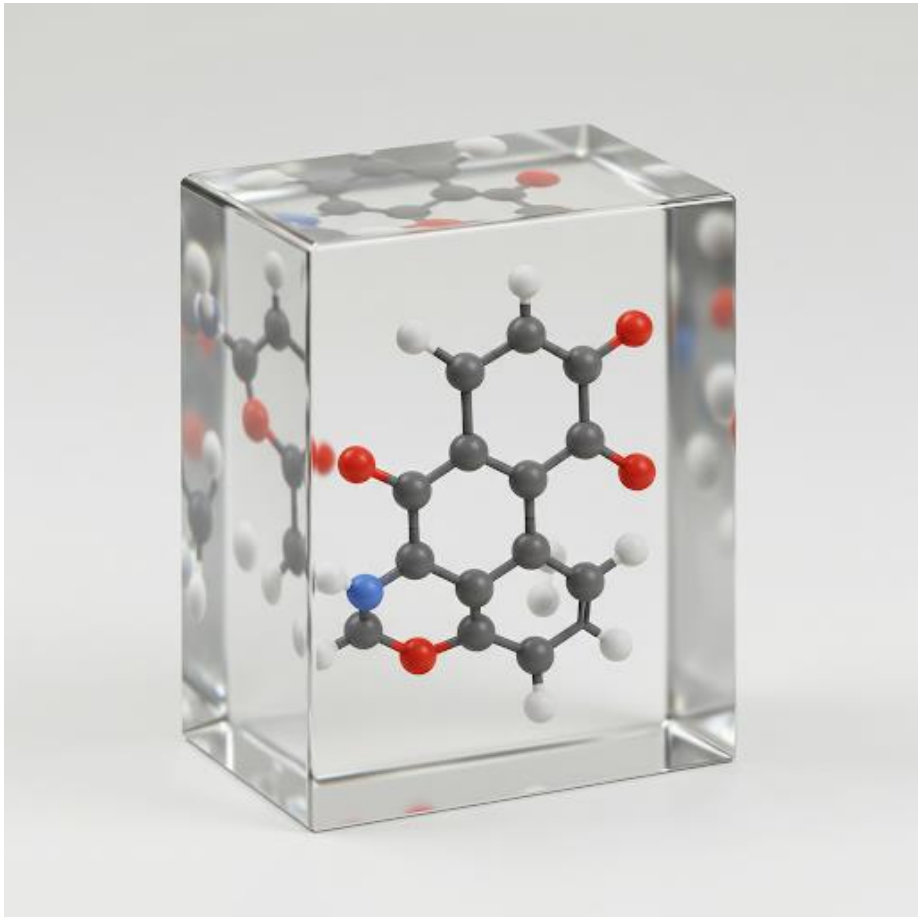
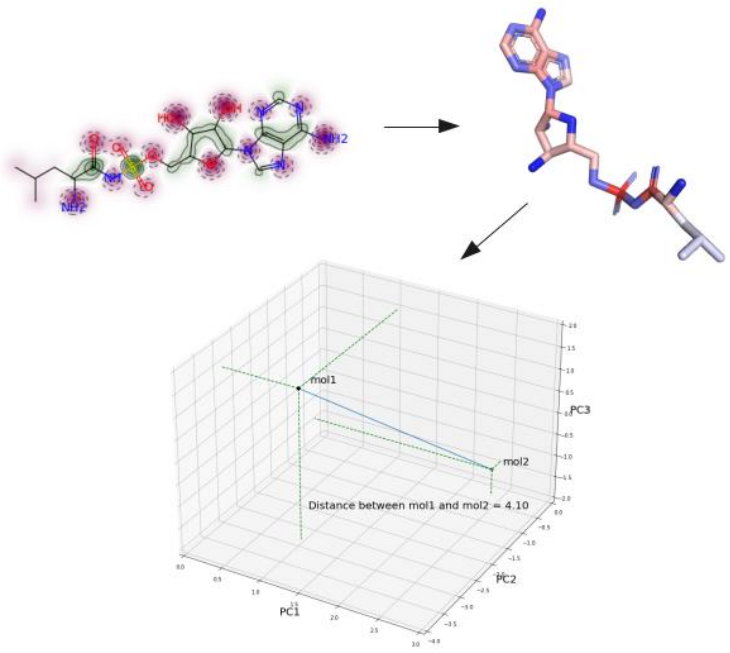
SOME CHALLENGES AND OPPORTUNITIES FOR AI IN PRECLINICAL DRUG DISCOVERY

