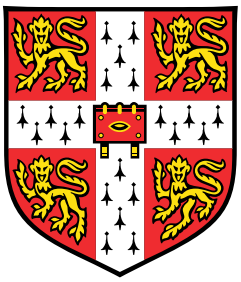
# Accelerating the Design-Make-Test cycle of Drug Discovery with Machine Learning



**William McCorkindale**

Cavendish Laboratory, Department of Physics
University of Cambridge

Supervisor: Dr. Alpha Lee

St. John's College

January 2023

# Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

William McCorkindale
January 2023

# Acknowledgements

TODO - finish acknowledgements And I would like to acknowledge ...

- Alpha
- Collaborators
- DeepMind?
- Cambridge friends
- Family

# Abstract

Drug discovery follows a design-make-test cycle of proposing drug compounds, synthesising them, and measuring their bioactivity, which informs the next cycle of compound designs. The challenges associated with each step leads to the long timeline of preclinical pharmaceutical development. This thesis focuses on how we can use machine learning tools to accelerate the design-make-test cycle for faster drug discovery.

We begin with the design of new compounds, looking at the initial stage of fragment-based hit finding where only the 3D coordinates of fragment-protein complexes are available. The standard approach is to "grow" or "merge" nearby fragments based on their binding modes, but fragments typically have low affinity so the road to potency is often long and fraught with false starts. Instead, we can reframe fragment-based hit discovery as a denoising problem – identifying significant pharmacophore distributions from an "ensemble" of fragments amid noise due to weak binders – and employ an unsupervised machine learning method to tackle this problem. We construct a model that screens potential molecules by evaluating whether they recapitulate those fragment-derived pharmacophore distributions. We show that this approach outperforms docking on distinguishing active compounds from inactive ones on historical data. Further, we prospectively find novel hits for SARS-CoV-2 Mpro and the Mac1 domain of SARS-CoV-2 non-structural protein 3 by screening a library of 1B molecules.

After identifying hit compounds, we enter the the hit-to-lead stage where we wish to optimise their molecular structures to improve bioactivity. Framing bioactivity modelling as classification of active/inactive would not allow us to rank compounds based on predicted bioactivity improvement, while the low number of active compounds and the measurement noise make a regression approach challenging. We overcome this challenge with a learning-to-rank framework via a classifier that predicts whether a compound is more or less active than another using the difference in molecular descriptors between the molecules as input. This allows us to make use of inactive data, and threshold the bioactivity differences above measurement noise. Validation on retrospective data for Mpro shows that we can outperform docking on ranking ligands, and we prospectively screen a library of 8.8M molecules and arrive at a potent compound with a novel scaffold.

After designing a drug candidate one needs to find a synthesis route to actually make the molecule in the real world. An exciting approach is to use deep learning models trained on patent reaction databases, but they suffer from being opaque black-boxes. It is neither clear if the models are making correct predictions because they inferred the salient chemistry, nor is it clear which training data they are relying on to reach a prediction. To address this issue, we developed a workflow for quantitatively interpreting a state-of-the-art deep learning model for reaction prediction. By analysing chemically selective reactions, we show examples of correct reasoning by the model, explain counterintuitive predictions, and identify Clever Hans predictions where the correct answer is reached for the wrong reason due to dataset bias.

Testing a drug candidate typically involves obtaining a pure sample of the molecule, and then measuring its bioactivity in solution via an assay. While necessary for maximum accuracy, compound purification can be time-consuming and costly. We investigated whether we needed compound purification at all for training machine learning bioactivity models by assaying crude reaction mixtures instead of pure samples. This approach allowed us to obtain bioactivity data in higher throughput and train useful models for identification of false negative assay measurements, as well as prospective screens.

The research presented in this thesis highlights the promise of applying machine learning in accelerating the design-make-test cycle of drug discovery. This thesis concludes by outlining promising research directions for applying machine learning within drug discovery.

# Table of contents

# Preface

Chapter 1 talks about ... Chapter 2 talks about ...

In Chapter 3 we discuss ... This work resulted in the publication of the following article:

> William McCorkindale, Ivan Ahel, Haim Barr, Galen J. Correy, James S. Fraser, Nir London, Marion Schuller, Khriesto Shurrush, Alpha A. Lee. Fragment-Based Hit Discovery via Unsupervised Learning of Fragment-Protein Complexes.

Specify who did what

Chapter 4

> Aaron Morris, William McCorkindale, The COVID Moonshot Consortium, Nir Drayman, John D. Chodera, Savaş Tay, Nir London and Alpha A. Lee. Discovery of SARS-CoV-2 main protease inhibitors using a synthesis-directed de novo design model, *Chem. Commun.*, 2021,57, 5909-5912

Chapter 5

> Dávid Péter Kovács, William McCorkindale and Alpha A. Lee. Quantitative interpretation explains machine learning models for chemical reaction prediction and uncovers bias. *Nature Communications* volume 12, Article number: 1695 (2021)

D.P.K. and W.M. implemented the algorithms, D.P.K. trained the models, designed the experiments and analysed the model attributions for the various reaction classes. W.M. implemented Tanimoto splitting and applied reaction templates for counting statistics and artificial dataset generation. A.A.L. supervised and directed the project. All authors discussed the results and approved the manuscript. D.P.K. and W.M. contributed equally to this study.

Chapter 6

The final chapter ...

# Chapter 1

# Introduction

The discovery of new pharmaceuticals traditionally follows the design-make-test paradigm, where molecules are repeatedly proposed, synthesized, and assayed. Drug candidates are designed based on some hypothesis relating chemical structure to drug activity, which gets updated in light of new activity results. This cycle repeats as the molecular search space narrows down until a candidate molecule satisfies the necessary activity/selectivity/toxicity criteria.

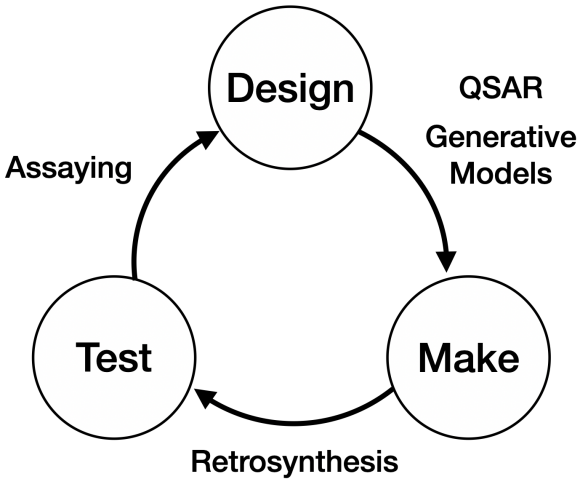



Fig. 1.1 An overview of the design-make-test cycle in drug discovery.

While computational methods have long been used in various stages of the cycle, there has been a recent surge in applying artificial intelligence to drug discovery following its success in various other fields, most notably computer vision and natural language processing. Since molecular assaying is largely an automated process the application focus has been on 'design' and 'make' [14], for example in modelling quantitative structure-activity relationships (QSAR),

designing generative models for proposing drug candidates, and planning retrosynthesis routes (Fig 1.1).

As the field of data-driven drug discovery matures beyond merely adapting the latest state-of-the-art machine learning (ML) methods, the present challenge is to tailor ML models specifically for the unique problems and situations faced in pharmaceutical chemistry. This report summarizes my efforts over the past year to play a part in this challenge with intuitions based on physical science. These consist of three separate tasks, one on 'Make' and two on 'Design':

- **Interpreting learnt chemical principles from Molecular Transformer:** a state-of-the-art reaction prediction model (Molecular Transformer) was investigated with input and data attribution methods to discern whether the model had learnt chemically reasonable patterns of reactivity, or had simply succumbed to hidden bias in the datasets.
- **Exploiting molecular shape for property prediction:** a descriptor of atomic positions known as SOAP, which has seen widespread use in condensed matter physics due to its symmetry-invariance properties, was utilized in a Gaussian Processes model and shown to be competitive with other state-of-the-art models on predicting bioactivity. It was also demonstrated that ensembling models with diverse representations led to further predictive power.
- **Designing Sars-CoV-2 MPro inhibitors:** An initiative known as COVID Moonshot [PostEra Inc.] was established to search for inhibitors of the Sars-CoV-2 main protease (MPro), crowd-sourcing drug candidate designs from the scientific community. In the early stages of the project, I utilised a genetic algorithm with SOAP descriptors for combining disparate fragment hits; in the most recent stage, I implemented a graph siamese network to learn how to rank the activity of assayed molecules, which was then used to suggest new candidates via computational screening of a constructed library.

The lessons learnt from these projects are used to inform possible avenues of future research, which are discussed in the final chapter of this report.

# Chapter 2

# Background

Detailed description of many things

- SMILES
- scaffolds
- fingerprints
- pharmacophores
- measuring affinities
- machine learning
  - random forest
  - gaussian process?
  - deep learning
  - transformer
  - train-test-splitting

# Chapter 3

# Fragment-Based Hit Discovery via Unsupervised Learning of Fragment-Protein Complexes

The process of finding molecules that bind to a target protein is a challenging first step in drug discovery. Crystallographic fragment screening is a strategy based on elucidating binding modes of small polar compounds and then building potency by expanding or merging them. Recent advances in high-throughput crystallography enable screening of large fragment libraries, reading out dense ensembles of fragments spanning the binding site. However, fragments typically have low affinity thus the road to potency is often long and fraught with false starts. Here, we take advantage of high-throughput crystallography to reframe fragment-based hit discovery as a denoising problem – identifying significant pharmacophore distributions from a fragment ensemble amid noise due to weak binders – and employ an unsupervised machine learning method to tackle this problem. Our method screens potential molecules by evaluating whether they recapitulate those fragment-derived pharmacophore distributions. We retrospectively validated our approach on an open science campaign against SARS-CoV-2 main protease (Mpro), showing that our method can distinguish active compounds from inactive ones using only structural data of fragment-protein complexes, without any activity data. Further, we prospectively found novel hits for Mpro and the Mac1 domain of SARS-CoV-2 non-structural protein 3. More broadly, our results demonstrate how unsupervised machine learning helps interpret high throughput crystallography data to rapidly discover of potent chemical modulators of protein function.

## 3.1 Introduction

Hit detection is a key step in the early stages of the drug discovery process following the identification of a biological target of interest [33]. A 'hit' compound acts as the starting point for the drug design process where the chemical structure of the hit is progressively optimised towards a candidate drug. Approaches towards hit detection generally involve screening large libraries of compounds, both experimentally and computationally.

Approaches towards hit detection generally involve the screening of libraries of compounds. For example, in high throughput screening (HTS) often hundreds of thousands of chemical compounds are synthesised and tested, requiring substantial resources as well as complex logistics. While experimental techniques such as DNA-Encoded libraries are being developed to increase the efficiency of large-scale compound screening [25], there has been a growing push towards conducting hit detection computationally instead to decrease the cost and accelerate this step of the drug discovery process [**?** ].

In this approach, known as virtual screening, a computational scoring function is used to estimate the potency of a compound. After computing the scores for all of the compounds in a library, only those ranked highly by the scoring function are chosen for synthesis and experimental validation. Currently the predominant scoring function used to conduct a virtual screen is molecular docking. In molecular docking, the 3D conformation of a ligand and the target are explicitly modelled and a physics-based simulation of the binding process is conducted, with the calculated energy of the bound ligand as the score. Although this approach has yielded success [50, 4], correctly performing molecular docking is non-trivial and the deficiencies of molecular docking for bioactivity prediction are well-documented [48, 52].

One of these methodologies is fragment-based drug design (FBDD). In this approach, very low molecular weight compounds ('fragments' with typically less than 18 nonhydrogen atoms [17]) are screened at high concentrations against the target protein with X-ray crystallography. A fragment screening approach is more likely to deliver hits than screening larger drug-like molecules because low molecular complexity compounds are more likely to possess good complementarity with the target protein [29]. Structures of these fragment-protein complexes can then inspire the design of potent binders, either by expanding a fragment to pick up new intermolecular interactions with active site residues, or merging together different spatially proximal fragments [35, 83]. However, despite showing up in X-ray crystallography, the binding affinity of the fragments themselves is typically low. Therefore, gaining potency by fragment expansion or merging is typically a long journey fraught with false starts.

Recently, advances in X-ray crystallography such as automatic crystal mounting robots, fast detectors, as well as increased accessibility to beamtime are enabling high throughput fragment screens. One can routinely go from screening a small fragment library and detecting a handful

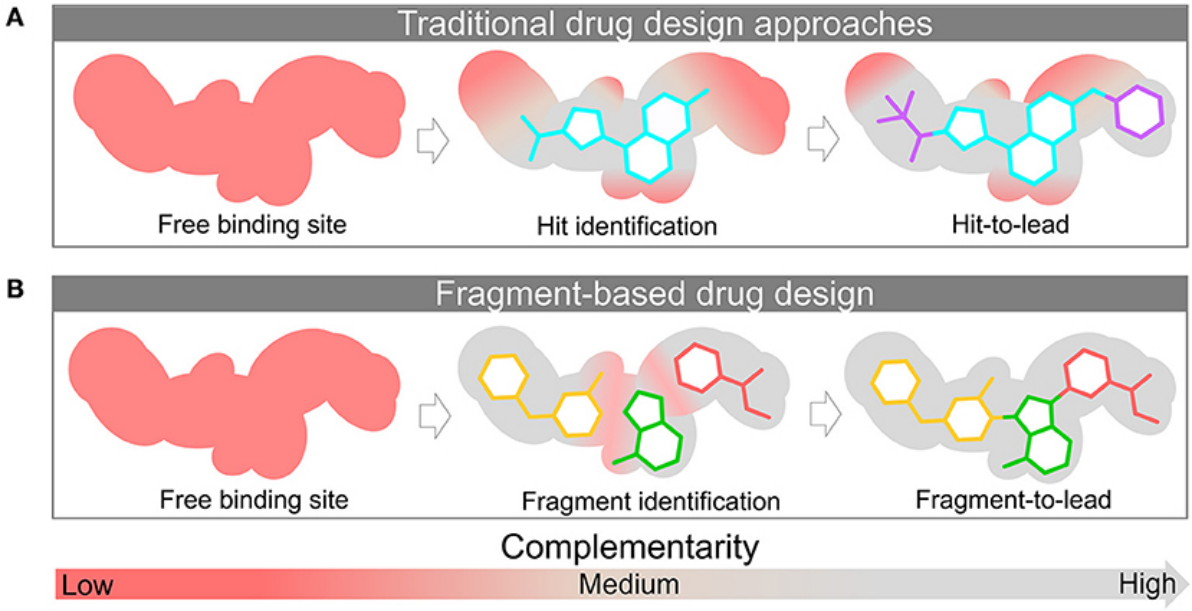



Fig. 3.1 An illustration comparing fragment-based drug discovery to traditional approaches.

of hits, to screening 1000s of fragments with ensembles of 100s of fragments hits spanning the binding site [65, 19]. This substantial increase in data enables a systematic data-driven approach for fragment-based hit discovery.

Although there exist some computational approaches for supporting FBDD, for example hot spot analysis and pocket druggability prediction [18], at present the main procedure of selecting which fragments to merge and how to do so remains largely intuition-based and human-driven. (and fraught to error? citation needed [**?** ])

Our key insight is to reframe fragment-based drug design as signal extraction from noisy data by seeking persistent pharmacophore correlations within a fragment ensemble, rather than looking at individual fragments. This is because a fragment itself has low affinity, thus we need the presence of multiple fragments with the same pharmacophore at a particular region of the binding site to provide statistical confidence.

In this work, we employ unsupervised machine learning to learn the spatial distribution of fragment pharmacophores in the binding site. We then use the trained model as a scoring function for virtual screening, picking out molecules with matching pharmacophores. We will first retrospectively validate our model on a dataset of SARS-CoV-2 main protease (Mpro) ligands from COVID Moonshot [75]. We then present prospective results on identifying hits against Mpro and the Mac1 domain of SARS-CoV-2 non-structural protein 3 (nsp3-Mac1) by performing a virtually screen a library of 1.4 billion purchasable compounds from EnamineREAL.

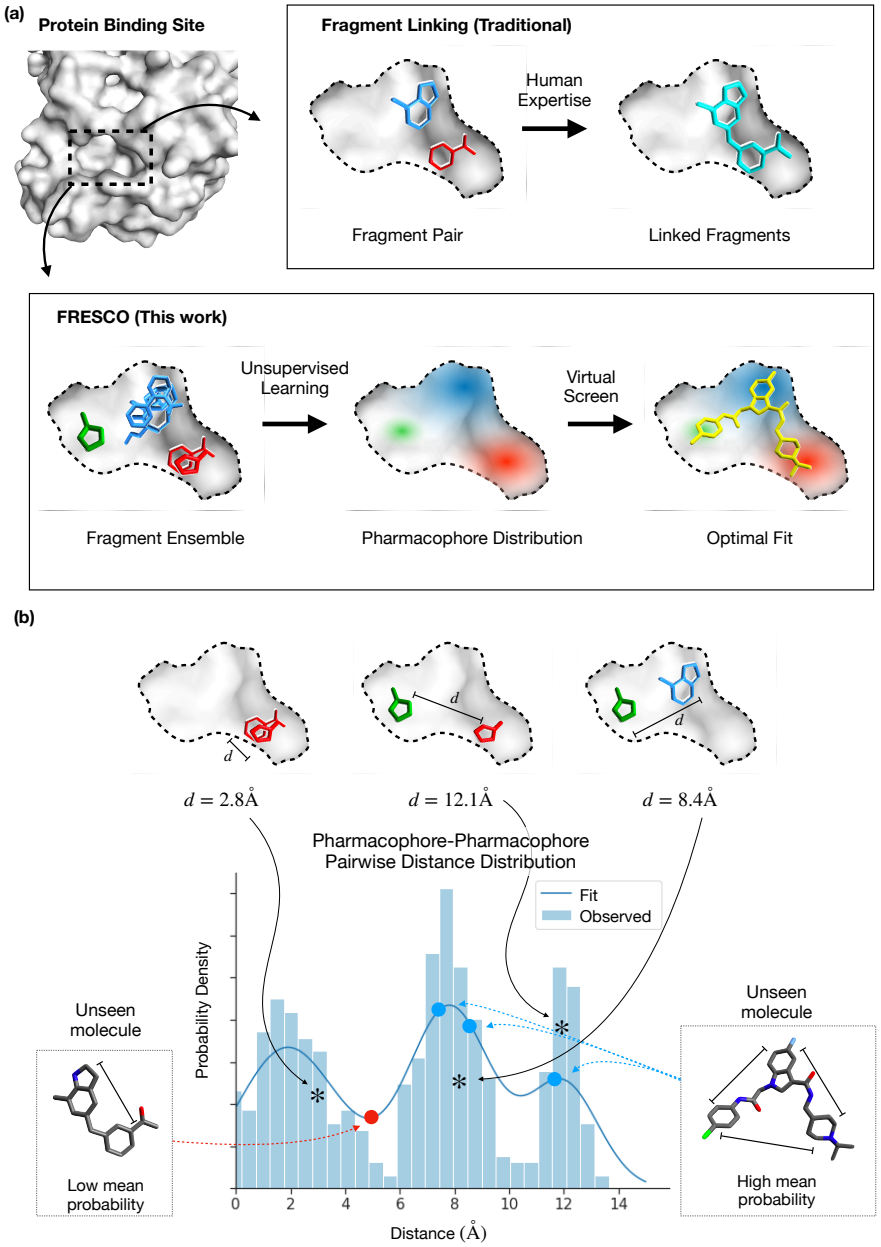



Fig. 3.2 (a) A visual illustration of how FRESCO differs from traditional fragment linking approaches. (b) A visual illustration of how we apply unsupervised learning to fragment ensembles and perform virtual screening of unseen molecules.

