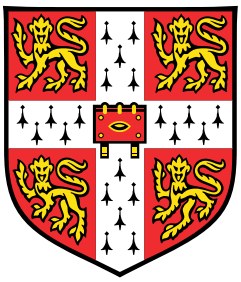
Accelerating the Design-Make-Test cycle of Drug Discovery with Machine Learning



William McCorkindale

Cavendish Laboratory, Department of Physics
University of Cambridge

Supervisor: Dr. Alpha Lee

St. John's College

March 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 Preface and Acknowledgements. This dissertation contains fewer than 65,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

The research described in this thesis was performed between October 2019 and December 2022, and was supervised by Dr Alpha A. Lee.

William McCorkindale
March 2023

Acknowledgements

TODO - finish acknowledgements

First and foremost, I would like to sincerely thank my advisor Alpha Lee for his help, guidance and for the freedom I was granted throughout these years.

In alphabetical order, I would also like to thank my colleagues who . . . : Dávid, Emma, Felix, Janosh, Kadi, Penny, Rokas, Rhys.

David, Michelle, Brian. Gates scholars. CUHKPGSA?

I would finally like to warmly thank my dear housemates Dávid and Eszti for their help in moments of doubt.

WM acknowledges the support of the Gates Cambridge Trust.

- 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 introduces the design-make-test cycle in drug discovery and the promise of machine learning (ML) for accelerating the process.

Chapter 2 gives an overview of molecular featurisation and computational methods which are used in this thesis.

In Chapter 3 starts at the hit-finding stage of drug discovery, and we discuss the usage of unsupervised learning for modelling the 3D distribution of pharmacophores in fragment-protein complexes. This work resulted in the following preprint (manuscript under review):

> 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.

In this work, I implemented the model and conducted the computational validation and virtual screening. Dr Ivan Ahel, Dr Haim Barr, Dr Khriesto Shurrush, and Prof Nir London performed bioactivity assays of ligands against SARS-CoV-2 Mpro. Dr Galen Correy and Prof James Fraser obtained X-ray crystallographic structures of ligand-bound structures to SARS-CoV-2 nsp3-Mac1. Dr Marion Schuller performed bioactivity assays of ligands against nsp3-Mac1. Dr Alpha A. Lee supervised the work.

Chapter 4 brings us to the hit-to-lead stage where modelling bioactivity becomes possible, and we discuss using a model that learns to rank molecules pairwise by activity. This work resulted in the following publication:

> 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

In this work, I developed the ranking model and constructed the screening library. Aaron Morris evaluated the model and generated compound synthesis routes. Dr John D. Chodera performed docking calculations. Prof Nir London performed bioactivity assays of ligands

against SARS-CoV-2 Mpro. Dr Nir Drayman and Prof Savaş Tay performed OC43 live virus assays. Dr Alpha A. Lee supervised the work.

In Chapter 5 we quantitatively explain predictions from deep learning models used for chemical reactions prediction, revealing model biases due to shortcomings in the training data. This work resulted in the following publication:

> 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)

I worked jointly with Dávid Kovács on this work which he completed as part of his MPhil research project under Dr Alpha A. Lee. We contributed equally to model development. Dávid Kovács trained the models, and analysed the model attributions for various reaction classes. I applied reaction templates for data analysis and artificial dataset generation, and investigated model performance under Tanimoto splitting. Dr Alpha A. Lee supervised the work.

Chapter 6 discusses the training of ML models on high-throughput bioactivty measurements from crude reaction mixtures instead of purified compounds. This research was carried out in collaboration with Dr Emma King-Smith, Mihajlo Filep, Prof Nir London, and Dr Alpha A. Lee. In this work, I implemented the random forest model and constructed the screening library. Dr Emma King-Smith implemented the gaussian process model and cleaned the experimental data. Mihajlo Filep performed bioactivity assays against SARS-CoV-2 Mpro. Dr Alpha A. Lee and Prof Nir London supervised the work.

The final chapter summarises the research presented and discusses promising directions for future research.

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.

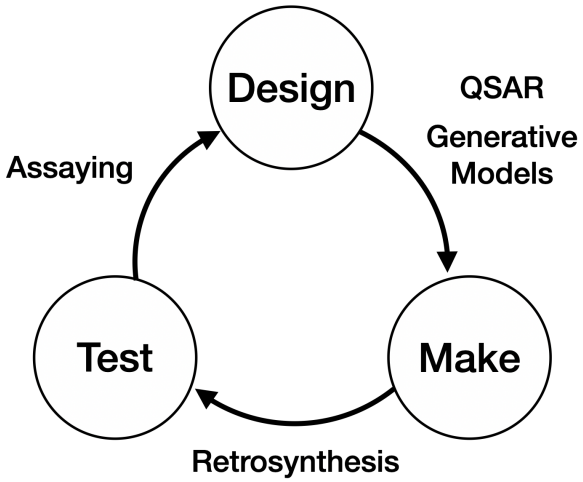


Fig. 1.1 An overview of the design-make-test cycle in drug discovery.

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' [**?**], for example in modelling quantitative structure-activity relationships (QSAR),

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

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 [**?**] 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.

End on influence of ML on drug discovery workflow?

During the course of this thesis, several fruitful collaborations have also led to the following publications. These are not discussed in detail within this dissertation.

> Kadi L. Saar, William McCorkindale, Daren Fearon, Melissa Boby, Haim Barr, Amir Ben-Shmuel, The COVID Moonshot Consortium, Nir London, Frank von Delft, John D. Chodera and Alpha A. Lee. Turning high-throughput structural

biology into predictive inhibitor design. *Proceedings of the National Academy of Sciences* volume 120 (11), Article number: e2214168120 (2023)

Ryan-Rhys Griffiths, Jake L Greenfield, Aditya R Thawani, Arian R Jamasb, Henry B Moss, Anthony Bourached, Penelope Jones, William McCorkindale, Alexander A Aldrick, Matthew J Fuchter and Alpha A. Lee. Data-driven discovery of molecular photoswitches with multioutput Gaussian processes. *Chemical Science* volume 13 (45), Article number: 13541-13551 (2022)

Chapter 2

Molecular Representations and Computational Methods in Drug Discovery

2.1 Molecular Representation

2.1.1 SMILES

The simplified molecular-input line-entry system (SMILES) [**?** **?**] is a widely-used text-based description of molecular structure. In SMILES strings, atoms are represented with their chemical symbols and aromatic atoms are denoted in lowercase (see Table **??** for examples). Single and aromatic bonds are omitted while for double and triple bonds the specials characters = and # are used. Branches are specified by enclosing them into parentheses. To encode cyclic structures a single bond in the ring is broken and the matching atoms are denoted by numbers. @ characters are used to denote chirality while \ and / characters specify local double bond configurations. Following these rules, a SMILES string is constructed by traversing the nodes of the molecular graph. Depending on the choice of starting node and traversal route there are often multiple valid SMILES representations per molecule, especially for larger molecules. In order to define a single unique SMILES representation for a molecule, known as the 'canonical' SMILES, a deterministic algorithm is used to choose the starting node and traversal route.

Reaction SMILES are a simple extension of SMILES for specifying chemical reactions. Reaction SMILES strings constructed by placing a > character between the SMILES strings of reactants, reagents, and products. If multiple molecules participate in the reactio, their SMILES strings are separated by a period (.) character.

The text-based nature of SMILES strings as well as its expressiveness in encoding the molecular graph alongside stereochemistry results in its widespread use for storing chemical data. In the context of machine learning, the vast majority of molecular datasets where ML

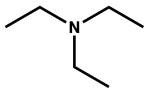
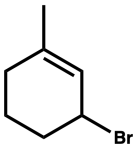
SMILES	Structure
C	CH_4
[Fe2+]	Fe^{2+}
C=O	CH_2O
C#N	HCN (cyan)
CCN(CC)CC	
CC1=CC(CCC1)Br	

Table 2.1 Demonstration of the SMILES language

models are used will have molecules represented as SMILES strings. For example, the ESOL dataset consists of 1128 SMILES strings alongside the measured solubility value for each molecule [**?**], while USPTO consists of 480k reaction SMILES strings [**?**]. For text-based ML models such as the Molecular Transformer (see chapter **??**), the SMILES strings are directly input to the model, while for other types of models the SMILES strings will be further processed to generate the necessary input features.

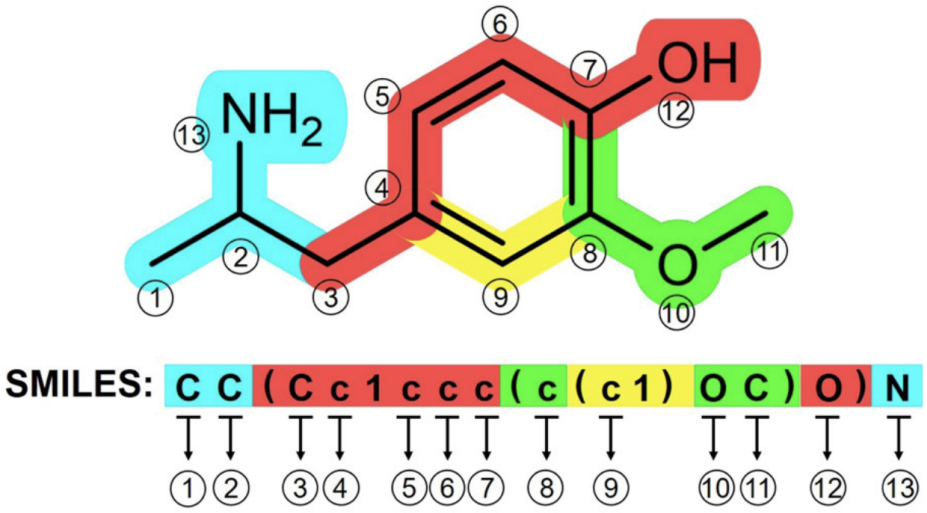


Fig. 2.1 **Illustration of the mapping from chemical structure to SMILES.** Adapted from [**?**].

While SMILES is by far the most widely used text-based representation of molecules, other representations have been developed and are in use to address some shortcomings in SMILES. For example, the International Chemical Identifier (InChI) [**?**] string representation, which

has a hierarchical construction for specifying tautomeric/stereochemical/charge states, allows greater precision and flexibility in querying molecules from large chemical databases. Another example is SELF-referencIng Embedded Strings (SELFIES) [**?**] which is constructed such that every SELFIES string, including random combinations of characters, is a valid molecule. This property is useful for the application of ML models that generate text as output - using SELFIES as the molecular representation, the model always output valid molecules whereas with SMILES that is not guaranteed.

2.1.2 Molecular Substructures

Given a dataset of molecules or chemical reactions encoded with SMILES, we often want to identify molecules or reactions that contain a specific substructure. For example, we may want to identify molecules that contain a specific functional group or reaction that contains a specific reaction center. The standard tool for performing these substructure queries is via SMILES Arbitrary Target Specification (SMARTS) notation [**?**]. The SMARTS line notation is expressive and allows extremely precise and transparent substructural specification and atom typing.

Using many of the same symbols as SMILES, it also allows specification of wildcard atoms and bonds, which allows expressive and precise definitons of substructures and atomic environments for searching chemical databases. One common misconception is that SMARTS-based substructural searching involves matching of SMILES and SMARTS strings. When performing a SMARTS query on a SMILES string, both SMILES and SMARTS strings are first converted to internal graph representations which are then searched for subgraph isomorphism.

SMARTS	Substructure
[C;R]	An aliphatic carbon in a ring
[#6]@[#6]	Two carbons connected by a ring bond
[N;$(NC=[O,S])]	amide or thioamide nitrogen
[N:1][C:2](=[O:3])[N:4]»[N:1][C:2](=[O:3])[C:4]	urea group transforming into an amide

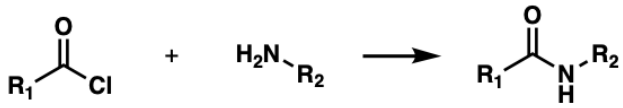
Table 2.2 Examples of SMARTS patterns

The precise substructure specification of SMARTS is useful in many aspects in the drug design process. For example, a common step in assessing the quality of a proposed drug candidate is to perform a SMARTS query to identify if the hit contains any substructures that are likely to produce artifacts in biochemical or cellular assays. These substructures are typically functional groups with a marked propensity to bind to multiple targets, so-called nuisance compounds, which are of little value in drug discovery. Many different sets of these

filters have been compiled in the literature such as REOS (rapid elimination of swill) [**?**] and PAINS (Pan Assay Interference Compounds) Filters [**?**]. Similarly, SMARTS queries are used to design 'structural alerts' which flag molecules containing reaction chemical substructures which may lead to undesirable toxicity in the compound itself or its metabolites [**?**].

Another use of SMARTS is in the labelling of pharmacophores in a molecule. Pharmacophores are an abstract description of the molecular features involved in ligand binding - typical examples of such features are hydrogen bond acceptors/donors and aromatic rings. SMARTS strings are used to map different molecular substructures to particular pharmacophore features and then to query for the presence of these features in a molecule. As with substructure filtering, different companies in the pharmaceutical industry have different/propritary sets of SMARTS strings tailored for their particular use cases. (see section **??** for details)

Beyond substructures for individual molecules, SMARTS can also be applied to reaction SMILES to capture transformation in substructures. These SMARTS strings for chemical reactions are often referred to in the literature as 'reaction templates' (Fig **??**). Beyond querrying for matching reactions from a dataset, reaction templates can also be directly applied to a set of molecules to computationally generate a 'reaction product'. This approach is used to generate virtual libraries [**?** **?**] *in silico* both for more specific usecases such as probing the SAR of a compound as well as the construction of ultra-large commercially available compound libraries eg ZINC [**?**], EnamineREAL.




```
[*:1][C:2](=[O:3])Cl.[N;H2:4][*:5]>>[*:1][C:2](=[O:3])[N;H1:4][*:5]
```

Fig. 2.2 An example of a reaction template for the synthesis of an amide from an acid chloride and a primary amine.

In addition to virtual library construction, reaction templates can be used for organic reaction prediction by framing the problem as trying to predict the correct reaction template for a given set of reactants. Starting from a catalogue of possible reaction templates, the best matching general template in the catalogue can be found utilising subgraph searching or machine learning, and applied on the input to obtain the predicted outcome of the reaction. This approach was originally proposed for the reverse problem of retrosynthesis [**?**] and has had success in forward reaction prediction for the design of synthetic pathways to drug-like molecules [**?**].

One major limitation of template-based approaches is scalability, as the template library needs to be maintained and updated every time a new reaction is reported. A further problem is that it is often not obvious which parts of the molecule are crucial for a given reaction. This means that given a reaction one can derive a smaller more general template or a larger one that is more specific for the particular reaction. This results in either too many templates matching a particular input resulting in many equally possible reaction outcomes or in the case of larger more specific templates the library will grow very big which results in very slow predictions.

2.1.3 Pharmacophores

A pharmacophore is an abstract description of molecular features that are necessary for molecular recognition of a ligand by a biological macromolecule [**?**]. A collection of pharmacophores in a geometric configuration is known as a 'pharmacophore model' and it is a representation of the interactions between ligands and the binding site. By their coarse-grained nature, pharmacophore model can explain structurally diverse ligands can bind to a common receptor site, and can be used to identify novel ligands that will bind to the same receptor.

Typical pharmacophore features, characterised by SMARTS strings, include hydrophobic groups, aromatic rings, hydrogen bond acceptors or donors, and charged functional groups.

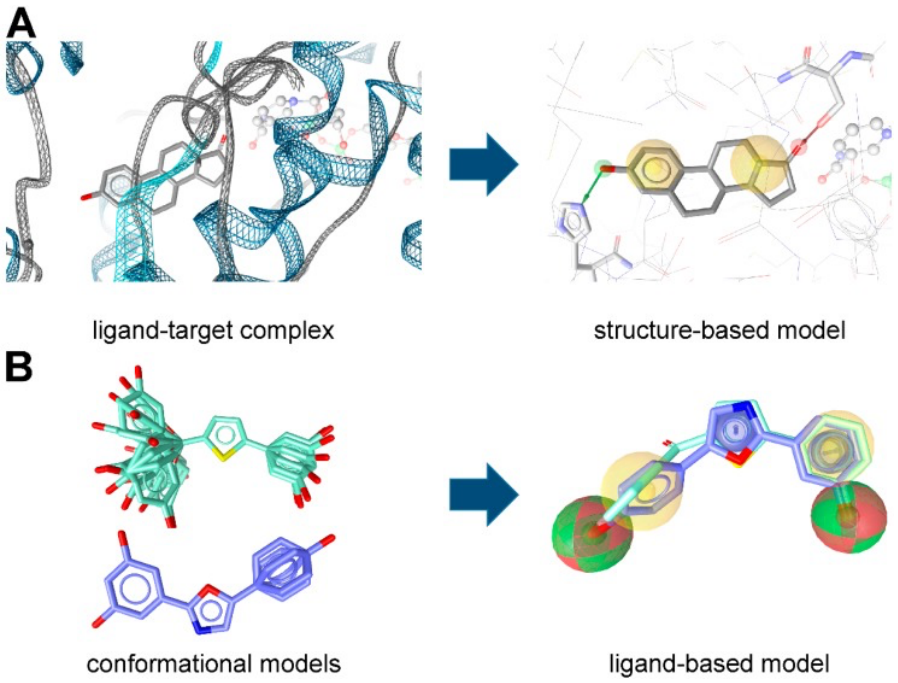


Fig. 2.3 An illustration of structure-based or ligand-based pharmacophore models. Reproduced from [**?**].

Pharmacophore models can either be constructed using structural data or from a set of active compounds [**?**]. In the structure-based approach, the pharmacophores are directly inferred from the observed interactions of a molecule and the binding site in experimentally determined ligand-protein complexes. (Fig **??**A).

In the ligand-based approach, the three-dimensional (3D) structures of known active molecules are aligned and pharmacophores that are found to overlap in space are extracted as the pharmacophore model (Figure **??**B). Although this approach circumvents the need for structural data which might not be possible to obtain, a downside is that all of the extracted pharmacophore have to be presumed as essential for protein-ligand binding, whereas in the structure-based approach it is possible to identify and discard non-important pharmacophores.

After obtaining a pharmacophore model, it can be used to virtually screen ligands from a database by identifying molecules that share similar pharmacophore features. This approach is known as 'pharmacophore-based virtual screening' and is a useful tool on its own as well as for complementing molecular docking and machine learning approaches [**? ? ? ?**].

2.1.4 Fingerprints

In order to apply powerful machine learning methods, we require a (typically fixed-length) vector representation of molecules konwn in the literature as 'molecular fingerprints'. The most popular molecular fingerprint is the Morgan fingerprint [**?**], also known as the Extended-Connectivity FingerPrint (ECFP) [**?**]. ECFPs are a particular example of 'topological' fingerprints that encode the presence of substructures in a molecule by traversing the molecular graph.

ECFPs are generated via a recursive hashing algorithm that numerically hashes the representation of each atom with those of its neighbours, and again with it's next-nearest neighbours etc until a pre-defined 'radius' is reached. The resulting hash values are then used to generate a fixed-length binary vector of 1/0 bits. The length of the vector is pre-determined and each 1-bit represents a unique substructure that is encountered during the traversal. The radius of the graph traversal is also a pre-defined parameter that controls the size of the substructures that are represented in the fingerprint. The radius is typically set to 2 or 3, and the length of the fingerprint is typically in the range of 1024-4096.

The popularity of the Morgan fingerprint owes to its usefullness in calculating molecular similarity [**?**]. Intuitively, we would expect two molecules that have 'similar' molecular fingerprints to have similar chemical structures. Numerically, we can quantify the similarity between two molecular fingerprints by the Tanimoto coefficient [**?**]:

$$\text{Tanimoto}(A,B) = \frac{A \cap B}{A \cup B} = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B} \tag{2.1}$$

where A and B refer to the bit-vector molecular fingerprints of two molecules. The numerator $A \cdot B$ represents the number of bits shared between the two fingerprints, while the denominator represents the total number of unique bits covered by the finerprints. Two structures are usually considered dissimilar if the tanimoto similarity is < 0.4 [**?**]. Alternate similarity measures exist but a number of comparison studies [**? ?**] have shown the Tanimoto similarity to be generally robust and consistently perform well in a variety of applications.

The combination of the Morgan fingerprint with Tanimoto similarity is useful for clustering [**?**] datasets of similar compounds as well as performing similarity-based virtual screening. Similarity-based virtual screening relies on the similarity property principle (SPP) [**?**], which states that similar compounds should have similar biological activity. As a guiding strategy this means one could search a database for similar compounds to a known active molecule, and expect those compounds to retain and perhaps have improved biological activity against a target. Although this hypothesis is not always valid in cases known as 'activity cliffs' where small changes in structure cause large changes in biological activity [**?**], empirically it has been shown that structurally similar compounds are much more likely to be actives compared to dissimilar ones [**?**]. Performing similarity-based virtual screening in practice involves calculating molecular similarities between known active compounds and unkonwn molecules from a database, then selecting those with the highest similarities; ECFP4 is a consistently well-performing fingerprint for this task [**? ?**].