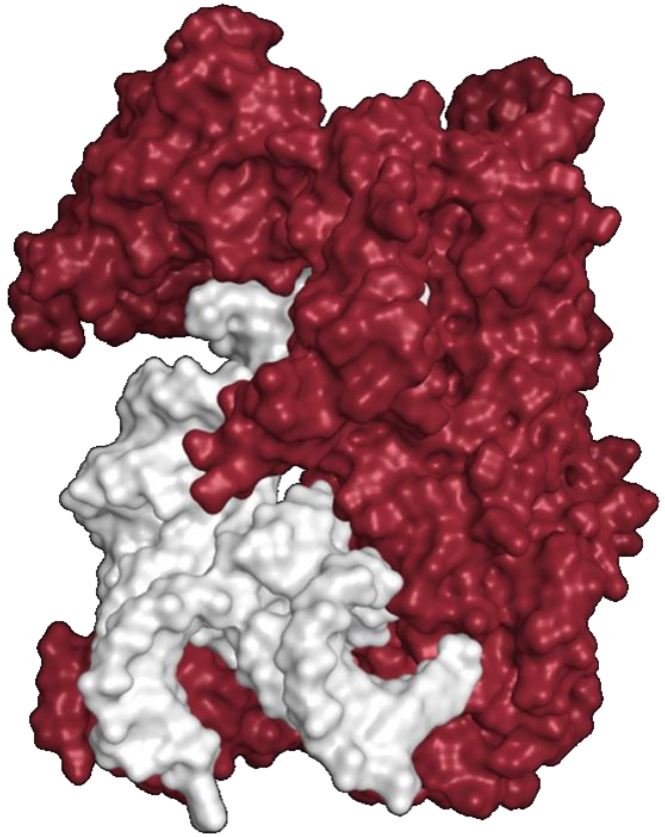
RICHARD COOPER, & ARAS ASAAD

OXFORD DRUG DESIGN LTD



OXFORD DRUG DESIGN: COMPUTATIONAL DRUG DISCOVERY TOOLS WITH A FOCUS ON T-RNA SYNTHETASE INHIBITORS





COMPUTATIONAL METHODS TEAM WITH BACKGROUND IN BIOLOGY, CHEMISTRY, BIG PHARMA AND MATHEMATICS

Dr Anthony Chubb

Software Engineer

Dr Paul Finn

Chief Scientific Officer

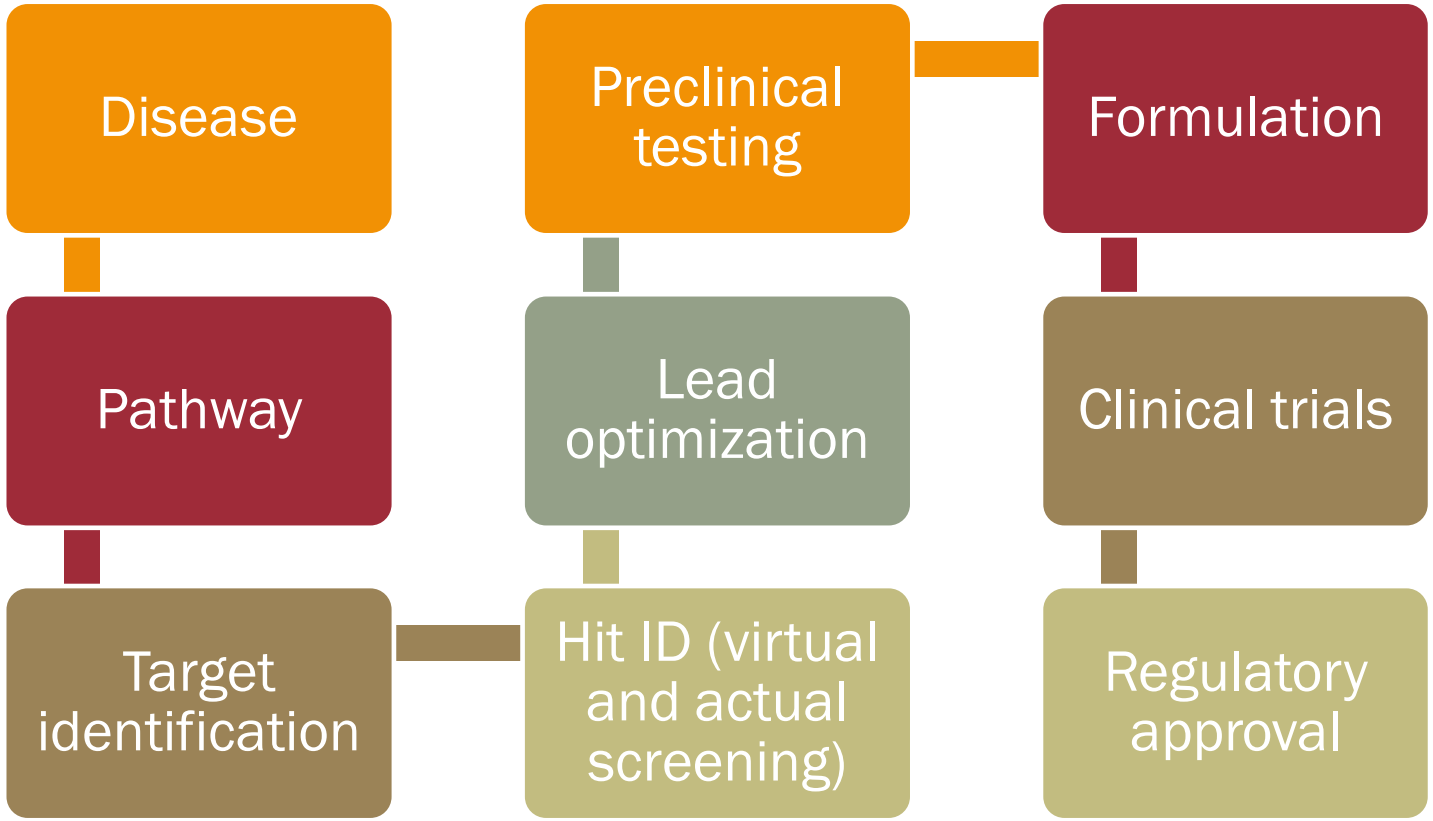
Dr Richard Cooper

Head of ML and Methods Development

Dr Aras Asaad

Machine Learning Scientist

DRUG DISCOVERY IS A HARD PROBLEM (BUT AN IMPORTANT ONE)



THE IMMENSE POTENTIAL OF AI LEADS TO A RISK OF OVER-PROMISING RESULTS TO CUSTOMERS AND INVESTORS

Recent builder.ai insolvency: reported problems include using hundreds of real programmers pretending to be AI models.

Builder.ai

Spies, spinners, solicitors: Builder.ai's 'perfectly normal' creditor list in full

Due Diligence

Spies, spin and an AI hype story

Premium

Builder.ai

Microsoft-backed Builder.ai collapsed after finding potentially bogus sales

Builder.ai

Microsoft-backed UK tech unicorn Builder.ai collapses into insolvency

Lack of expertise in either statistical methods or drug discovery can lead to mistakes.

We can only make progress by bridging this gap, and avoiding hype.