## 3.2 Results

### 3.2.1 Unsupervised Learning of Pharmacophore Distributions

To turn fragment hits into a model that predicts whether an unknown ligand will bind potently to the binding site, we employ an interpretation inspired by statistical physics. There are

multiple chemical motifs that can engage residues on the binding site. These different modes of engagement can be considered as a statistical distribution. Each interaction between a chemical motif on the fragment and a binding site residue corresponds to an instance of this statistical distribution. We assume that the fragment library broadly covers chemical space, and anticipate that stronger interactions will be sampled and therefore observed more often amongst fragment hits than weaker interactions. Note that an individual fragment is a weak binder – fragment screens are done at a high concentration which forces the equilibrium towards forming fragment-protein complexes enabling detection via crystallography. Therefore, we analyse the statistical distribution of fragment-protein interactions formed by the dense fragment hits, rather than any individual fragment (Figure 3.2a).

To numerically approximate this distribution, we quantify binding interactions by coarse-graining the fragment molecules into hydrogen-bond donor, hydrogen-bond acceptor, and aromatic ring "pharmacophores" (Figure S1). These are a simple abstractions of molecular features that can make potent interactions with binding site residues, and is a commonly used tool to interpret the biological activity of ligands [41]. The distribution which we then choose to approximate is the pair-wise distance between these pharmacophores. Computational screening of compounds based on pharmacophore distances is a commonly used technique in medicinal chemistry, though here we are extending this concept to enable a statistical interpretation of fragment hit. We consider pharmacophore features, rather than specific protein-ligand interactions, so that the downstream model takes the ligand as the input rather than having to perform the additional step of computationally placing the ligand in the binding site.

We utilise kernel density estimation [60] to estimate this spatial distribution of pair-wise pharmacophore distances (Figure S2). We then score unseen molecules by evaluating pharmacophore distances within that molecule against the probability distribution of pharmacophore distances derived from the fragment ensemble (Figure 3.2b). We take the mean probability over all of the distances between all possible pharmacophore-pharmacophore pairs as the score for the molecule. This is an unsupervised approach – starting from the results of a crystallographic fragment screen, without any bioactivity data, we can build a model that computationally screens unseen molecules. We term our approach Fragment Ensemble Scoring (FRESCO).

FRESCO conceptually departs from machine learning approaches in the literature for fragment-based hit discovery. These approaches, such as DeLinker [36], SyntaLinker [82], and Develop [37]), as well as data-mining methods such as Fragment Network [28], attempt to grow single fragments or merge only a pair of fragments. They all require expert insights in choosing which fragments to merge, or what pharmacophoric constraints need to obeyed, instead of leveraging all of the information from an ensemble of fragment hits in a data-driven manner.

FRESCO also closes a gap in the burgeoning literature on machine learning for bioactivity prediction [58]. These models cannot be used when no training data exists, as is the case in the hit-finding phase. Thus a new modelling approach – here we employed unsupervised learning – is needed to tackle the "zero-to-one" problem. Although physics-based model of ligand-protein binding such as docking [51, 5, 22] can be used in the absence of any bioactivity data, FRESCO crucially incorporates information on preferential interactions between regions of the binding site and pharmacophore on the fragments.

We validate our approach by performing a retrospective study on historical data, as well as embarking on prospectively campaigns on two different protein targets. Retrospective tests or benchmarks, typically the only method used to compare machine learning models, are insufficient for measuring the impact of incorporating the model in the decision-making process of compound selection in drug discovery [42]. Thus we go beyond typical model development and undertake a prospective search for hit molecules using only FRESCO to obtain a more realistic measure of its performance.

### 3.2.2 Computational Retrospective Study

To validate FRESCO, we evaluate how our method compares against the computational approach of docking, as well as the human expertise of medicinal chemists. Specifically, we wish to estimate the extent to which FRESCO could have accelerated hit identification in a fragment-based drug discovery campaign. This requires a dataset that is explicitly exhibiting structure-activity data from the fragment-to-lead phase of a campaign to accurately reflect the degree of structural diversity and distribution of molecular activity. Use of data from an early-stage high-throughput screen would exaggerate the diversity of structures explored, while data from the lead-optimisation phase of a campaign would artificially contain many potent molecules.

For this reason, we choose to study the COVID Moonshot campaign [75] which is targeting the SARS-CoV-2 main protease (Mpro). Mpro is a target of interest for antiviral drug design as inhibition of Mpro inhibits viral replication, as shown by the recent clinical successes of Paxlovid and Ensitrelvir [59, 78]. COVID Moonshot is, to our knowledge, the only openly available dataset of fragment-to-lead drug discovery, driven by a community of medicinal chemists, where every structure and associated activity is disclosed. This unique dataset allows us to perform a time-split analysis, focusing on the fragment-to-lead phase.

In addition, molecular docking studies have also been done extensively on molecules from the Moonshot campaign [57, 64]. For our analysis we utilise the same docking protocols as those reported previously for consistency, the details of which can be found in the methods section.

In the hit identification phase of drug discovery, relatively little is known about what ligand-protein interactions are feasible, thus most proposed molecules are unlikely to be active. A meaningful metric for comparing methods in this regime is the top-$N$ "hit rate", which measures the percentage of the top-$N$ predictions which are active. We expect the curve from plotting the hit rate against $N$ of an informative method to be consistently higher than that of a less informative method. For the Moonshot data we set an IC50 (concentration of inhibitor required to inhibit 50% of protein activity) threshold of 5$\mu$M for defining a "hit". This threshold is relatively arbitrary and we also repeat the analysis for both lower and higher IC50 thresholds (Figure S3).

The baseline hit rate in the dataset i.e. the percentage of compounds with IC50 $< 5\mu$M , is 6.0%. This represents the hit rate of medicinal chemists using traditional and computational tools at their disposal to design compounds for the Moonshot drug discovery campaign. The hit rate for docking is computed by choosing the top-$N$ molecules with the best score. To calculate the hit rate for FRESCO, we first fit a FRESCO model on 23 publicly reported crystallographic structures of non-covalent fragments bound to the SARS-CoV-2 Mpro protein [19] and score the whole dataset using the fitted FRESCO model.

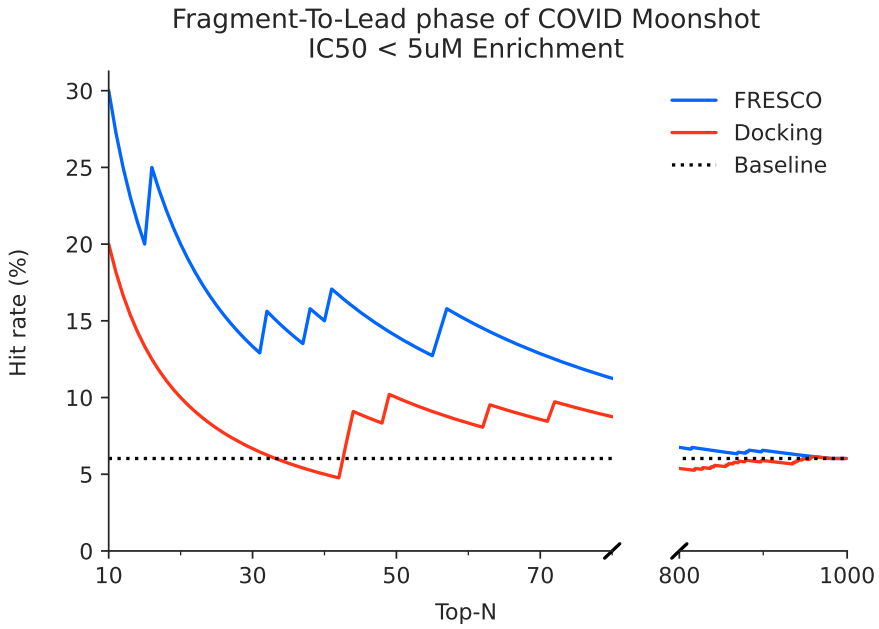



Fig. 3.3 FRESCO is able to retrospectively perform hit detection. High hit rates are achieved relative to docking and the human expert baseline when ranking molecules from the fragment-to-lead phase of COVID Moonshot.

### 3.2.3 Hit finding against SARS-CoV-2 Mpro

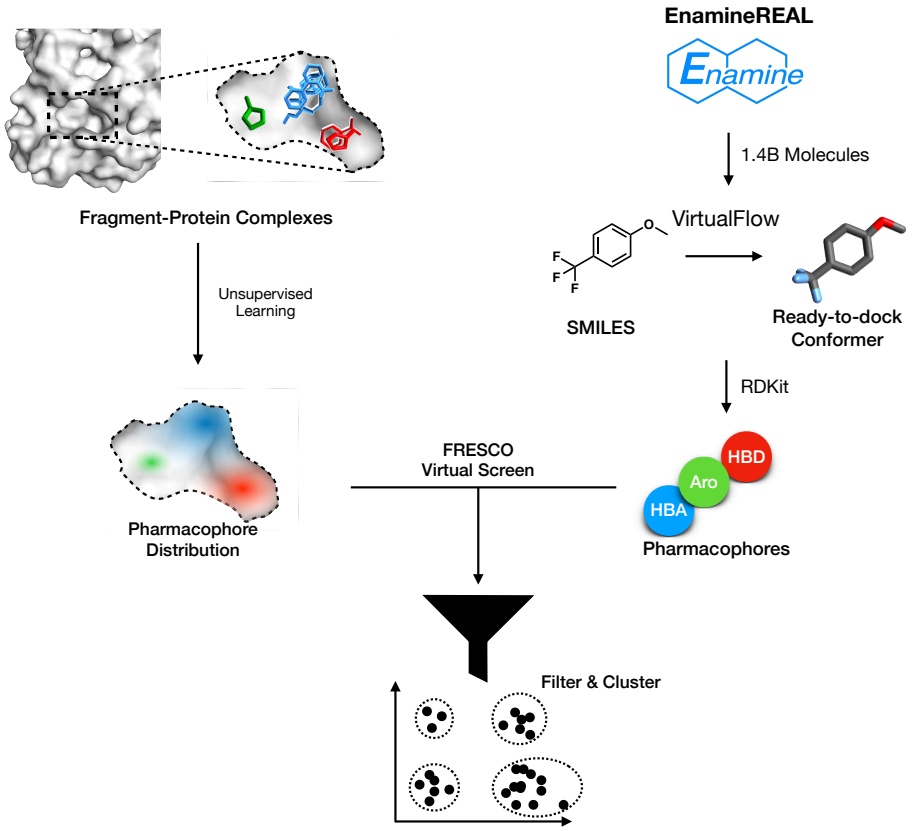



Fig. 3.4 A schematic of the FRESCO screening workflow.

Building on the results of the retrospective evaluation, we performed a prospective study on Mpro. Rather than rescreening Moonshot compounds, we instead deployed the model to screen the whole Enamine REAL database of 1.4 billion molecules implemented in VirtualFlow [27] library of commercially available compounds. We then focused the top predicted compounds, filtered them by their physical properties to maximise "drug-likeness", and selected diverse compounds by clustered hit by structural similarity and picking centroids of the most populous clusters (Figure 3.4).

The cluster centroids favored by FRESCO are typically 2 aromatic moieties connected via an amide or amide isostere. This scaffold is exhibited by three of the initial fragment hits (x0434, x0678, x1093), with the most of the other fragment hits possessing an aromatic group bound at similar locations (Figure S4). We successfully synthesised and assayed 38 of the compounds identified by FRESCO (see SI for the whole library). The most promising compound, WIL-UNI-d4749f31-37, has an IC50 of 25.8$\mu$M measured via fluorescence assay while the remaining compounds were found to be weak-to-negligible activity.

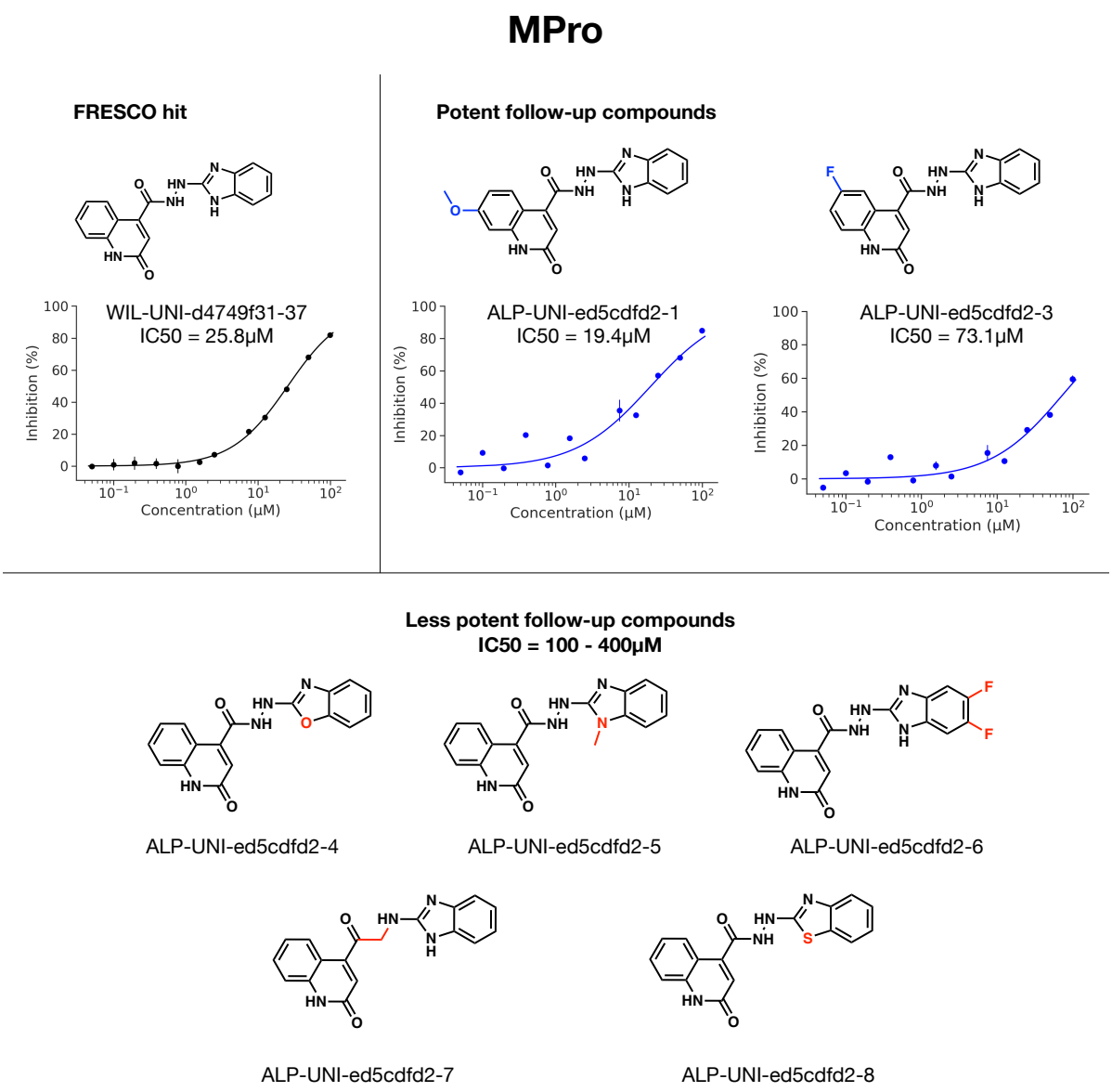



Fig. 3.5 Compound WIL-UNI-d4749f31-37 is identified as a hit against Mpro, with hit confirmation via follow-up compounds demonstrating SAR. Perturbations to the 2-hydroxyquinoline substructure of WIL-UNI-d4749f31-37 led to increased potency while changes to the benzimidazole group consistently decreased potency. Structural differences between the follow-up compounds and WIL-UNI-d4749f31-37 are highlighted in blue/red.

To validate compound activity, we synthesized 8 close analogues to demonstrate the existence of responsive Structure-Activity Relationship [31, 56] (Figure 3.5). 3 of those compounds, which contained modifications to the 2-hydroxyquinoline substructure of WIL-UNI-d4749f31-37, retained relatively high potency of IC50 $< 100\mu$M with one of them (ALP-UNI-ed5cdfd2-1) exhibiting a lower IC50 of 19.4$\mu$M . The remaining 5 compounds

which perturbed the benzimidazole functional group of WIL-UNI-d4749f31-3 exhibit decreased potency, with only 20-50% inhibition at a concentration of 99.5$\mu$M .

### 3.2.4 Hit finding against SARS-CoV-2 nsp3-Mac1

We then turn to SARS-CoV-2 nsp3-Mac1, a structurally unrelated protein target, to demonstrate generalisability of FRESCO in performing hit detection. nsp3-Mac1 is a viral ADP-ribosylhydrolase which counteracts host immune response by cleaving ADP-ribose that is transferred to viral proteins by host ADP-ribosyltransferases. Unlike Mpro, there is no potent chemical matter against nsp3-Mac1. As such, this is a novel first-in-class biological target.

Repeating the FRESCO workflow on a fragment screen against Mac1 [67], we find that the molecules favored by FRESCO tend to contain a HBA-HBD pair that is spatially proximal within a hetrocyclic motif. This mimics adenosine, a core in the natural substrate, and this motif is shared in many of the initial fragment hits (Figure S5). We successfully ordered and assayed 52 of the compounds identified by FRESCO (see SI for the whole library). Two of the compounds show non-negligible activity at high concentration - at 250$\mu$M , compound Z5551425673 (as a racemic mixture) has an inhibition of 30.1%, while compound Z1102995175 has 24.8%.

In addition, an X-ray crystallographic screen was also run on the compounds revealing the structure of Z5551425673 (as the S-stereoisomer) bound to the active site (Figure 3.6). Crystal structures of 9 other compounds chosen via the FRESCO workflow were also obtained though they did not show notable inhibition via HTRF assay (Figure S6). The orthogonal experimental assay and crystal structure results confirm that Z5551425673 is a hit.

As with Mpro, 11 close analogues to Z5551425673 were ordered to explore the structure-activity relationship of the hit and ensure that the compound is not a singleton. 4 compounds perturbing the aliphatic tail substructure had relatively negligible effect while the remaining compounds perturbing the purine group led to a large drop in activity (Figure 3.7). These sets of molecules, still weak in potency, are potentially promising starting points for a hit expansion campaign.