In addition, Morgan fingerprints have also been shown to be versatile as a molecular descriptor for machine learning (ML). Machine learning models learn statistical patterns from data and can be used to make predictions on new data (see section **??**). In the context of drug discovery, ML can be used associate patterns in the molecular fingerprints of the molecules in a dataset with experimentally measured properties. For example, fingerprint-based models have demonstrated success in predicting physical/chemical properties such as solubility [**?**], biological activities [**?**] as well as yields and stereoselectivities for chemical reactions [**?**]. Non-fingerprint based deep learning methods have recently been developed that learn molecular representations directly from molecular graphs, however ECFP-based shallow learning techniques continue to provide a strong, robust baseline to compare against.

2.2 Computational Approaches

2.2.1 Docking

Molecular docking is the process of predicting the binding mode of a small molecule to a protein target, and one of the most frequently used methods in structure-based drug design. [**? ?**] The binding mode is the relative orientation of the small molecule in a particular binding site of the protein, which is determined by the shapes of the binding site and molecule, and the physical interactions between the two. The binding mode has a large effect on the strength of the interaction between the small molecule and the protein, known as the binding affinity. The binding affinity, in turn, is a key determinant of the biological activity of the small molecule. The philosophy of structure-based drug design is to experimentally obtain binding modes of molecules and use this information to guide the design of new molecules by docking and choosing molecules that may have more optimal protein-ligand interactions and hence binding affinity.

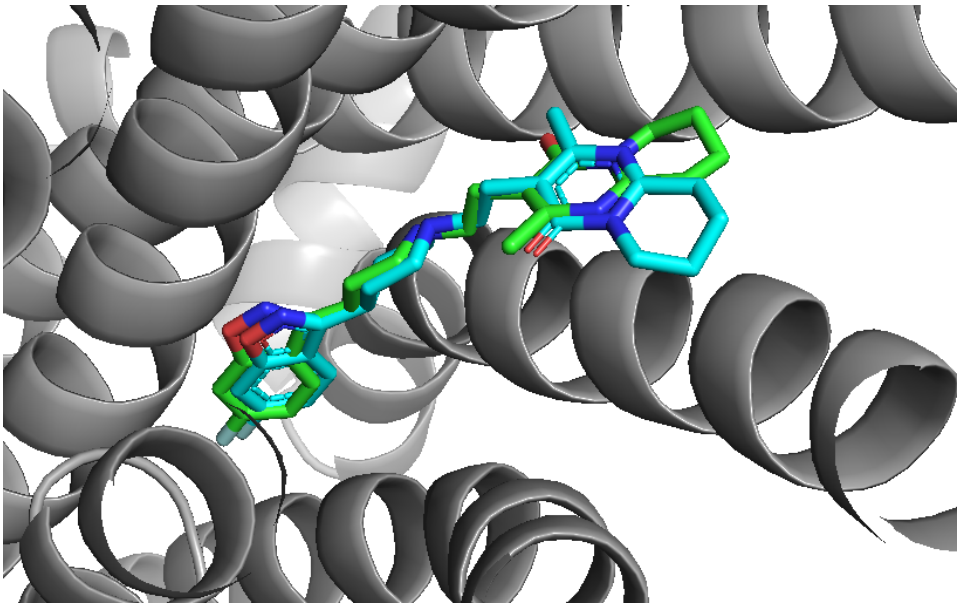


Fig. 2.4 **Example of a docked molecule.** The experimental structure of the ligand risperidone bound to the serotonin 2A receptor is shown in green, with the protein in grey (PDB: 6A93). The structure of the same ligand docked using GOLD is shown in cyan.

The physics of ligand-protein binding are complex, and in reality each ligand will have an ensemble of binding modes. Attempting to accurately simulate the binding process is computationally intractable, and so the goal of molecular docking is to predict the most likely binding mode. In practice this is approached as an opimisation problem, where the coordinates of the ligand and/or protein atoms are adjusted until the 'best-fit' is acheived.

An essential preliminary step to performing molecular docking is obtaining a structure of the protein of interest. Traditionally this means using biophysical techniques such as X-ray crystallography, NMR spectroscopy or cryo-electron microscopy (cryo-EM), but recent development in computational protein structure prediction [**?** **?**] open the door to performing fully *in-silico* structure-based drug design.

Every docking approach is essentially composed of two parts - conformation scoring and conformation searching. Potential ligand poses are ranked using a scoring function, which are typically physics-based molecular mechanics force fields that estimate the energy of the pose within the binding site. The scoring function can be composed of many components, such as electrostatic interactions, solvent and steric effects, hydrogen bonding, as well as knowledge-based potentials derived from observed interactions from databases of protein-ligand structures [**?**]. While the accuracy of scoring functions has to be good enough to distinguish good poses from bad ones, major emphasis is put on computational efficiency due to the large number of evaluations required during docking. Thus, scoring functions often involve many assumptions and simplifications to reduce computational costs.

The search space of conformations is imposible to exhaustively explore as in theory it consists of all possible orientations and configurations of the protein paired with the ligand. In practice, usually the whole conformational space of the ligand is searched, while the protein is often treated rigidly. Exploration of the conformational search space is often done using stochastic methods such as Monte Carlo or genetic algorithms which randomly sample the space of conformation parameters (e.g. torsion angles) towards a minimisation of the scoring function.

The wide range of design choices for the scoring function and conformation search results in a large number of different docking algorithms that are in use in the field, such as DOCK [**?**], Glide [**?**], AutoDock Vina [**?**], GOLD [**?**], and FRED [**?**]. The relative performance between these docking algorithms are typically retrospectively evaluated by directly comparing predicted binding poses to known crystal structures of ligand-protein complexes. The benchmark datasets used for this purpose are typically high-quality structures of drug-like molecules such as PDBbind [**?** **?**]. There are also community assessments on the relative prospective performance of different docking approaches [**?**] and scoring functions [**?**].

Besides the structural focus of binding pose prediction, increasingly in recent years docking has been used to directly virtually screen large databases of molecules in silico to identify molecules that are likely to bind to protein target of interest [**?**]. This approach puts the focus on the scoring function, with the rationale that molecules with high docking scores are much more likely to be active than those with low scores. In this scenario, success is defined by

the enrichment of active compounds in the top ranks of a docking screen, measured via the enrichment factor:

$$\mathrm{EF}(n) = \frac{\mathrm{Hit\,rate(predicted\,top}{-}n)}{\mathrm{Hit\,rate(baseline)}} \tag{2.2}$$

where the baseline hit rate is the proportion of actives in the dataset overall, representing the performance of simple random ordering. Different methods are benchmarked by retrospectively evaluating the enrichment factor of known ligands from a large database of presumed non-binding, "decoy" molecules for multiple protein targets - the classic benchmark dataset for this is the Directory of Useful Decoys (DUD) [**?** **?**].

Prospectively, large-scale virtual screening with molecular docking has had notable successes. A review specifically looking at G protein-coupled receptors (GPCRs) [**?**], which are the target of more than 30% of all marketed drugs, showed 62 successful virtual screens for 22 unique protein targets belonging to 14 different receptor families in the past decade. Of particular note is that increasing availability of computational resources, together with increases in the sizes of commercially-available make-on-demand compound libraries, has made possible ultra-large virtual screening campaigns against libraries of >100 million compounds [**?** **?** **?**]. At the same time, limitations in the accuracy of scoring functions and in the modelling of protein flexibility [**?** **?**] restrict the ability of docking to reliably distinguish active molecules from inactive ones [**?** **?**], leading to false positives which are exacerbated when screening large libraries [**?**].

In the absence of exisiting ligand bioactivity measurements for a protein target, virtual screening with molecular docking remains the only computational method of choice and is a default starting point for beginning drug discovery against a brand new target. However, there has been recent results claiming that deep learning models which use neural networks to directly generate ligand binding poses [**?** **?**] outperform docking algorithms in terms of accuracy. These results are promising and, coupled with continued advancements in protein structure prediction to account for protein flexibility, suggest that deep learning may be a viable alternative to molecular docking for binding pose prediction and virtual screening in the coming years.

2.2.2 Machine Learning

Machine learning (ML) refers to the use of algorithms that 'learns' how to make predictions or decisions based on observed data. By designing models that can learn patterns directly from input data, it is found that ML methods can often surpass manually created algorithms by humans on a wide variety of tasks. In the following paragraphs, we provide an overview

of the ML concepts and ideas needed to understand the research presented in this thesis. We will focus on supervised learning and neglect many important sub-fields such as reinforcement learning and generative models which are more described in greater detail in refs [**? ?**].

Very broadly, the aim of machine learning is to learn the parameters θ of a predicative model $y = f(x, \theta)$ that minimise a given cost function, $\mathscr{L}(y, \hat{y})$, where x is a given input, y is the target variable and $\hat{y}$ is the predicted value, i.e. to find the solution

$$\hat{\theta} = \arg\min_{\theta} \mathscr{L}(\theta) \tag{2.3}$$

For regression, which is the modelling of a continuous variable, the most common loss function choice is the squared residuals,

$$\mathscr{L}(y, \hat{y}) = \sum_i (y_i - \hat{y}_i)^2 \tag{2.4}$$

while for binary classification, which is the task of predicting which class an input x belongs to, the most common loss function is the binary cross-entropy loss:

$$\mathscr{L}(y, \hat{y}) = -\sum_i y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \tag{2.5}$$

where the target variable y can be either 0 or 1 while the predicted value $\hat{y}_i$ is the predicted probability of class 1 and $1 - \hat{y}_i$ is the predicted probability of class 0.

To find the solution $\hat{\theta}$ for a dataset in practice, we would first divide the input data into training, validation and test sets. The model is initially fit on the training data set, where the data-dependent parameters of the model (e.g. the coefficients of a polynomial regression model) are optimised to minimise the loss function. Afterwards, the fitted model is used to make predictions on the validation data set. The validation data set provides an unbiased estimate of the model's performance on the training data set for the purpose of tuning the non-data-dependent parameters, known as 'hyperparameters', of a model (e.g. the number of degrees to include in a polynomial regression model). This process may be repeated multiple times, with the model's performance on the validation data set being used to select the best hyperparameters. This overall process is known as 'training a model'.

After a model has been trained we use the test dataset, which has never been seen by the model during training, to evaluate the performance of the model. It is important to use the same training and test datasets for a fair comparison of different models, and curated datasets from the literature are commonly used as a benchmark for evaluating the performance of new models.

For regression models, the most common metric used to evaluate the performance of a model is the root mean squared error (RMSE) or the pearson correlation coefficient (PCC). For binary classification models, the most common metric used is the area under the receiver operating characteristic curve (AUC). The receiver operating characteristic (ROC) curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) as the discrimination threshold for classifying one class over the other is varied **??**. The AUC is the area under the ROC curve, and is a measure of the model's ability to distinguish between the two classes. AUC values range from 0 to 1, with 0.5 indicating a model that is no better than random guessing, and 1 indicating a perfect model.

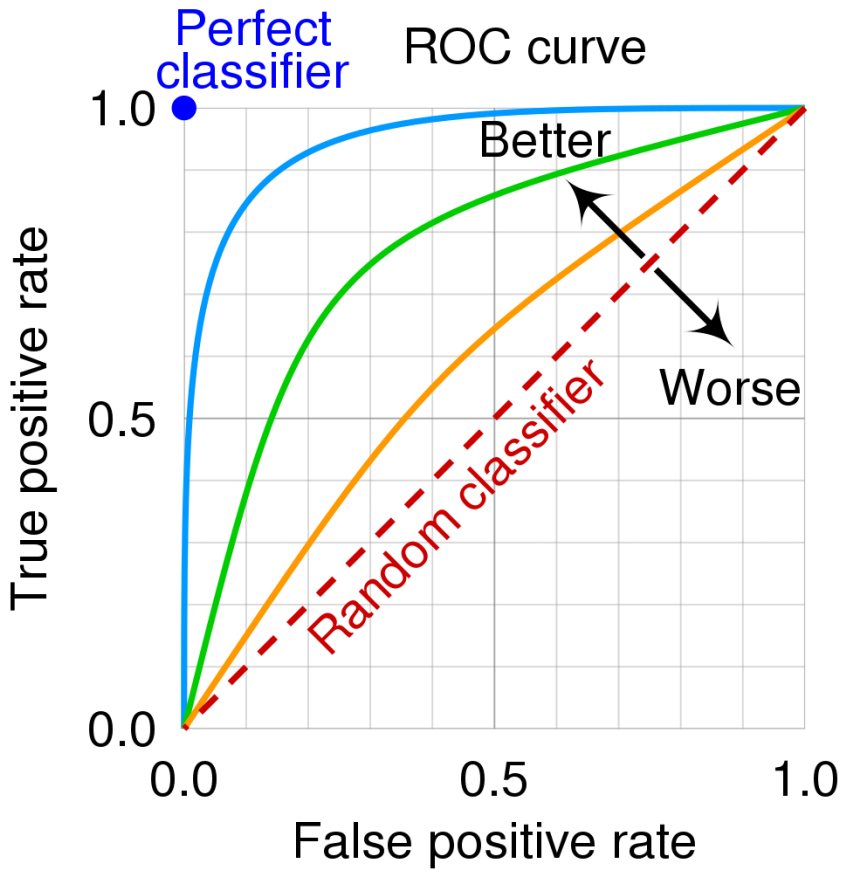


Fig. 2.5 An example Receiver Operating Characteristic (ROC) curve. The diagonal shows the performance of a random classifier. Three example classifiers (blue, orange, green) are shown. Reproduced from [**?**].

A large number of machine learning techniques have been described and applied for drug discovery, and an overview of them can be found in [**?** **?** **?**]. In this thesis, two common techniques are used: Random forests and Gaussian processes.

Random Forest

Random Forest models are ensemble learning methods that utilise a large number of decision trees for making predictions [**?**]. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the output is the mean of the predictions by the individual trees:

$$f(x) = \frac{1}{N}\sum_{i}^{N} f_i(x) \tag{2.6}$$

where f_i is the ith tree in the forest and N is the total number of trees in the forest. Decision trees are constructed to succerssively split the data into branches via 'decision boundaries' (e.g. $x > 1.5$). Decision boundaries are chosen to minimise the square deviations (for regression) or information entropy (for classification) between the samples and the sample mean in each branch or leaf of the tree.

Although extremely computationally efficient and interpretable, individual decision trees are very prone to over-fitting. By using a large number of decision trees each trained on different random subsets of the data (a process known as bootstrap aggregation), random forest models can acheive a lower variance and hence improved performance. To further reduce correlation between the decision trees, random forests use random feature selection, where only a random subset of the features are considered for each decision boundary.

Random forests are relatively easy to use, require little tuning of hyperparameters, and are robust to over-fitting, and thus are a popular general machine learning technique. In particular for drug discovery, random forests have been extensively used with molecular fingerprints features for prediction of properties such as solubility [**?**], biological activity [**? ?**], and toxicity [**?**].

Gaussian Process

Gaussian Process models are a kernel-based method that utilises what is known as the 'kernel trick' to calculate high-dimensional weighted averages [**?**]. The kernel trick is the use of a kernel function to calculate the inner product between the features vectors of two datapoints in a high-dimensional feature space without explicitly computing higher-dimensional feature vectors. A typical kernel function is the squared exponential kernel:

$$\mathscr{K}(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2l^2}\right) \tag{2.7}$$

where l is the length scale of the kernel, a hyperparameter that determines how quickly the function can change, and x and x' are the feature vectors of the datapoints in the initial

low-dimensional feature space. Mathematically, the kernel trick enables the construction of models that are, in theory, infinitely complicated with a finite amount of computation [**?**]. Gaussian Processes utilise the kernel function to calculate the covariance matrix of the Gaussian distribution over the function values. The covariance matrix is then used to calculate the posterior distribution over the function values given the training data, which can then be used to make predictions on new data with associated uncertainty values.

The fact that GPs have few hyperparameters to tune and maintain uncertainty estimates over property values have also led to their use for molecular property prediction [**? ? ?**], in particular incorporating the Tanimoto similarity in the kernel function [**? ?**].

Deep Learning

In contrast to methods which use hand-crafted features as input ('shallow learning'), deep learning revolves around learning representations directly from the raw data using neural networks. Neural networks are composed of layers of 'neurons' that successively perform non-linear transformations on their inputs, mimicking the way that biological neurons transfer signals to one another. These transformations are typically of the form:

$$\mathbf{h} = \sigma(\mathbf{W} \cdot \mathbf{x} + \mathbf{b}) \tag{2.8}$$

where $\mathbf{W}$ is a weight matrix, $\mathbf{x}$ is a vector of inputs, $\mathbf{b}$ is a vector of biases and σ is an optional non-linear activation function. The output of the layer is the vector $\mathbf{h}$, which is either input to the next layer or taken as the output of the model. The weights $\mathbf{W}$ and biases $\mathbf{b}$ from all of the layers collectively are the parameters θ of the neural network that are learned by fitting on data.

Deep learning has found remarkable success in a wide range of applications, including computer vision [**?**], natural language processing [**?**], speech recognition [**?**], and bioinformatics [**? ?**]. This is because of the ability of neural networks to learn complex, non-linear relationships between inputs and outputs in the presence of large amounts of data. 'Big data' domains are computationally intractable for shallow learning methods, but deep learning can be successfully applied as neural networks can be optimised effectively using gradient-based approaches, such as gradient descent:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathscr{L} \tag{2.9}$$

where at each step t the model's parameters are updated according to the learning rate η, a hyperparameter of the optimiser. Instead of calculating the gradient of the loss $\nabla_\theta \mathscr{L}$ on the full training set, standard practice is to use a stochastic approximation of the gradient

that is calculated from a randomly sampled batch of training data. This significantly speeds up the optimisation process and allows neural networks to be trained on large datasets that would otherwise be intractable. These stochastic gradient steps are iterated repeatedly over the training set until the value of the loss has satisfactorily converged.

The gradient of the loss function with respect to the model parameters can be obtained efficiently by applying the chain rule via a process called 'backpropogation'. Using automatic differentiation frameworks that can be carried out on hardware accelerators, such as graphical processing units (GPUs), the time needed to train neural networks are dramatically reduced [**?**]. Further improvements to model optimisation can be achieved by incorporating more sophisticated optimisation algorithms, such as Adam [**?**], as well as the use of regularisation techniques such as dropout [**?**] and batch normalisation [**?**], and remains a significant area of active research.

The only constraint on the design of a neural network is that the mathematical operations in the model must have defined derivatives so that the gradient of the loss function can be calculated with backpropagation for efficient training. This results in a zoo of different neural network designs (referred to as 'architectures') that use differentiable building blocks with specific inductive biases tailored to the task at hand. For example, convolutional neural networks (CNNs) utilise built-in translational invariance for computer vision applications (e.g. AlexNet [**?**], and ResNets [**?**]) while recurrent neural networks (RNNs) designed for learning temporal dependencies are applied to sequential data such as text [**?**] and speech [**?**] (e.g. Gated-Recurrent-Unit (GRU) [**?**] and Long-Short-Term-Memory (LSTM) networks [**?**]).

In the context of drug discovery, the challenge of modelling molecular inputs has led to the devlopment of graph-based neural network architectures known as 'graph neural networks' (GNNs). These models are designed to learn representations of molecular graphs that are invariant to the order of the nodes and edges. GNNs have been successfully applied to a range of molecular tasks, including molecular property prediction [**? ? ? ?**], and predicting reaction templates for input reactants [**?**]. More standard architectures such as Transformer models developed for natural language processing have also been applied to SMILES, as well as three-dimensional voxel-based CNNs which can be trained on protein-ligand complexes for the prediction of binding affinity [**? ? ?**].

Neural networks can also be used with pre-computed features such as molecular fingerprints. Example applications include the use of bioactivity prediction [**?**], reaction prediction [**? ?**], and the prediction of docking scores [**?**].

Despite their success in certain molecular tasks, deep learning still has several limitations when it comes to drug discovery. Chief among these is the need for large amounts of training data for strong performance, which can often be costly and time-consuming to obtain. When

only a small amount of data is available, which is typical in the early stages of drug discovery, neural networks may perform worse than simpler shallow learning models [**?**].

Additionally, neural networks may struggle to generalize to new molecules that are substantially different from the molecules in the training set. This is known as the 'generalization gap' and model performance with typical random-split cross-validation procedures do not accurately reflect the true generalization performance of the model [**?**]. This has led to a growing movement towards measuring model performance using 'scaffold split' [**?** **?**] or 'time split' [**?**] cross-validation, which splits the data into disjoint sets of molecules that are similar in structure or time of data acquisition, respectively. This allows for a more accurate assessment of the generalization gap, but this remains a challenge for applying deep learning and machine learning models in general in drug discovery.