1. Assessing new methods
2. Awareness of dataset biases
3. Do we have sufficient information?

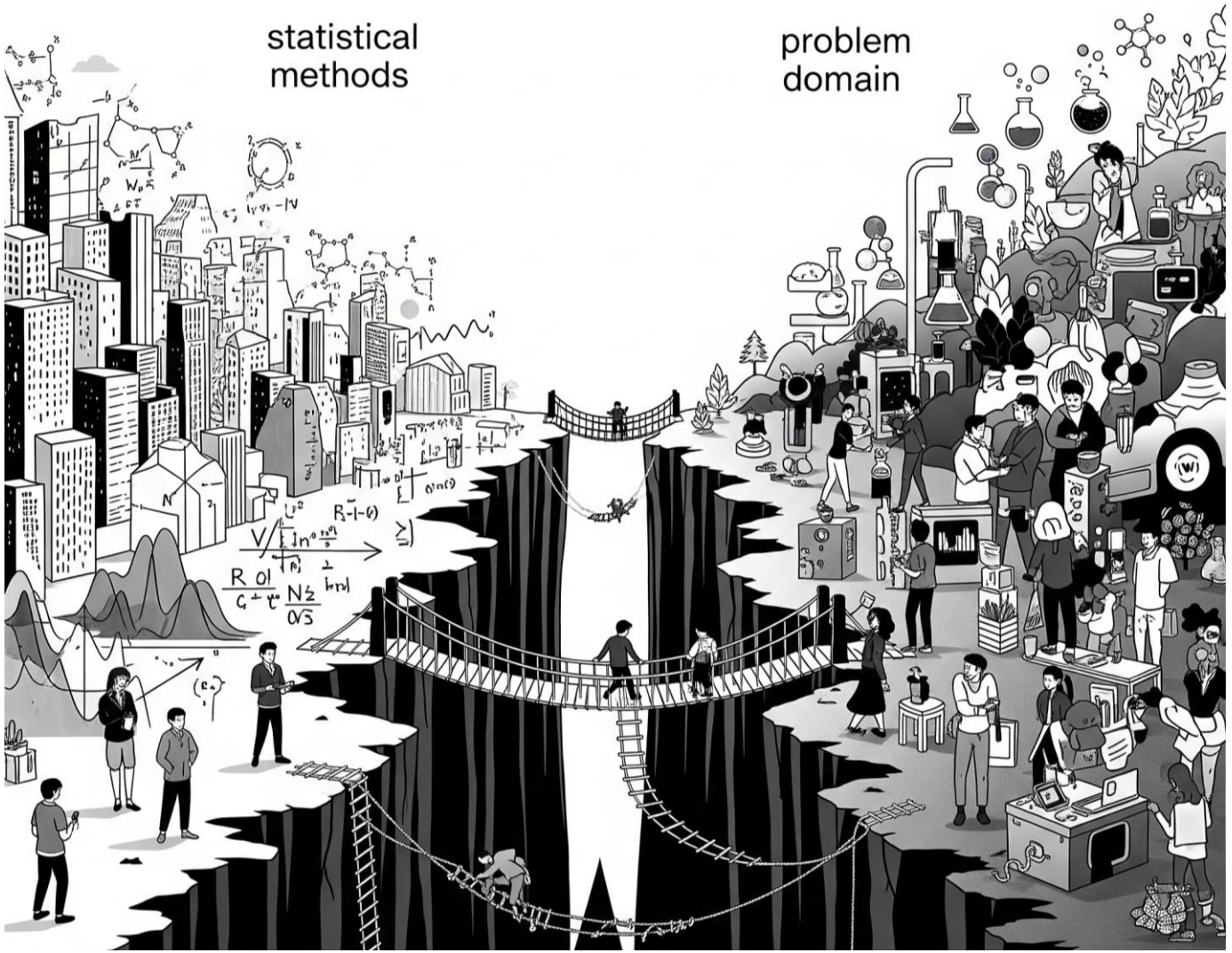
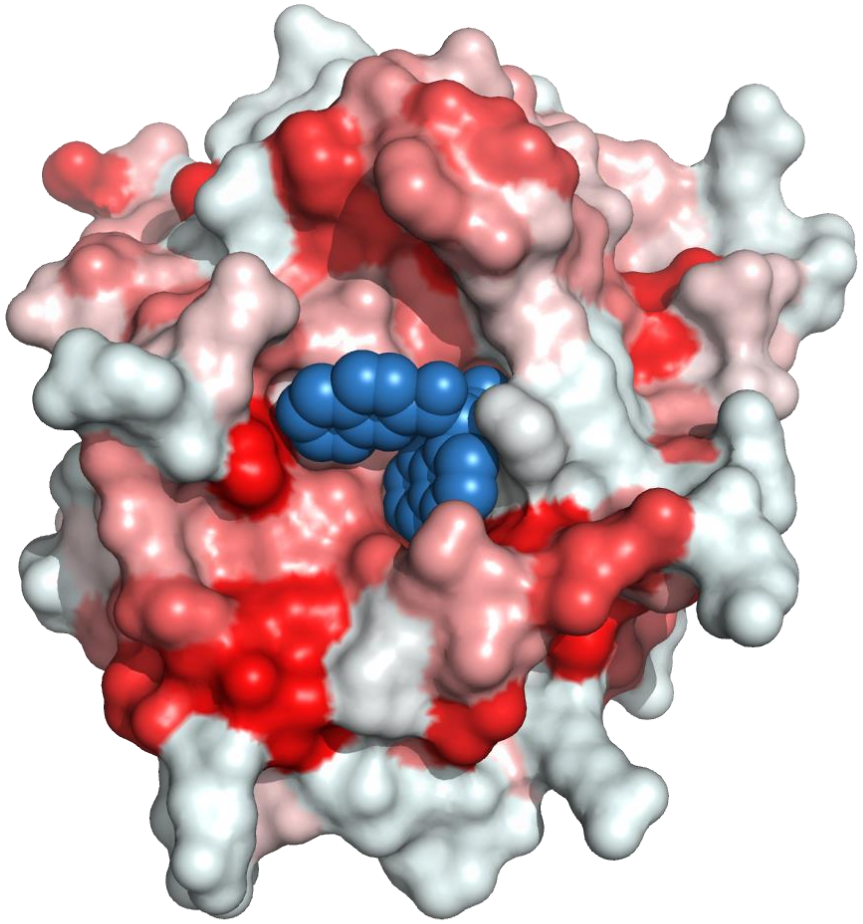


Image source: Google Gemini

DATA REPRESENTATION

- Preclinical models tend to use representations of the structure or connectivity of atoms in a drug molecule. For ‘structure based’ methods, the structure or sequence of amino acids in a protein is also required.
- We usually need to **end up with a vector representation** of our physical system, *e.g.*
  - one-hot encoded atom environments: “fingerprints”
  - amino acid sequence or interaction counts or fingerprints
  - description of shape and property distributions: inertial tensors, spherical harmonics, persistent homology approaches
  - embeddings of graphs, latent space of autoencoders, principal components



Crystal structure of 1KYN protein (red/white) and small molecule inhibitor (blue)

1. ASSESSING THE EFFECTIVENESS OF NEW METHODS

1.1 PoseBusters: AI-based docking fails to generate valid poses or generalise to novel sequences.

1.2 Are the results too good to be true? MUV

1.3 Are the results too good to be true? Co-crystallization

1.1 POSEBUSTERS: SOME DL MOLECULAR DOCKING METHODS FROM THE LITERATURE

Buttenschoen, Morris and Deane ***Chem. Sci.***, 2024,**15**, 3130-3139

Method	Description
DeepDock	Learns a statistical potential based on the distance likelihood between ligand atoms and points of the mesh of the surface of the binding pocket (protein).
DiffDock	Equivariant graph neural networks in a diffusion process for blind docking.
EquiBind	Equivariant graph neural networks for blind docking.
TankBind	Blind docking method using a trigonometry-aware neural network for docking in each pocket predicted by a binding pocket prediction method.
Uni-Mol	Docking with SE3-equivariant transformers

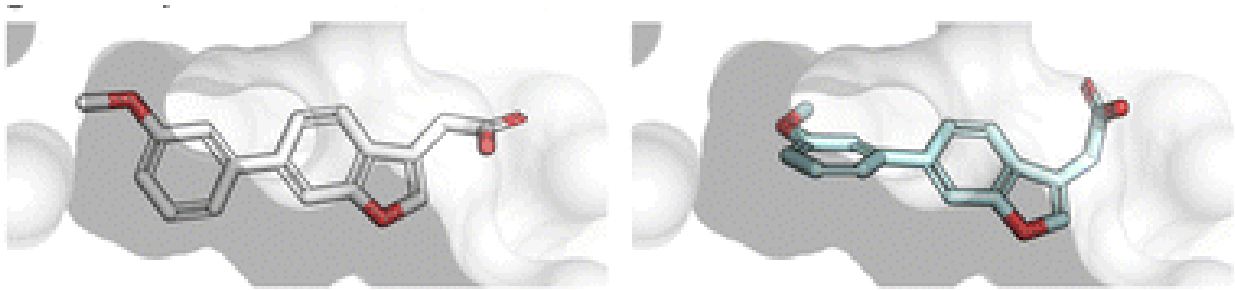
“state-of-the-art performance in terms of RMSD of atom positions in a docked molecule to experimentally measured positions in a crystal structure”

1.1 POSEBUSTERS: THE PROBLEM

"However, despite claims of state-of-the-art performance … it has become apparent that they often produce physically implausible molecular structures."

Martin Buttenschoen, Garrett M. Morris and Charlotte M. Deane
Chem. Sci., 2024,**15**, 3130-3139

(f) Double bond not flat. TankBind prediction for ligand DBQ of protein-ligand complex 1U4D. RMSD 1.7 Å.



(h) Clash with protein. DiffDock prediction for ligand XQ1 of protein-ligand complex 7L7C. RMSD 1.6 Å.