## 3.3 Discussion and Conclusion

Here we show that the combination of computational statistics with high-throughput structural biology and large libraries of purchasable fragment-like molecules unlocks a powerful tool in hit discovery. Going beyond classical fragment-based drug design, which involves merging or expanding a small set of fragments, we derived a statistical framework that leverages

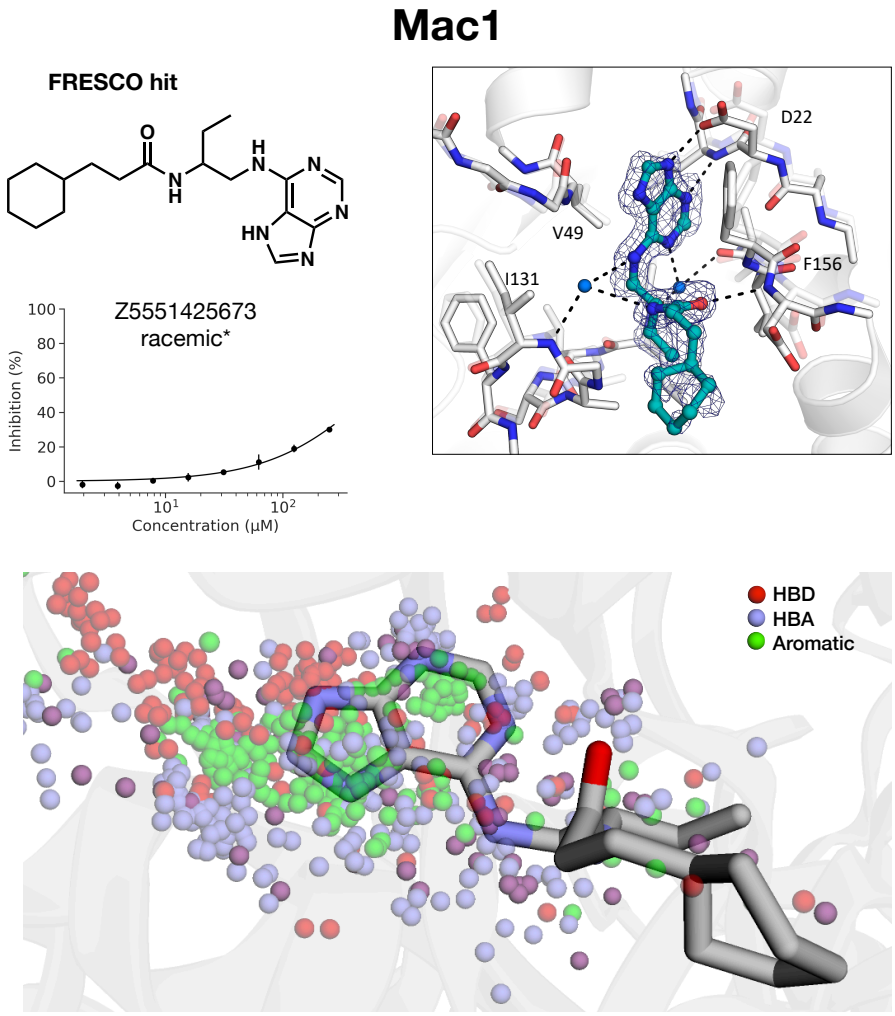



Fig. 3.6 (a) Compound Z5551425673 is identified as a hit against Mac1 via HTRF assay, with (b) hit confirmation via resolution of a crystal structure of Z5551425673 (colored in cyan) bound to the Mac1 active site. (c) The pharmacophores of Z5551425673 match those exhibited by the fragment hits as highlighted by overlaying the bound structure of Z5551425673 (PDB 7FR2) on the distribution of pharmacophores from the fragment ensemble. Note that some functional groups can be regarded as both hydrogen-bond acceptor (blue) and hydrogen-bond donor (red) pharmacophores and hence they are illustrated as purple.

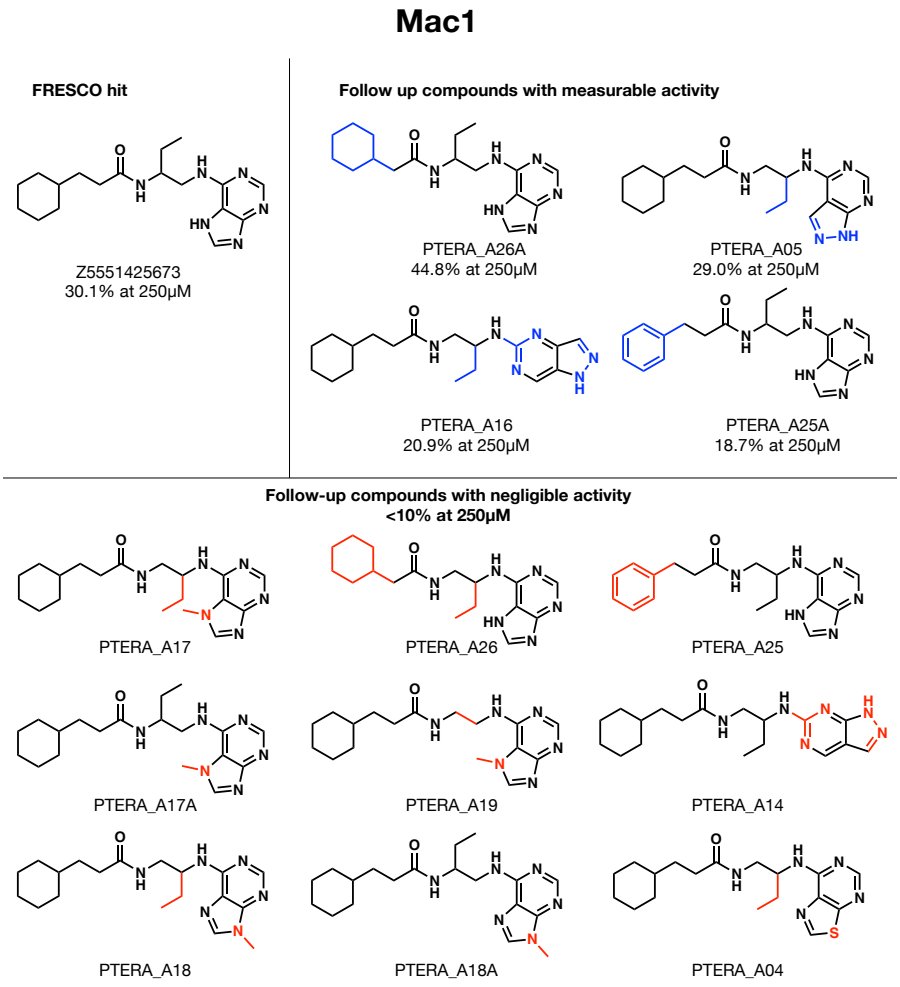



Fig. 3.7 Close analogues around the hit compound identified by FRESCO, Z5551425673, reveals structure-activity relationship which derisks singleton artefacts.

dense fragment hits to build potent inhibitors. Whilst individual fragments are weak binders, our key insight is that a fragment-protein interaction is likely to be significant if there are multiple fragments making similar interactions. Therefore, by picking out these persistent interactions, we can discern the salient chemical motifs which make favourable interactions with the binding site. Specifically, we coarse-grained fragments into pharmacophores, and infer the distribution of pairwise distances between pharmacophores using Kernel Density Estimation. We then screen large libraries of purchasable compounds against this fragment-derived pharmacophore distribution. We retrospectively validated our method using data from The COVID Moonshot, an open science drug discovery campaign against the SARS-CoV-2 main protease, and prospectively discovered new hits against SARS-CoV-2 main protease and nsp3-Mac1.

More generally, we note that our method does not require the observation of affinity data in order to infer potency. This is done by employing an unsupervised machine learning approach on unlabelled structural biology data. As the throughput of structural biology increases, we hope that an unsupervised approach may unlock novel ways of overcoming data limitations in the protein-ligand affinity prediction problem.

Finally, although prospective studies demonstrated FRESCO's ability to identify hits, we note that the hit rate and potency of the identified hits are both lower than the retrospective experiments. This highlights the importance of prospective validation in machine learning – retrospective studies are biased by the fact that the model is rescoring "reasonable" design from medicinal chemists, whereas in prospective evaluations, the model is used to score the large chemical space without further inductive biases. Future efforts to improve FRESCO should seek to include further inductive biases, for example incorporating physics-based constrains such as docking to filter FRESCO outputs, as well as solidifying a human-in-the-loop approach to select top hits.

## 3.4 Methods

### 3.4.1 Datasets

Fragment crystal structures for model training were downloaded from `Fragalysis`. For Mpro, non-covalent fragments from the XChem fragment screen [19] were used while for Mac1 both XChem and UCSF fragment data were used [67].

The Moonshot activity data for the retrospective study was accessed in Mar 22nd 2021. The IC50 values in that dataset, as well as in the prospective study on Mpro were measured from a fluorescence based enzyme activity assay, the details of which are described below. To narrow

down the data to moleucles during the fragment-to-lead stage of the Moonshot campaign, we only selected molecules which were designed before September 1st, 2020, which gave us a dataset of 979 compounds.

For Virtual Screening, we utilize VirtualFlow, a published dataset of more than 1.4 billion commercially available molecules from EnamineREAL & ZINC15 in a ready-to-dock format [27].

### 3.4.2 Model Construction

The model used in this work takes as input the 3D pharmacophore distribution of a candidate molecule, and evaluates the log-probability that the distribution matches that of the fragment screen on the target site.

The 3D pharmacophore distribution of a molecule is obtained by extracting pharmacophores from the molecular SMILES and their corresponding conformer coordinates, and then evaluating the pairwise distance matrix between all possible pharmacophore pairs (eg Donor-Donor & Aromatic-Acceptor). SMARTS pattern matching following default pharmacophore definitions in `RDKit` were used to extract pharmacophores from the fragment SMILES. The pharmacophores considered are hydrogen bond donors, hydrogen bond acceptors, and aromatic rings. The coordinates of each pharmacophore are defined as the average over the atoms in the pharmacophore (eg the position of an aromatic pharmacophore from a benzene ring would be the mean of the coordinates of the 6 carbon atoms in the ring).

For some fragments, multiple crystallographic poses are recorded. To account for this, we weigh the contribution of each fragment structure to the overall fragment pharmacophore distribution by $\frac{1}{n}$ where $n$ is the number of conformations recorded for each conformer. In addition, we exclude the counting of correlations between pharmacophores from the same fragment - only correlations between different fragments are measured. This is to avoid spurious intra-fragment correlations that are unrelated to binding to the binding site - strong correlations in pharmacophore distribution between multiple independent fragments are indicative of useful binding interactions and these are what we hope to capture with this methodology.

The bandwidth for KDE fitting was chosen for each system using the Improved Sheather-Jones algorithm [8] (implemented in `KDEpy`). KDEs of the systems are then constructed using the chosen bandwidths with `scikit-learn` for technical ease of use in evaluating probabilities. The `scikit-learn` implementation relies on a relatively slow tree-based algorithm that searches over the training datapoints - to increase the efficiency of virtual screening, computationally fast approximations of the KDEs are made using the `scipy interp1d` function.

Virtual screening of molecular libraries is done by evaluating the probability of the pharmacophore distribution of each molecule using the KDEs in Python. We utilised the VirtualFlow

library which provided molecular conformers calculated via ChemAxon and OpenBabel in PDBQT format, converted the conformers into SDF format, and generated the pharmacophore distributions with RDKit as described above. The pharmacophore distributions for each molecule are saved as a pickled dictionary of numpy arrays. Each entry of the dictionary contains the flattened pairwise distance distribution for a particular pharmacophore pair. The KDE functions for each pharmacophore combination, trained on the fragment ensemble, are applied onto the dictionary and outputs a mean log-probability for the corresponding pharmacophore combination. The overall score for the molecule is returned as the mean log-probability over all of the pharmacophore combinations.

### 3.4.3 Compound Selection

After conducting a virtual screen, the top-500k predictions were selected and filtered to remove undesirable properties. A series of successive filtering steps were performed: first, only molecules with physical properties in well-understood "lead-like" chemical space [Che] were kept. Secondly, the sum of the number of hydrogen bond donors and hydrogen bond acceptors were constrained to an upper limit of 8. Then, we remove molecules that match known filters for pan-assay interference compounds (PAINS) [7] as well as filters for moieties that are undesirable for medicinal chemistry (eg furan, thiophene, nitro groups). Duplicate tautomers for each molecule are also removed. Finally, for ease of synthetic accessibility, we only consider molecules with less than two chiral centers.

The top-50k molecules remaining from the filtering were then clustered via Butina Clustering [10] with a Tanimoto distance threshold of 0.2. This resulted in 24748 and 22358 clusters for Mpro and Mac1, respectively. For both targets the centroids of the 50 most populous clusters (or the closest purchasable analogue if it wasn't available) were chosen as the candidate compounds. These compounds were ordered for synthesis from Enamine which resulted in 38 and 52 successfully made molecules for Mpro and Mac1, respectively.

### 3.4.4 Docking against SARS-CoV-2

All molecules synthesised by the COVID Moonshot Consortium were docked against structure x2908 reported by Diamond XChem [19]. We use the "Classic OEDocking" floe v0.7.2 as implemented in the Orion 2020.3.1 Academic Stack (OpenEye Scientific). Omega was used to enumerate conformations (and expand stereochemistry) with up to 500 conformations. FRED was used for docking in HYBRID mode using the x2908 bound ligand. The docked poses of the ligands were scored using the Chemgauss4 scoring function.

### 3.4.5 Homogeneous Time Resolved Fluorescence assay

The experimental procedure for measuring Mpro inhibition is the same as that previously reported by COVID Moonshot[75], which is repeated below.

Dose response assays were performed in 12 point dilutions of 2-fold, typically beginning at 100$\mu$M . Highly active compounds were repeated in a similar fashion at lower concentrations beginning at 10$\mu$M or 1$\mu$M . Reagents for Mpro assay were dispensed into the assay plate in 10$\mu$l volumes for a final volume of 20$\mu$L.

Final reaction concentrations were 20mM HEPES pH7.3, 1.0mM TCEP, 50mM NaCl, 0.01% Tween-20, 10% glycerol, 5nM Mpro, 37nM fluorogenic peptide sybstrate ([5-FAM]-AVLQSGFR-[Lys(Dabcyl)]-K-amide). Mpro was pre-incubated for 15 minutes at room temperature with compound before addition of substrate and ex/em filter set. Raw data was mapped and normalized to high (Protease with DMSO) and low (No Protease) controls using Genedata Screener software. Normalized data was then uploaded to CDD Vault (Collaborative Drug Discovery). Dose response curves were generated for IC50 using nonlinear regression with the Levenberg-Marquardt algorithm with minimum inhibition = 0% and maximum inhibition = 100%.

Inhibition of SARS-CoV-2 nsp3-Mac1 (aa residues 206–379 of nsp3) was assessed by the displacement of an ADP-ribose conjugated biotin peptide from His6-tagged protein using a HTRF-technology-based screening assay which was performed as previously described [67]. Compounds were dispensed into ProxiPlate-384 Plus (PerkinElmer) assay plates using an Echo 525 liquid handler (Labcyte). B inding assays were conducted in a final volume of 16 $\mu$l with 12.5 nM SARS-CoV-2 nsp3-Mac1 protein, 400 nM peptide ARTK(Bio)QTARK(Aoa-RADP)S (Cambridge Peptides), 1:20000 Anti-His6-Eu3+ cryptate (HTRF donor, PerkinElmer) and 1:125 Streptavidin-XL665 (HTRF acceptor, PerkinElmer) in assay buffer (25 mM HEPES pH 7.0, 20 mM NaCl, 0.05% bovine serum albumin and 0.05% Tween-20). Assay reagents were dispensed manually into plates using a multichannel pipette while macrodomain protein and peptide were first dispensed and incubated for 30 min at room temperature. This was followed by addition of the HTRF reagents and incubation at room temperature for 1 h. Fluorescence was measured using a PHERAstar microplate reader (BMG) using the HTRF module with dual emission protocol (A = excitation of 320 nm, emission of 665 nm, and B = excitation of 320 nm, emission of 620 nm). Raw data were processed to give an HTRF ratio (channel A/B $\times$ 10,000), which was used to generate IC50 curves. The IC50 values were determined by nonlinear regression using GraphPad Prism v.9 (GraphPad Software, CA, USA).

### 3.4.6 Crystallographic Screening

Crystallographic screening of compounds was performed using Mac1 crystals grown in the P43 space group, following the previously described protocol (PMID: 33853786). Compounds synthesized by Enamine/WuXi were prepared in DMSO to 100 mM and were added to crystallization drops using an Echo 650 liquid handler (Labcyte) (PMID: 28291760). Crystals were soaked at either 10 or 20 mM for 2-4.5 hours, before being vitrified in liquid nitrogen using a Nanuq cryocooling device (Mitegen). Soak times and concentrations are listed in Table S1. Diffraction data were collected at beamlines 12-1 and 12-2 of the Stanford Synchrotron Radiation Lightsource (SSRL). The data collection strategy and statistics are listed in Table S1. Compound binding was detected using the PanDDA algorithm (PMID: 28436492) as described previously (PMID: 35794891). PanDDA was initially run using a background map calculated with 34 datasets collected from crystals soaked only in DMSO (annotated as dmso_34 in Table S1). PanDDA was rerun with a background map calculated using two sets of 35 datasets where no compound binding was detected (annotated as either ssrl_1 or ssrl_2 in Table S1). This procedure led to the identification of an additional nine hits (Table S1).

Compounds were modeled into PanDDA event maps using COOT (PMID: 20383002) with coordinates and restraints generated by phenix.elbow from SMILES strings (PMID: 19770504). Duplicate soaks were performed for most compounds: where the same compound was identified in multiple datasets, the highest occupancy compound was modeled. Both the compound-bound and compound-free coordinates were refined together as a multi-state model following the protocol described previously (PMID: 28436492). Compound occupancy was set based on the background density correction (BDC) value (PMID: 28436492). Refinement statistics are presented in Table S1. Coordinates and structure factor amplitudes have been deposited in the protein data bank (PDB) with the group deposition code G_1002254. PanDDA input and output files have been uploaded to Zenodo (DOI: 10.5281/zenodo.7231822), and the raw diffraction images are available at https://proteindiffraction.org/.

## 3.5 Author Contributions

WM and AAL designed the study. WM and AAL devised the predictive model and WM implemented it. WM and AAL wrote the original draft, all authors commented on it.

WM acknowledges the support of the Gates Cambridge Trust. AAL acknowledges the Winton Programme for the Physics of Sustainability.

All data and code used for this work can be found in the GitHub repo https://github.com/wjm41/frag-pcore-screen. Supplementary figures and tables can be found in an accompanying file.

# Chapter 4

# Discovery of SARS-CoV-2 main protease inhibitors using a synthesis-directed de novo design model

Electronic Supplementary Information (ESI) contains experimental and assay details. Our training set, de novo design method and generated molecules are available on https://github.com/wjm41/mpro-rank-gen.

**The SARS-CoV-2 main viral protease ($M^{pro}$) is an attractive target for antivirals given its distinctiveness from host proteases, essentiality in the viral life cycle and conservation across coronaviridae. We launched the COVID Moonshot initiative to rapidly develop patent-free antivirals with open science and open data. Here we report the use of machine learning for *de novo* design, coupled with synthesis route prediction, in our campaign. We discover novel chemical scaffolds active in biochemical and live virus assays, synthesized with model generated routes.**

Coronaviruses are a family of pathogens that is frequently associated with serious and highly infectious human diseases, from the common cold to the SARS-CoV pandemic (2003, 774 deaths, 11% fatality rate), MERS-CoV pandemic (2012, 858 deaths, 34% fatality rate) and most recently the COVID-19 pandemic (ongoing pandemic, 1.7 million deaths up to Dec 2020).

The main protease ($M^{pro}$) is one of the best characterized drug targets for direct-acting antivirals [62, 11]. $M^{pro}$ is essential for viral replication and its binding site is distinct from known human proteases, thus inhibitors are unlikely to be toxic [40, 47]. Moreover, the high degree of conservation across different coronaviruses renders $M^{pro}$ targeting a fruitful avenue towards pan-cornavirus antivirals [77]. To date, most reported $M^{pro}$ inhibitors are

peptidomimetics, covalent, or both [11]. Peptidomimetics are challenging to develop into oral therapeutics, and covalent inhibitors incur additional idiosyncratic toxicity risks. We launched the COVID Moonshot consortium in March 2020, aiming to find oral antivirals against COVID-19 in an open-science, patent-free manner [12].