Another challenge is the lack of interpretability of neural network models, which make it difficult to understand the underlying reasons for a model's predictions. Without explainations of model predictions, it becomes difficult to avoid correct predictions for the wrong reasons (the so-called clever Hans effect) [**?**], avert unfair biases, and gain potentially useful insights from the model. This is a challenge in general for deep learning, but is particularly difficult in drug discovery due to the domain-specific complication of projecting 'explanations' onto molecule representations [**?**].

Accounting for these limitations are critical for realizing the full potential of applying both shallow and deep learning models to accelerate the design-make-test cycle in drug discovery.

Chapter 3

Hit Discovery via Unsupervised Learning of Fragment-Protein Complexes

Hit detection is a key step in the early stages of the drug discovery process following the identification of a biological target of interest [**?**]. 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.

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

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 chapter, 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 [**?**]. 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.

3.1 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 **??**a).

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 **??**). 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 [**?**]. 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 (KDE) [**?**] to estimate this spatial distribution of pairwise pharmacophore distances. We then score unseen molecules by evaluating pharmacophore

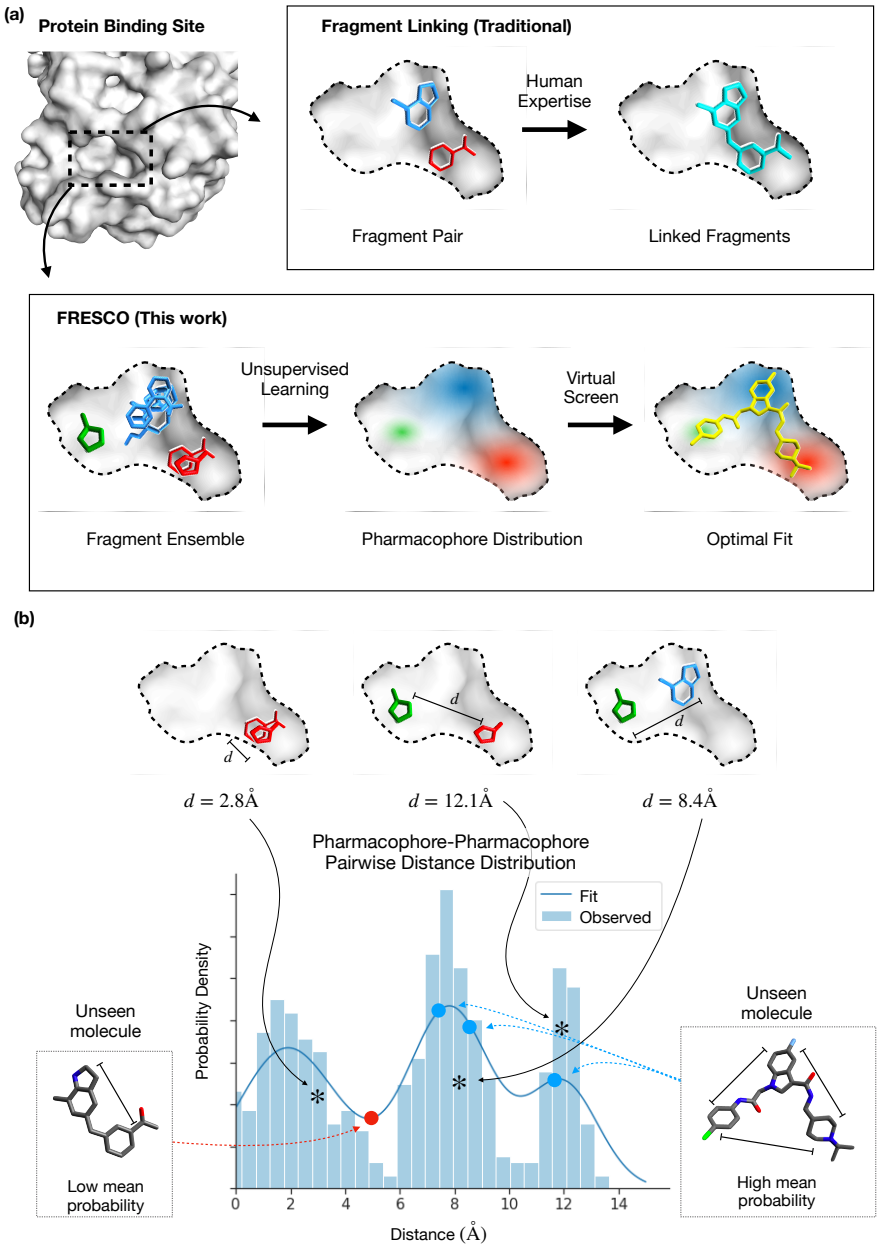


Fig. 3.1 (a) A visual illustration of how FRESCO differs from traditional fragment linking approaches. (b) A visual illustration of how we apply unsupervised learning to fragment ensembles and perform virtual screening of unseen molecules.

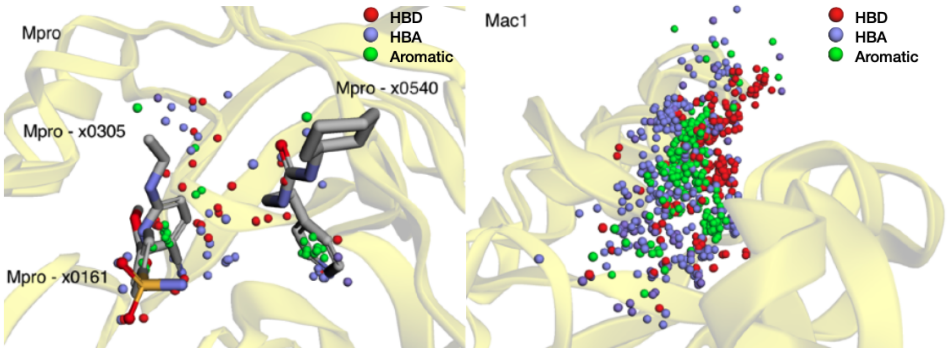


Fig. 3.2 The pharmacophores of the fragment ensemble shown in the 3D binding sites of (a) Mpro, and (b) nsp3-Mac1. The red, blue, and green spheres depict hydrogen bond donors, acceptors, and aromatic pharmacophores respectively. Several of the Mpro fragments are drawn to illustrate the 'origin' of some pharmacophores. None are drawn for nsp3-Mac1 due to the density of pharmacophores in the binding site.

distances within that molecule against the probability distribution of pharmacophore distances derived from the fragment ensemble (Figure **??**b). 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 [**?**], SyntaLinker [**?**], and Develop [**?**]), as well as data-mining methods such as Fragment Network [**?**], 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 [**?**]. 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 [**? ? ?**] can be used in the absence of any bioactivity data, FRESCO crucially incorporates information from the fragment screen on preferential interactions between regions of the binding site and the fragment pharmacophores.

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 [**?**]. 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.

3.1.1 Model Implementation

To train our model, we first process a set of experimental fragment-protein complexes. In this particular work, we downloaded structures from the `Fragalysis` platform [**?**]. For Mpro, non-covalent fragments from the XChem fragment screen [**?**] were used while for Mac1 both XChem and UCSF fragment data were used [**?**].

We then extract the pharmacophore features from the fragment molecules and their corresponding conformer coordinates. Specifically, we used SMARTS pattern matching following default pharmacophore definitions in `RDKit` to extract pharmacophores from the fragment SMILES. The pharmacophores considered are hydrogen bond donors, hydrogen bond acceptors, and aromatic rings. The corresponding coordinates for 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). We then compute the pairwise distance matrix between all possible pharmacophore pairs (eg Donor-Donor & Aromatic-Acceptor) between different fragment molecules.

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.

With the processed 3D pharmacophore distributions, we can then fit a FRESCO model by learning the probability distribution of the pairwise distances using kernel density estimation (KDE). The bandwidth for KDE fitting was chosen for each pariwise distribution using the Improved Sheather-Jones algorithm [**?**] (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 computational efficiency of inference for virtual screening, computationally fast approximations of the KDEs are made using the `scipy` `interp1d` function.

With a trained FRESCO model, we can then score unseen molecules by evaluating the probability of the pharmacophore distribution of each molecule. Given a set of input molecular conformers (eg from docking), the same processing workflow is used to obtain the 3D pharmacophore distributions of each molecule, and the probability of the distributions are evaluated using the KDEs. The overall score for the molecule is returned as the mean log-probability over all of the pairwise pharmacophore combinations.

3.2 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 [**?**] 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 [**?** **?**]. 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.

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.

In addition, molecular docking studies have also been done extensively on molecules from the Moonshot campaign [**?** **?**]. 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 5μM for defining a "hit". This threshold is relatively arbitrary and repeated analysis for both lower and higher IC50 thresholds (1-15μM) show similar results.

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 [**?**] and score the whole dataset using the fitted FRESCO model.

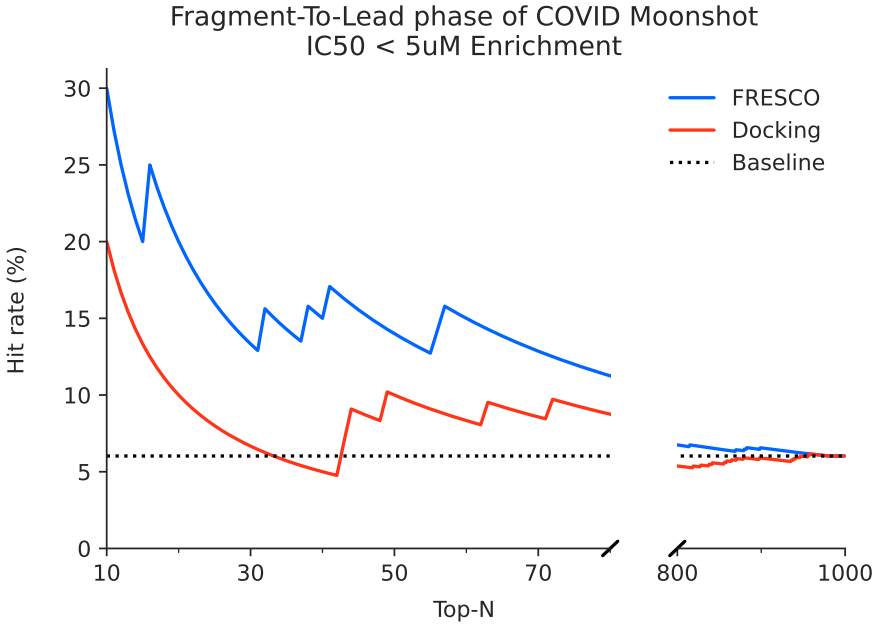


Fig. 3.3 FRESCO is able to retrospectively perform hit detection. High hit rates are achieved relative to docking and the human expert baseline when ranking molecules from the fragment-to-lead phase of COVID Moonshot.

Figure **??** shows that FRESCO achieves higher hit rates compared to both computational docking and the medicinal chemists. Looking at the top-5% of the molecules ($N < 50$), FRESCO has a hit rate of 12-30%, roughly 2-5 times that of the medicinal chemists. Hit rates are also higher than baseline for both lower and higher IC50 thresholds (Figure S3).

This shows that it is possible to correlate bioactivity with unsupervised learning of fragment pharmacophore distributions, and that FRESCO could accelerate hit detection in a real-world drug discovery campaign. In this retrospective study, FRESCO is standing on the shoulders of medicinal chemists – it is used to rescore compounds that are designed by chemists. Therefore, we next turn to interrogate the performance of FRESCO in a real-world context when it is used to score a large unbiased library of compounds via a series of prospective studies.

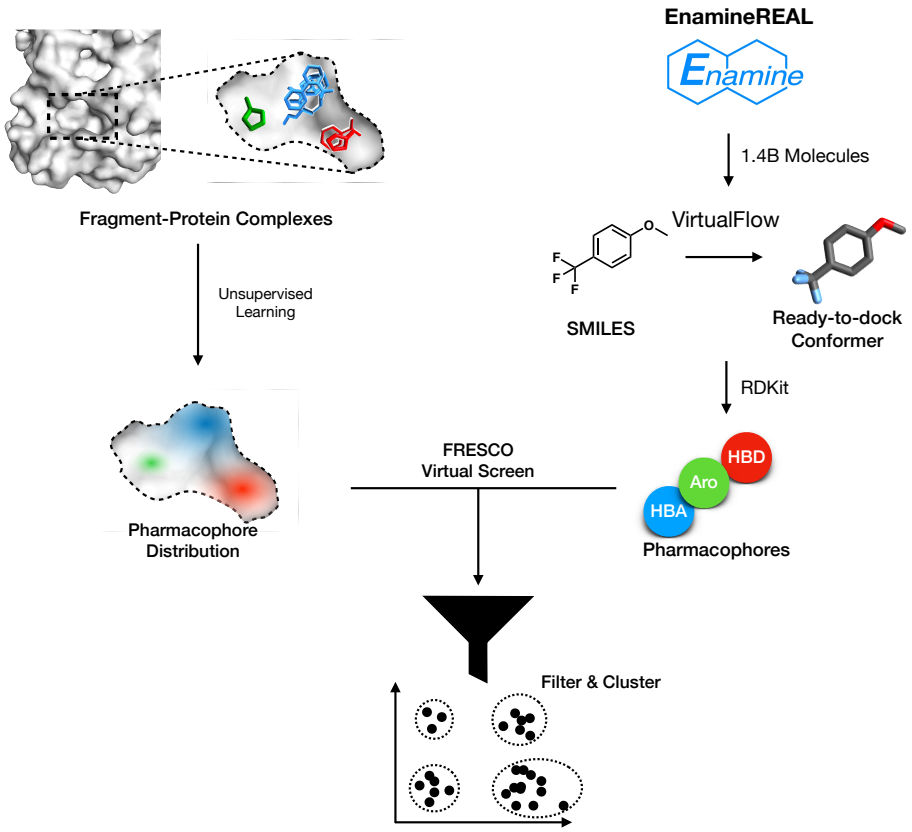


Fig. 3.4 A schematic of the FRESCO screening workflow.

3.3 Prospective hit finding

Building on the results of the retrospective evaluation, we performed a prospective study on Mpro. Rather than rescreening Moonshot compounds, we instead deploy the model to virtually screen a library of commercially available compounds. By synthesising and assaying the top-ranked compounds, we can evaluate the performance of FRESCO in a real-life use case of hit discovery.

The computational workflow we follow to perform the virtual screening is shown in Figure **??**. Using a FRESCO model trained on the fragment-protein complexes, we score the library and rank the compounds by score. The top-ranked compounds are then filtered 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.

The library we screen is VirtualFlow, a published dataset of more than 1.4 billion commercially available molecules from EnamineREAL & ZINC15 with pre-generated molecular conformers in a ready-to-dock format [**?**]. 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 [**?**] 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) [**?**] 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 [**?**] with a Tanimoto distance threshold of 0.2. This resulted in 24748 for Mpro. 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 successfully made molecules.

Inspecting the cluster centroids favored by FRESCO, we observe 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 **??**a). The most promising compound, WIL-UNI-d4749f31-37, has an IC50 of 25.8μM measured via fluorescence assay while the remaining compounds were found to be weak-to-negligible activity.

To validate compound activity, we synthesized 8 close analogues to demonstrate the existence of responsive Structure-Activity Relationship [**?** **?**] (Figure **??**b). 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 19.4μM . 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 99.5μM .

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-

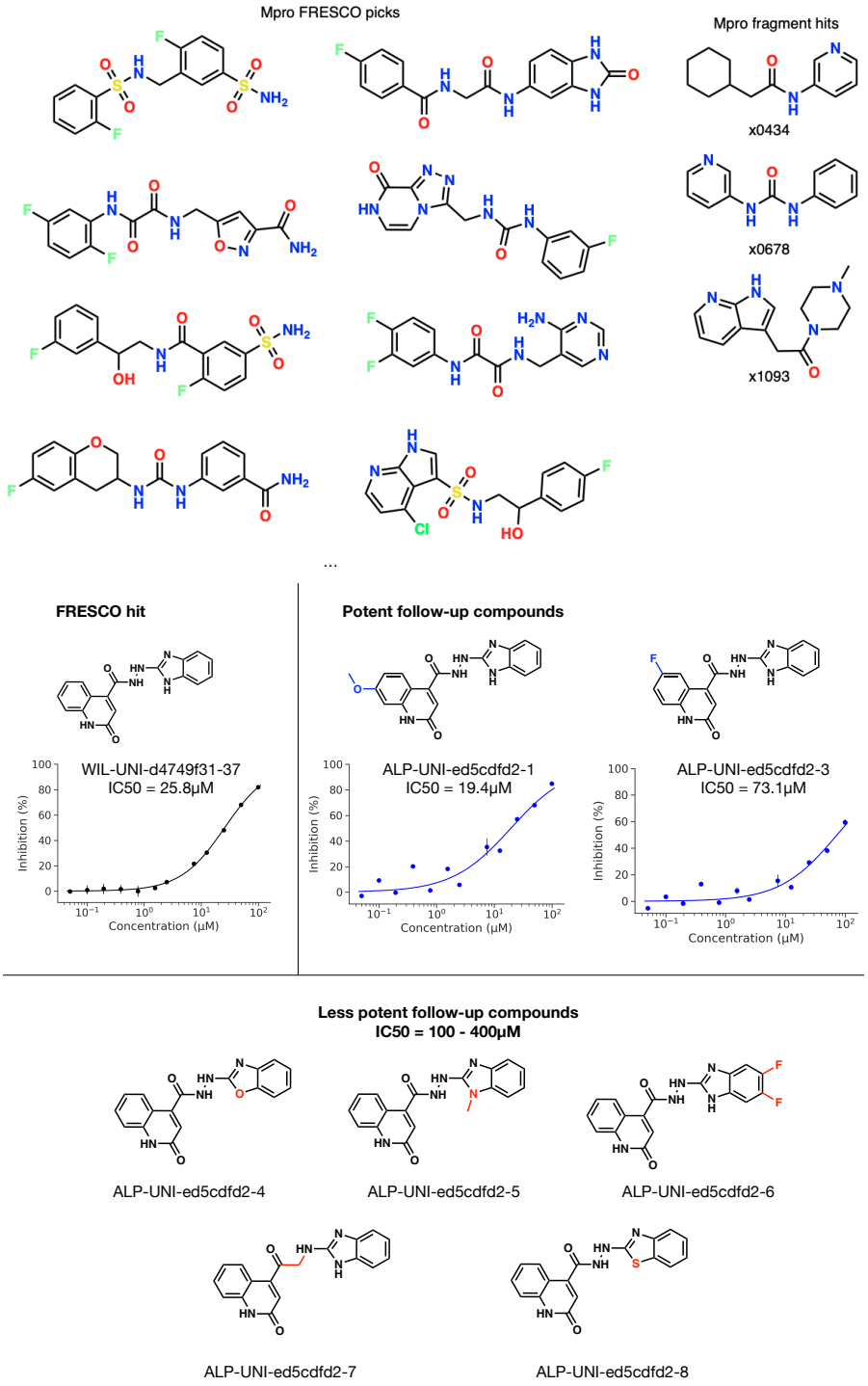


Fig. 3.5 (**a**) Example structures of cluster centroids after executing the FRESCO screening workflow on Mpro. The molecules favoured by FRESCO tend to have 2 aromatic moieties connected via an amide or an amide isostere, similarly exhibited by three of the initial fragment hits whose structures are also shown. (**b**) 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.