1.1 POSEBUSTERS: THE RESULTS

“We show that both in terms of physical plausibility and the ability to generalise to examples that are distinct from the training data, ***no deep learning-based method yet outperforms classical docking tools***. In addition, we find that molecular mechanics force fields contain docking-relevant physics missing from deep-learning methods.”

“It is vital, particularly for deep learning-based methods, that they are also evaluated on steric and energetic criteria.”

1.2 DEEP LEARNING THE MUV DATASET: PUBLISHED RESULTS

A deep learning method with convolutional layers, biformer network and novel training strategy was published in a journal in the *Nature* portfolio recently. This caught Alex's attention because its performance on a well known 'medium difficulty' drug molecule activity classification dataset (MUV) has a ROC AUC of > 0.99. This places it "**among the top-performing models**" in the field. (Next best is Trimnet with ROC AUC of 0.851).

Alex Tanaka
Doctoral Student and dataset sleuth

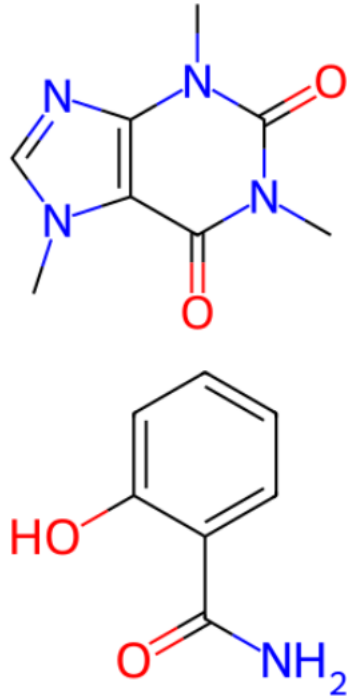
1.2 DEEP LEARNING THE MUV DATASET: THE PROBLEM

The code and data is available. Alex determined that the metric used to compute ROC AUC was “not standard” for a binary classification problem, enabling apparently amazing performance from a model with little to no predictive skill at all.

Investigation is ongoing. The correct ROC AUC for several targets tested so far is not so impressive.

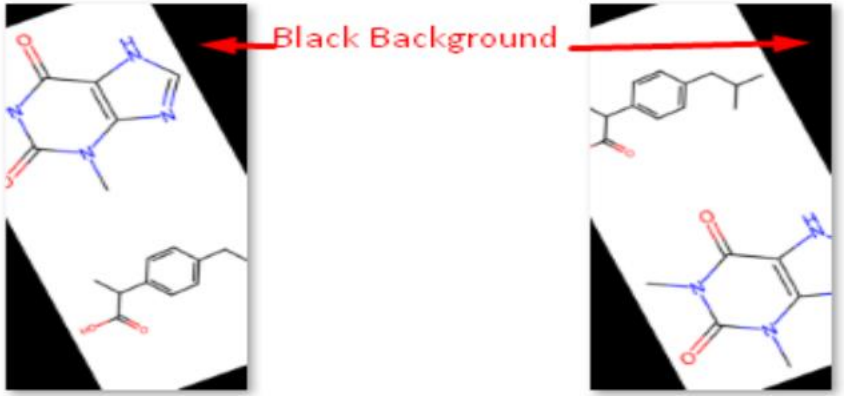
1.3 DRUG FORMULATION VIA CO-CRYSTALLIZATION

- Published CNN model
- Model trained to classify likelihood of pairs of molecules forming a co-crystal, in which both molecules are incorporated, using experimental results
- Novel data representation: Data presented as images containing 2D chemical diagrams of both potential component molecules
- ROC AUC :0.925



1.3 DRUG FORMULATION VIA CO-CRYSTALLIZATION – SOME PROBLEMS

Data augmentation was used to correct a class imbalance. Rotated copies of images were used to augment the positive class. Image (right) is typical. Black background is easy to detect.



Data processing script did 10-fold cross validation (good) but did not clear out working folder in-between folds (bad). First fold always worse than the rest (due to leakage of test set information).

1. CAREFUL ASSESSMENT OF RESULTS IS CRITICAL ESPECIALLY RELATIVE TO EXISTING BASELINES

Posebusters paper shows that optimising and assessing on RMSD alone can lead to unphysical results. Problem domain knowledge required. ‘Correct’ results no better than classical methods.

The deep learning results for MUV show the importance of reviewing code, and also of demanding a high standard of proof for surprisingly good results. Methods knowledge required.

The co-crystallization analysis revealed data leakage during model training (intermediate files left behind) and unintentional biasing of the positive class due to the data augmentation procedure. Methods knowledge required.

2. WHAT IS THIS DATASET FOR? AWARENESS OF BIAS

2.1 The Directory of Useful Decoys (extended) (DUD-E): A frequently misused dataset for docking

2.2 CASF2016: Training bias: Have you seen this protein before?

2.1 THE DIRECTORY OF USEFUL DECOYS – ENHANCED (DUD-E)

- DUD-E is designed to help benchmark ***molecular docking programs*** by providing challenging decoys.
- **22,886 active molecules** and their affinities against **102 targets**, an average of 224 ligands per target
- **50 decoy molecules** for each active, having similar physico-chemical properties but dissimilar 2-D topology

Latest version is DUDE-Z (DUD-E for Gen Z)

2.1 DUD-E IS A TEMPTING CHOICE FOR ML/AI CLASSIFICATION TASKS BECAUSE THERE IS SO MUCH DATA

And – it gives results that appear amazing:

"Using a clustered cross-validation on DUD-E, we achieve an average AUC ROC of 0.92 and a 0.5% ROC enrichment factor of 79."

2.1 BUT … DUD-E USED MOLECULAR FINGERPRINTS TO SELECT ITS "DECOY" MOLECULES, WHICH ENABLES EASY CLASSIFICATION USING MOLECULAR FINGERPRINT METHODS

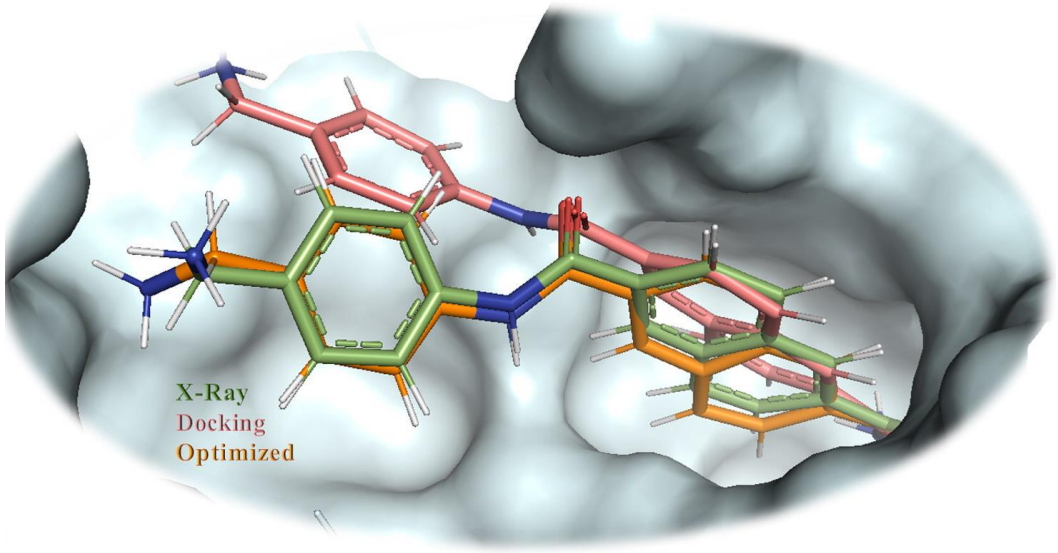
- The DUD-E and DUDE-Z (DUD-E for Gen Z) websites say: "designed for 3D molecular screening methods such as docking and ***not for 2D methods such as molecular similarity based methods***"
- See "Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening" Lieyang Chen *et al.* PLOS, 2019 https://doi.org/10.1371/journal.pone.0220113. A good number of examples of DUD-E misuse – primarily CNNs, but principle applies to any models that capture atomic environment in any way.