Here we report the prospective use of a simple model to rapidly expand hits. Starting from 42 compounds with $\mathrm{IC}_{50}$ within assay dynamic range ($< 100\mu\mathrm{M}$) and 515 inactives, our model designed 5 new compounds predicted to have higher activity, together with predicted synthetic routes. All designs were were chemically synthesized and experimentally tested, and 3 have measurable activity against $\mathrm{M}^{\mathrm{pro}}$. The top compound has comparable $\mathrm{M}^{\mathrm{pro}}$ inhibition to the best in the training set, but with a different scaffold, and is active against the OC43 coronavirus in a live virus assay.

Algorithmic *de novo* design aims to automatically generate compounds that are chemically diverse, synthetically accessible and biologically active [66]. Classic approaches apply heuristics to fragment and modify known active compounds, with the region of chemical space explored and synthetic accessibility constrained by those rules [9, 61, 30]. Recent machine learning approaches explore chemical space in more abstract molecular representation space [26, 70], but this often comes at the expense of synthetic accessibility [24]. Our approach builds on rule-based fragmentation and molecule generation, but employs a method that combines regression and classification amid noisy data, and use of machine learning to predict synthesis routes. Our model comprises two parts: compound prioritisation and chemical space exploration.

Our compound prioritisation model aims to predict whether a designed compound is likely to be an improvement in activity over the incumbent. However, as is typical in the hit-expansion stage, bioactivity modelling is hindered by insufficient data where the majority of compounds are inactive, and noisy data as measurement variability increases for lower affinity compounds. Thresholding the data and framing the problem as classification of active/inactive would not allow us to rank compounds based on predicted improvement over the incumbent, yet the amount of measured bioactivity data and the measurement noise makes a regression approach challenging.

To overcome both challenges, we develop a learning-to-rank framework [20, 2]. Rather than training a regression model to predict the $\mathrm{IC}_{50}$ of a compound, we instead train a classifier to predict whether a compound is more or less active than another compound, with the input to the model being the *difference* in molecular descriptors between the molecules (see Figure 4.1 for a schematic). This model accounts for both compounds with $\mathrm{IC}_{50}$ measurements and compounds that are simply inactive – active compounds are ranked by their $\mathrm{IC}_{50}$, all inactives with no measurable $\mathrm{IC}_{50}$ are considered less active than active compounds, and inactive-inactive pairs are ignored. Further, we account for noise by only considering $\mathrm{IC}_{50}$

| **Percentile** | 1% | 2.5% | 10% |
|---|---|---|---|
| **Enrichment Factor** | 1.7 | 2.3 | 1.7 |

Table 4.1 Enrichment factor for the time-split dataset, where we consider model performance on data arriving after the model has been deployed to generate compounds for synthesis and testing.

differences amongst actives above 5 $\mu M$. We use the FastAI Tabular model [32], with input features generated from concatenated Morgan, Atom Pair, and Topological Torsion fingerprints implemented in RDkit [Landrum], and dataset was randomly split into training (80%) and testing (20%); details about model implementation can be found in ESI and source code.

Figure 4.1 shows that our binary ranking model achieves an AUC of 0.88 (95% CI: [0.83,0.96]) in ranking ligands within the test set, and AUC for 0.94 (95% CI: [0.91,0.98]) where we compare a ligand in the training set against another ligand in the test set; the latter is more relevant as our goal is finding ligands more active than the best incumbent. The 95% confidence interval is computed using bootstrapping. We also compare our model against OpenEye's FRED hybrid docking mode as implemented in the "Classic OEDocking" floe, a physics-based docking algorithm, on the Orion online platform, which achieves AUC of 0.72; 95% CI: [0.722,0.723] (see ESI for implementation details). Note that docking does not require ligand bioactivity as training data, thus is not a directly comparison to machine learning. In the Supplementary Material, we discuss that our model ranks ligands better than a model that directly learns $\mathrm{IC}_{50}$ (AUC = 0.86; 95% CI: [0.71,0.95]).

Beyond train-test split, model performance can be evaluated from a time-split. Five months have elapsed from the time we deployed our model to select compounds to writing up the manuscript. During that time, the COVID Moonshot Consortium (a team of expert medicinal chemists) has independently designed, synthesised and tested 356 compounds [74], out of which 15% were better than the top 2 compounds (having $\mathrm{IC}_{50}$ comparable within error) in our dataset. Table 4.1 shows that our model has an enrichment factor of $\sim$2, i.e. if we rescore the 356 compounds synthesized by the medicinal chemistry team using our model, and pick the top 1%-10% percentile, the proportion of molecules that would be better than the top 2 compounds would be $\sim$2x higher than human selection.

Having demonstrated the accuracy of our ranking model, we now turn to chemical space exploration. We first consider a set of chemically reasonable perturbations (e.g. amide to retroamide, amide to urea), which is applied to the whole set of active molecules. We then fragment along synthetically accessible bonds (e.g. amides and aromatic C-C and C-N), and reconnect the synthons to generate an exhaustive library. The resulting library of 8.8 million

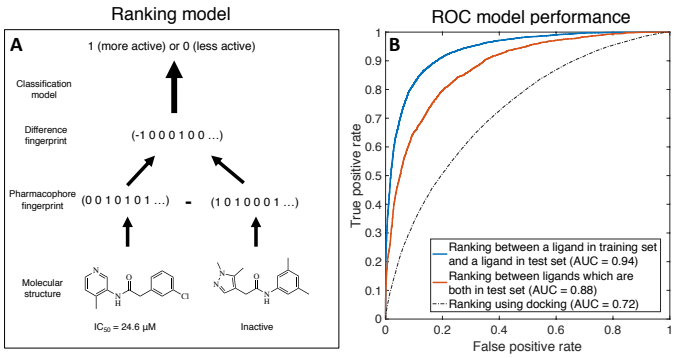



Fig. 4.1 Relative ranking of ligands can be predicted by our learning-to-rank machine learning model. (A) A schematic of the model setup. A classifier takes the difference in pharmacophore fingerprint between two molecules and predicts where one molecule is more or less active than the other. (B) The Receiver Operating Characteristic curve of classifying whether a molecule is more/less active than the other. AUC 95% CI reported in main text.

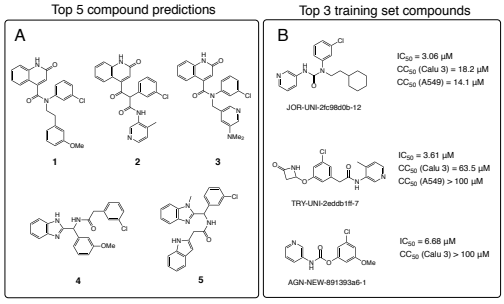



Fig. 4.2 Our synthesis-driven design model prioritises molecular scaffold that are not in the top hits. (A) The 5 compounds selected by our methodology for synthesis and testing. (B) The top 3 compounds from the training set, with potency and cytotoxicity measurements.

generated molecules is scored using our ranking model by the probability of having a higher potency compared to the most potent molecule in the dataset.

Although virtual "reactions" were used to generate new molecules, the synthons are not necessarily off-the-shelf nor the reactions optimal. As such, we use a retrosynthesis predictor to triage based on synthetic accessibility. We fed top hits into Manifold, our platform for synthesis route prediction (https://postera.ai/manifold). Manifold searches for synthetic routes starting from purchasable molecules. The underlying technology is based on Molecular Transformer, a machine learning model for reaction prediction using sequence-to-sequence translation [46, 69]. The top 5 molecules with predicted routes <4 steps were synthesised and tested (Figure 4.2A). For comparison, the most potent molecules from the training set are shown in Figure 4.2B; $\mathbf{1}-\mathbf{5}$ have Tanimoto similarity <0.48 (1024-bit ECFP6) to every molecule in the training set.

Figure 4.3 shows that for Compounds **1**, **2**, **4** and **5** our retrosynthesis algorithm generates successful routes, thus provides a reasonable estimate of synthetic complexity. The syntheses were carried out at the Wuxi AppTec and compounds were assayed as received. Minor variations in building blocks were employed depending on what was readily available. We note that our algorithm failed to estimate the synthetic complexity of Compound **3**. The final amide formation step was unexpectedly challenging, and no desired product was seen despite significant efforts in condition screening. Compound **3** was furnished via an alternative strategy, employing an Ullmann coupling to arylate the amide, which was not predicted by our approach.

Compounds **1**-**5** were tested for Mpro activity using a fluorescence assay. Figure 4.4 shows that Compounds **1**-**3** have $\mathrm{IC}_{50}$ within assay dynamic range ($< 100\mu M$), and Compound **1** has $\mathrm{IC}_{50} = 4.1\mu M$. Compound **1** is further assayed in live virus assays, with the less pathogenic OC43 coronavirus, showing $\mathrm{EC}_{50} = 13\mu M$ and is not cytotoxic ($\mathrm{CC}_{50} > 100\mu M$ against A549 cell line; $CC_{50}$ is the concentration required to cause 50% cell death). We employ OC43 as a rapid surrogate assay for SARS-CoV-2 as the former can be done in a BSL-2 rather than BSL-3 lab. Interestingly, the top non-cytotoxic hit of the training set (TRY-UNI-2eddb1ff-7) does not show OC43 activity, showcasing the utility of using generative models to suggest new scaffolds with complementary physicochemical properties.

In summary, we demonstrated the utility of a *de novo* design model, guided by estimation of synthetic complexity, for generating ideas in hit expansion. At the time of writing, the quinolone series is undergoing optimisation by the COVID Moonshot initiative (https://postera.ai/covid). Data for Compound **1**-**5** is registered as the `ALP-POS-ddb41b15` series on the Moonshot platform.

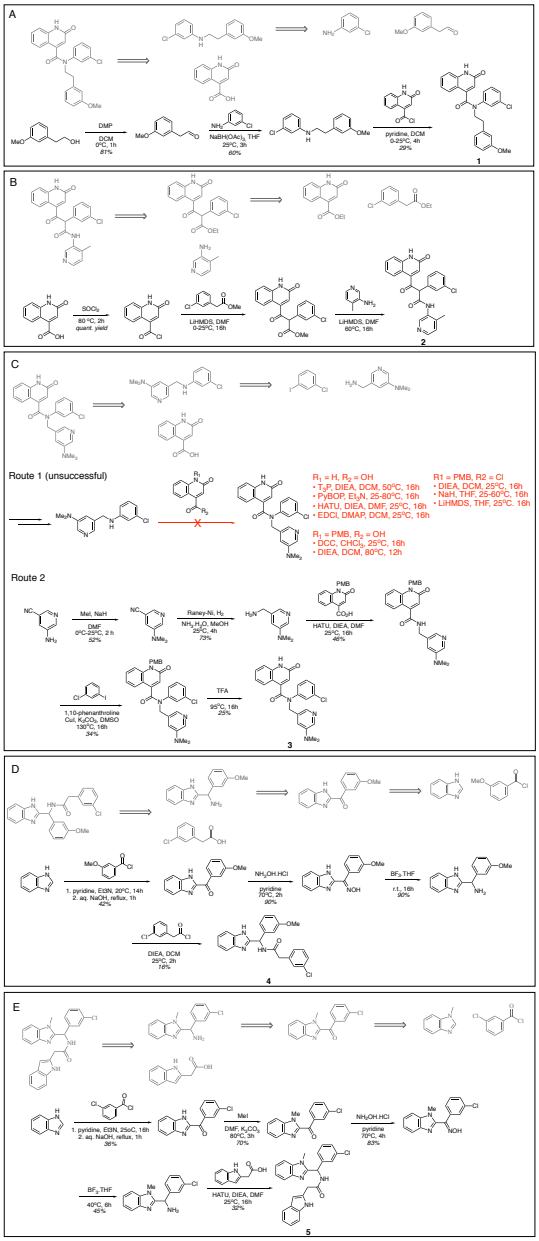



Fig. 4.3 Model generated synthetic schemes that are experimentally validated. Schemes (A)-(E) show the synthesis schemes generated by our model (grey) and experimental schemes for Compounds **1**-**5**. The ESI contains experimental procedures provided by our contract research organisation.

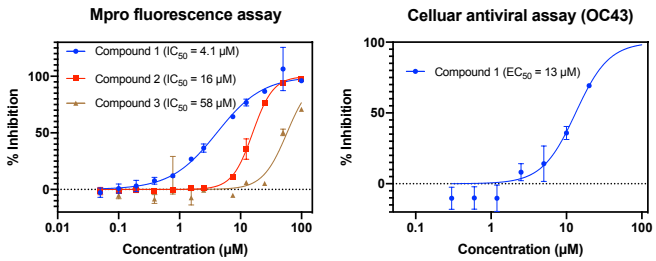



Fig. 4.4 Three compounds generated using our synthesis-directed model exhibit Mpro activity. Our most active compound has measurable antiviral activity against the OC43 coronavirus and no measurable cytotoxic effect ($CC_{50}(A549) > 100\mu M$). 95% CI: IC50 (Mpro) – Compound 1 [3.42,4.86] $\mu$M, Compound 2 [15.1,16.5] $\mu$M, Compound 3 [48.8,69.4] $\mu$M; EC50 (OC43) – Compound 1 [10.1, 18.4] $\mu$M. See ESI for assay details.

# Chapter 5

# Make - Understanding the Molecular Transformer

## 5.1 Introduction

Although the design of drug candidates is exhaustingly difficult, it is in fact the 'make' part of the design-make-test cycle which is the most costly, time consuming and labour intensive. The key to streamlining molecular synthesis is in improving route planning, developing faster ways of designing shorter reaction paths from basic molecular building blocks to the desired molecule, reducing the number of steps and hence the risk of failure.

Once a synthesis route is designed it is important to validate each step of the plan. Forward chemical reaction prediction is concerned with predicting the (major) product of an organic reaction given the reactants, reagents and preferably the conditions like solvent, temperature, concentrations etc. By having the ability to predict the product of reactions with reliable uncertainties it is possible to design clever synthesis plans where the reactions with higher uncertainty are put first. This way if a synthesis protocol fails it does so fast and cheap instead of in the later stages of the route where substantial time and cost would go to waste.

Route planning and reaction prediction have traditionally been done by expert chemists relying on experience, as well as reaction databases like Reaxys [Elsevier]. Nowadays, Computer Assisted Synthesis Planning tools are increasingly being used [16], as these tools can memorise libraries of commercially available building blocks and quickly evaluate large numbers of possible bond disconnections via efficient algorithms such as Monte Carlo Tree Search [71]. Unsurprisingly, machine learning methods have also entered into the fray [15, 14] and have recently emerged as the most successful approach [16, 68].

ML reaction prediction models are trained on reaction data that is extracted from patents and publications. In these documents usually the metadata about reactions like the temperature, concentrations and solvents are found in the synthesis protocol section making it very challenging to extract this information in an automated manner. Therefore these models are usually trained only on the reactants and reagents with all of the context information missing. In spite of this there are reported models achieving remarkably high near 90% Top-1 prediction accuracy on these datasets, even outperforming quantum mechanics-based approaches [68].

The natural question that arises is: how is the model able to achieve such high accuracy on often rather challenging reactions from such limited source of data? Has the model learnt the well-established underlying mechanistic drivers of reactivity purely from data? It is of utmost importance to validate these models to see if they are able to generalize and predict the outcome of reactions reliably or if they are merely learning hidden biases in the datasets which results in the seemingly strong performance.

One way to accomplish this is with ML interpretability methods [6]. Interpretability methods can help uncover the reasoning of model predictions in simple well understood cases where the physical or chemical cause for certain outcomes is well established. For chemical reaction prediction our understanding of mechanisms and selectivities serves as good guides for the observed reactivities.

In this work, we use a well-known ML interpretation method called Integrated Gradients (IGs) to probe the understanding of the Molecular Transformer (MT), the current state-of-the-art machine learning model for chemical reaction prediction. Our approach builds on the work of McCloskey et. al. [54] who used IGs to understand binding prediction models on artificial datasets. We extend the method to Transformer architectures, and use it in the context of reaction predictions on real experimental data. We also present a novel method for attributing the predictions of neural network models to training set datapoints. With these tools we show that MT often fails to learn the mechanistic reasoning behind chemical selectivity and hypothesize that this is due to hidden biases in the dataset. We justify this claim by creating biased synthetic datasets and demonstrating selectivity bias in the model predictions, suggesting that it is the quality of training data rather than the particulars of model architectures that is constraining the potential for ML reaction prediction.

The work in this chapter was done collaboratively with Dávid Péter Kovács. We did the code development together and discussed all of the results of the work. He executed the code, analysed the model attributions, and designed the majority of the experiments and all of the adversarial examples. I created the SMARTS templates for counting statistics from the patent datasets as well as for dataset generation in the synthetic experiments. Preliminary

results from this work were presented at the ICML 2020 'ML Interpretability for Scientific Discovery' Workshop [55]. All code including a `README` with the usage can be found in the GitHub repo `MTExplainer` [Kovacs et al.].

## 5.2 Methods

### 5.2.1 Molecular Transformer

The Molecular Transformer [68] is a tailored version of the Transformer architecture [79] which was designed for machine translation and has had wide-ranging success in many Natural Language Processing tasks. It has an encoder-decoder structure, where both the encoder and the decoder are made up of so called transformer blocks. These blocks process the inputs by applying a multi-head scaled dot-product attention mechanism followed by layer normalization and some fully connected feed forward layers. Mathematical details can be found in (somewhere).

The string input to the model is broken down into individual tokens with a learnt embedding that is fed into the encoder layer with positional encoding. The encoder is composed of 4 identical attention blocks each containing a multi-head self-attention layer and a 2-layer fully connected feed-forward neural network. The decoder is very similar to the encoder with the only difference being that the multi-head attention uses the output of the encoder as the keys and the values with the output of the previous decoder layer being the query. The predictions are generated in an autoregressive way meaning that the decoder predicts one token at a time and the previously generated tokens are fed into the decoder when generating the next tokens. The prediction is considered final when an <end> token is generated or the maximum length is reached. Through this process each translation gets assigned a probability score:

$$P(\text{tgt} \mid \text{src}) = \prod_{i=1}^{N} P(\text{tok}_i \mid \text{tok}_1, \cdots, \text{tok}i-1, \text{src}) \tag{5.1}$$

where $\text{tok}_i$ is the $i$-th predicted token and $N$ is the length of the prediction.

Our implementation of the work was based on the `OpenNMT` package [43].

### 5.2.2 Data

We trained the model on a publicly available dataset of organic reactions mined from the US patent office [49] which has been filtered [39]. The data contains reactants, reagents, and products represented as SMILES (without including stereochemical information) which is a

text based representation of molecules[80, 81]. The training set was made up of 377 419 reactions which we augmented by an equal number of identical reactions made up of random equivalent SMILES. This augmentation is done to help the model to learn the underlying molecular graph from the SMILES sequence.There were 23 589 reactions for the validation set and 70 765 reactions in the hold-out test set, neither of which were augmented. The SMILES strings were tokenized following [68].

The trained model achieved 88.8% Top-1 accuracy on the test set. This model was used throughout the interpretability experiments and is referred to as USPTO Transformer.

The second dataset used was the commercial Pistachio dataset [53]. This dataset contains over 9 million reactions text mined from US and EPO patents. This dataset was filtered similarly to USPTO to remove erroneous and and a large number of duplicate reactions. The final dataset consisted of 2 375 385 reactions, of which 2 019 078 were used for training, 118 770 for validation and 237 537 for testing.