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 [**?**], we obtained 22358 clusters of top-ranked compounds and successfully made 52 molecules. 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 **??**). 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 250μM , 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 **??**). Crystal structures of 9 other compounds chosen via the FRESCO workflow were also obtained though

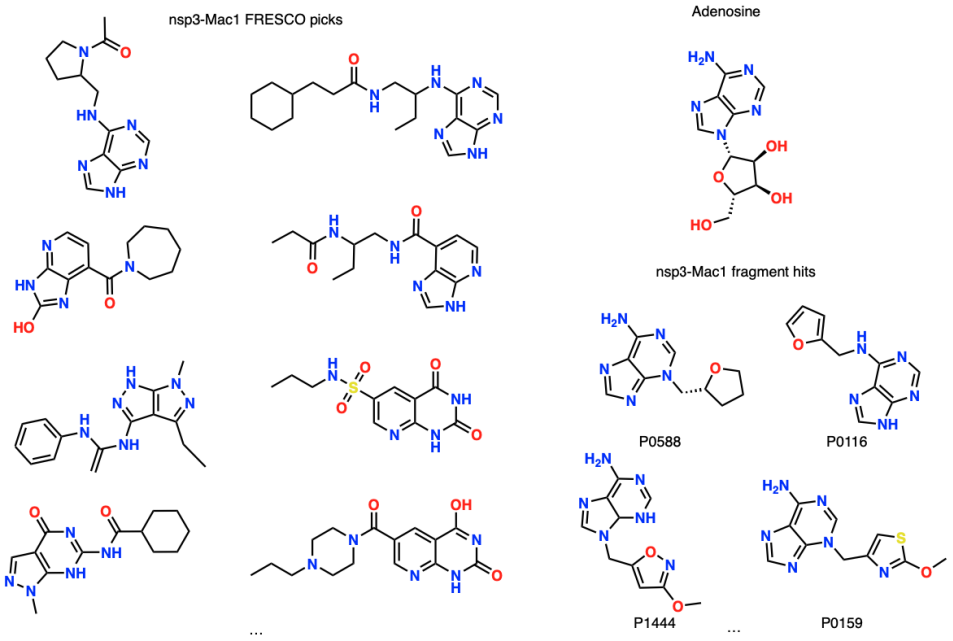


Fig. 3.6 Example structures of cluster centroids after executing the FRESCO screening workflow on nsp3-Mac1. The molecules favoured by FRESCO tend to contain an acceptor-donor pair spatially proximal to a hetrocyclic motif. This mimics adenosine, a core in the natural substrate.This motif is also shared in many of the initial fragment hits, with some example structures shown in the figure.

they did not show notable inhibition via HTRF assay. 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 **??**). These sets

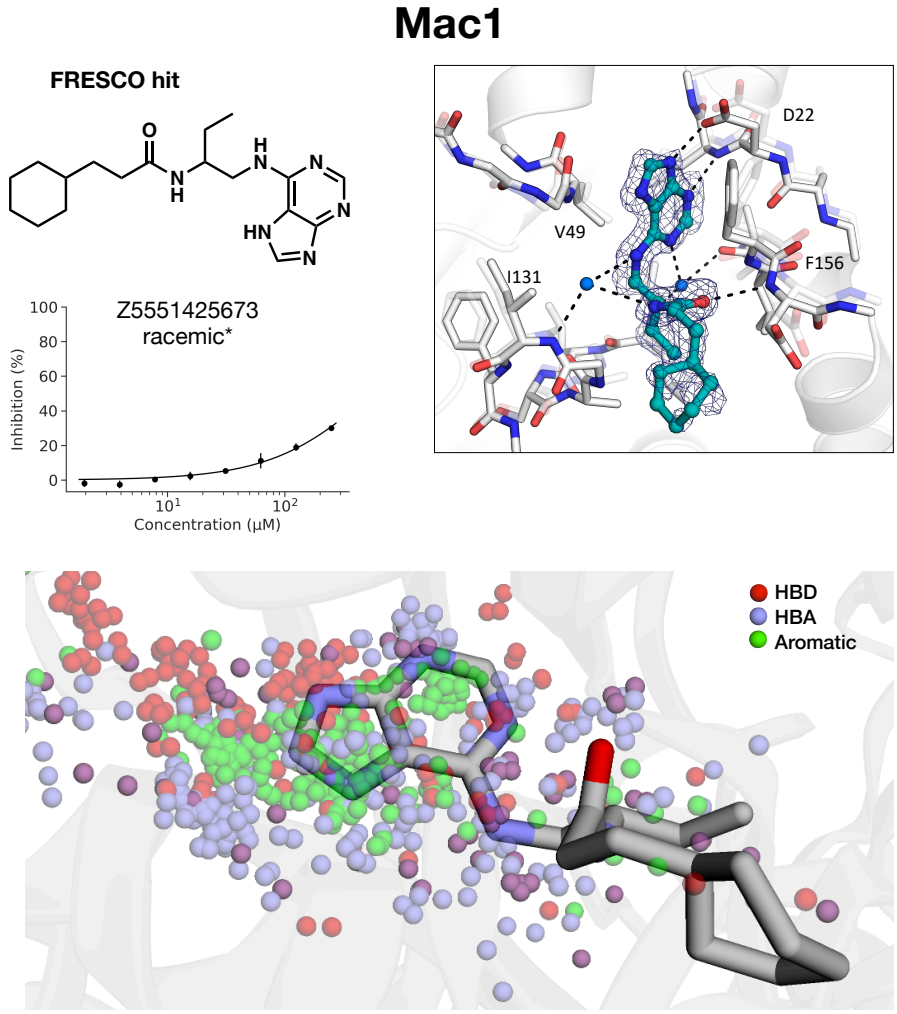


Fig. 3.7 (a) Compound Z5551425673 is identified as a hit against Mac1 via HTRF assay, with (b) hit confirmation via resolution of a crystal structure of Z5551425673 (colored in cyan) bound to the Mac1 active site. (c) The pharmacophores of Z5551425673 match those exhibited by the fragment hits as highlighted 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.

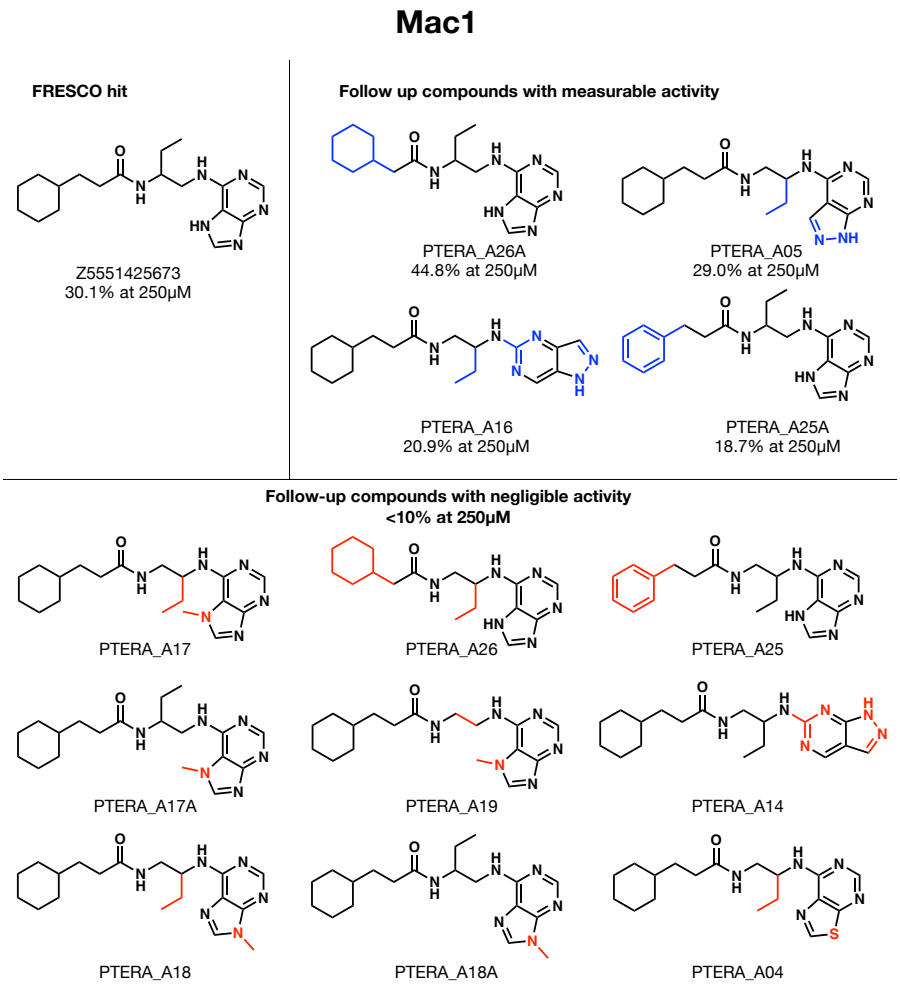


Fig. 3.8 Close analogues around the hit compound identified by FRESCO, Z5551425673, reveals structure-activity relationship which derisks singleton artefacts.

of molecules, still weak in potency, are potentially promising starting points for a hit expansion campaign.

3.4 Discussion

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 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.

Chapter 4

Discovery of SARS-CoV-2 main protease inhibitors via synthesis-directed de novo design

This chapter is based on 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

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 (Mpro) is one of the best characterized drug targets for direct-acting antivirals [**?** **?**]. Mpro is essential for viral replication and its binding site is distinct from known human proteases, thus inhibitors are unlikely to be toxic [**?** **?**]. Moreover, the high degree of conservation across different coronaviruses renders Mpro targeting a fruitful avenue towards pan-cornavirus antivirals [**?**]. To date, most reported Mpro inhibitors are peptidomimetics, covalent, or both [**?**]. 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 [**?**].

Here we report the prospective use of algorithmic *de novo* design to rapidly expand hits utilising machine learning (ML) models for ranking compounds by bioactivity as well as synthesis route prediction. Starting from 42 compounds with IC_{50} within assay dynamic range ($< 100\mu$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 Mpro. The top compound has comparable Mpro 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.

4.1 Learning to rank compounds

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 relatively small number of quantitative potency measurement bioactivity data and the measurement noise makes a regression approach challenging.

Instead of predicting IC50 values directly, we focused our attention on a learn-to-rank approach [**?** **?**] that predicts the pairwise comparison of ligands - given a pair of molecules (A,B), the model predicts whether A is more active than B. This approach allows us to assimilate both coarse (active/inactive) and fine (quantitative potency measurements) data into a single model, effectively combining 'easy' classification and 'difficult' regression into a 'moderate' task of ranking input pairs. With a trained ranking model, we can screen for more potent inhibitors by ranking new molecules against the most potent active compounds in the dataset.

To implement this ranking model (Figure **??**), we use the difference in molecular fingerprints between two molecules, $f_A - f_B$ as input to the model, and the output is the whether the molecule A is more or less potent than molecule B. The descriptor we use for representing a molecule is a concatenation of 3 512-bit fingerprint representations (Morgan, Atom, Topological Torsion) into one 1536 representation. A multilayer perceptron, implemented via FastAI Tabular framework [**?**] with default hyperparameters, is used for modelling the data. The choice of molecular descriptor and model was based on empirical performance (Details on the model implementation can be found in Appendix **??**).

To construct a suitable dataset for training the model, the activity data must be reformated into pairs of molecules. This is done by pairing all actives with all inactives, as well as pairing

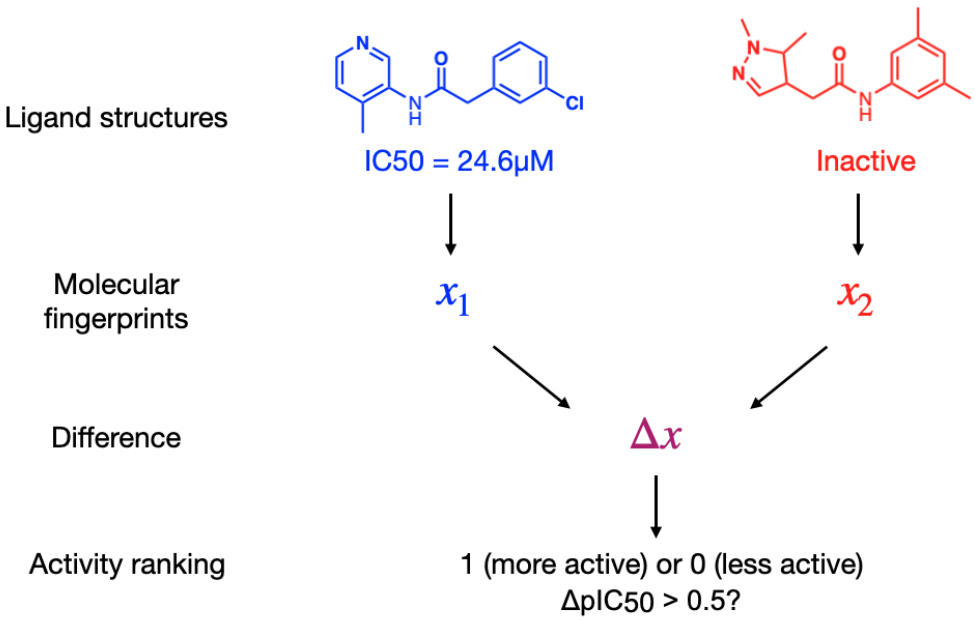


Fig. 4.1 A schematic of the model setup. A classifier takes the difference in molecular fingerprint between two molecules and predicts where one molecule is more or less active than the other.

up actives where there is a significant different in bioactivity (ΔpIC50 > 0.5). This threshold was chosen to match typical assay error, forcing the model to ignore irrelevant experimental noise by ensuring that it is ranking only ligands with demonstrably different bioactivity. The inactive molecules were not paired up between each other as the ranking of inactivity is not relevant to the task at hand and potentially noisy/misleading.

A flipped pairing is included for all pairs so that the dataset is antisymmetric since the model would just predict ones otherwise. Theoretically this method is appealing because it is a natural way of oversampling the low proportion of actives, addressing the problem of dataset imbalance commonly seen in drug discovery classification tasks. Additionally, creating pairs between the actives allows the exploitation of activity information without the noise/difficulty of trying to learn accurate pIC50 values.

For performance evaluation, the dataset was randomly split into training (80%) and testing (20%) sets (with roughly the same active/inactive proportion) before the molecules are paired up independently within each set. This ensures that there is no cross-talk between the train/test sets where the model could simply memorize the activity of certain compounds. We train the model on pairs of compounds within the training set and evaluate the model on both pairs within the test set as well as pairs between the training and test set.

Figure **??** 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])

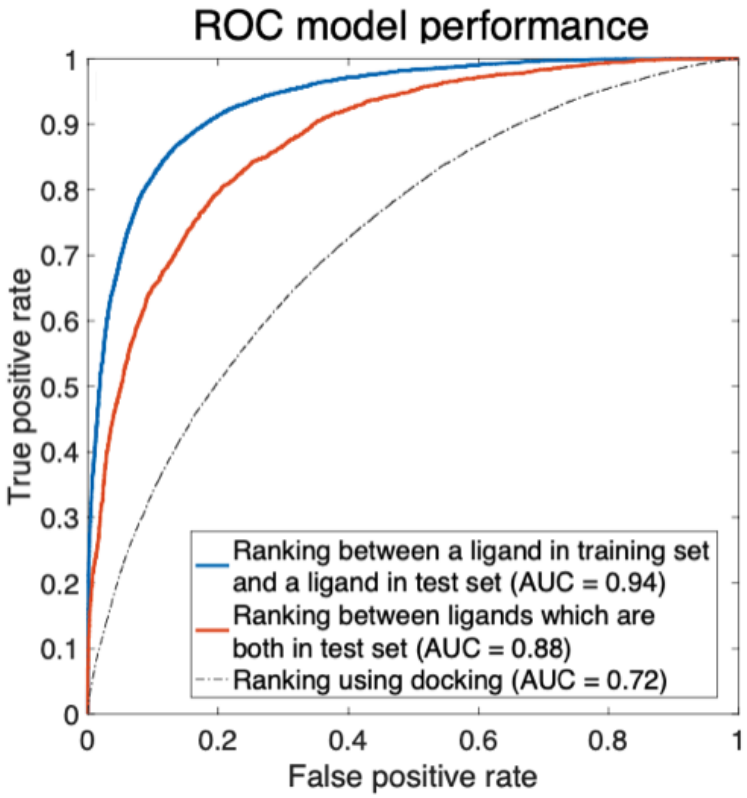


Fig. 4.2 Relative ranking of ligands can be predicted by our learning-to-rank machine learning model. The Receiver Operating Characteristic curve of classifying whether a molecule is more/less active than the other.

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 ranking compounds using docking scores generated with OpenEye's FRED docking algorithm, which achieves an AUC of 0.72 (95% CI: [0.722,0.723]) (Details of the docking procedure can be found in Appendix **??**). Note that docking does not require ligand bioactivity as training data, thus is not a directly comparison to machine learning.

Beyond train-test split, model performance can be evaluated from a time-split. Five months have elapsed from the time we deployed our model for prospective compound selection to writing up the this work. During that time, the COVID Moonshot Consortium (a team of expert medicinal chemists) has independently designed, synthesised and tested 356 compounds [**?**], out of which 15% were better than the top 2 compounds (having IC_{50} comparable within error) in our dataset. Table **??** 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 ∼2x higher than human selection.

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.

These retrospective results illustrate that a learning-to-rank approach can leverage bioactivity data from both active and inactive molecules for the enrichment of potent compounds in a real-world drug discovery campaign. In the next section, we deploy our model to discover new Mpro inhibitors in a prospective experiment.

4.2 Prospective chemical space exploration

After designing a well-founded ML scoring model, we must decide on a virtual library of compouds to explore. While one could screen an ultra-large library of make-on-demand compounds as in the previous chapter, it is only feasible for a relatively cheap computational model which is not the case for the ML model developed in this work.

Instead, we consider a more targeted approach by exploring the chemical space of chemical substructres contained within our initial dataset. Building on rule-based fragmentation methods such as BRICS [**?**] and CReM [**?**], our general approach is to decompose the existing molecules into a large set of distinct substructure components before enumerating all components with one another, generating a large number of novel and diverse molecules.

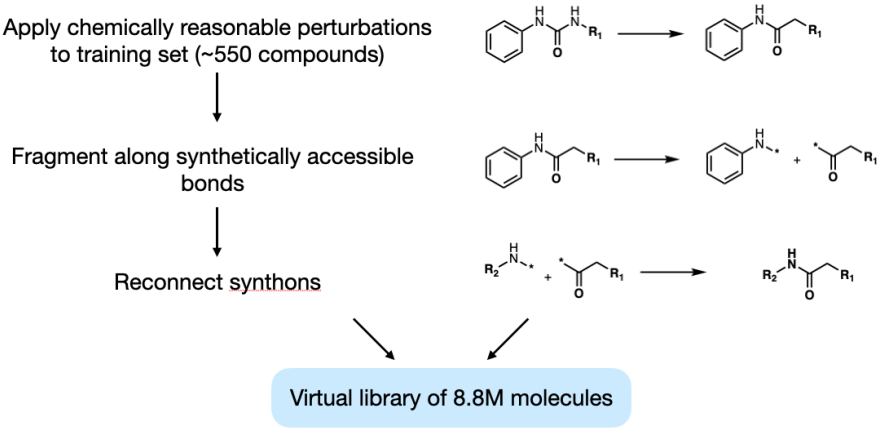


Fig. 4.3 A schematic of the methodology for the library generation process.

Specifically, we first introduce a set of chemically reasonable perturbations (linker and chemotype swaps, e.g. amide to retroamide, amide to urea, swapping N-aryl groups), 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 (Figure **??**). These operations are defined using SMARTS rules.

The resulting library of 8.8 million generated molecules is then scored against the top 3 compounds in the training set using the learning-to-rank framework, and the mean score is taken as the final score for each compound.

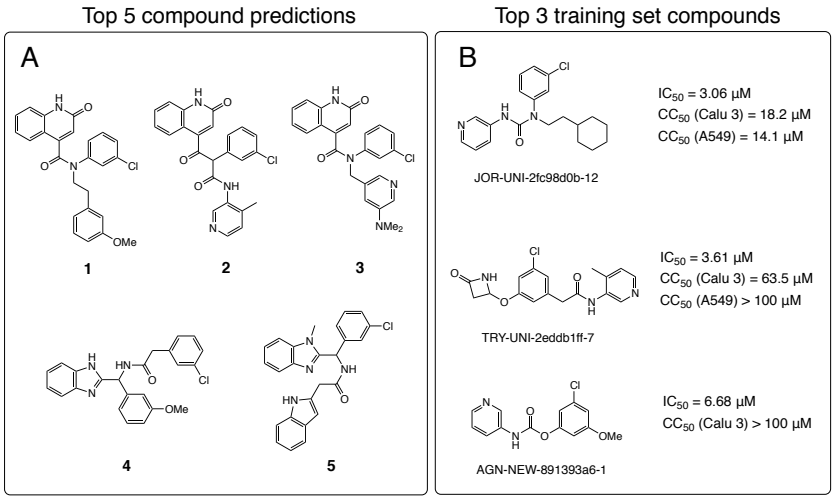


Fig. 4.4 Our synthesis-driven design model prioritises molecular scaffold that are not in the top hits. (A) The 5 compounds selected by our methodology for synthesis and testing. (B) The top 3 compounds from the training set, with potency and cytotoxicity measurements.

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 used Manifold, a platform for synthesis route prediction (https://postera.ai/manifold), to generate synthetic routes for the model's top-ranked molecules starting from purchasable building blocks. The underlying technology is based on Molecular Transformer, a machine learning model for reaction prediction using sequence-to-sequence translation [**?** **?**]. The top 5 molecules from the screening library with <4 steps in their predicted routes were synthesised and tested (Figure **??**A). For comparison, the most potent molecules from the training set are shown in Figure **??**B. All five compounds have Tanimoto similarity <0.48 (1024-bit ECFP6) to any molecule in the training set, indicating that the model is not merely reproducing molecules similar to the most potent actives but are exploring novel scaffolds.