2.2 CASF* 2016 TASK: PREDICT THE AFFINITY OF A SMALL MOLECULE LIGAND BINDING TO A PROTEIN STRUCTURE

* comparative assessment of scoring functions

Binding affinity can be estimated from the 3D coordinates of a ligand bound to a protein using a scoring function. This estimate can guide lead discovery and optimization tasks. Scoring functions are usually simple functions of physical interactions but may still be competitive with more rigorous physics-based methods.

CASF-2016 defines 285 curated protein-ligand complexes as the 'test set'



Improving protein-ligand docking results using the Semiempirical quantum mechanics: testing on the PDBbind 2016 core set Zainab Mohebbinia *et al.* https://doi.org/10.1080/07391102.2023.2299742

2.2 CASF 2016 PROBLEMS: MANY METHODS USE THE REST OF THE PDBBIND REFINED SET AS TRAINING DATA

After some filtering, and removal of the test set complexes, **4638** complexes remain in the PDB 'refined' set for use as training data, but…

- 102 complexes (2.1%) have identical small molecule inhibitors
- 994 complexes (21%) have identical protein sequence to test set proteins
- 1498 complexes (32%) have > 90% sequence similarity

Knowledge of the approximate binding affinity in one third of cases gives a huge boost in model performance.

2.2 OXFORD DRUG DESIGN DE-DUPLICATED CASF TRAINING DATA IS MUCH MORE CHALLENGING*

- Complexes with protein sequence identity > 90% to any sequence in test set removed
- Complexes containing ligands identical to ligands in test set were also removed.
- 2500 training complexes remain in de-biased training set

Affinity prediction using OnionNet implementation	RMSE	Pearson's R
Refined set (test set removed)	1.28	0.82
Our de-biased data set	1.68	0.63

Dr Marco Albanese

Computational Chemist

* "More challenging" = more opportunity for genuinely measuring progress

2. DATASET BIASES

DUD-E is not designed to be used for machine learning / deep learning methods. Domain knowledge required.

CASF2016 the test set has high similarity to much of the default training set. Careful filtering is required to de-bias.

De-biasing to produce more challenging datasets produces worse headline results. This is unfortunate in a short attention-span world, however no progress can be made without testing models against even modestly difficult datasets.

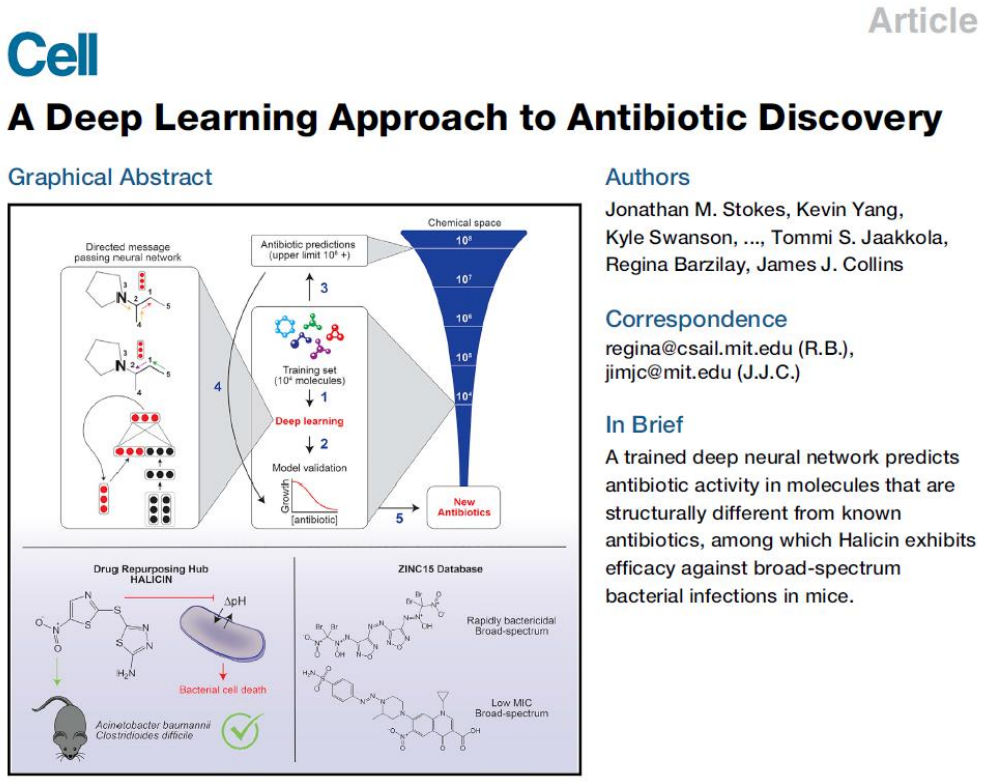
3. BE SCEPTICAL OF THE HYPE. AND OF LEARNING THINGS THAT CAN'T BE LEARNED.

3.1 AI discovers novel antibiotic

3.2 What is GenAI doing for discovery?

3.3 How to help the model: Physical constraints enforce reality

3.1 AI DISCOVERS NOVEL ANTIBIOTIC?



Article

Cell

A Deep Learning Approach to Antibiotic Discovery

Graphical Abstract

Authors

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, ..., Tommi S. Jaakkola, Regina Barzilay, James J. Collins

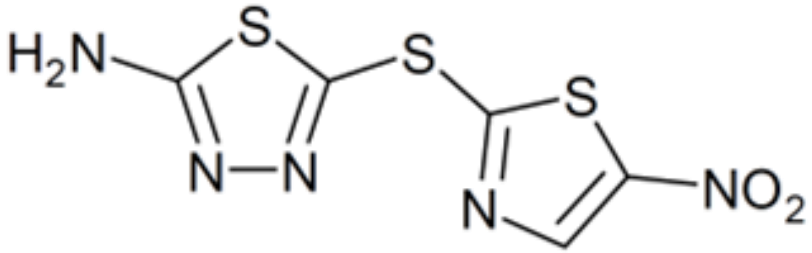
Correspondence

regina@csail.mit.edu (R.B.), jimjc@mit.edu (J.J.C.)