The model trained as described above achieved 76.4% Top-1 accuracy on the test set. Even though this looks like a substantially lower performance in reality the two models perform similarly well on new reactions. The possible reasons for the large difference in the measured performance on the held-out test sets are described in detail below. This model obtained was also used in the interpretability experiments to test the effect of increased training set size on the models understanding of chemistry and is referred to as Pistachio Transformer from here onwards.

### 5.2.3 Integrated Gradients

To understand the predictions of MT with respect to the input features, we use the Integrated Gradients [73] attribution method. Integrated Gradients (IGs) is a principled model-agnostic feature attribution method adapted from game theory which obeys certain axioms of fairness. It can be used for any model where gradients are available, which is the case for all neural networks that are trained by some variant of gradient based optimization.

In general, the attribution of feature $i$ for input $x$ is given by

$$IG_i(x) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \tag{5.2}$$

where $x_i$ is the vector of feature $i$ for the input $x$, and $x_i'$ is a vector corresponding to a non-informative baseline input, $F(x)$ represents the model prediction for input $x$, and the integral is taken over the straight-line path from the baseline to the input of interest.

It has been discussed before that the choice of baseline can have a large effect on the values of the attributions [72]. While we could have chosen unreactive molecules as our

baseline, it is important to select baselines which are completely non-informative to avoid any ambiguity. This would traditionally be the black image in the case of image recognition. In this work we use the embedding vector of the SMILES ' . ' token which is used to separate different molecules and hence does not contain any chemical information on its own.

For MT, we take great care to define $F(x)$ as it is not an appropriate question to ask what part of the reactant-reagent input is most important for predicting a given product. All of the input tokens contain crucial information that are used by the decoder to generate the entire target structure correctly. To eliminate this effect we define $F(x)$ as the difference in predicted probability of two possible products. Since the inert parts of the input are the same for the two products they should not substantially contribute to their predicted probability difference. This method is especially suited for examining reactions with selectivities. In other words we attribute the selectivity between two products to the inputs, ideally highlighting the chemically important groups driving this selectivity (by summing the attributions of the tokens comprising these groups).

If it is found that the correct product is predicted for the wrong reason i.e. the attribution on the chemical important group is low, it can be confirmed that model has not been able to learn the underlying chemistry through the construction of adversarial examples. In these examples we only change the parts that are chemically important, but not according to the model. This way the model can be fooled into incorrect predictions if the interpretation is correct, or the interpretation can be falsified if the model is able to predict the correct product. This is a crucial element of our method as any interpretation that cannot be falsified would be no more than speculation.

When talking about the size of an attribution we always compare it to the amount of attribution the group would get if the probability difference would be distributed uniformly across the input tokens. This serves as a way of normalizing the attributions by the size of the different substructures. We consider the parts of the reactant that get substantially higher attribution than expected to be 'important'.

### 5.2.4 Data Attribution

In cases when a model predicts something very unexpected to humans attributions to parts of the input can be difficult to make sense of. Sometimes it can be much more illustrative to attribute to data instead and see a couple of example inputs that the model finds similar, which can reveal biases that the model has learnt.

To successfully attribute to data, we must understand how 'similar' two input datapoints are according to the model by defining a similarity metric. For the Molecular Transformer which has an encoder-decoder architecture we use the output of the encoder layers as a basis

for comparing data points. The challenge lies in the fact that these encoder hidden states have a non-fixed length $256 \times N$ where $N$ is the length of the input sequence. To overcome this we average these vectors over the sequence dimension $N$, obtaining a representation of fixed size 256. We hypothesized that averaging can work because of the relatively large dimensionality and hence sparsity of the embedding space, allowing the averaged vector to retain most of the information about the structure, reagents and reactivity.

We generated these averaged encoder state vectors for all of the reactions in the training sets. When a new example input is given it is passed through the Transformer encoder and the average hidden state vector of it is calculated. The similarity score of this vector to the training set vectors is calculated by

$$score = \frac{1}{1+D} \tag{5.3}$$

where $D$ is the Euclidean distance between the vectors. This can be implemented in a vectorized way resulting in very quick computation even in the case of dataset sizes like Pistachio made up of over 2 million training examples. The top-$n$ most similar reactions are returned where $n$ is defined by the user. A similar approach is used in [3] to measure the model-learned similarity between molecules for a graph neural network trained on toxicity prediction. These similarities are also used as evidence for judging the reliability of the model predictions, but only for unseen molecules not in the training set. In this work we go beyond assessing reliability into explaining the failures of the model by explicitly examining the training data itself to reveal hidden biases.

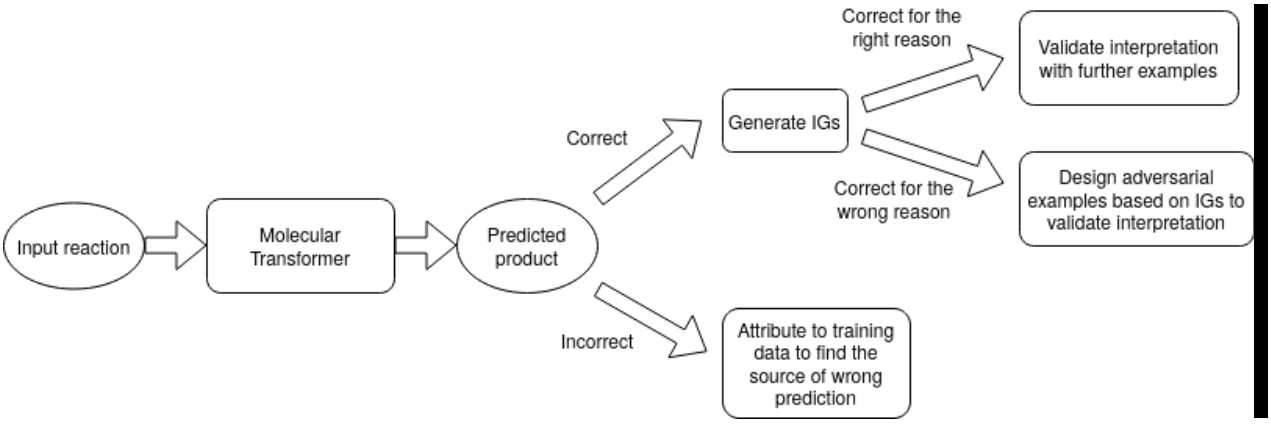



Fig. 5.1 An overview of the workflow for interpreting the Molecular Transformer.

## 5.3 Results

To interpret the predictions of the Molecular Transformer we follow an analysis workflow (Fig 5.1) and examine a number of reaction types that are commonly used in synthetic organic chemistry using both input and data attribution techniques.

### 5.3.1 Diels-Alder reactions

The Diels-Alder reactions transform a conjugated diene and an alkene (called dienophile) to a six membered ring with a double bond [13]. A typical example is shown in Fig 5.2. Diels-Alder reactions are regioselective meaning that the methoxy and nitrile group can be opposite or one carbon apart on the ring formed as shown in Fig 5.2. The major product is the one marked TRUE on the figure because of more favourable HOMO-LUMO interactions. Due to the large number of possible products and complicated rules determining the major product the Diels-Alder reaction can serve as a challenging test for any reaction prediction model.

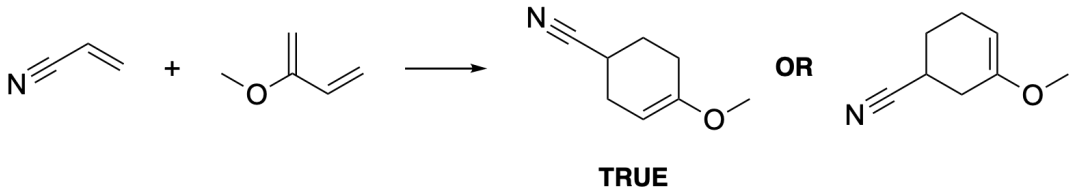



Fig. 5.2 A typical example of a Diels-Alder reaction with challenging selectivity.

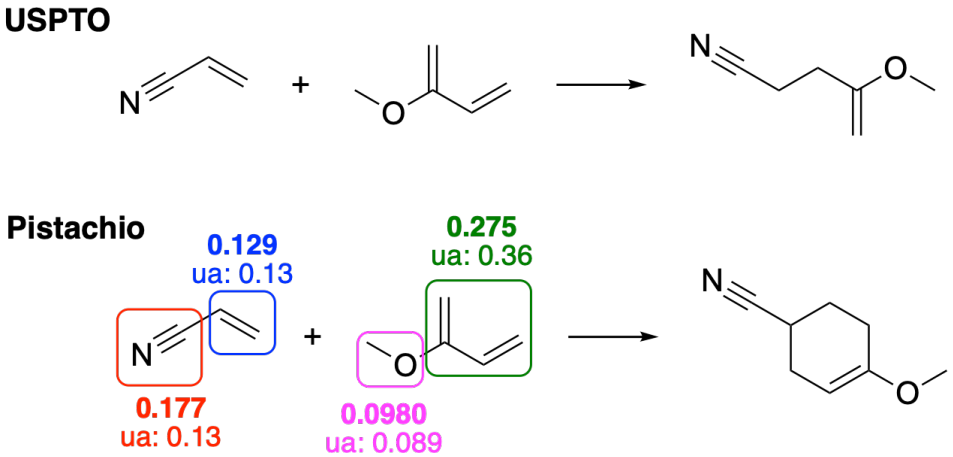



Fig. 5.3 The USPTO transformer makes a completely incorrect prediction, while the Pistachio model correctly predicts the product and recognises the importance of the nitrile group. For the Pistachio model the IG attributions are shown together with the corresponding uniform attribution (ua) values.

Fig 5.3 shows the Top-1 prediction of the USPTO and Pistachio models. The USPTO model does not seem to recognize the Diels-Alder reaction and gets the prediction wrong, indicating its own uncertainty by assigning a very low score of 0.300 to the prediction. To find the reason for the wrong prediction we attributed to training data (Fig 5.4). The first reaction seems to be an erroneous datapoint whereas the other two are different carbon-carbon bond formation reactions. This indicates that either the model has not learnt to recognize Diels-Alder reactions or the dataset did not contain any of them.

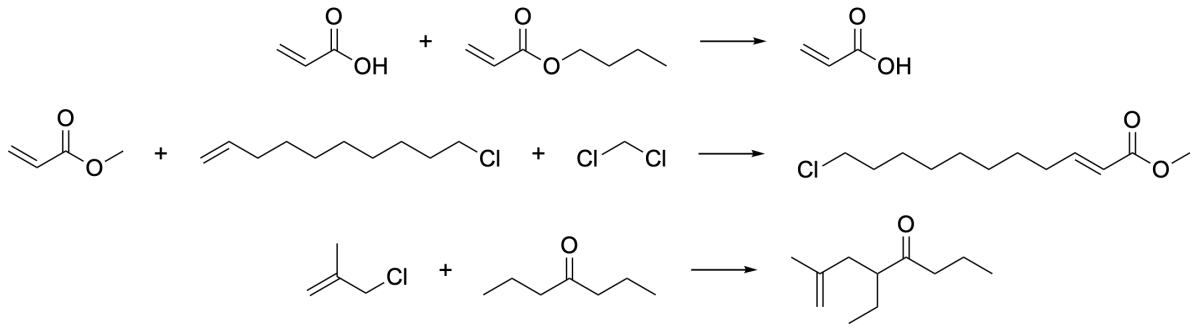



Fig. 5.4 Attribution to the USPTO training data shows that the USPTO transformer either completely fails to recognize Diels-Alder reactions or that no Diels-Alder reactions exist in the dataset.

To check this we devised a simple reaction template of Diels-Alder reactions and ran a template matching algorithm on the training data. We validated the template by ensuring it was able to identify the reactions where a diene and a double bond participate in cyclo-addition, which made up 80% of the ∼2.3k labelled Diels-Alder reactions in Pistachio. This template matched only 7 reactions in the USPTO dataset confirming that it contains very few instances of this type of reaction. Furthermore this suggests that the model was not able to generalize across chemical space to infer this reactivity from different types of reactions. This level of generalization could only be expected from physics based models that have direct access to quantum mechanical information driving the reactions.

For the Pistachio model the Top-1 prediction is correct, as shown in Fig 5.3, and it has a confidence score of 0.819 indicating that it is fairly certain in the prediction. We also generated the IGs for this reaction to see if the selectivity is caused by the relevant nitrile and methoxy groups. The probability difference between the correct major and the minor products was 0.77 and it was distributed on the compounds as shown in Fig 5.3 alongside the uniform attribution values. It can be seen that the nitrile group received a higher than uniform attribution indicating that the model recognises its importance. The same cannot be said unambiguously about the methoxy group whose attribution is only slightly more than the corresponding uniform value. Based on this example we can conclude that the model has

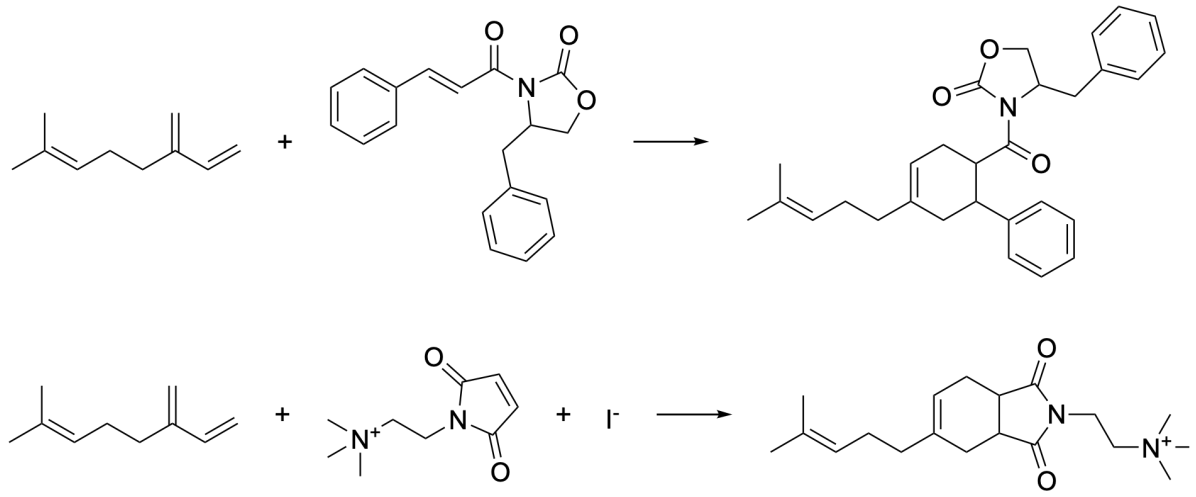


Fig. 5.5 Further test reactions correctly predicted by the Pistachio model, validate that Pistachio correctly understands Diels-Alder reactions. [34, 76]

learnt to recognize Diels-Alder reactions, and the IGs point towards the fact that it has learnt the regioselectivity causes too. To confirm this a couple of reactions taken from publications were tested (Fig 5.5) and the Pistachio transformer is able to predict the correct products.

### 5.3.2 Friedel-Crafts acylation reactions

Friedel-Crafts acylation reactions are an example of electrophilic aromatic substitution reactions [13, 23] where a hydrogen on an aromatic ring is substituted to an acyl group. In the case of a benzene ring with a single substituent on it there are three different hydrogen positions where this substitution can happen. The electronic and steric character of the substituent on the ring will determine the selectivity of these reactions. An example of a selective Friedel-Crafts reaction is shown in Fig 5.6.

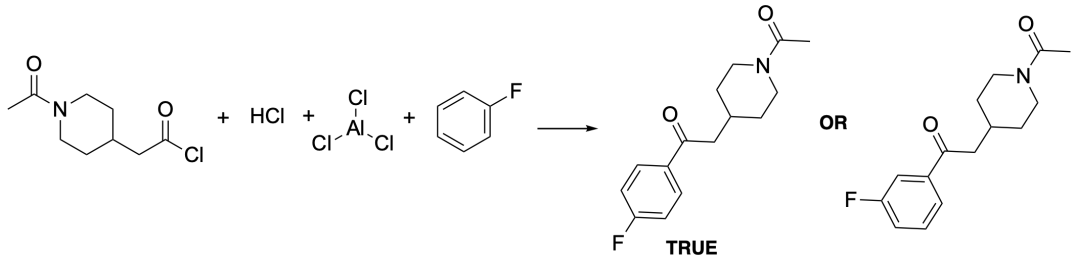



Fig. 5.6 Friedel-Crafts acylation reaction taken from the USPTO training set showing para selectivity.

The predictions of the two models are shown in Fig 5.7. Both models predict the para selectivity correctly with confidence scores close to 1.0. When inspecting the IG attributions

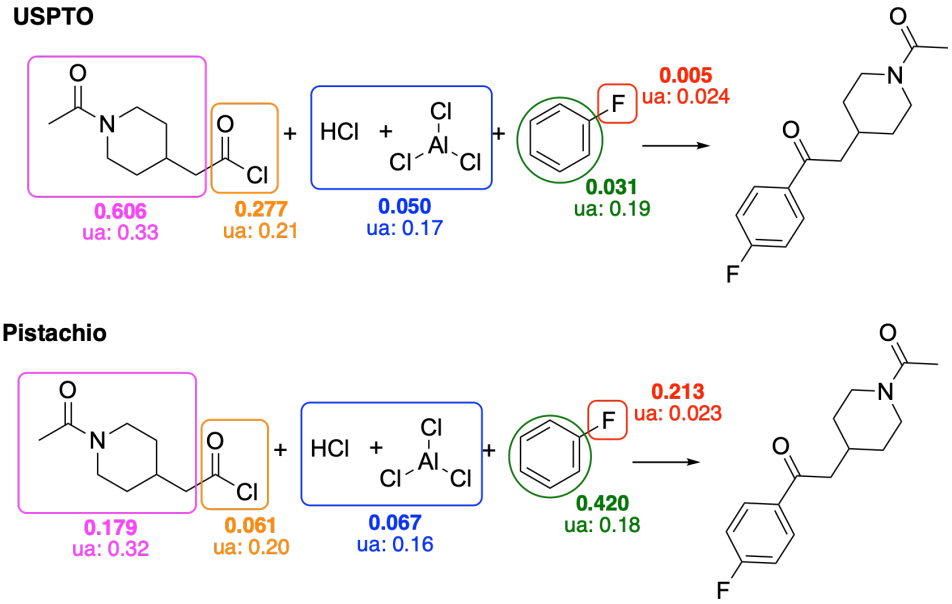



Fig. 5.7 Both models predict the correct Friedel-Crafts acylation product but only the Pistachio model recognizes the importance of the -F atom in determining selectivity. The Integrated Gradients attributions are also shown along with the uniform attribution (ua) values.

it can be seen that the USPTO model puts a very small weight on the Fluorine, only a fifth of the uniform attribution value. There is a large attribution given to the reagent though which does not affect the selectivity of this reaction at all. The attributions indicate that the USPTO transformer has not learnt the importance of F as the cause of para selectivity in these reactions. On the other hand the Pistachio transformer assigns a very high attribution value to F, suggesting that it has recognized the reason for the selectivity.

Guided by the attributions we designed a number of adversarial examples where we have changed only the Fluorine part of the reactant-reagent input. This choice was motivated by the fact that according to the USPTO model the selectivity was driven by the reagents instead of the substituent on the benzene ring. If our interpretation is correct the model should keep predicting the para product even if a meta directing substituent is attached to the ring. The predictions of the models and the IG attributions of the meta directing groups are shown in Fig 5.8.

It can be seen that both transformer models fail in terms of predicting the meta directing effect of the substituents on the rings. In this case negative attributions favour the meta and positive the para product. There seems to be no correlation between the attribution values and the directing effect of the substituents, and even the Pistachio transformer is struggling with identifying the chemically important parts of the input.