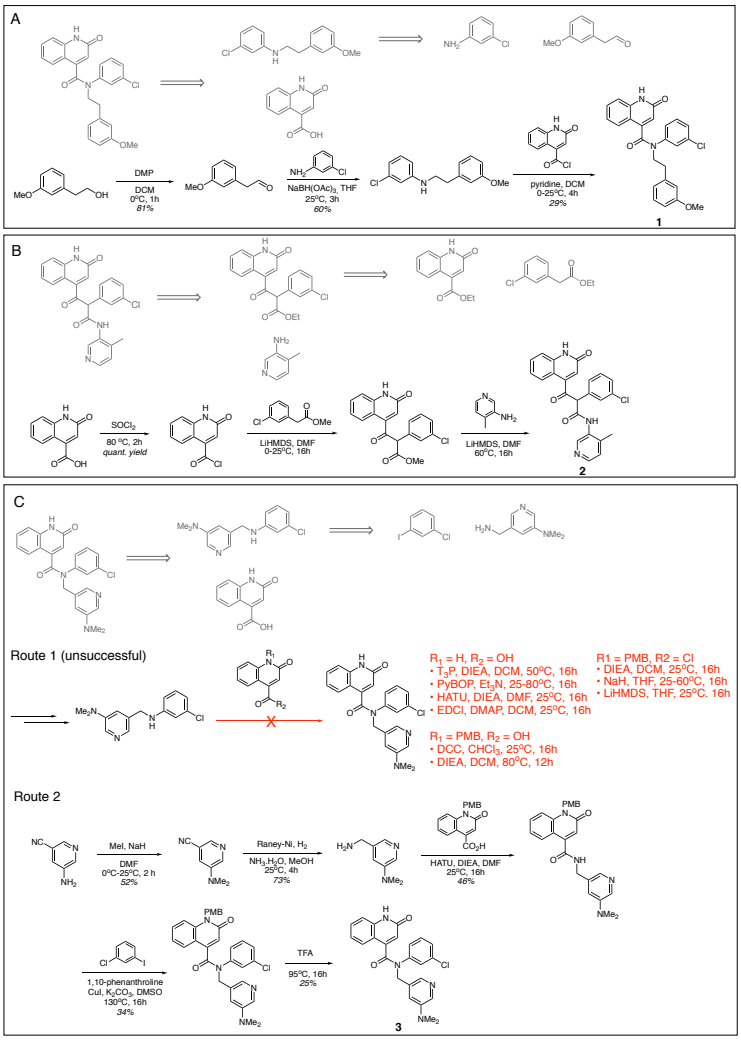


Fig. 4.5 Model generated synthetic schemes that are experimentally validated. Schemes (A)-(C) show the synthesis schemes generated by our model (grey) and experimental schemes (black) for Compounds **1**-**3**. The schemes for compounds **4** and **5** can be found in Appendix **??**.

Figure **??** 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.

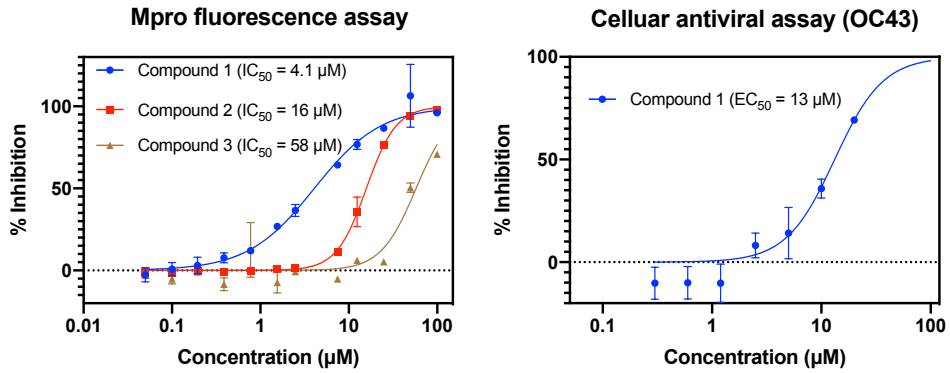


Fig. 4.6 Three compounds generated using our synthesis-directed model exhibit Mpro activity. Our most active compound has measurable antiviral activity against the OC43 coronavirus and no measurable cytotoxic effect.

Compounds **1**-**5** were tested for Mpro activity using a fluorescence assay. Figure **??** shows that Compounds **1**-**3** have IC_{50} within assay dynamic range ($< 100\mu M$), and Compound **1** has $IC_{50} = 4.1\mu M$ (95% CI: [3.42,4.86]). Compound **1** is further assayed in live virus assays, with the less pathogenic OC43 coronavirus, showing $EC_{50} = 13\mu M$ (95% CI: [10.1, 18.4]) and is not cytotoxic ($CC_{50} > 100\mu M$ 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 (See Appendix **??** for assay details).

4.3 Discussion

In summary, we demonstrated the utility of a *de novo* design framework that learns to rank bioactivity and estimates synthetic complexity, for generating ideas in hit expansion. At the time of writing, optimisation of a quinolone-based scaffold is ongoing in 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.

The learning-to-rank approach presented in this chapter is a promising technique for maximally utilising data from inactive compounds as well as accounting for noise in the experimental values. Further work extending this approach to utilise information from ligand-protein complexes has shown great success in hit-to-lead optimisation [**?**]. By using docking-based structural descriptors as the input to their ranking model, the authors were able to utilise crystal structures of inactive compounds and outperform docking as well as a fingerprint-based ranking model on ranking ligand activity. Applying this model on a docked structures from a virtual library, the authors were able to greatly improve the potency from their starting point and

extended the ligand into unknown regions of the binding site. This result shows the potential for applying ranking-based approaches for modelling bioactivity.

A caveat for pair-based ranking is that even for moderately sized datasets the number of molecular pairs (which is proportional to the square of the dataset size) may become very large and therefore unfeasible to train on. Engineering approaches for efficient sampling of molecular pairs, such as using Tanimoto similarity to constrain pair selection, may be necessary and should be the subject of future work.

The AI-based estimation of synthetic feasibility played a critical role in compound selection in this work and is a promising enabler for automated design workflows in drug discovery [**?**]. Beyond the capability to triage libraries by synthesisability and to guide the design of synthetic routes, ML synthesis models could play a key part in an automated drug design workflow. Utilising an automated workflow for the automatic generation of molecular designs as well as decision-making could potentially reduce iteration cycle time, require fewer compounds and iterations to produce a candidate, and scale to more programs [**?** **?**]. The framework presented in this chapter is an example of an automated design workflow albeit only with a single iteration - utilisation of multiple iterations via an optimisation [**?**] or reinforcement learning feedback loop [**?**] will be a challenging but exciting area of future work.

Chapter 5

Make - Understanding the Molecular Transformer

This chapter is based on 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)

Organic synthesis remains a a major challenge in drug discovery. Although a plethora of machine learning models have been proposed as solutions in the literature, 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. This opaqueness hinders both model developers and users. In this chapter, we quantitatively interpret the Molecular Transformer, the state-of-the-art model for reaction prediction. We develop a framework to attribute predicted reaction outcomes both to specific parts of reactants, and to reactions in the training set. Furthermore, we demonstrate how to retrieve evidence for predicted reaction outcomes, and understand counterintuitive predictions by scrutinising the data. Additionally, we identify Clever Hans predictions where the correct prediction is reached for the wrong reason due to dataset bias. We present a new debiased dataset that provides a more realistic assessment of model performance, which we propose as the new standard benchmark for comparing reaction prediction models.

5.1 Introduction

Organic synthesis remains a challenge in small molecule drug design, sinking time in the design-make-test cycle and potentially limiting the complexity of chemical space being explored [**? ?**]. The challenge of synthesis planning lies in searching through myriad of possible reactions to find optimal routes, and in predicting whether each possible reaction is indeed feasible and high yielding for the particular substrate in question. The problem of efficient search in synthesis has been recently addressed, inspired by innovations in computer science on searching and gameplay [**? ? ? ? ?**]. However, accurately predicting the outcome of chemical reactions remains a hurdle [**? ? ?**].

The current state-of-the-art in reaction prediction is the Molecular Transformer [**?**], which employs the transformer neural network architecture that was first introduced for neural machine translation [**?**]. The input to the model is a text representation of the chemical structures of the reactant and reagent, and the model performs machine translation to predict most likely output molecule with a probability score. The Molecular Transformer achieves a 90% Top-1 accuracy on the USPTO dataset of organic reactions that was text mined from US patents [**?**] and filtered [**?**]. Recent work shows that thorough dataset augmentation improves model performance by allowing it to consider different equivalent SMILES representations [**?**].

However, a key stumbling block in the Molecular Transformer is the lack of interpretability. Why the Molecular Transformer predicts one reaction outcome over another, and which training set reactions it finds most similar when reaching a particular prediction, are both unclear. Quantitative interpretability is crucial to both model users and model developers.

For model users, interpretability is important because chemical reactions are highly contextual, with important anthropomorphic metadata that the model overlooks. For example, reactants, reagents and products are only a part of the reaction. The reaction conditions, the scale of a particular reaction (e.g. discovery chemistry or scale up), and scientific focus of the project (e.g. total synthesis, medicinal chemistry or methods development) are some of the context that a skilled chemist can employ to interpret and understand the reaction.

For model developers, physical organic chemistry principles explain chemical reactivity and selectivity. As such, probing whether rationales outputted by the Molecular Transformer are congruent with physics allows developers to interrogate whether the Molecular Transformer is getting the correct prediction for the right reasons, and design model improvements based on those insights.

In this chapter, we develop a suite of methods that quantitatively interprets the Molecular Transformer by attributing predictions to the input chemical structure and the training data. We illustrate our two-prong approach via a series of examples, showing how we uncovered what the model is learning, what it finds difficult, and explains its failure modes. Our method

discovers hidden biases in the training data that hinder generalization performance and masks model shortcomings, which we resolved by introducing a new unbiased train / test split.

5.2 Molecular Transforer

The Molecular Transformer [**?**] is a tailored version of the Transformer architecture [**?**] 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. An in-depth review of the attention mechanism can be found in [**?**].

In Transformer models for text, we first break down the string input to the model into individual tokens and generate a learnt vector embedding for each token. To include the relative order of the tokens and thus distinguish tokens of the same type at different positions, an additional vector is generated based on the token positions (known as the positional encoding) and added to the token embeddings. This sequence of token embeddings is then input into the encoder and decoder layers. The encoder layer is responsible for encoding the input sequence into an informative vector representation to input into the decoder layer. The decoder layer reads the output from the encoder layer and generates the output sequence. During model training, we train the decoder layer to predict the next token in the sequence given the previous tokens in the sequence. During model inference, 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 a pre-specified 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 tok_i is the i-th predicted token and N is the length of the prediction.

The Molecular Transformer (Fig **??**) uses this approach to perform reaction prediction by using tokenized SMILES of reaction reagents and reactants as input, and the tokenized SMILES of the reaction product as output. Typically, the SMILES of the reaction product are canonicalized while both canonical as well as non-canonical SMILES are used for the reactants and reagents as that has been shown to improve model performance [**?**] compared to only

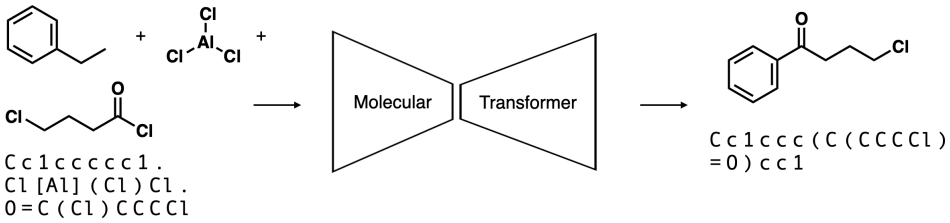


Fig. 5.1 **Schematic illustration of the Molecular Transformer.** The inputs to the model are tokenized SMILES of the reactants and reagenst, and the model performs machine translation to predict the most likely product molecule with a probability score.

using canonical SMILES. In addition, no distinction is made between "reactant" and "reagent" and the model is only trained on reactions with a single reaction product.

5.2.1 Training data

The training data used in this study was obtained by the text mining work of Lowe [**?**], where organic reactions from US patents filed between 1976 and 2016 were extracted. The dataset was filtered by Jin et. al. [**?**] to remove duplicates and some of the erroneous reactions which resulted in a set of ca 480 000 organic reactions. We found that this dataset, though much cleaner, still contained a large number of erroneous reactions whose sole product were halogen ions, nitric, sulphuric or phosphoric acids etc. We verified that the Molecular Transformer indeed learns these reactions if they are present in the training, resulting in catastrophic overfitting and unphysical predictions in some cases. To eliminate this effect we deleted a further 8 000 reactions to obtain a dataset of 471 791 reactions. From this number we used 377 419 for training, 23 589 for validation and 70 765 as a hold-out test set. The training set was augmented by an equal number of random equivalent SMILES strings following the protocol of Schwaller et. al. [**?**]. We trained a Molecular Transformer model as described in the original paper and were able to achieve 88.8% Top-1 accuracy on the test set, similarly to the original paper. This model was used throughout the interpretability experiments.

An important aspect of the training data is that since it was extracted from patented reactions, it naturally contains a number of biases. Firstly there are no negative results included meaning that any combination of reactants and reagents in the dataset leads to a well defined product. This is in contrast to reality where often there is no reaction, or the product is a mixture of many different compounds. This bias will always be reflected in the machine learning models predictions. A further bias stems from the distribution of reaction types in the dataset. Most of the patented reactions come from the medicinal chemistry community leading to reactions popular amongst medicinal chemists being over-represented. This bias can be useful since

the model learns the kind of reactions medicinal chemists like using [**?**] but it also hinders generalization because popular reactions are not necessarily better as has recently been shown in the case of inorganic chemical reactions [**?**].

5.3 Quantitative Interpretation methods

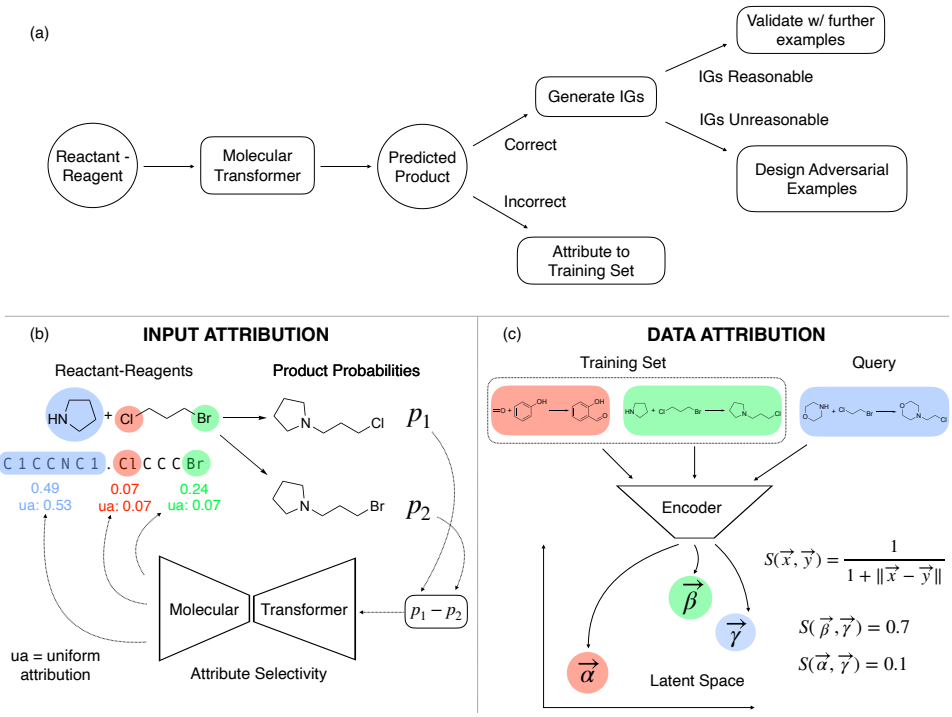


Fig. 5.2 **Schematic illustration of the attribution workflow.** (**a**) Overview of our workflow to interpret the Molecular Transformer. (**b**) Schematic of how the predicted probability difference between two products are attributed back to the reactant-reagent string in order to interpret the model's understanding of selectivity. The IG attributions below the reactant SMILES are compared to the uniformly distributed probability difference (ua) below. (**c**) Schematic of how the latent space encoding of reactant-reagent strings are used to infer the learnt similarity between query reactants and those from the training set.

There are three key factors determining the prediction of a machine learning model: the architecture, the training data and the input. Neural network models are often considered as black-boxes because of the complex ways these three factors interact to yield a prediction.

To interpret model prediction, we first need to define what interpretability means. We suggest interpretability is the ability to discover associations and counterfactuals between input and output, and the ability to query evidence in the data supporting a certain outcome. Our approach follows the accepted scientific process: A scientific theory usually identifies factors

that are related to a certain outcome and conversely how the absence of those factors is related to absence of outcome. Furthermore the investigator needs to show pieces of evidence that support the theory.

We employ Integrated Gradients [**?**] as a rigorous method for attributing the predicted probability difference of two plausible products of a selective chemical reaction to parts of the input. The attributions show how much each substructure is contributing to the predicted selectivity of the model. This is illustrated on Figure **??**b. The values of the attributions are compared to the value each subgroup would receive if the probability difference would be distributed evenly across the input. The parts of the structures getting higher IGs than the uniform attribution (ua) are considered important. Details of the methods can be found in Section **??**.

Attributing the predictions of neural networks to most similar training data points is less widely researched. To achieve this goal we developed a new method based on the latent space similarity of the reactions. We used the outputs of the Molecular Transformer encoder averaged over the tokens to achieve a fixed length vector representation of the reactions. The most similar training reactions according the model were than identified using the Euclidean distance of these latent space vectors. A schematic overview of our method is shown on Figure **??**c. Details of the methods can be found in Section **??**.

We validate our interpretations in two ways. The first is via falsification. If the integrated gradients attributions are chemically unreasonable, i.e. predictions are correct for the wrong reasons, we design adversarial examples that force the model into wrong predictions. The second is by identifying causes for the prediction in the training data. If a prediction is wrong, we interrogate whether a similarly incorrect entry is in the training data.

5.3.1 Input attribution

To unpack the Molecular Transformer we decided to focus our efforts on reactions containing selective chemical transformations which means that they have multiple plausible outcomes. These reactions are most fit for identifying if the model is making the predictions on true chemical basis because the underlying chemical causes are well established. Our general framework of interpreting chemical reactions is shown on Figure **??**a.

Once a suitable chemical reaction with two possible target molecules is chosen the Molecular Transformer probability scores of the products are generated. The difference in probability score between the true and the incorrect but plausible products is than attributed back to the reactant reagent inputs.

Recently there were many methods developed and applied successfully for attributing the predictions of neural networks to parts of the input. Some of the most notable examples are

LIME, SHAP, Layer-wise Relevance Propagation (LRP) and Integrated Gradients [**?** **?** **?** **?**]. These methods are designed to propagate back the output of the models in a fair way to determine the contribution (importance) of each of the input features to the prediction. There are several methods that have their roots in cooperative game theory and are proven to yield fair attributions as defined by the axioms of fairness [**?**]. For machine learning models where the gradients are not readily available, there are so called Shapley-values and the closely related SHAP method [**?**]. For models such as the Transformer where the gradients are easy to evaluate the Integrated Gradients (IGs) method is a more natural choice [**?**] though other methods such as LRP have also been applied successfully [**?**]. The IGs method has also been applied previously for interpreting language models in natural language processing applications and for designing adversarial examples in the context of question answering [**?**]. A graphical illustration of IGs is shown on Figure **??**b. Our approach builds on the work of McCloskey et. al. [**?**] who used IGs to understand binding prediction by graph neural networks on artificial datasets. We extend the method to Transformer architectures, and use it in the context of reaction predictions on real experimental data.

IGs are calculated by evaluating the path integral of the gradient of the output with respect to the input along a straight line path in the input space from a non-informative baseline to the input of interest.

Given a neural network denoted by the function $F : \mathbb{R}^n \to [0,1]$, the input $x \in \mathbb{R}^n$ and the baseline input $x' \in \mathbb{R}^n$ the IG attribution of feature i is given by

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

In the case of the Molecular Transformer x is the $N \times 256$ dimensional embedding of the input SMILES string of length N and x' is the embedding of the '.' token taken N times. This token is used in the SMILES language to separate different molecules and hence on its own bears no chemical information making it an ideal baseline choice. To obtain the total contribution of each of the input tokens the attributions are summed along the 256 dimensional embedding vectors.

Finally to make the attributions easier to interpret we devised a few simple rules to map the token level attributions to chemically meaningful substructures. Reagents like sulphuric acid or meta-Chloroperoxybenzoic acid (mCPBA) are fed into the model by their full SMILES strings but in reality they act as single units as far as the reaction is concerned. Their attributions are more meaningful to look at as a whole rather than token by token. A related problem is with the attributions corresponding to special characters in SMILES like numbers or parentheses. To resolve this we consider rings as single units and their attribution is calculated by summing over the ring atoms and numbers. This way the information about the relative positions of the

ring substituents will also be included in the attribution of this part of the structure. Branches are also considered as single units and their attribution is the sum over their atoms and the parentheses specifying them.