In Brief

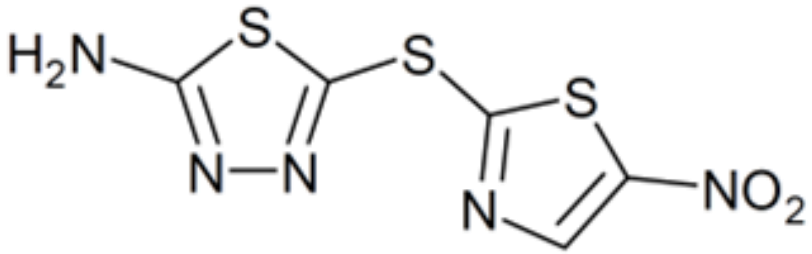
A trained deep neural network predicts antibiotic activity in molecules that are structurally different from known antibiotics, among which Halicin exhibits efficacy against broad-spectrum bacterial infections in mice.

A deep learning approach to antibiotic discovery, Stokes at el., Cell 2020, **180**, 688



1. Message passing NN
2. Trained on antibacterial activity of FDA approved drugs and natural product library
3. Predicted activity of compounds in the Drug Repurposing Hub database
4. 51 actives identified, one compound, **Halicin**, selected for detailed study

3.1 AI DISCOVERS NOVEL ANTIBIOTIC? OR AI GETS LUCKY?



Halicin

- Novel mode of action
- No related chemotypes (similar chemical structure)

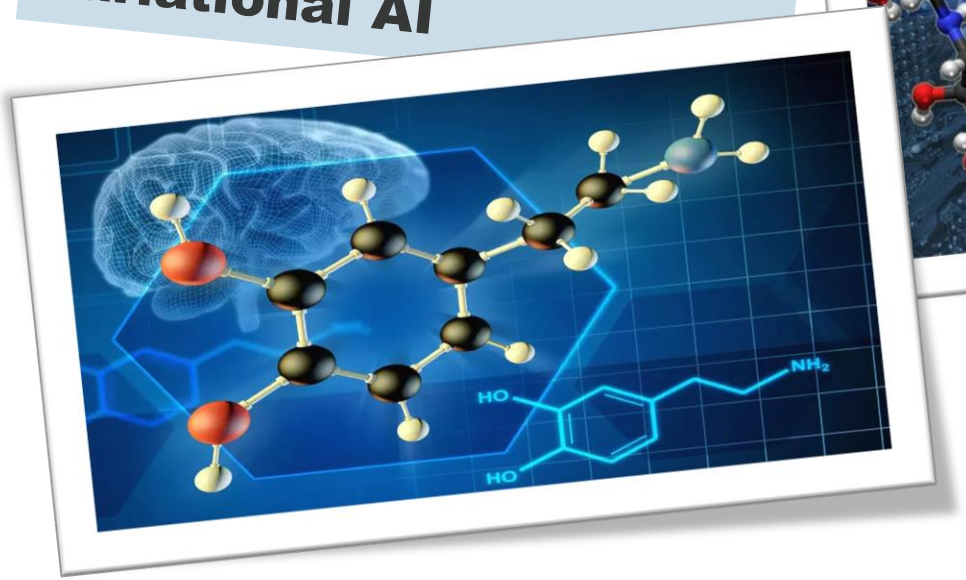
- Sulfur is present in 31% of actives and 14% of inactives in training data
- If β-lactams are removed from training, sulfur is only present in 10% actives and 13% inactives and **Halicin is no longer predicted to be active**

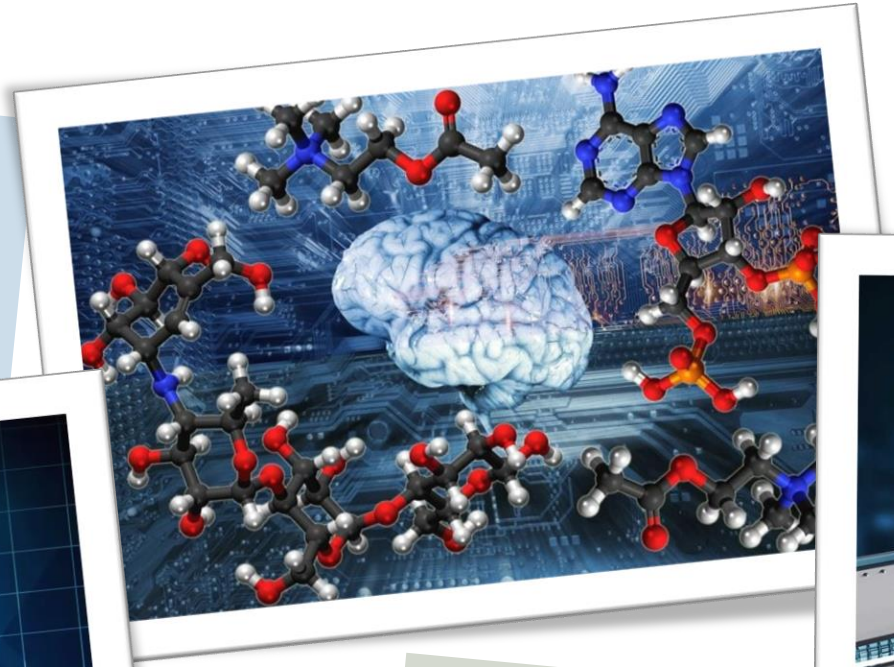
- *Although Halicin is active, its classification as such seems to be artefactual.* **There is no information in the dataset to support the activity of the proposed structure.**

See Jagdev, Madsen & Finn, *J Mol Model.* 2022, 29, 22

3.2 THERE IS A TENDENCY TO FRAME RESULTS AS “DISRUPTIVE” OR “REVOLUTIONARY”

100 AI-generated molecules are worth a 1,000,000 molecule high-throughput screen – Variational AI



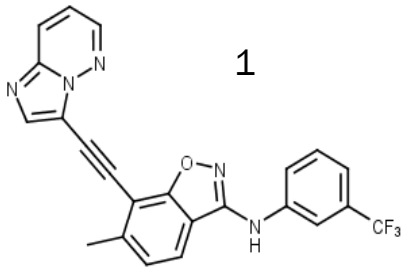


Deep learning enables rapid identification of potent DDR1 kinase inhibitors - InSilico

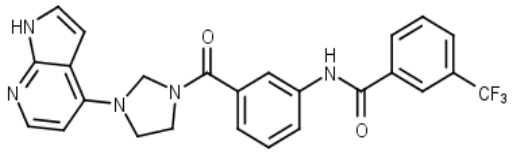
3.2 GENERATIVE AI CREATES NOVEL MOLECULES SAMPLED FROM THE CHEMICAL SPACE OF ITS TRAINING DATA

E.g. GENTRL: Deep generative model, generative tensorial reinforcement learning. "GENTRL optimizes **synthetic feasibility**, **novelty**, and **biological activity**".