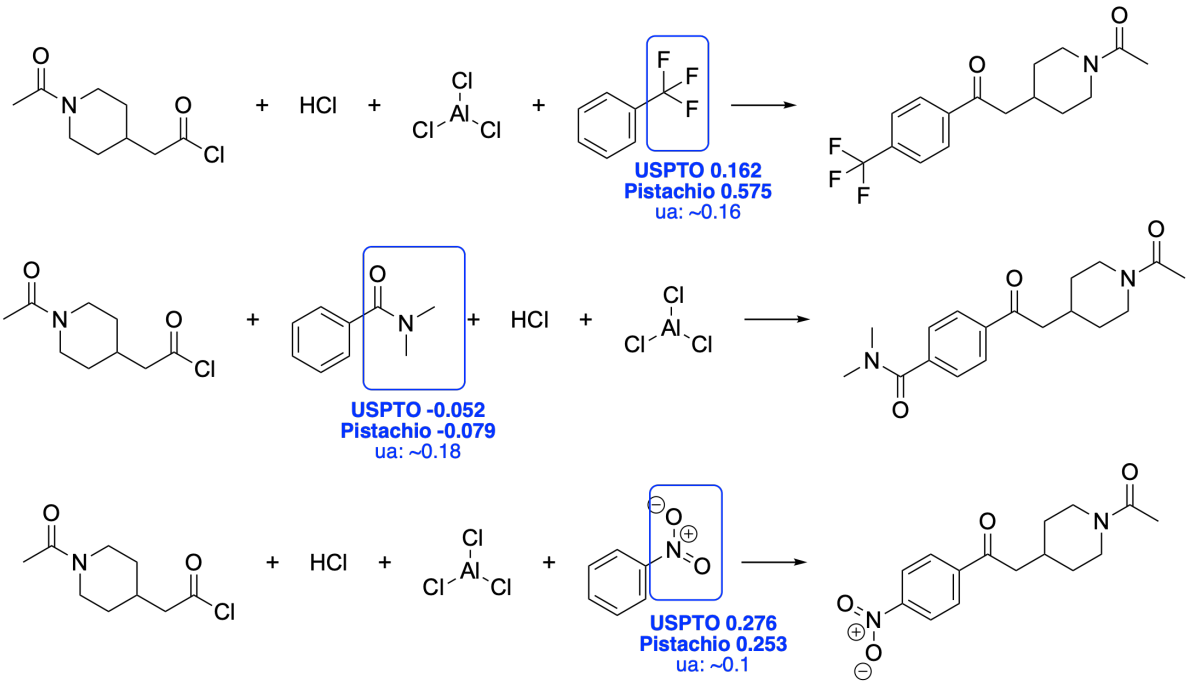



Fig. 5.8 Adversarial examples designed using Fig 5.7 reveal that both models can easily fail to predict the correct meta product. The uniform attribution (ua) values for the IGs are also shown.

A suggestive observation is that in the second example the attributions on the meta directing group are negative, meaning that according to the models the amide group (correctly) favours the formation of the meta product. This agrees with chemical principles, but the model is still predicting the para to be the major product. We hypothesized that this might be due to biases in the training data, because if there are many more para substitution reactions than meta, the model could become biased towards predicting para substitutions even in the presence of meta directing groups.

To check if this hypothesis was correct we counted the number of ortho, meta and para Friedel-Crafts acylations in the training dataset using reaction templates. There was a large number of reactions matching multiple templates because often the benzene rings had multiple substituents on them. The results are summarized on Fig 5.9.

The overall number of meta substitutions was 680 and 896 for the USPTO and Pistachio datasets respectively compared to the 952 and 1534 para substitutions. However, these numbers do not reveal the true extent of the bias as in our test case the benzene ring was only singly substituted. The number of reactions where there is only a single meta directing group on the ring is only 7 and 23 for the two datasets, which are extremely small numbers

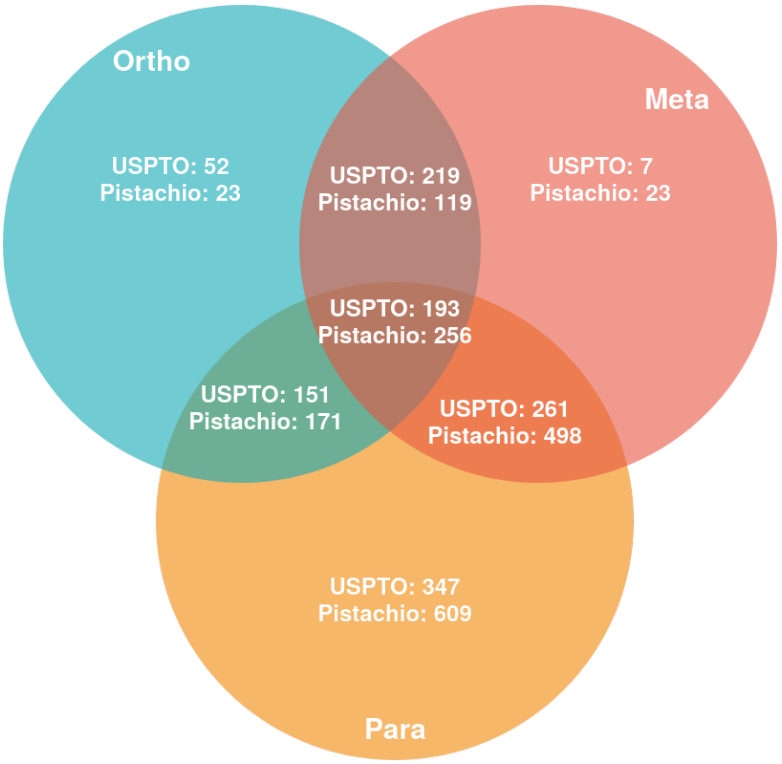



Fig. 5.9 Counting the number Friedel-Crafts acylation reactions in the training sets with ortho, meta and para selectivities reveals an alarming bias – the number of para reactions far outweigh those of meta or ortho reactions.

compared to those for a single para directing substituent which are 347 and 609, a ratio of 25-50 times.

This may result in the models not being able to learn meta directing substitution reactions because it can already achieve very high ( 98%) accuracy on the training set by always predicting the para product. The inclusion of further meta substitution reactions could considerably increase the models performance in real tasks. To confirm that the model would be able to learn the selectivities if the dataset was not biased, we probe the model with a synthetic dataset of para and meta Friedel-Crafts reactions in Sec. 5.3.4.

### 5.3.3 Selective reduction of aldehydes and ketones

Reduction of esters and aldehydes follow very well defined selectivity that is determined by the reducing agent. It is possible to reduce selectively an aldehyde or a ketone to alcohol in the presence of an ester. In this example the reduction of aldehydes using sodium-borohydride is examined [13]. If the Na is replaced by Li the reduction stops being selective to aldehydes and esters get reduced as well. The question is whether the models were able to learn the role of the cations in driving this subtle selectivity. An example reaction containing this selectivity is shown in Fig 5.10.

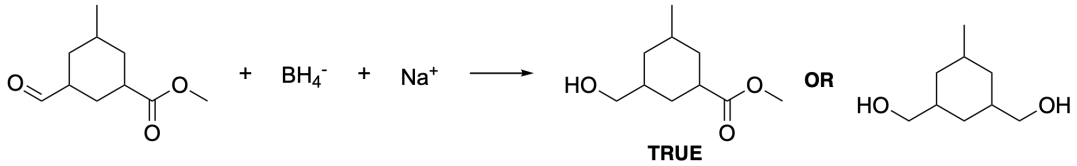



Fig. 5.10 An example of a reaction showing how $NaBH_4$ reduces the aldehydes selectively in the presence of an ester.

Both models are able to predict the product correctly with very high confidence (score > 0.95). In this case it is not immediately obvious what an interpretable attribution would be. One could argue that the selectivity is caused by $Na^+$ because if we swap it to $Li^+$ the other product would become the true product. The IG attributions on the `[Na+]` token are 0.013 and 0.017 for the USPTO and Pistachio models respectively, less than what the uniform attribution would be. This suggests that the models have not identified the importance of $Na^+$ ion.

To better understand the reliability of the predictions we attribute the reaction to the training data. For both models the most similar reactions (Fig 5.11) are $BH_4$ reductions of molecules containing both a ketone and an ester group. These examples suggest that both models have learnt this selectivity correctly, but they are not helping in understanding the role of the $Na^+$ion in the reaction.

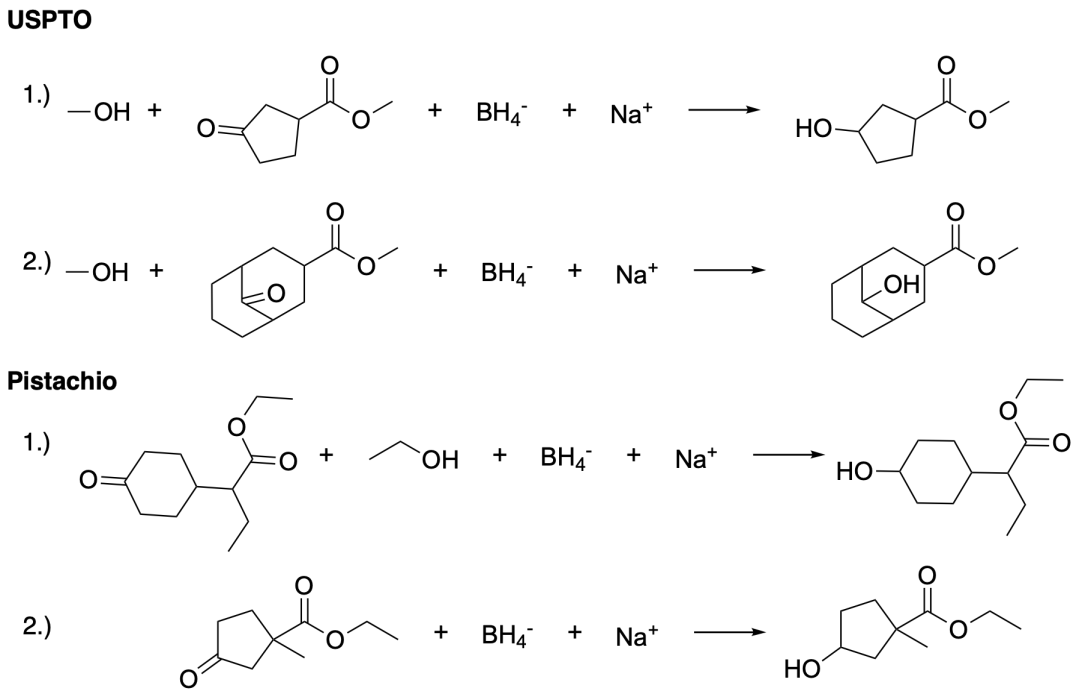



Fig. 5.11 Data attribution to the reaction in Fig 5.10 suggests that both models have correctly learnt the selectivity of reduction.

To investigate further if the models have learnt the importance of the cation in these reactions we designed an adversarial example where the only difference from the reaction on Fig 5.10 was the replacement of $Na^+$ with $Li^+$. From Fig 5.12 we see that the USPTO transformer keeps predicting the aldehyde being selectively reduced whereas the Pistachio model recognizes the change and predicts the correct product with both groups reduced.

To prove that one model was able to identify the importance of the cation and the other was not the IG attributions were generated and are shown in Fig 5.12. Here positive attributions favour the selective reduction, and the negative attributions favour the correct product. It is immediately obvious from the attributions that the USPTO model did not take into account the $Li^+$ ion as it was given an attribution score that is an order of magnitude smaller than the uniform attribution value.

For the Pistachio model the probability score difference between the products was -0.21 and there was a lot of variation in attribution across different parts of the structure. The `[Li+]` token was given a very large negative attribution meaning that the model was strongly relying on it when making the correct prediction. Overall comparing the attributions it can be concluded that the USPTO model did not learn the chemistry of $LiBH_4$, but the Pistachio one did. This can be due to the fact that the Pistachio model has seen more than 6 times more examples with this reagent.

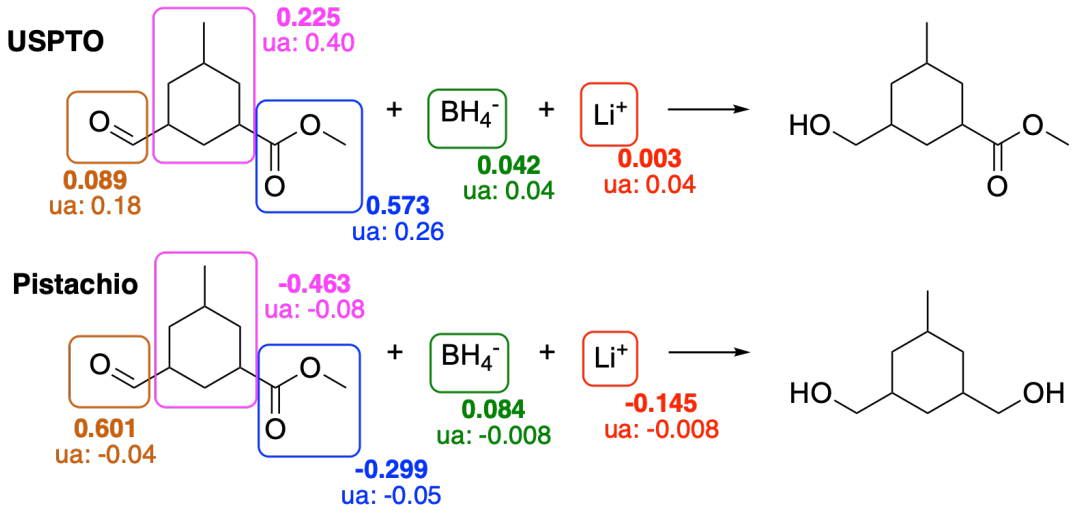



Fig. 5.12 Adversarial example for the borohydride reduction in Fig 5.10 where the $Na^+$ ion was replaced by $Li^+$ ion. The USPTO model continues predicting the selective reduction of the aldehyde wrongly whereas the Pistachio model predicts the correct product – the understanding of the models is reflecting in the IG attribution on the $Li^+$ ion.

### 5.3.4 Exploring the model with artificial data

One of the limitations of learning chemistry from patented and published reactions is that these reactions were designed by trained chemists who avoid transformations that have non-obvious selectivities. This makes it difficult for the models to infer the order of reactivity of functional groups. A further point is the effect of bias in the datasets on the models performance. To better understand these effect with full control over the experimental parameters, we have designed two artificial tasks where the training reaction data is generated using explicit SMARTS templates.

In the first experiment we test whether the transformer model is able to learn selective chemistry if given enough data. We assembled a synthetic dataset of 90 000 reduction reactions. Carbon scaffolds were randomly selected from the ZINC database of drug-like molecules [38]. To each scaffold we added an aldehyde group, an ester group, or both. Finally the 90 000 reactions, summarized in Table 5.1, were obtained by applying one of three reduction templates shown in Fig 5.13 to them.

When training the transformer on this dataset we found that the only limiting factor in the performance of the model was its ability to reconstruct the sometimes tricky backbones of the molecules from the ZINC dataset. The selective chemistry was learnt (99% top-1 accuracy) by the model in less than 10 000 steps. This shows that given sufficient data the model is able to learn selective transformations. To investigate how these findings are reflected in

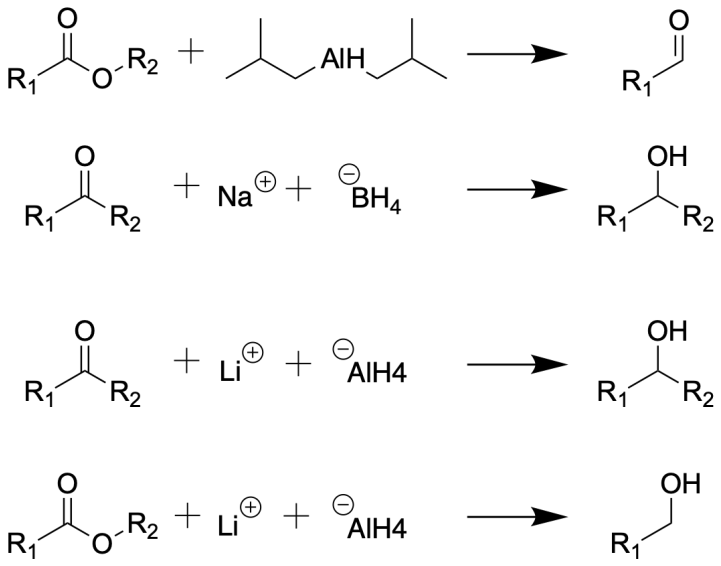



Fig. 5.13 Three uniquely selective reduction templates are chosen to challenge the transformer's ability to learn selectivities if given enough data.

Table 5.1 Number of reactions in the synthetic reduction datasets

| Subset | Aldehyde | Ester | Both |
|---|---|---|---|
| $NaBH_4$ | 20 000 | 0 | 10 000 |
| DIBAL | 0 | 20 000 | 10 000 |
| $LiAlH_4$ | 10 000 | 10 000 | 10 000 |

the IG attributions and verify that they are able to capture these empirical observations we generated the attributions for the reaction shown in Fig 5.14.

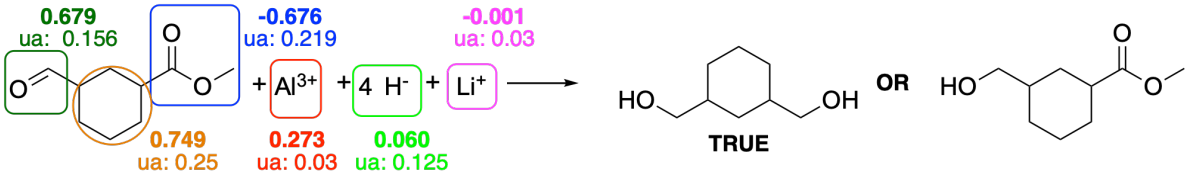



Fig. 5.14 The model trained on artificial reduction data correctly attributes importance to the $Al^{3+}$ ion.

From the attributions we see that the model rightly gives high attribution to the $Al^{3+}$ ion and small attribution to $Li^{+}$. This suggests that the model has learnt that it should only attend to one of the two tokens because they always appear together in the reactions. The $H^{-}$ ions get a low attribution as expected. The attributions also verify that the model is finding the backbone part of the input very important, as reconstruction of the backbone is the most

challenging aspect of making a correct prediction for this dataset of diverse backbones but restricted chemistry.

In the second more challenging synthetic experiment we investigated the models ability to learn the selectivity of aromatic electrophilic substitution reactions. By constructing balanced and biased datasets of Friedel-Crafts acylation reactions we hope to recover the observed behaviour in Sec. 5.3.2. In the balanced and biased datasets we chose 10 para and 10 meta directing substituents which were placed on a benzene ring. For the last 'super-biased' dataset, we only included three of the 10 meta directing substituents. These made the initial set of 20 reactant molecules which were reacted with a set of acyl chlorides generated by enumerating all possible straight carbon chains up to 8 carbons with a maximum of one double bond. This way we obtained 310 acyl chlorides that were reacted with the substituted benzenes to yield the meta or the para product. We use small carbon chains rather than the ZINC scaffolds to facilitate the learning of the backbones compared to learning the chemistry, which is partly why the dataset size was reduced as well and SMILES augmentation was applied. A summary of the training sets is shown in Table 5.2.