For the attributions to be meaningful it is important to look at reactions where there are two possible products which have non-zero probability scores according to the model. This is crucial since for the prediction of a single product every token of the reactant is important, since missing a remote carbon would also result in a wrong prediction. By looking at the probability difference of two plausible products this effect can be eliminated and the attributions highlight the groups driving the chemical selectivity (according to the model). In particular, canonical SMILES for both products should be used to ensure the probability scores are non-negligible.

Finally, to determine if a particular group is important according to the model we compare its attribution to the attribution that would fall onto it, if the probability difference was distributed evenly across the input tokens. Substructures that get substantially higher attribution than uniform are most important for the model when it favours one product over the other.

5.3.2 Training data attribution

Attributing the predictions of neural networks to training data can serve as a tool for explaining predictions as well as gaining understanding of the models inner workings [**?**]. 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 see a couple of example inputs that the model finds similar. Usually seeing a number of similar examples can help humans identify patterns that may serve as the basis of the model's prediction. This can either result in the discovery of new trends or laws in the scientific domain or it can reveal biases that the model has learnt. In the latter case this information can be used to improve the model or the dataset.

To create a successful method for attribution to data the most crucial element is the careful design of a similarity measure. The similarity should be defined such that it measures how similar two input datapoints are according to the model. For different neural network architectures different choices of similarity measures can be appropriate. In the case of feed-forward or convolutional architectures a natural choice is to define a fingerprint vector for each data point that consists of the neural networks layer outputs (activations) concatenated together. This similarity measure has been shown to be useful for judging the reliability of toxicity models predictions by comparing molecules not in the training set [**?**].

In the case of the Molecular Transformer which has an encoder-decoder architecture the output of the encoder layers can be used as a basis for comparing data points. Since the encoder hidden states have a non-fixed length we take the average of them across the input tokens to

obtain a fixed-length 256 dimensional vector representation for each of the reactions. Averaging is expected to work because of the relatively large dimensionality of the latent space. The size of the vocabulary of the USPTO dataset is 288 so there are almost as many orthogonal directions in the latent space as there are possible different input tokens. This is expected to lead to minimal loss of information upon averaging. For each reaction in the training set the 256 dimensional hidden state vector is generated and the matrix of training set reaction hidden states is saved as a binary. When a new example input is given to the model it is passed through the Transformer encoder and the average hidden state vector of it is calculated. A schematic diagram depicting the method is shown in (Figure **??**b). The similarity score of the input reaction vector $\mathbf{u}$ to a training set vector $\mathbf{v}$ is calculated by

$$\text{score}(\mathbf{u},\mathbf{v}) = \frac{1}{1+\|\mathbf{u}-\mathbf{v}\|} \tag{5.3}$$

In this work we inspect the top-5 most similar reactions from the training set for comparison with the input reaction.

5.4 Investigation of Specific Reaction Classes

We investigate in detail three reaction classes that are commonly used in medicinal chemistry. Through these examples we illustrate each of the three branches in Figure **??**(a). We first examine the selective epoxidation of alkenes where the Molecular Transformer produces the right prediction for the right reason.

We then turn to the Diels-Alder reaction, which is a scaffold-building transformation widely used in synthesis. We show that the Molecular Transformer is not able to correctly predict this reaction. Following the bottom branch of Figure **??**a we investigate it using Data attribution and find that the USPTO dataset contains very few instances of Diels-Alder reactions, likely explaining why the model is not able to predict the outcome correctly.

Finally we consider the Friedel-Crafts acylation reactions of substituted benzenes. We show that the Molecular Transformer predicts the right product for the wrong reason and validate our interpretation using a number of adversarial examples.

5.4.1 Epoxidation

The oxidation of alkenes to form epoxides is an important intermediate reaction in many synthesis plans [**?**]. The common oxidant in these reactions are peroxy compounds. The most widely used example of them is mCPBA, which is a versatile reagent appearing 2052 times in

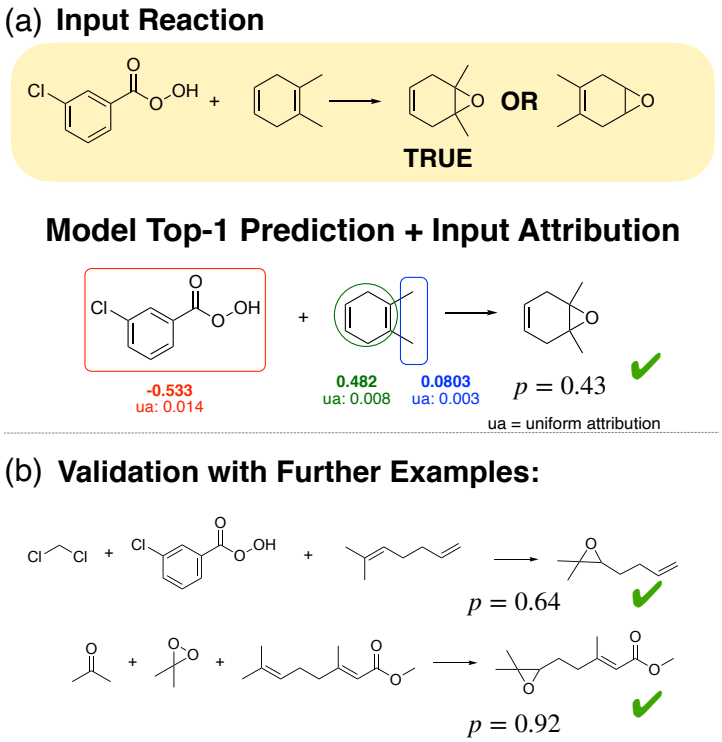


Fig. 5.3 **IG attributions highlight correct model reasoning.** (**a**) The model correctly predicts the product of a typical epoxidation reaction, and shows significant positive attributions to the two methyl group that are responsible for the selectivity. (**b**) We validate the model's knowledge on two unseen epoxidation reactions from chemical literature [**?**]

the USPTO dataset. This is in the high data regime where we would expect the model to do well due to the large number of different training examples available.

Epoxidation reactions can be regioselective, with more substituted alkenes reacting faster because they are more electron-rich [**?**]. A typical example reaction showing this type of selectivity is shown on Figure **??**a.

The Molecular Transformer is able to predict the product with the correct selectivity, giving it a probability score of 0.43. The probability score of the alternative incorrect product was less only by 0.025. This is a case where the model predicts two similarly plausible outcomes, so IGs can help to judge whether or not a prediction can be trusted. Since the probability difference is close to 0, the sign of the attributions at different parts of the input is in itself interesting and contains information regarding the favoured outcome.

Figure **??**a shows the IG attributions of the different parts of the input. In this case the positive attributions favour the correct product while the negative attributions favour the incorrect product. The IGs show that the two methyl substituents circled with blue are significantly contributing to the correctly predicted selectivity. The attributions on the other

parts of the molecule are harder to interpret. This can be the result of the model being uncertain in the prediction leading to larger gradients along the path integral during the calculation of the attributions.

To validate the interpretation that the model has learnt this selectivity we generated the Molecular Transformer predictions for two further examples from the literature as shown on Figure **??**b. The first example is very similar to the one examined in detail above and the model is consistently predicting the correct product. The second example is more challenging for the model for a number of reasons. First the reagent is not mCPBA but dimethyldioxirane which appears much less frequently, only 14 times in the training data, secondly both double bonds are substituted, and the difference is made by a more subtle chemistry, the ester group being electron withdrawing. The model is able to predict the correct outcome here as well confirming that the predictions are correct for the right reason.

5.4.2 Diels-Alder

The Diels-Alder reaction transforms a conjugated diene and an alkene (called dienophile) to a six membered ring with a double bond [**?**]. There are very few limitations on the character of the diene. It only has to be flexible enough to take up an s-cis conformation. The dienophile, on the other hand, should have carbon-carbon double bonds conjugated preferably with an electron withdrawing group. A typical example of a Diels-Alder reaction used as a test-case is shown on Figure **??**a.

The Molecular Transformer was unable to predict the regioselectivity of this reaction, and in fact the predicted product was clearly wrong with the actual possible products getting 0 probability scores. Since the prediction is obviously wrong, we followed the bottom branch of the workflow at Figure **??**a and generated the most similar training reactions to see what causes this erroneous prediction.

Figure **??**b shows the Top-3 most similar reactions from the training set based on the model encoder output similarities. The most similar training reaction (i) is an erroneous reaction, whilst the second and third are carbon-carbon bond formations, but via Grubbs methathesis [**?**] rather than cycloadditions. This means that the model has not learnt a good representation of Diels-Alder reactions in the latent space.

To investigate if the cause of this was a lack of training data we devised a reaction template corresponding to the [4+2] cycloaddition and found that there were only 7 reactions matching it in the entire USPTO database. This example illustrates how attribution to data can be useful for identifying erroneous predictions caused partly due to erroneous data and partly due to the scarcity of training examples.

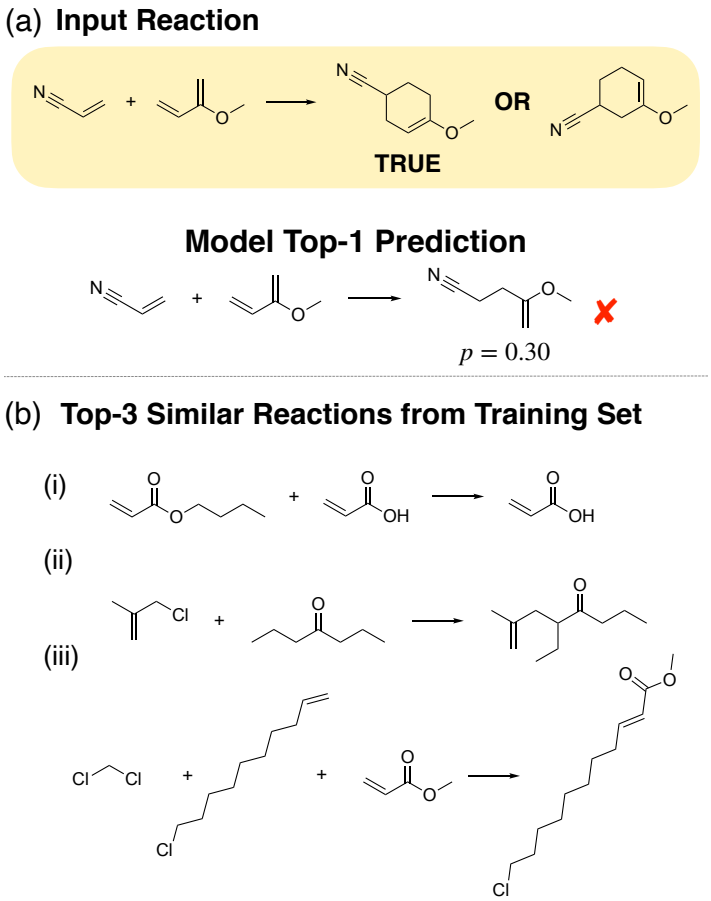


Fig. 5.4 **Data attribution explains erroneous predictions.** (**a**) The model makes an obviously incorrect prediction on a typical example of a Diels-Alder reaction with challenging selectivity. (**b**) Attribution to the USPTO training data shows that the model either completely fails to recognize Diels-Alder reactions or that no Diels-Alder reaction is present in the dataset.

5.4.3 Friedel-Crafts Acylation

Friedel-Crafts acylation reactions are an example of electrophilic aromatic substitution [**?**]. In these reactions a hydrogen on an aromatic ring is substituted by an acyl group. In the case of a benzene ring with a single substituent, there are three different hydrogen positions where this substitution can happen. The electronic and steric character of the substituent on the ring determine the selectivity of these reactions. An example of a selective Friedel-Crafts reaction is shown on Figure **??**(a) where according to the patent the para product is formed with a yield of 90% [**?**]. In this reaction that acyl group is primarily substituting the hydrogen in the para position compared to the -F substituent. The transformation is correctly predicted by the Molecular Transformer.

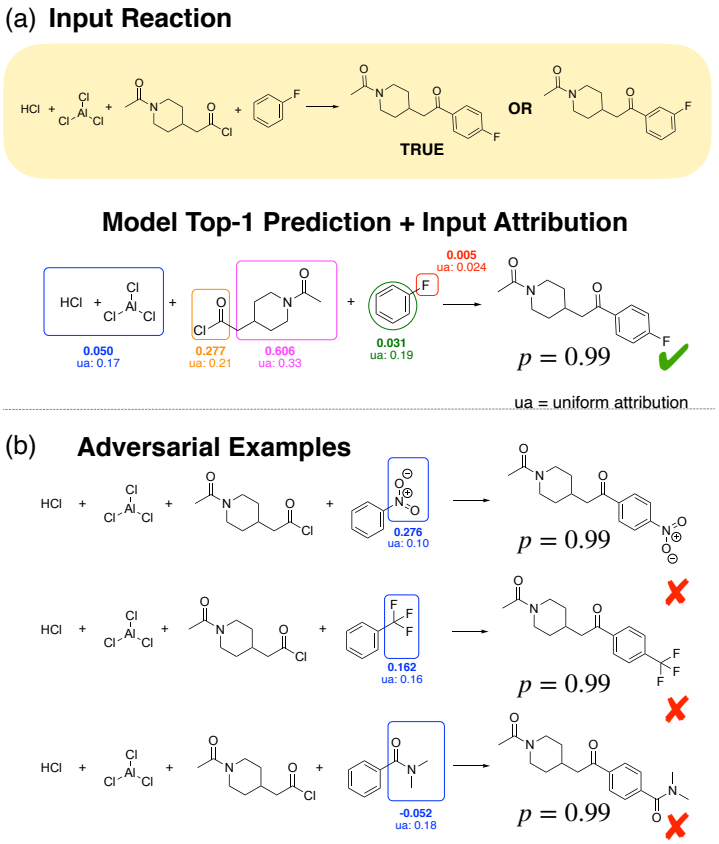


Fig. 5.5 **IG attributions reveal incorrect reasoning and guide the design of adversarial examples.** (**a**) The model correctly predicts the major para product of a typical Friedel-Crafts acylation, but low attribution is given to the para-directing -F group. (**b**) The model is fooled into incorrectly predicting the para product when the -F is replaced by meta-directing groups. The low attributions given to the directing groups indicate that the model has not learnt their importance.

The IG attributions indicate that the importance of the fluorine (-F) for this reaction is completely neglected by the model. A much larger attribution is given to the reagent suggesting that the model attributes this selectivity to the reagent rather than the true directing group. Guided by the attributions we replaced the fluorine by a number of typical meta directing groups to create adversarial examples. We observe that the model (wrongly) predicts the para product. In this case negative attributions favour the meta product and positive attributions the para product. We do not find any correlation between the attribution values and the directing effect of the substituent. From this we can conclude that the model has not learnt the selectivity in the case of Friedel-Crafts acylation reactions on substituted benzene rings.

5.5 Revealing the Effect of Bias through Artificial Datasets

Interestingly in one of the adversarial examples the attribution on the meta directing group is negative, meaning that according to the model the amide group (correctly) favours the formation of the meta product. This agrees with chemical principles, but the model is nonetheless still predicting the para to be the major product. We hypothesize that this might be due to biases in the training data – using template analysis to count the number of para/meta/ortho Friedel-Crafts Acylations in USPTO (Figure **??**), we find that there are many more para substitution reactions than meta in the training dataset. This could result in the model being biased towards predicting para substitutions even in the presence of meta directing groups, as the model can achieve very high ($\sim 98\%$) accuracy on the training set by always predicting the para product.

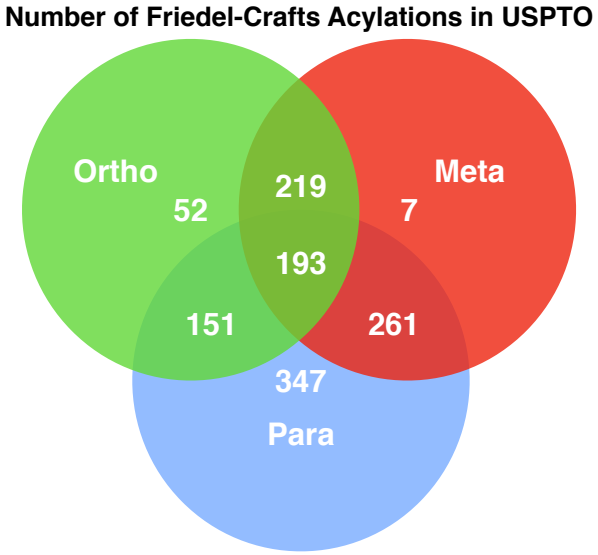


Fig. 5.6 **The number of para Friedel-Crafts acylation reactions in USPTO far outweigh those of meta or ortho reactions.** Overlaps in the Venn diagram denotes cases where the benzene has more than 1 substituent.

To more quantitatively investigate how this imbalance in training data affects the model predictions, we train the Molecular Transformer on artificial datasets with varying proportions of meta and para Friedel-Crafts substitutions. By comparing the performance of the trained models, we demonstrate that the unchemical para-favouring behaviour of the USPTO-trained Molecular Transformer is the result of dataset bias.

5.5.1 Artificial dataset construction

We generate three sets of artificial training data and one held-out artificial test set of electrophilic aromatic substitution reactions using SMART templates. Each reaction consists of a benzene

ring singly substituted with a directing group reacting with an acyl chloride to form either a para- or meta- acylated product.

Ten para directing groups (fluorobenzene, chlorobenzene, isopropylbenzene, tert-butylbenzene, N-phenylacetamide, N-phenylpropionamide, phenol, ethoxybenzene, isopropoxybenzene, sec-butylbenzene) and ten meta directing groups (N,N,N-trimethylbenzenaminium, (trifluoromethyl)benzene, benzaldehyde, acetophenone, methyl benzoate, ethyl benzoate, benzonitrile, nitrobenzene, methyl benzenesulfonate, ethyl benzenesulfonate) were used.

The -R groups for the acyl chlorides were generated by enumerating straight carbon chains of length 2-8 with 0-1 C=C double bonds also using SMARTS templates. Acyl chlorides were obtained by placing an acyl chloride group onto a random sp3 carbon on each of the -R groups. The acyl chlorides are enumerated with the benzyl compounds to generate valid chemical reactions.

To investigate the effect of dataset bias, we vary the proportion of para:meta reactions in the training dataset and observe how the Molecular Transformer performs on a test set with an 1:1 proportion of para:meta reactions (Table **??**). We first construct a 'Balanced' dataset which has a 1:1 ratio of para:meta reactions (3100:3100) by enumerating all acyl chlorides with all benzyl compounds. We also create a 'Biased' dataset which has a 9:1 para:meta ratio (2790:310) by performing a 10:1 random split on the acyl chlorides so that less meta reactions are present. Finally we generate a 'Severely Biased' dataset with 100:1 para:meta ratio (3000:30), which is closest to the observed ratio in USPTO, by performing a 33:1 random split on the acyl chlorides and also only keeping three meta-directing benzyl compounds (benzaldehyde, (trifluoromethyl)benzene, and nitrobenzene).

Table 5.1 **Number of meta-/para-directing reactions in the artificial datasets.**

	Meta	**Para**
Balanced (training)	3100	3100
Biased (training)	310	2790
Severely-Biased (training)	30	3000
Test set	177	177

The test set has an equal proportion of para and meta reactions generated using the three meta directing benzyl compounds from the 'Severely Biased' training set and three para directing ones (Fluorobenzene, N-phenylpropionamide, and ethoxybenzene), together with -R groups from enumerating straight carbon chains of length 9-10 with no double bonds. This resulted in a test set with 177 para and 177 meta reactions.

5.5.2 Model performance on artificial datasets

On USPTO the Molecular Transformer was trained for ∼300 epochs, so for fair comparison this is the regime we wanted to investigate with the artificial datasets. We trained 10 transformer models on each of three training sets and saved checkpoints from the beginning of model training up to 256 epochs.