- Four compounds were active in biochemical assays
- Two were validated in cell-based assays



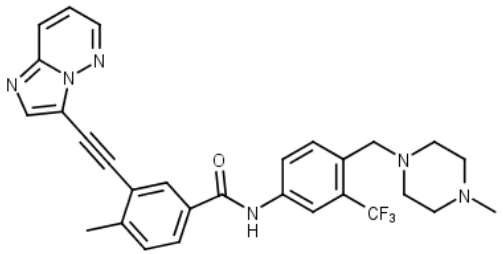
1



2

IC_{50} (nM)		
Compound	DDR1	DDR2
1	10	234
2	21	76

Workflow item	Compounds
GENTRL	30,000
Property filters	12,147
Med Chem filters	7,912
Tanimoto diversity filter	5,542
Commercial analogue filter	4,642
Kinase SOM filter	1,951
Pharmacophore VS	848
Sammon mapping	40
Synthetic accessibility	6



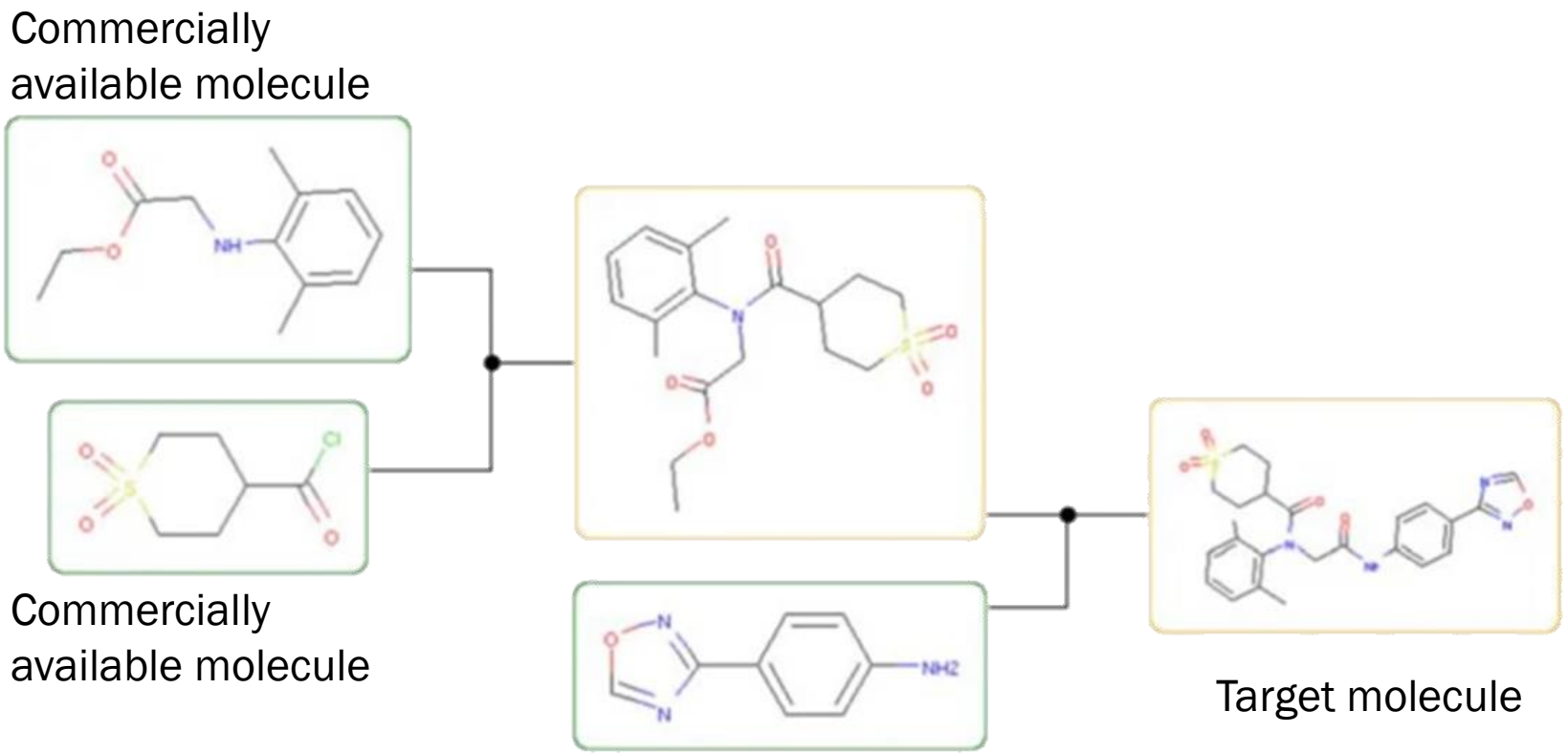
ponatinib

[Aside: The similarity of Compound 1 with ponatinib has been noted…]

3.2 BUT SYNTHETIC ACCESSIBILITY IS NOT A PRIORITY FOR GEN AI OUTPUTS

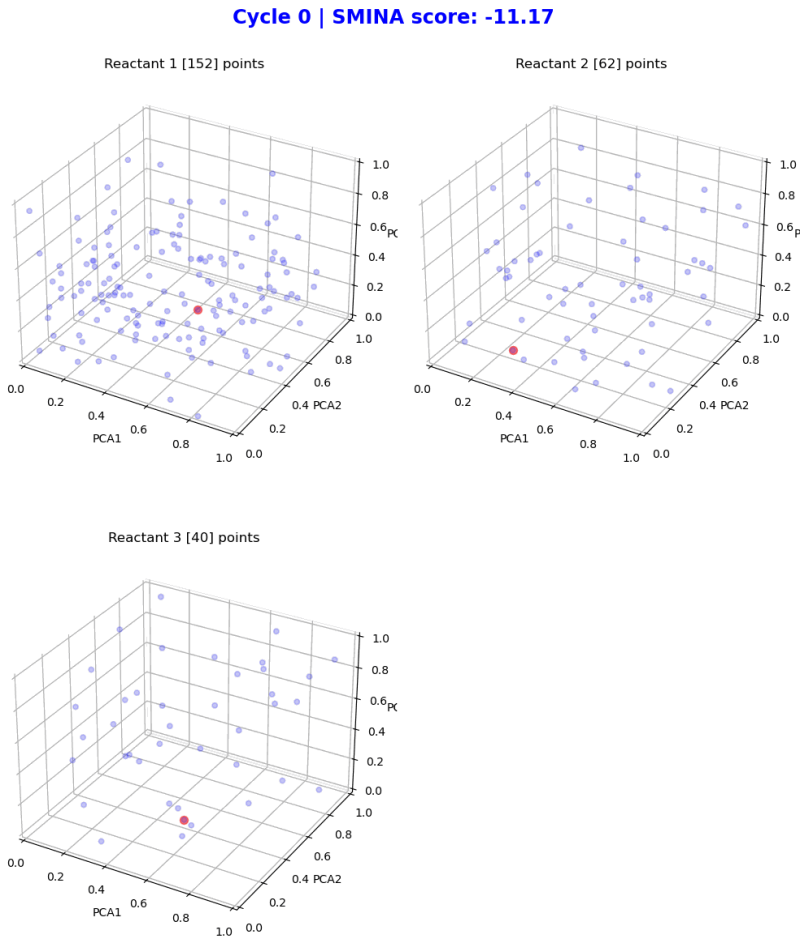
Generative AI publication	Molecules generated	Molecules synthesized
"Deep learning enables rapid identification of potent DDR1 kinase inhibitors" Zhavoronkov et. al., Nature Biotechnology 2019, **37**, 1038-1040	30,000	6
"Discovery of Pyrazolo[3,4-*d*]pyridazinone derivatives as Selective DDR1 Inhibitors *via* Deep Learning Based Design, Synthesis, and Biological Evaluation" Tan et al., J Med Chem 2022, **65**, 1, 103-119	19,929	2
"Design and Synthesis of DDR1 Inhibitors with a Desired Pharmacophore Using Deep Generative Models" Yoshimori et al., ChemMedChem 2020, **16**, 6, 955-958	570,542	9
"PCW-A1001, AI-assisted *de novo* design approach to design a selective inhibitor for FLT-3(D835Y) in acute myeloid leukemia" Jang et al., Front Mol Biosci 2022, **9**	10,416	1
"AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor" Ren et al., Chem Sci. 2023, **14**, 1443-1452	8,918	7
"Accelerating drug target inhibitor discovery with a deep generative foundation model" Chenthamarakshan et al., Sci Advances 2023, **9**, 25	875,000	4