Table 5.2 Number of reactions in the synthetic Friedel-Crafts training datasets

| | **Meta** | **Para** |
|---|---|---|
| Balanced | 3100 | 3100 |
| Biased | 310 | 2790 |
| Super-Biased | 30 | 3000 |

The USPTO and Pistachio transformers were trained for ∼300 and ∼100 epochs respectively, so this is the regime we wanted to investigate. We trained 10 transformer models on each of the datasets and saved checkpoints regularly. We created a test set using three meta directing and three para directing substituted benzenes combined with acyl chlorides not in the training sets, resulting in a balanced test set made up of 177 meta and 177 para substitution reactions. Using SMARTS template matching, we tested what proportion of the model predictions (with valid SMILES) are meta and para as a function of the number of epochs for different dataset biases. The results are shown in Fig 5.15.

We see that the balanced dataset converges quickly close to the correct ratio of 1:1 between meta and para predictions. On the other hand the bias in the training set is reflected in the predictions of the other two models, in the case of the super-biased model to the extent that it does not predict any meta products. This is particularly revealing because that is the training set whose ratio is closest to that found in the USPTO and Pistachio datasets. This serves as empirical proof that the observed failure to predict the meta substituent in Sec. 5.3.2 is the result of biases in the dataset.

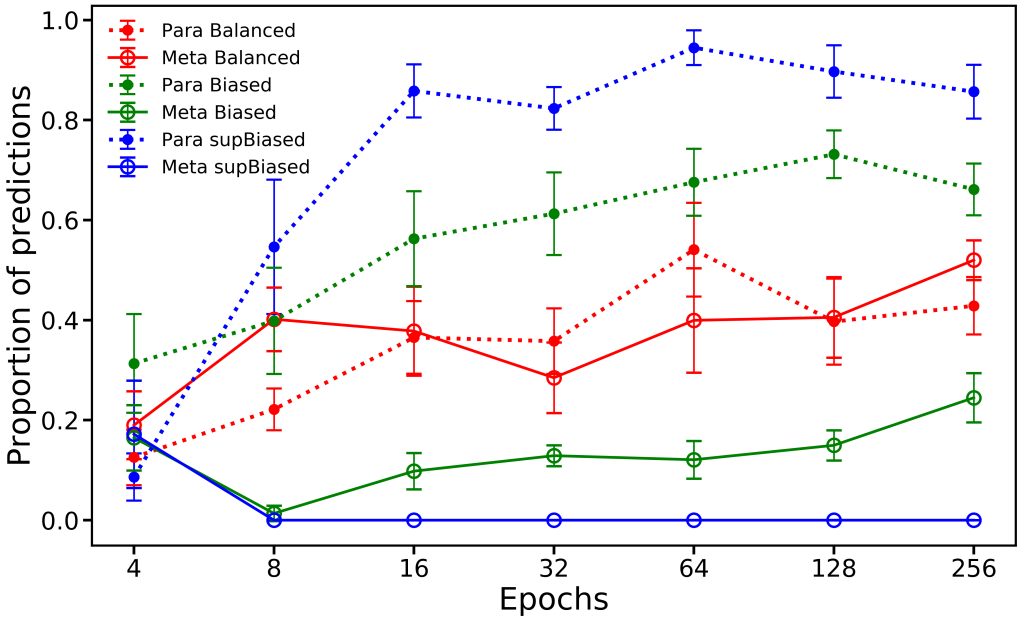



Fig. 5.15 Dataset bias is reflected in the model predictions. The figure shows the proportion of para (solid line) and meta (dashed line) predictions on a balanced test set as a function of the number of training epochs for different biased training sets.

Finally we ran the models to convergence to see if eventually they are able to predict the correct structures. After ∼4 000 epochs the ratio of meta to para was exactly 1:1 for the balanced dataset and about 3:5 on both the biased and super-biased datasets. This shows that by training longer the effect of dataset bias can be mitigated, but it cannot be removed altogether.

### 5.3.5 Outlook

Chemical reaction prediction models have undergone a revolution driven by innovations in the field of machine learning. This large increase in accuracy came at the expense of interpretability as expert crafted rules and reaction mechanisms gave way to black-box deep learning models.

The predictions of machine learning models depend on two essential components. One of them is the training data which acts as an upper limit to the performance of the model. Any machine learning model can be only as good as the data it was trained on. The other ingredient is the input that is processed and turned into the prediction. We have developed two robust methods for interpreting and testing reaction prediction models focusing on each of these two ingredients, and applied them to analyse the Molecular Transformer which is the current state-of-the-art model.

The first method builds on the Integrated Gradients method [73] for attributing the prediction of neural network models to parts of the input. This method has been used to identify which parts of the inputs to the model are important when predicting typical selective chemical reactions. It has been found that often the model does not identify chemically important substructures, demonstrated by the design of adversarial examples based on the attributions which fool the model.

The other method we developed attributes the predictions of the model to training data. We averaged the vector outputs from the last encoder layer of the Molecular Transformer to define a similarity metric between different reactions, as understood by the model. Attributing back to training data serves multiple purposes in the case of reaction prediction. It can either support or invalidate a prediction by telling the user which are the most similar training reactions according to the model. Furthermore this can be used to identify unknown trends or biases in the dataset or sometimes even to identify erroneous training examples.

Using evidence from these attribution methods, we hypothesized that many of the erroneous predictions of the Molecular Transformer model stem from data biases. We have validated this hypothesis by designing an artificial dataset of Friedel-Crafts acylation reactions where we could show how biases in the dataset manifest in the predictions of the model. In addition, we observe that the model trained on the Pistachio dataset had in general better

predictions and much better calibrated uncertainty scores, in spite of the fact that this model only achieved 76% test-set accuracy. This suggests that Pistachio is not as biased as USPTO, and hints that the addition of training data can substantially improve model performance.

From these results we believe that for reaction prediction, the Top-$N$ accuracy from testing on randomly chosen held-out test sets do not provide an adequate measure of the models true performance and generalization ability. We believe that this is partly results from the fact that publications and patents often contain reaction carried out on a series of analogous reactants. Therefore there is a high chance that essentially identical reactions end up in the training and test sets. A more honest measurement of the model's true generalizability could be realized by only including reactions in the test-set whose products have a low similarity to the products in the training set. The exploration of this idea is the subject of further work.

Overall, from this work we believe that improvements to the training data can be just as impactful as improving the machine learning models themselves. By demonstrating the power of interpretability methods when rigorously applied to scientific questions, we have shown that these methods can be useful beyond just giving explanations of predictions by exposing dataset biases. Applying our approach for data and input interpretation beyond chemical reaction prediction to other fields will likely be equally constructive, illuminating the path to improved training data and hence improved artificial intelligence models.

# Chapter 6

# Augmenting Nanomolar High-Throughput Screening with Machine Learning for Lead Optimisation

Typically, testing a drug candidate involves obtaining a pure sample of the molecule, and then mixing it in solution with the protein target under study to measure its bioactivity via an assay. While necessary for maximum accuracy, compound purification can be time-consuming and costly, particularly for chiral molecules. In collaboration with the London Lab at The Weizmann Institute of Science, we investigated whether we needed compound purification at all for training machine learning bioactivity models by using non-purified compound assays. Focusing on a particular scaffold synthesised with an amide coupling as the final step, we added the acid and amine reactants directly in solution with the protein to obtain an assay reading from the crude reaction mixture. By skipping the purification step, this allowed us to quickly screen a library of $< 300 >$ amines with the same acid in high-throughput which we used to train RF and GP models. Leave-one-out validation on the training data correctly identified false negatives, and a prospective virtual screen of EnamineREAL with the trained models returned top hits with similar potency and better pharmacokinetic properties.

Machine learning (ML) has seen great advances over the past two decades in drug discovery and development, from protein-ligand interactions to novel scaffold generation to virtual screening of pharmacokinetic properties and toxicity.1-5 Generally, ML algorithms are tasked with prediction of specific target(s) with the endeavor that the faster predictions in silico will lead to faster pharmaceutical development. Hastening drug development is a key factor in reducing the overall cost of pharmaceutics, where it can reach up to $2 billion to bring a compound to market.1 However, one overlooked area in this development cycle is ML applied as a filtering protocol for initial lead discovery, despite reports that ML

methods often implicitly identify false positives and false negatives.6,7 Crude activity screening (assuming some level of introduction by Mihajlo is given previously) is a logical area to apply such techniques as noise and false hits/misses play a substantial role in obscuring valuable data. We hypothesized that combining two robust ML methodologies, Gaussian Processes (GPs) and Random Forests (RFs), could be used to identify hidden gems (false negatives) and overlooked molecules (low activity positives). Both GPs and RFs have been utilized in numerous chemoinformatics tasks, with several precedents in pharmaceutics development, making both ideal for predicting activity of novel compounds.8-12 Given the difference in approach to modeling for GPs and RFs, it was hypothesized that a combination of the two would lead to a highly robust framework; a compound predicted to have low activity from both a GP and a RF is likely to be inactive and likewise a compound with high predicted activity from both the GP and RF is likely potent.

Figure 1 - schematic

Thus, we separately trained a GP and RF on the crude inhibition data, identifying 5 compounds which had predicted activity from both the GP and RF but no crude activity. We suspected that these were false negatives and re-synthesized, purified, and re-tested them with full dose-response curves to obtain IC50 inhibition values. This revealed that one of them was active with IC50 = 0.113$\mu$M (ASAP-0000204). (needs a nice sentence to round it off).

Figure 2 - leave-one-out

Looking forward, we test the ability of the trained models to extrapolate to novel compounds by prospectively screening an external library of amides. We virtually enumerate primary and secondary amine building blocks from Enamine with the same carboxylic acid substructure from the crude activity screening. This results in a library of 62,800 amides which were scored by the trained GP and RF models, and we select the top 20 compounds with high predicted activity for both the GP and RF for synthesis and assaying to obtain IC50 values. Gratifyingly, the top 2 ML compounds showed promising average IC50 values of 0.0525$\mu$M (ASAP-0000169) and 0.075$\mu$M (ASAP-0000211), respectively. The top 2 most potent molecules based off of the crude inhibition values were compounds that, whilst active at IC50 = 0.034$\mu$M (ASAP-0000221) and IC50 = 0.064$\mu$M (ASAP-0000164), contained the toxic benzene-1,4-diamine motif that is generally avoided.13 The top 2 compounds without the aforementioned motif derived from only crude inhibition values had similar pure compound IC50 values to our framework's identified compounds, 0.046$\mu$M (ASAP-0000155) and 0.064$\mu$M (ASAP-0000225), respectively. This result highlights ML's ability to identify promising yet overlooked scaffolds without compromising potency.

Figure 3 - top 2 crude, false negative, overall strip plot?

References:

1 Selvaraj, C., Chandra, I. & Singh, S. K. Artificial intelligence and machine learning approaches for drug design: challenges and opportunities for the pharmaceutical industries. Molecular Diversity 26, 1893-1913, doi:10.1007/s11030-021-10326-z (2022). 2 Göller, A. H. et al. Bayer's in silico ADMET platform: A journey of machine learning over the past two decades. Drug discovery today 25, 1702-1709 (2020). 3 Lavecchia, A. Machine-learning approaches in drug discovery: methods and applications. Drug Discovery Today 20, 318-331, doi:https://doi.org/10.1016/j.drudis.2014.10.012 (2015). 4 Lipinski, C. F., Maltarollo, V. G., Oliveira, P. R., Da Silva, A. B. & Honorio, K. M. Advances and perspectives in applying deep learning for drug design and discovery. Frontiers in Robotics and AI 6, 108 (2019). 5 Vamathevan, J. et al. Applications of machine learning in drug discovery and development. Nature Reviews Drug Discovery 18, 463-477, doi:10.1038/s41573-019-0024-5 (2019). 6 Ardila, D. et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature medicine 25, 954-961 (2019). 7 Ryu, J. Y., Lee, M. Y., Lee, J. H., Lee, B. H. & Oh, K.-S. DeepHIT: a deep learning framework for prediction of hERG-induced cardiotoxicity. Bioinformatics 36, 3049-3055 (2020). 8 Jiménez-Luna, J., Grisoni, F. & Schneider, G. Drug discovery with explainable artificial intelligence. Nature Machine Intelligence 2, 573-584, doi:10.1038/s42256-020-00236-4 (2020). 9 Reker, D. & Schneider, G. Active-learning strategies in computer-assisted drug discovery. Drug discovery today 20, 458-465 (2015). 10 Kapsiani, S. & Howlin, B. J. Random forest classification for predicting lifespan-extending chemical compounds. Scientific Reports 11, 13812, doi:10.1038/s41598-021-93070-6 (2021). 11 Svetnik, V. et al. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. Journal of Chemical Information and Computer Sciences 43, 1947-1958, doi:10.1021/ci034160g (2003). 12 Kang, B., Seok, C. & Lee, J. Prediction of Molecular Electronic Transitions Using Random Forests. Journal of Chemical Information and Modeling 60, 5984-5994, doi:10.1021/acs.jcim.0c00698 (2020). 13 Kumar, M. S., Tamilarasan, R. & Sreekanth, A. 4-Salicylideneamino-3-methyl-1,2,4-triazole-5-thione as a sensor for aniline recognition. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 79, 370-375, doi:https://doi.org/10.1016/j.saa.2011.03.030 (2011).

# Chapter 7

# Future Work

## 7.1 Short-Term: Continutation of ongoing work

Given the relative success of work so far, the immediate plan is to continue ongoing research to completion. After developing a more appropriate benchmark dataset for evaluating the performance of reaction prediction models, the intention is to write up the results of chapter 5 within the next month, potentially following up with an attempt to 'fix' the model which may also be a contribution to the field of NLP in particular regarding the modelling of long sequences. The investigation into SOAP descriptors for QSAR should hopefully be concluded shortly after resubmitting the results to JMedChem. If the drug candidates proposed by the Siamese GNN prove potent, then there is a strong incentive to refine and retrospectively validate the model on historical data and publish the methodology.

In addition, the COVID Moonshot project will likely continue for another ∼8 months and I will continue my participation of the project given the obvious urgency of the pandemic. No doubt this enterprise will remain a fruitful source of interesting problems with real experimental data, which will hopefully lead to innovative solutions. Research will probably be on continuing optimisation of existing molecular series, or searching for alternative backup series while the most promising series undergo *in vivo* toxicity screening.

## 7.2 Long-Term: Investigation of new modalities

Although using artificial intelligence for optimising the small-molecule drug discovery process is undoubtedly a difficult task, there has been extensive interest from both academia and private industry with many breakthroughs having been made already. The field is maturing to the extent that ML algorithms are already beginning to become part of the commercial design-

make-test workflow, such that the remaining challenges are arguably merely an engineering problem.

The therapeutic space beyond small-molecules, however, is relatively unexplored territory for data-driven techniques. Applying machine learning to this area will likely present even more complex challenges, but the potential impact of developing new modalities far outweigh that of 'just' improving small-molecule QSAR modelling. While the potential areas of research are numerous, thus far two topics of interest have been identified:

- functionalisation of flexible biomolecules (glycans, peptides),
- understand/design self-assembled nanostructures for drug delivery.

Both of these topics involve structures that are larger and less well-understood by medicinal (bio)chemists because of their energetic/entropic complexity. This is a promising area where I could combine physics-based intuition for modelling interactions, as well as pragmatic ML for designing models that would be useful in a drug discovery setting.

While no ML in involved in the Test of design-make-test it is nonetheless vital to retain this part of the cycle for proper validation of ML drug discovery methods, for taking the important step from mere concept to real-life data-driven-drugs. Therefore there is an expectation that some form of experimental work will be carried out, likely alongside more experienced collaborators, in the latter stages ($2^{nd}$-$3^{rd}$ year) of the PhD irrespective of the ultimate direction of research.

# References

[Che] Chemspace: Lead-like compounds.

[2] Agarwal, S., Dugar, D., and Sengupta, S. (2010). Ranking chemical structures for drug discovery: a new machine learning approach. *Journal of chemical information and modeling*, 50(5):716–731.

[3] Allen, T. E. H., Wedlake, A. J., Gelžinytė, E., Gong, C., Goodman, J. M., Gutsell, S., and Russell, P. J. (2020). Neural network activation similarity: a new measure to assist decision making in chemical toxicology. *Chem. Sci.*, 11:7335–7348.

[4] Alon, A., Lyu, J., Braz, J. M., Tummino, T. A., Craik, V., O'Meara, M. J., Webb, C. M., Radchenko, D. S., Moroz, Y. S., Huang, X.-P., Liu, Y., Roth, B. L., Irwin, J. J., Basbaum, A. I., Shoichet, B. K., and Kruse, A. C. (2021a). Structures of the

$\sigma$

2 receptor enable docking for bioactive ligand discovery. *Nature*, 600(7890):759–764.

[5] Alon, A., Lyu, J., Braz, J. M., Tummino, T. A., Craik, V., O'Meara, M. J., Webb, C. M., Radchenko, D. S., Moroz, Y. S., Huang, X.-P., Liu, Y., Roth, B. L., Irwin, J. J., Basbaum, A. I., Shoichet, B. K., and Kruse, A. C. (2021b). Structures of the

$\sigma$

2 receptor enable docking for bioactive ligand discovery. *Nature*, 600(7890):759–764.

[6] Alvarez-Melis, D. and Jaakkola, T. S. (2018). On the Robustness of Interpretability Methods.

[7] Baell, J. B. and Holloway, G. A. (2010). New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry*, 53(7):2719–2740. PMID: 20131845.

[8] Botev, Z. I., Grotowski, J. F., and Kroese, D. P. (2010). Kernel density estimation via diffusion. *The Annals of Statistics*, 38(5):2916 – 2957.

[9] Brown, N., McKay, B., Gilardoni, F., and Gasteiger, J. (2004). A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *Journal of chemical information and computer sciences*, 44(3):1079–1087.

[10] Butina, D. (1999). Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39(4):747–750.

[11] Cannalire, R., Cerchia, C., Beccari, A. R., Di Leva, F. S., and Summa, V. (2020). Targeting sars-cov-2 proteases and polymerase for covid-19 treatment: State of the art and future opportunities. *Journal of medicinal chemistry*.

[12] Chodera, J., Lee, A. A., London, N., and von Delft, F. (2020). Crowdsourcing drug discovery for pandemics. *Nature Chemistry*, 12(7):581–581.

[13] Clayden, J., Greeves, N., and Warren, S. (2012). *Organic Chemistry*. Oxford University Press, 2nd edition.

[14] Coley, C. W., Eyke, N. S., and Jensen, K. F. (2019a). Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angewandte Chemie - International Edition*, pages 2–38.

[15] Coley, C. W., Eyke, N. S., and Jensen, K. F. (2019b). Autonomous Discovery in the Chemical Sciences Part II: Outlook. *Angewandte Chemie - International Edition*, pages 2–25.

[16] Coley, C. W., Green, W. H., and Jensen, K. F. (2018). Machine Learning in Computer-Aided Synthesis Planning. *Accounts of Chemical Research*, 51(5):1281–1289.

[17] Davis, B. J. and Roughley, S. D. (2017). Chapter eleven - fragment-based lead discovery. In Goodnow, R. A., editor, *Platform Technologies in Drug Discovery and Validation*, volume 50 of *Annual Reports in Medicinal Chemistry*, pages 371–439. Academic Press.

[18] de Souza Neto, L. R., Moreira-Filho, J. T., Neves, B. J., Maidana, R. L. B. R., Guimarães, A. C. R., Furnham, N., Andrade, C. H., and Silva, F. P. (2020). In silico strategies to support fragment-to-lead optimization in drug discovery. *Frontiers in Chemistry*, 8.