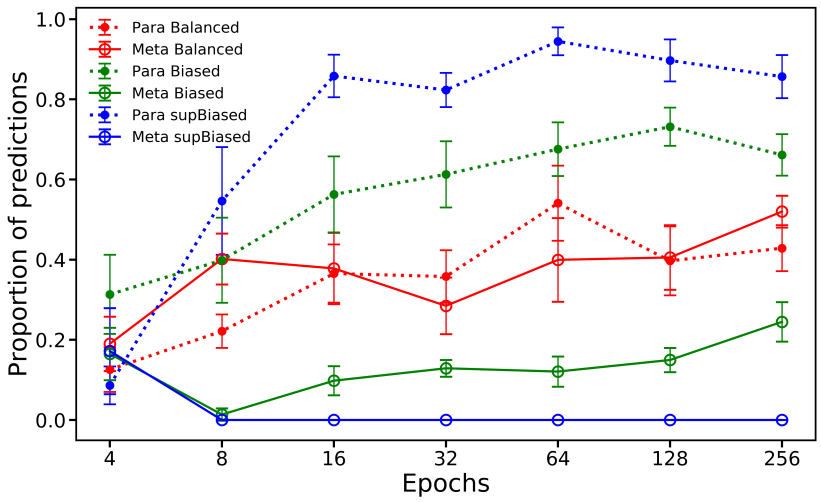


Fig. 5.7 **Biased training data leads to biased predictions from the Molecular Transformer.** The figure shows the proportion of para (solid line) and meta (dashed line) predictions on a balanced test set as a function of the number of training epochs for different biased training sets. The error bars shown indicate the standard deviation in the results from training an ensemble of 10 randomly initialized models. The proportion of meta and para predictions does not always add up to 1, because it takes a number of iterations for the model to learn the SMILES syntax and we discount invalid predictions.

Using SMARTS template matching, we measured the proportion of model predictions (with valid SMILES) that are meta and para as a function of the number of epochs for different dataset biases (Figur **??**). The results show that the Molecular Transformer is highly susceptible to learning dataset bias. When the model is trained on the balanced dataset, it rapidly converges to predicting equal amounts of para and meta substitution reactions, confirming that the bias is not caused by neural network architecture limitations. The model trained on the biased dataset containing only 10% meta reactions in the training set is not able to get rid of the bias. For the severely biased training set (where the proportion of para/meta reactions are closest to the observed ratio in USPTO) the model does not predict any meta products at all.

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 severely-biased datasets. This shows that by training longer the effect of dataset bias can be mitigated, but it cannot be removed altogether.

This numerical experiment confirms that the Molecular Transformer is guilty of the Clever Hans effect – it appears to know chemical reactivity only because it learns hidden bias in the dataset. This is analogous to the bias observed in neural machine translation, where a pronoun indicates the gender of a word, but the model disregards it when making the translation due to the presence of gender stereotypes in the training data [**?**].

5.6 Uncovering Scaffold bias

Our case study of Friedel-Crafts acylation reveals the sensitivity of the Molecular Transformer to dataset bias. We turn to examine another source of bias – compound series bias, or scaffold bias [**?**]. This is the phenomena where very similar molecules appear in both the training and the test set. This leads to ML models achieving high accuracy on the held-out set which does not necessarily correlate with the true generalization performance of the models. This is particularly acute for drug discovery datasets as medicinal chemists typically design molecular 'series' by adding various functional groups to a central chemical 'scaffold'. In chemical reaction datasets, scaffold bias manifest itself as similar molecules undergoing very similar transformations.

To gain further insight into this phenomenon, we apply a 50:50 random train/test split to the full USPTO dataset and inspect reactions from one set that have structurally similar products to those from the other set. We define the 'structural similarity' of two molecules by calculating the Tanimoto similarity σ between the Morgan fingerprints of the respective molecules [**?**]. Figure **??** reveals that many training and test set reactions are remarkably similar as measured by both σ as well as the Tanimoto similarity of the reaction difference fingerprints of the reaction [**?**].

We find that 57% to 93% of reactions from the test set contain a structurally similar product to a reaction from the training set. This would not be problematic if the datapoints involved different reactants and reagents reacting via different mechanisms to form the same product. However this is not the case – reactions with similar products often also share reactants and undergo similar chemical changes. This means that using a random train / test split to assess the performance of reaction prediction models could be a misleading indicator of their ability to generalize. Indeed, this reconciles the seeming contradiction between the reported 90% top-1 accuracy of the Molecular Transformer and our findings above regarding the model's fragility to reactions involving chemical selectivity.

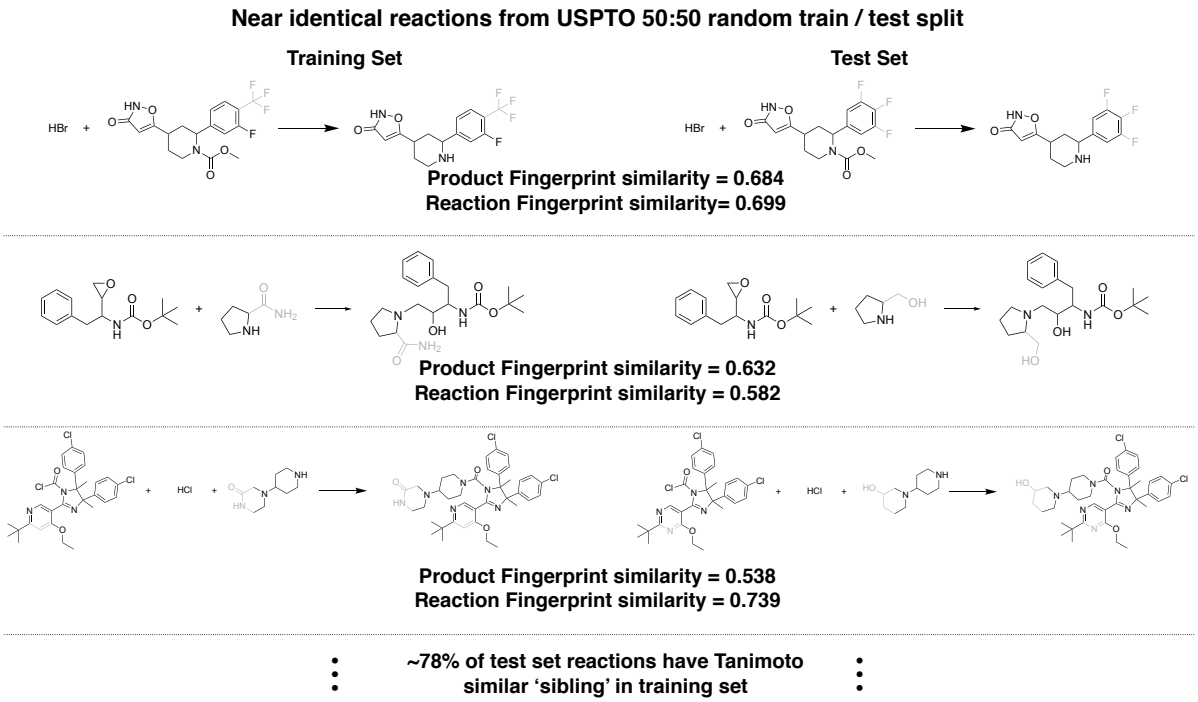


Fig. 5.8 **Randomly splitting USPTO results in a large number of near-identical reactions shared between train/test sets.** 78% of reactions in the test set have products that are within Tanimoto similarity 0.5 of a product in the training set following a 50:50 random split. By eye it can be seen that many reactions with similar products (differences are highlighted by shading) have similar reagents and follow near-identical reaction mechanisms. This intuition is confirmed by the similarly high similarity of the reaction difference fingerprints from the reactions. The equivalent proportions are 93% and 57% for Tanimoto similarity > 0.4 and > 0.6, respectively.

5.6.1 Tanimoto-Splitting USPTO

To account for this drastic scaffold bias, we propose that datasets for training machine learning reaction prediction models should be split by the Tanimoto similarity of the molecular fingerprints of the reaction products. In other words, it should be ensured that no reactions in the test set have a product that is within Tanimoto similarity σ of any product from a training set reaction.

We implement this by first conducting a random split of the dataset, and then transferring all reactions that violate the Tanimoto similarity criteria from the test set to the training set – the proportion of the initial random split is adjusted until the desired final train/test ratio is obtained. For USPTO with Tanimoto threshold $\sigma = 0.6$, the dataset was randomly split 70%:30% and the ratio after Tanimoto splitting was 89.1%:10.9%. For the $\sigma = 0.4$, the initial dataset was randomly split 30%:70% and the ratio after Tanimoto splitting was 91.7%:8.3%.

The intent of such a dataset split is to remove structural bias but we must also make sure that the distribution of different reaction types in the train and test sets is still similar. This is important because we would like the test set score to reflect how well the model learnt the chemistry contained in the training set and we are less interested in extrapolation to unseen reaction types. To characterize the new Tanimoto-split dataset we inspected the distribution of reaction types in the training and test sets for both the random and Tanimoto-split datasets (Fig **??**).

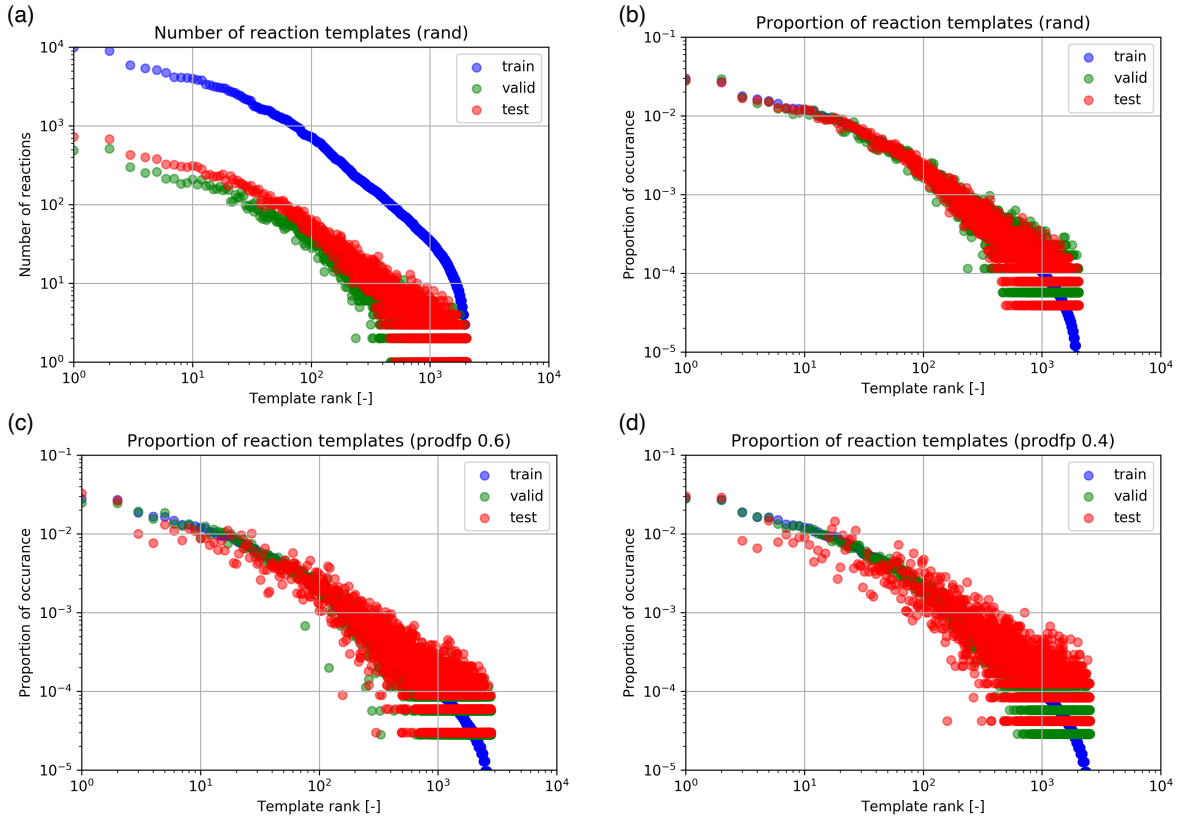


Fig. 5.9 **Tanimoto splitting minimally affects reaction template distribution.** (**a** - **b**) The absolute occurrence (**a**) and fractional occurrence (**b**) of reaction templates in train/valid/test sets of USPTO from a random split. The distribution of test set reactions closely resembles that of the validation set. (**c** - **d**) The fractional occurrence of reaction templates in train/valid/test sets of USPTO from two different Tanimoto splits using the Morgan fingerprint of the reaction product. As the Tanimoto similarity threshold value is tightened from 0.6 (**c**) to 0.4 (**d**), the deviation in frequency of the test set reactions from the training set increases.

In order to inspect how the distribution of reaction types changes when using fingerprint similarity-based splitting, open-source template extraction code [**?**] was applied on the training, validation, and test sets from different dataset splitting methods (Fig **??** (a)). Reaction SMARTS describing bond changes of radius 1 were used to classify reactions to particular

templates. The frequency of occurrence for each reaction template is divided by the size of the training/validation/test set to obtain the fractional occurrence of the template, which is plotted in decreasing order of frequency in the training set. For rare templates (ie low frequency reaction types) floating point errors are encountered; however these do not affect the qualtitative trends obvserved.

These graphs show that distribution of templates in the test set closely follows that of the training set in all cases (Fig **??** (b - d)). As increasingly strict fingerprint similarity-based splitting is applied, the fractional occurrence of rare templates deviates more and more from that of the training set. In addition, for both random and Tanimoto-splits there are no reaction templates present in the test set that are not contained in the training set, i.e. all reaction types in the test set are 'seen' by the model during training. In fact, Tanimoto-splitting increases the number of unique templates in the test set from $\sim$3k to $\sim$4.9k, suggesting that this splitting method can produce test sets that better represents the distribution of reaction types from the full dataset ($\sim$26k templates) compared to a random split. This is similar to an importance sampling scheme that helps sampling the tails of the distribution as well.

We also inspected using the fingerprint difference between the product and reactant molecules for calculation of Tanimoto similarity calculation, which led to qualitatively similar changes in the reaction template distribution. However, we have concerns that noise present in the data-mining of reactants and reagents (presence/absence of salt/catalysts/solvents etc) could cause unintentional effects on similarity calculation using reaction fingerprints and lead to additional hidden biases within the split. Together with the relative interpretability of the product fingerprint, we believe it is most practical to the community to simply use the molecular fingerprints of the reaction products for similarity calculation in splitting datasets.

5.6.2 Model performance on Tanimoto-split USPTO

We train and evaluate the Molecular Transformer on Tanimoto-split USPTO with $\sigma = 0.6$ and $\sigma = 0.4$, as well as the WLDN5 model of Coley et. al. [**?**] which is a widely-used graph-based machine learning reaction prediction model. This model explicitly represents molecules as graphs and considers reactions as series of graph edits instead of the Molecular Transformer's text-based translation of SMILES strings.

Table **??** shows that the model performance of both the graph-based model and the Molecular Transformer significantly decrease upon debiasing the dataset, but Molecular Transformer continues to outperform WLDN5. These results show that scaffold bias affects both graph-based and sequence-based models, confirming that this bias is intrinsic to data and independent of model architecture. Importantly, this demonstrates that there is significant scope for im-

provement in the performance of reaction prediction, and that the 90% accuracy obtained for a randomly split dataset does not necessarily translate to real-life applications.

Table 5.2 **Reaction prediction models are strongly affected by scaffold bias.** The performance of the Molecular Transformer and WLDN5 on various USPTO train/test splits are shown, with the accuracy of the best-performing model highlighted in bold.

Model	**Top-1[%]**	**Top-3[%]**	**Top-5[%]**
		Original	
Molecular Transformer	**90.4%**	**94.6%**	**95.3%**
WLDN5	85.6%	92.8%	93.4%
		Tanimoto Similarity <0.6	
Molecular Transformer	**80.9%**	**88.2%**	**89.6%**
WLDN5	75.9%	86.2%	88.8%
		Tanimoto Similarity <0.4	
Molecular Transformer	**74.6%**	**82.9%**	**84.5%**
WLDN5	69.3%	80.9%	84.1%

5.7 Discussion

In this chapter we developed a framework for quantitatively interpreting the predictions of Molecular Transformer, a state-of-the-art model for predicting the outcome of chemical reactions. We show that the model makes predictions based on patterns it recognizes and the statistics of the training data, but this does not necessarily coincide with the underlying chemical drivers of reactivity. This can result in erroneous predictions. Attributing the predicted probability to parts of the input allowed us to foresee these failure modes.

Through this interpretation framework, we discover that the model is susceptible to the Clever Hans effect, where the correct outcome is reached by learning bias. For instance, the dataset contains orders of magnitude more para than meta electrophilic aromatic substitution reactions, and the Molecular Transformer frequently arrived at correct test set prediction by simply memorising this fact. We believe that the inclusion of additional physical insight into models, as done in recent work incorporating explicit reaction mechanisms for reaction prediction [**?**] and machine-learning regio-selectivity prediction [**?**], could be an effective way of increasing model robustness against dataset bias. A possible way to accomplish this in Transformer models is via the augmentation of token embeddings with physical descriptors.

Moreover, future efforts should focus on benchmarking other graph-based synthesis prediction tools such as the recent MEGAN architecture as well [**?**].

We have also shown that incorrect predictions can be the result of erroneous training data points. This can be revealed using our method to attribute model predictions to training data. This method can also aid experimental chemists using the Molecular Transformer. The references corresponding to the most similar training reactions can be used to impute experimental conditions. This principle can be used in many scientific machine learning applications where the training data is generated via text-mining which is known to lead to loss of important meta data, like reaction conditions.

Finally we have shown that scaffold bias is a phenomena present in the published literature on reaction prediction. Many of the reactions in the test set have almost identical twins in the training set. This leads to an overestimation of the generalization performance of the models as reported in the literature. We have re-trained two of the leading models the Molecular Transformer and the graph-based WLDN5 model on our new Tanimoto-split dataset and found that the Top-1 accuracy of the models dropped significantly.

Our work highlights the importance of understanding and evaluating scientific machine learning models beyond looking at their accuracy on standard benchmark datasets. By rigorously applying interpretability techniques, we reveal how systematic weaknesses of the models can be uncovered, proving insights that facilitate the work of model developers. We believe further work into the use of input attribution and interpretability tools to critically analyse machine learning models for retrosynthesis, as well as other areas of computational science, is vital and necessary for continued refinement of predictive models.

Chapter 6

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

Testing a drug candidate requires compound synthesis and purification followed by preparation of a solution of known concentration to measure compound activity via an assay. While necessary for maximum accuracy, compound purification can be time-consuming and costly, bottlenecking the throughput of compound screening. This is a challenge particularly when exploring large and diverse sets of analogues for an intermediate hit or lead compound in order to derive its structure-activity relationship (SAR).

Recent work in developing nanomolar-scale high-throughput chemistry seeks to address this issue [**?** **?** **?**]. (TODO - not quite connected to the first paragraph). Adopting techniques from plate-based biological-assay screening, reacting one reagent with a different second reagent in each well of the plate, these approaches enable commonplace medicinal chemistry reactions (e.g., amide couplings and Suzuki reactions) to be conducted in a high-throughput manner with minimal starting material (<300 nmol). Utilising this method at the end of the synthesis route allows high-throughput generation of analogues for SAR exploration. In addition to higher throughput, nanomolar-scale chemistry also reduces costs by lowering solvent usage and conserving advance intermediates in the synthesis route. The drawback, however, is that it is rarely possible to perform purification for reactions conducted at the nanomole scale.

nanosar [**?**].

A related approach for synthesizing large and diverse combinatorial libraries are DNA-encoded libraries (DEL) [**?**]. DEL workflows also do not require purification of every compound, and there has been recent success in training ML models on DEL bioactivity screen

for hit-finding [**?**]. The success of this approach despite the inherent uncertainty in compound yields suggest that a similar approach with nanomolar screening data

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 peptide 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.

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.

Mixture of reactants and products from crude assay.

6.1 Identifying False Negatives in Experimental Data

Long discussion on the regressed performance. Disagreements between RF and GP model?

< Figure of the false negative >

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

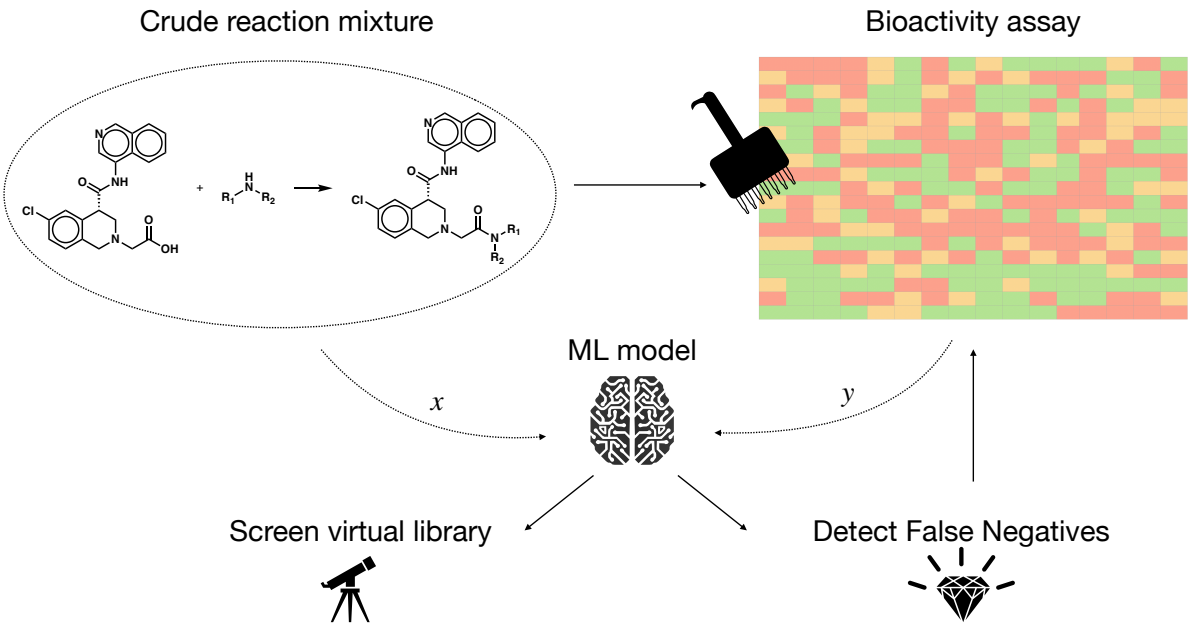


Fig. 6.1 Schematic of workflow.