3.3 USE PHYSICAL CONSTRAINTS TO ENFORCE REALITY (OXFORD DRUG DESIGN INTERNAL PROJECT)



1. Generate multiple **synthesis routes** to a **target molecule**
2. Get lists of available reactants that can be substituted at each node
3. Search the reactant space to optimize a property of the product (Bayesian Optimization)

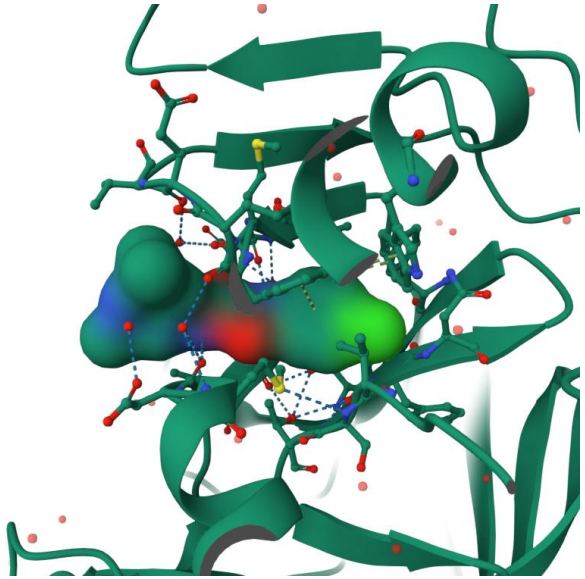
3.3 SEARCH THE REACTANT PROPERTY SPACE

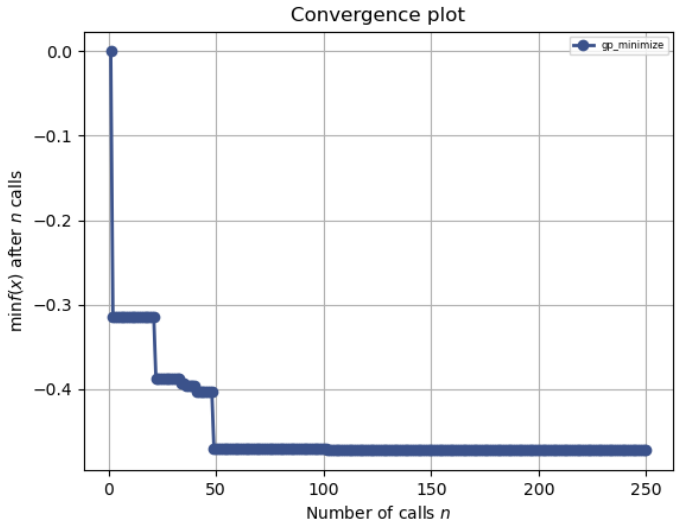


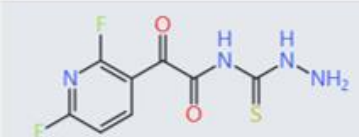
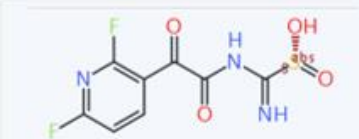
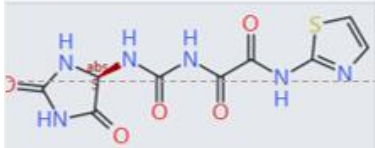
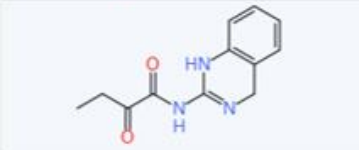
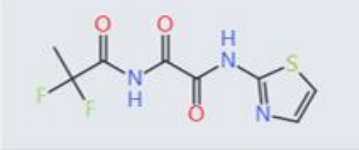
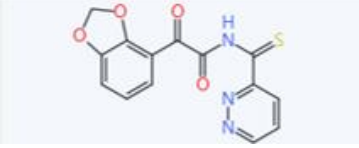
- Basis vectors of **reactant chemical space** derived from in-house **shape descriptor**
- Reactant **molecules** with similar **shape and charge** distributions are **close together**
- **Bayesian Optimization *exploits* similarity and *explores*** to find **new** active **clusters**
- GP model retrained at each step
- ALL PRODUCTS CAN BE SYNTHESIZED

3.3 CONSTRAINED BAYESIAN SEARCH CONVERGES RAPIDLY IN LARGE COMBINATORIAL CHEMICAL SPACE

pdb 3TGS: **HIV-1 clade C strain C1086 gp120 core in complex with NBD-556**





Structure	LE	AFFINITY
	-0.3889113	-6.61149168
	-0.3884022	-6.99124002
	-0.3878654	-8.14517307
	-0.3844684	-6.53596306
	-0.3817651	-6.49000597
	-0.3692331	-8.12312794

Rapid docking using Lin_F9 empirical scoring function. Secondary objectives / pareto front calculation to be determined.

3. HYPE AND LEARNING THE UNLEARNABLE

The Halicin example shows that it is still possible to stumble upon good molecules without robust predictive models. An “anecdote vs data” problem.

Generative methods are impressive, but may have a synthesizability problem. How to compare 6 molecules filtered from GenAI output to 6 molecules filtered from a library of millions of commercially available molecules?

One solution is to design the search process around the physical task – synthesizing molecules – to guarantee that molcules can be made.

SUMMARY

- Don’t believe the hype (without verifying it for yourself)
- Suggest domain specialists / AI specialists to balance peer review
- Insist on reporting results against more challenging datasets
- Be suspicious of missing code or data
- Check for unexpected biases
- Always compare to baseline models
- Use explainable models (and take time to look at them)

OPPORTUNITIES FOR AI IN PRECLINICAL DRUG DISCOVERY AND DEVELOPMENT

To make progress we need:

- Inter-disciplinary collaboration
- Transparent approaches to dataset preparation and splitting
- Challenging (and problem appropriate) datasets
- Rigorous model intercomparison projects
- Cross-validation
- Explainable models
- Improved representations of molecular flexibility and intermolecular interactions

CHAT OR COLLABORATE WITH OXFORD DRUG DESIGN

We are always keen to discuss and collaborate on methods development and projects in preclinical drug discovery. We have in-house resources to help:

- de-biased public datasets for virtual screening and binding affinity prediction
- (confidential) data from real in-house drug development programmes
- computational tools and molecule catalogues

Contacts: richard.cooper@oxforddrugdesign.com
and aras.asaad@oxforddrugdesign.com