[19] Douangamath, A., Fearon, D., Gehrtz, P., Krojer, T., Lukacik, P., Owen, C. D., Resnick, E., Strain-Damerell, C., Aimon, A., Ábrányi-Balogh, P., Brandão-Neto, J., Carbery, A., Davison, G., Dias, A., Downes, T. D., Dunnett, L., Fairhead, M., Firth, J. D., Jones, S. P., Keeley, A., Keserü, G. M., Klein, H. F., Martin, M. P., Noble, M. E. M., O'Brien, P., Powell, A., Reddi, R. N., Skyner, R., Snee, M., Waring, M. J., Wild, C., London, N., von Delft, F., and Walsh, M. A. (2020). Crystallographic and electrophilic fragment screening of the sars-cov-2 main protease. *Nature Communications*, 11(1):5047.

[20] Duffy, N. P. (2010). Molecular property modeling using ranking. US Patent 7,702,467.

[Elsevier] Elsevier. Reaxys Database. https://www.reaxys.com/.

[22] Fink, E. A., Xu, J., Hübner, H., Braz, J. M., Seemann, P., Avet, C., Craik, V., Weikert, D., Schmidt, M. F., Webb, C. M., Tolmachova, N. A., Moroz, Y. S., Huang, X.-P., Kalyanaraman, C., Gahbauer, S., Chen, G., Liu, Z., Jacobson, M. P., Irwin, J. J., Bouvier, M., Du, Y., Shoichet, B. K., Basbaum, A. I., and Gmeiner, P. (2022). Structure-based discovery of nonopioid analgesics acting through the

-adrenergic receptor. *Science*, 377(6614):eabn7065.

[23] Friedel, C. and Crafts, J. (1877). Sur une nouvelle méthode générale de synthèse d'hydrocarbures, d'acétones, etc.

[24] Gao, W. and Coley, C. W. (2020). The synthesizability of molecules proposed by generative models. *Journal of Chemical Information and Modeling*, 60(12):5714–5723.

[25] Gironda-Martínez, A., Donckele, E. J., Samain, F., and Neri, D. (2021). Dna-encoded chemical libraries: A comprehensive review with succesful stories and future challenges. *ACS Pharmacology & Translational Science*, 4(4):1265–1279.

[26] Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276.

[27] Gorgulla, C., Boeszoermenyi, A., Wang, Z.-F., Fischer, P. D., Coote, P. W., Padmanabha Das, K. M., Malets, Y. S., Radchenko, D. S., Moroz, Y. S., Scott, D. A., Fackeldey, K., Hoffmann, M., Iavniuk, I., Wagner, G., and Arthanari, H. (2020). An open-source drug discovery platform enables ultra-large virtual screens. *Nature*, 580(7805):663–668.

[28] Hall, R. J., Murray, C. W., and Verdonk, M. L. (2017). The fragment network: A chemistry recommendation engine built using a graph database. *Journal of Medicinal Chemistry*, 60(14):6440–6450.

[29] Hann, M. M., Leach, A. R., and Harper, G. (2001). Molecular complexity and its impact on the probability of finding leads for drug discovery. *Journal of chemical information and computer sciences*, 41(3):856–864.

[30] Hartenfeller, M., Zettl, H., Walter, M., Rupp, M., Reisen, F., Proschak, E., Weggen, S., Stark, H., and Schneider, G. (2012). Dogs: reaction-driven de novo design of bioactive compounds. *PLoS Comput Biol*, 8(2):e1002380.

[31] Hermann, J. C., Chen, Y., Wartchow, C., Menke, J., Gao, L., Gleason, S. K., Haynes, N.-E., Scott, N., Petersen, A., Gabriel, S., Vu, B., George, K. M., Narayanan, A., Li, S. H., Qian, H., Beatini, N., Niu, L., and Gan, Q.-F. (2013). Metal impurities cause false positives in high-throughput screening campaigns. *ACS Medicinal Chemistry Letters*, 4(2):197–200.

[32] Howard, J. et al. (2018). fastai. https://github.com/fastai/fastai.

[33] Hughes, J. P., Rees, S., Kalindjian, S. B., and Philpott, K. L. (2011). Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249.

[34] Husinec, S., Savic, V., Simic, M., Tesevic, V., and Vidovic, D. (2011). Annulations of isoquinoline and $\beta$-carboline ring systems: Synthesis of 8-oxoprotoberberine derivatives. *Tetrahedron Letters*, 52:2733–2736.

[35] Ichihara, O., Barker, J., Law, R. J., and Whittaker, M. (2011). Compound design by fragment-linking. *Molecular Informatics*, 30(4):298–306.

[36] Imrie, F., Bradley, A. R., van der Schaar, M., and Deane, C. M. (2020). Deep generative models for 3d linker design. *Journal of Chemical Information and Modeling*, 60(4):1983–1995.

[37] Imrie, F., Hadfield, T. E., Bradley, A. R., and Deane, C. M. (2021). Deep generative design with 3d pharmacophoric constraints. *Chem. Sci.*, 12:14577–14589.

[38] Irwin, J. J. and Shoichet, B. K. (2005). ZINC - A free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, 45(1):177–182.

[39] Jin, W., Coley, C. W., Barzilay, R., and Jaakkola, T. (2017). Predicting organic reaction outcomes with weisfeiler-lehman network. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):2608–2617.

[40] Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., Zhang, B., Li, X., Zhang, L., Peng, C., et al. (2020). Structure of m pro from sars-cov-2 and discovery of its inhibitors. *Nature*, 582(7811):289–293.

[41] Kaserer, T., Beck, K. R., Akram, M., Odermatt, A., and Schuster, D. (2015). Pharmacophore models and pharmacophore-based virtual screening: Concepts and applications exemplified on hydroxysteroid dehydrogenases. *Molecules*, 20(12):22799–22832.

[42] Kearnes, S. (2021). Pursuing a prospective perspective. *Trends in Chemistry*, 3(2):77–79.

[43] Klein, G., Kim, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT.

[Kovacs et al.] Kovacs, D. P., McCorkindale, W., and Lee, A. A. Molecular Transformer Explainer. https://github.com/davkovacs/MTExplainer.git.

[Landrum] Landrum, G. RDKit: Open-source cheminformatics. http://www.rdkit.org.

[46] Lee, A. A., Yang, Q., Sresht, V., Bolgar, P., Hou, X., Klug-McLeod, J. L., Butler, C. R., et al. (2019). Molecular transformer unifies reaction prediction and retrosynthesis across pharma chemical space. *Chemical Communications*, 55(81):12152–12155.

[47] Liu, Y., Liang, C., Xin, L., Ren, X., Tian, L., Ju, X., Li, H., Wang, Y., Zhao, Q., Liu, H., et al. (2020). The development of coronavirus 3c-like protease (3clpro) inhibitors from 2010 to 2020. *European journal of medicinal chemistry*, page 112711.

[48] Llanos, M. A., Gantner, M. E., Rodriguez, S., Alberca, L. N., Bellera, C. L., Talevi, A., and Gavernet, L. (2021). Strengths and weaknesses of docking simulations in the sars-cov-2 era: the main protease (mpro) case study. *Journal of Chemical Information and Modeling*, 61(8):3758–3770. PMID: 34313128.

[49] Lowe, D. M. (2012). *Extraction of chemical structures and reactions from the literature*. Phd, University of Cambridge.

[50] Lyu, J., Wang, S., Balius, T. E., Singh, I., Levit, A., Moroz, Y. S., O'Meara, M. J., Che, T., Algaa, E., Tolmachova, K., Tolmachev, A. A., Shoichet, B. K., Roth, B. L., and Irwin, J. J. (2019a). Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229.

[51] Lyu, J., Wang, S., Balius, T. E., Singh, I., Levit, A., Moroz, Y. S., O'Meara, M. J., Che, T., Algaa, E., Tolmachova, K., Tolmachev, A. A., Shoichet, B. K., Roth, B. L., and Irwin, J. J. (2019b). Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229.

[52] Macip, G., Garcia-Segura, P., Mestres-Truyol, J., Saldivar-Espinoza, B., Ojeda-Montes, M. J., Gimeno, A., Cereto-Massagué, A., Garcia-Vallvé, S., and Pujadas, G. (2022). Haste makes waste: A critical review of docking-based virtual screening in drug repurposing for sars-cov-2 main protease (m-pro) inhibition. *Medicinal research reviews*, 42(2):744–769.

[53] Mayfield, J., Lowe, D., and Sayle, R. (2018). Pistachio 2.0.

[54] McCloskey, K., Taly, A., Monti, F., Brenner, M. P., and Colwell, L. J. (2019). Using attribution to decode binding mechanism in neural network models for chemistry. *Proceedings of the National Academy of Sciences of the United States of America*, 116(24):11624–11629.

[55] McCorkindale, W., Kovács, P., and Lee, A. (2020). Unpacking chemical reaction prediction models using integrated gradients. In *ML Interpretability for Scientific Discovery*, ICML'20.

[56] Morreale, F. E., Testa, A., Chaugule, V. K., Bortoluzzi, A., Ciulli, A., and Walden, H. (2017). Mind the metal: A fragment library-derived zinc impurity binds the e2 ubiquitin-conjugating enzyme ube2t and induces structural rearrangements. *Journal of Medicinal Chemistry*, 60(19):8183–8191.

[57] Morris, A., McCorkindale, W., Consortium, T. C. M., Drayman, N., Chodera, J. D., Tay, S., London, N., and Lee, A. A. (2021). Discovery of sars-cov-2 main protease inhibitors using a synthesis-directed de novo design model. *Chem. Commun.*, 57:5909–5912.

[58] Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., Oprea, T. I., Baskin, I. I., Varnek, A., Roitberg, A., et al. (2020). Qsar without borders. *Chemical Society Reviews*, 49(11):3525–3564.

[59] Owen, D. R., Allerton, C. M. N., Anderson, A. S., Aschenbrenner, L., Avery, M., Berritt, S., Boras, B., Cardin, R. D., Carlo, A., Coffman, K. J., Dantonio, A., Di, L., Eng, H., Ferre, R., Gajiwala, K. S., Gibson, S. A., Greasley, S. E., Hurst, B. L., Kadar, E. P., Kalgutkar, A. S., Lee, J. C., Lee, J., Liu, W., Mason, S. W., Noell, S., Novak, J. J., Obach, R. S., Ogilvie, K., Patel, N. C., Pettersson, M., Rai, D. K., Reese, M. R., Sammons, M. F., Sathish, J. G., Singh, R. S. P., Steppan, C. M., Stewart, A. E., Tuttle, J. B., Updyke, L., Verhoest, P. R., Wei, L., Yang, Q., and Zhu, Y. (2021). An oral sars-cov-2 m<sup>pro</sup> inhibitor clinical candidate for the treatment of covid-19. *Science*, 374(6575):1586–1593.

[60] Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065 – 1076.

[61] Patel, H., Bodkin, M. J., Chen, B., and Gillet, V. J. (2009). Knowledge-based approach to de novo design using reaction vectors. *Journal of chemical information and modeling*, 49(5):1163–1184.

[62] Pillaiyar, T., Manickam, M., Namasivayam, V., Hayashi, Y., and Jung, S.-H. (2016). An overview of severe acute respiratory syndrome–coronavirus (sars-cov) 3cl protease inhibitors: peptidomimetics and small molecule chemotherapy. *Journal of medicinal chemistry*, 59(14):6595–6628.

[PostEra Inc.] PostEra Inc. COVID moonshot. https://postera.ai/covid.

[64] Saar, K. L., Fearon, D., Consortium, T. C. M., von Delft, F., Chodera, J. D., and Lee, A. A. (2021). Turning high-throughput structural biology into predictive inhibitor design. *bioRxiv*.

[65] Schiebel, J., Krimmer, S. G., Röwer, K., Knörlein, A., Wang, X., Park, A. Y., Stieler, M., Ehrmann, F. R., Fu, K., Radeva, N., et al. (2016). High-throughput crystallography: reliable and efficient identification of fragment hits. *Structure*, 24(8):1398–1409.

[66] Schneider, P. and Schneider, G. (2016). De novo design at the edge of chaos: Miniperspective. *Journal of medicinal chemistry*, 59(9):4077–4086.

[67] Schuller, M., Correy, G. J., Gahbauer, S., Fearon, D., Wu, T., Díaz, R. E., Young, I. D., Martins, L. C., Smith, D. H., Schulze-Gahmen, U., Owens, T. W., Deshpande, I., Merz, G. E., Thwin, A. C., Biel, J. T., Peters, J. K., Moritz, M., Herrera, N., Kratochvil, H. T., null null, Aimon, A., Bennett, J. M., Neto, J. B., Cohen, A. E., Dias, A., Douangamath, A., Dunnett, L., Fedorov, O., Ferla, M. P., Fuchs, M. R., Gorrie-Stone, T. J., Holton, J. M., Johnson, M. G., Krojer, T., Meigs, G., Powell, A. J., Rack, J. G. M., Rangel, V. L., Russi, S., Skyner, R. E., Smith, C. A., Soares, A. S., Wierman, J. L., Zhu, K., O'Brien, P., Jura, N., Ashworth, A., Irwin, J. J., Thompson, M. C., Gestwicki, J. E., von Delft, F., Shoichet, B. K., Fraser, J. S., and Ahel, I. (2021). Fragment binding to the nsp3 macrodomain of sars-cov-2 identified through crystallographic screening and computational docking. *Science Advances*, 7(16):eabf8711.

[68] Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Bekas, C., and Lee, A. A. (2019a). Molecular Transformer - A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science*, 5(9):1572–1583.

[69] Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. (2019b). Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583.

[70] Segler, M. H., Kogej, T., Tyrchan, C., and Waller, M. P. (2018a). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science*, 4(1):120–131.

[71] Segler, M. H., Preuss, M., and Waller, M. P. (2018b). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555(7698):604–610.

[72] Sturmfels, P., Lundberg, S., and Lee, S.-I. (2020). Visualizing the impact of feature attribution baselines. *Distill*. https://distill.pub/2020/attribution-baselines.

[73] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. *34th International Conference on Machine Learning, ICML 2017*, 7:5109–5118.

[74] The COVID Moonshot Consortium (2020). Covid moonshot: open science discovery of sars-cov-2 main protease inhibitors by combining crowdsourcing, high-throughput experiments, computational simulations, and machine learning. *bioRxiv, doi:10.1101/2020.10.29.339317*.

[75] The COVID Moonshot Consortium, Achdout, H., Aimon, A., Bar-David, E., Barr, H., Ben-Shmuel, A., Bennett, J., Bilenko, V. A., Bilenko, V. A., Boby, M. L., Borden, B., Bowman, G. R., Brun, J., BVNBS, S., Calmiano, M., Carbery, A., Carney, D., Cattermole, E., Chang, E., Chernyshenko, E., Chodera, J. D., Clyde, A., Coffland, J. E., Cohen, G., Cole, J., Contini, A., Cox, L., Cvitkovic, M., Dias, A., Donckers, K., Dotson, D. L., Douangamath, A., Duberstein, S., Dudgeon, T., Dunnett, L., Eastman, P. K., Erez, N., Eyermann, C. J., Fairhead, M., Fate, G., Fearon, D., Fedorov, O., Ferla, M., Fernandes, R. S., Ferrins, L., Foster, R., Foster, H., Gabizon, R., Garcia-Sastre, A., Gawriljuk, V. O., Gehrtz, P., Gileadi, C., Giroud, C., Glass, W. G., Glen, R., Glinert, I., Godoy, A. S., Gorichko, M., Gorrie-Stone, T., Griffen, E. J., Hart, S. H., Heer, J., Henry, M., Hill, M., Horrell, S., Huliak, V. D., Hurley, M. F., Israely, T., Jajack, A., Jansen, J., Jnoff, E., Jochmans, D., John, T., Jonghe, S. D., Kantsadi, A. L., Kenny, P. W., Kiappes, J. L., Kinakh, S. O., Koekemoer, L., Kovar, B., Krojer, T., Lee, A., Lefker, B. A., Levy, H., Logvinenko, I. G., London, N., Lukacik, P., Macdonald, H. B., MacLean, B., Malla, T. R., Matviiuk, T., McCorkindale, W., McGovern, B. L., Melamed, S., Melnykov, K. P., Michurin, O., Mikolajek, H., Milne, B. F., Morris, A., Morris, G. M., Morwitzer, M. J., Moustakas, D., Nakamura, A. M., Neto, J. B., Neyts, J., Nguyen, L., Noske, G. D., Oleinikovas, V., Oliva, G., Overheul, G. J., Owen, D., Pai, R., Pan, J., Paran, N., Perry, B., Pingle, M., Pinjari, J., Politi, B., Powell, A., Psenak, V., Puni, R., Rangel, V. L., Reddi, R. N., Reid, S. P., Resnick, E., Ripka, E. G., Robinson, M. C., Robinson, R. P., Rodriguez-Guerra, J., Rosales, R., Rufa, D., Saar, K., Saikatendu, K. S., Schofield, C., Shafeev, M., Shaikh, A., Shi, J., Shurrush, K., Singh, S., Sittner, A., Skyner, R., Smalley, A., Smeets, B., Smilova, M. D., Solmesky, L. J., Spencer, J., Strain-Damerell, C., Swamy, V., Tamir, H., Tennant, R., Thompson, W., Thompson, A., Tomasio, S., Tsurupa, I. S., Tumber, A., Vakonakis, I., van Rij, R. P., Vangeel, L., Varghese, F. S., Vaschetto, M., Vitner, E. B., Voelz, V., Volkamer, A., von Delft, F., von Delft, A., Walsh, M., Ward, W., Weatherall, C., Weiss, S., White, K. M., Wild, C. F., Wittmann, M., Wright, N., Yahalom-Ronen, Y., Zaidmann, D., Zidane, H., and Zitzmann, N. (2022). Open science discovery of oral non-covalent sars-cov-2 main protease inhibitor therapeutics. *bioRxiv*.

[76] Tiamas, S. G., Audet, F., Samra, A. A., Bignon, J., Litaudon, M., Fourneau, C., Ariffin, A., Awang, K., Desrat, S., and Roussi, F. (2018). Asymmetric Total Synthesis and Biological Evaluation of Proapoptotic Natural Myrcene-Derived Cyclohexenyl Chalcones. *European Journal of Organic Chemistry*, 2018(42):5830–5835.

[77] Ullrich, S. and Nitsche, C. (2020). The sars-cov-2 main protease as drug target. *Bioorganic & Medicinal Chemistry Letters*, page 127377.

[78] Unoh, Y., Uehara, S., Nakahara, K., Nobori, H., Yamatsu, Y., Yamamoto, S., Maruyama, Y., Taoda, Y., Kasamatsu, K., Suto, T., et al. (2022). Discovery of s-217622, a noncovalent oral sars-cov-2 3cl protease inhibitor clinical candidate for treating covid-19. *Journal of Medicinal Chemistry*, 65(9):6499–6512.

[79] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):5999–6009.

[80] Weininger, D. (1988). SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36.

[81] Weininger, D., Weininger, A., and Weininger, J. L. (1989). SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *Journal of Chemical Information and Computer Sciences*, 29(2):97–101.

[82] Yang, Y., Zheng, S., Su, S., Zhao, C., Xu, J., and Chen, H. (2020). Syntalinker: automatic fragment linking with deep conditional transformer neural networks. *Chem. Sci.*, 11:8312–8322.

[83] Yu, H. S., Modugula, K., Ichihara, O., Kramschuster, K., Keng, S., Abel, R., and Wang, L. (2021). General theory of fragment linking in molecular design: Why fragment linking rarely succeeds and how to improve outcomes. *Journal of Chemical Theory and Computation*, 17(1):450–462. PMID: 33372778.