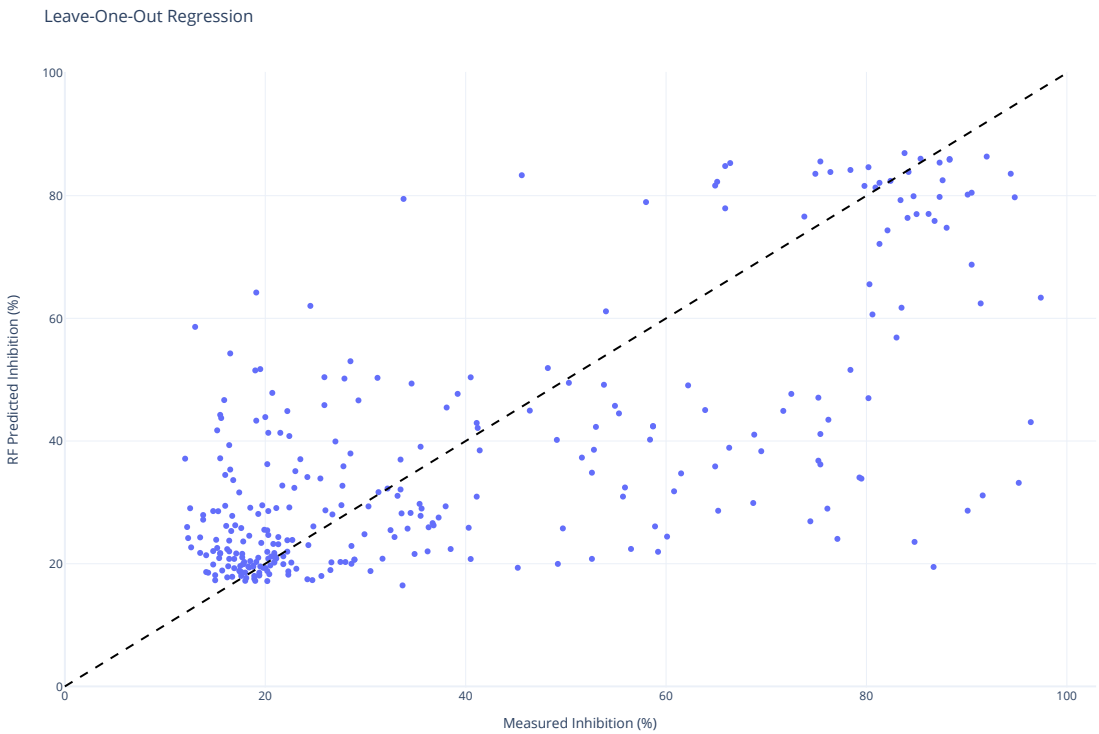


Fig. 6.2 Leave one out regression.

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 = 0.113μM (ASAP-0000204). (needs a nice sentence to round it off).

TODO - broad discussion of model predictions, correlation with yield? yields determind by integration of the UV spectra of each reaction.

Examples of two nanomole scale high throughput chemistry (HTC) campaigns to optimize the potency of intermediate binders. In one case we used the Chan-Lam reaction (Fig S7) and in another amide coupling (Fig S8). Direct biochemical screening of crude reactions identified candidates that were resynthesized and in both cases were able to improve the potency of the parent compound.

A complementary method for rapid SAR evaluation was the use of nanomole scale high-throughput chemistry(43, 44) (HTC), coupled with a 'direct to biology'(45-47) biochemical screening. The optimization of the amide coupling reaction to extend MAT-POS-4223bc15-21 (Figure 2D). The co-crystal structure of the parent compound (Fig S6) suggested vectors that could target the P4 pocket of Mpro. Optimization of the reaction conditions was performed for the starting building block with model amines (Figs S7-S8) and the optimal conditions were applied to HTC with a library of 300 amine building blocks. Yield estimation was performed and showed 151 of the library yielded >30% of the desired product. Nevertheless, the crude mixtures were subjected to a biochemical assay against Mpro (Experimental details can be found in Appendix **??** & **??**). 20 compounds were selected for resynthesis (Fig S9). In parallel to synthesis, the crude reaction mixtures were subjected to soaking and x-ray crystallography. The structures verified the extended compounds indeed adopt a similar binding mode to the parent, extending towards P3/P5 instead of the P4 pocket nevertheless forming new interactions with Mpro (Figure 2E). Upon resynthesis, several of the amide-coupling series were able to improve by up-to 300-fold on the parent acid-compound (up-to 3-fold on the corresponding methyl-amide) with the best inhibitor exhibiting an IC50 of 28nM.

6.2 Scaffold Exploration via Virtual Screening

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.

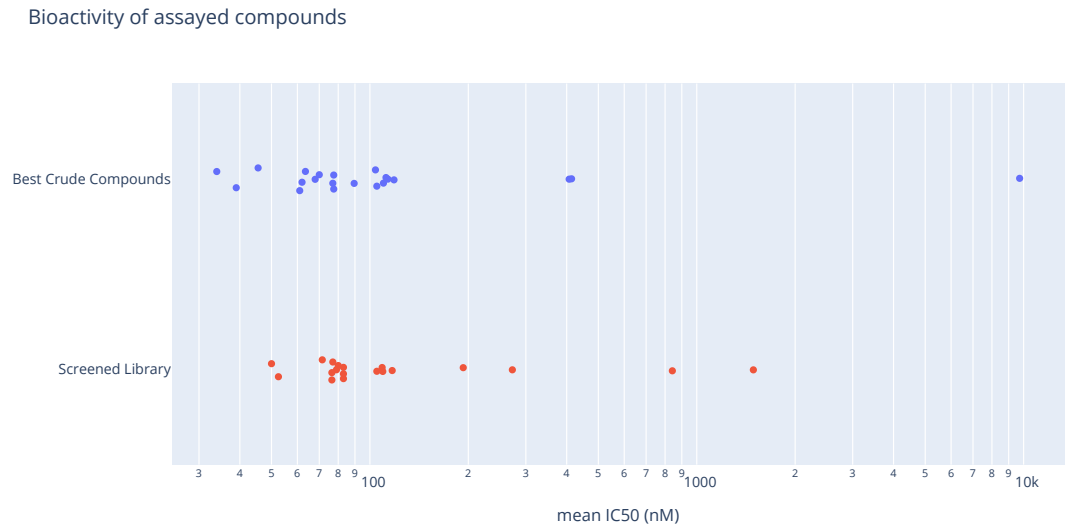


Fig. 6.3 Plot of mean IC50 values.

Gratifyingly, the top 2 ML compounds showed promising average IC50 values of 0.0525μM (ASAP-0000169) and 0.075μM (ASAP-0000211), respectively. The top 2 most potent molecules based off of the crude inhibition values were compounds that, whilst active at IC50 = 0.034μM (ASAP-0000221) and IC50 = 0.064μM (ASAP-0000164), contained the toxic benzene-1,4-diamine motif that is generally avoided [**?**]. 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, 0.046μM (ASAP-0000155) and 0.064μM (ASAP-0000225), respectively. This result highlights ML's ability to identify promising yet overlooked scaffolds without compromising potency.

< Figure of the top 2 shown in panel B >

6.3 Discussion

Usefulness of training models with crude reaction mixtures. Not only is it useful for nanoSAR, it's very potent for generating data and screening libraries - ideal for setting up automated feedback loop and/or bayesian optimisation workflow.

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 **??** 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 $\sim$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

[] 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.

[] 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.

[] 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. (2021). Structures of the

$$\sigma$$

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

[] Antunes, D. A., Devaurs, D., and Kavraki, L. E. (2015). Understanding the challenges of protein flexibility in drug design. *Expert Opinion on Drug Discovery*, 10(12):1301–1313.

[] 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.

[] Bajusz, D., Rácz, A., and Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1):20.

[] Baldi, P. and Nasr, R. (2010). When is chemical similarity significant? the statistical distribution of chemical similarity scores and its extreme values. *Journal of Chemical Information and Modeling*, 50(7):1205–1222.

[] Ballante, F., Kooistra, A. J., Kampen, S., de Graaf, C., and Carlsson, J. (2021). Structure-based virtual screening for ligands of g protein–coupled receptors: What can molecular docking do for you? *Pharmacological Reviews*, 73(4):1698–1736.

[] Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. (2018). Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–43.

[] Bender, A. and Cortés-Ciriano, I. (2021). Artificial intelligence in drug discovery: what is realistic, what are illusions? part 1: Ways to make an impact, and why we are not there yet. *Drug Discovery Today*, 26(2):511–524.

[] Bender, A. and Cortes-Ciriano, I. (2021). Artificial intelligence in drug discovery: what is realistic, what are illusions? part 2: a discussion of chemical and biological data. *Drug Discovery Today*, 26(4):1040–1052.

[] Bender, B. J., Gahbauer, S., Luttens, A., Lyu, J., Webb, C. M., Stein, R. M., Fink, E. A., Balius, T. E., Carlsson, J., Irwin, J. J., and Shoichet, B. K. (2021). A practical guide to large-scale docking. *Nature Protocols*, 16(10):4799–4832.

[] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

[] Bjerrum, E. J. (2017). Smiles enumeration as data augmentation for neural network modeling of molecules.

[] Blakemore, D. C., Castro, L., Churcher, I., Rees, D. C., Thomas, A. W., Wilson, D. M., and Wood, A. (2018). Organic synthesis provides opportunities to transform drug discovery. *Nature chemistry*, 10(4):383.

[] Born, J., Manica, M., Oskooei, A., Cadow, J., Borgwardt, K., and Martínez, M. R. (2019). Paccmann^{RL}: Designing anticancer drugs from transcriptomic data via reinforcement learning.

[] Boström, J., Brown, D. G., Young, R. J., and Keserü, G. M. (2018). Expanding the medicinal chemistry synthetic toolbox. *Nature Reviews Drug Discovery*.

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

[] Bradshaw, J., Kusner, M. J., Paige, B., Segler, M. H. S., and Hernández-Lobato, J. M. (2019). A generative model for electron paths.

[] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

[] 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.

[] 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*.

[] Chemspace (2023). Lead-like compounds.

[] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

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

[] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling.

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

[] Coleman, R. G., Carchia, M., Sterling, T., Irwin, J. J., and Shoichet, B. K. (2013). Ligand pose and orientational sampling in molecular docking. *PLOS ONE*, 8(10):1–19.

[] 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.

[] Coley, C. W., Eyke, N. S., and Jensen, K. F. (2020). Autonomous discovery in the chemical sciences part ii: Outlook. *Angewandte Chemie International Edition*, 59(52):23414–23436.

[] 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.

[] Coley, C. W., Jin, W., Rogers, L., Jamison, T. F., Jaakkola, T. S., Green, W. H., Barzilay, R., and Jensen, K. F. (2019b). A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.*, 10:370–377.

[] Commons, W. (2018). Receiver operating characteristic (roc) curve. https://commons.wikimedia.org/wiki/File:Roc_curve.svg.

[] Corey, E. J., Long, A. K., and Rubenstein, S. D. (1985). Computer-assisted analysis in organic synthesis. *Science*, 228(4698):408.

[] Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. S. (2023). Diffdock: Diffusion steps, twists, and turns for molecular docking. In *The Eleventh International Conference on Learning Representations*.

[] Cortés-Ciriano, I. and Bender, A. (2019). Reliable prediction errors for deep neural networks using test-time dropout. *Journal of Chemical Information and Modeling*, 59(7):3330–3339. PMID: 31241929.

[] 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.

[] Degen, J., Wegscheid-Gerlach, C., Zaliani, A., and Rarey, M. (2008). On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, 3(10):1503–1507.

[] Delaney, J. S. (2004). Esol: Estimating aqueous solubility directly from molecular structure. *Journal of Chemical Information and Computer Sciences*, 44(3):1000–1005.

[] Dixon, S. L., Smondyrev, A. M., and Rao, S. N. (2006). Phase: A novel approach to pharmacophore modeling and 3d database searching. *Chemical Biology & Drug Design*, 67(5):370–372.

[] 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.

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

[] Eberhardt, J., Santos-Martins, D., Tillack, A. F., and Forli, S. (2021). Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898.

[] Erickson, J. A., Jalaie, M., Robertson, D. H., Lewis, R. A., and Vieth, M. (2004). Lessons in molecular recognition: The effects of ligand and protein flexibility on molecular docking accuracy. *Journal of Medicinal Chemistry*, 47(1):45–55.

[] 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 α_{2A}-adrenergic receptor. *Science*, 377(6614):eabn7065.

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

[] Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., and Shenkin, P. S. (2004). Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749.

[] Gehrtz, P., Marom, S., Bührmann, M., Hardick, J., Kleinbölting, S., Shraga, A., Dubiella, C., Gabizon, R., Wiese, J. N., Müller, M. P., Cohen, G., Babaev, I., Shurrush, K., Avram, L., Resnick, E., Barr, H., Rauh, D., and London, N. (2022). Optimization of covalent mkk7 inhibitors via crude nanomole-scale libraries. *Journal of Medicinal Chemistry*, 65(15):10341–10356.

[] Gentile, F., Agrawal, V., Hsing, M., Ton, A.-T., Ban, F., Norinder, U., Gleave, M. E., and Cherkasov, A. (2020). Deep docking: A deep learning platform for augmentation of structure based drug discovery. *ACS Central Science*, 6(6):939–949.

[] Gesmundo, N. J., Sauvagnat, B., Curran, P. J., Richards, M. P., Andrews, C. L., Dandliker, P. J., and Cernak, T. (2018). Nanoscale synthesis and affinity ranking. *Nature*, 557(7704):228–232.

[] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1263–1272. JMLR.org.

[] 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.

[] Goldman, B., Kearnes, S., Kramer, T., Riley, P., and Walters, W. P. (2022). Defining levels of automated chemical design. *Journal of Medicinal Chemistry*, 65(10):7073–7087.

[] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[] 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.

[] Griffiths, R.-R., Greenfield, J. L., Thawani, A. R., Jamasb, A. R., Moss, H. B., Bourached, A., Jones, P., McCorkindale, W., Aldrick, A. A., Fuchter, M. J., and Lee, A. A. (2022). Data-driven discovery of molecular photoswitches with multioutput gaussian processes. *Chem. Sci.*, 13:13541–13551.

[] Guan, Y., Coley, C. W., Wu, H., Ranasinghe, D., Heid, E., Struble, T. J., Pattanaik, L., Green, W. H., and Jensen, K. F. (2021). Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chem. Sci.*, 12:2198–2208.

[] 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.

[] 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.

[] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

[] Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., and Pletnev, I. (2013). Inchi - the worldwide chemical structure identifier standard. *Journal of Cheminformatics*, 5(1):7.

[] 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.

[] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

[] Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171 – 1220.

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

[] Huang, N., Shoichet, B. K., and Irwin, J. J. (2006). Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry*, 49(23):6789–6801.

[] 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.

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

[] Imrie, F., Bradley, A. R., van der Schaar, M., and Deane, C. M. (2018). Protein family-specific models using deep neural networks and transfer learning improve virtual screening and highlight the need for more data. *Journal of Chemical Information and Modeling*, 58(11):2319–2330.

[] 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.

[] 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.

[] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift.

[] Irwin, J. J., Tang, K. G., Young, J., Dandarchuluun, C., Wong, B. R., Khurelbaatar, M., Moroz, Y. S., Mayfield, J., and Sayle, R. A. (2020). Zinc20—a free ultralarge-scale chemical database for ligand discovery. *Journal of Chemical Information and Modeling*, 60(12):6065–6073.

[] Jia, X., Lynch, A., Huang, Y., Danielson, M., Lang'at, I., Milder, A., Ruby, A. E., Wang, H., Friedler, S. A., Norquist, A. J., and Schrier, J. (2019). Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature*, 573:251–255.

[] Jiang, D., Wu, Z., Hsieh, C.-Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., and Hou, T. (2021). Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics*, 13(1):12.

[] Jiménez, J., Škalič, M., Martínez-Rosell, G., and De Fabritiis, G. (2018). Kdeep: Protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of Chemical Information and Modeling*, 58(2):287–296.

[] Jiménez-Luna, J., Grisoni, F., and Schneider, G. (2020). Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584.

[] 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.

[] 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 mpro from sars-cov-2 and discovery of its inhibitors. *Nature*, 582(7811):289–293.

[] Johansson, S., Thakkar, A., Kogej, T., Bjerrum, E., Genheden, S., Bastys, T., Kannas, C., Schliep, A., Chen, H., and Engkvist, O. (2020). Ai-assisted synthesis prediction. *Drug Discovery Today: Technologies*.

[] Johnson, M. A. and Maggiora, G. M. (1990). *Concepts and applications of molecular similarity*. Wiley.

[] Jorner, K., Brinck, T., Norrby, P.-O., and Buttar, D. (2021). Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. *Chem. Sci.*, 12:1163–1175.

[] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

[] Karpov, P., Godin, G., and Tetko, I. V. (2020). Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *Journal of Cheminformatics*, 12(1):17.

[] 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.

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

[] Kim, H., Na, J., and Lee, W. B. (2021). Generative chemical transformer: Neural machine learning of molecular geometric structures from chemical language via attention. *Journal of Chemical Information and Modeling*, 61(12):5804–5814.

[] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization.

[] Kishimoto, A., Buesser, B., Chen, B., and Botea, A. (2019). Depth-first proof-number search with heuristic edge cost and application to chemical synthesis planning. In *Advances in Neural Information Processing Systems*, pages 7224–7234.

[] Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery*, 3(11):935–949.

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

[] Klucznik, T., Mikulak-Klucznik, B., McCormack, M. P., Lima, H., Szymkuć, S., Bhowmick, M., Molga, K., Zhou, Y., Rickershauser, L., Gajewska, E. P., Toutchkine, A., Dittwald, P., Startek, M. P., Kirkovits, G. J., Roszak, R., Adamski, A., Sieredzińska, B., Mrksich, M.,

Trice, S. L., and Grzybowski, B. A. (2018). Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem*, 4(3):522–532.

[] Korovina, K., Xu, S., Kandasamy, K., Neiswanger, W., Poczos, B., Schneider, J., and Xing, E. P. (2019). Chembo: Bayesian optimization of small organic molecules with synthesizable recommendations.

[] Kovacs, D. P., McCorkindale, W., and Lee, A. A. (2021). Molecular Transformer Explainer. https://github.com/davkovacs/MTExplainer.git.

[] Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. (2020). Self-referencing embedded strings (selfies): A 100string representation. *Machine Learning: Science and Technology*, 1(4):045024.

[] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

[] Kumar, M. S., Tamilarasan, R., and Sreekanth, A. (2011). 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(2):370–375.

[] Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096.

[] 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.

[] Li, J., Fu, A., and Zhang, L. (2019). An overview of scoring functions used for protein–ligand interactions in molecular docking. *Interdisciplinary Sciences: Computational Life Sciences*, 11(2):320–328.

[] Limban, C., Nuţă, D. C., Chiriţă, C., Negreș, S., Arsene, A. L., Goumenou, M., Karakitsios, S. P., Tsatsakis, A. M., and Sarigiannis, D. A. (2018). The use of structural alerts to avoid the toxicity of pharmaceuticals. *Toxicology Reports*, 5:943–953.

[] Lipton, Z. C., Berkowitz, J., and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning.

[] 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.

[] Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., Nie, W., Liu, Y., and Wang, R. (2014). PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, 31(3):405–412.

[] 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.

[] Lluch, A. M., Sánchez-Baeza, F., Messeguer, A., Fusco, C., and Curci, R. (1993). Regio- and chemoselective epoxidation of fluorinated monoterpenes and sesquiterpenes by dioxiranes. *Tetrahedron*, 49(28):6299–6308.

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

[] Lundberg, S. M. and Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017-Decem(Section 2):4766–4775.

[] Lyu, J., Irwin, J. J., and Shoichet, B. K. (2023). Modeling the expansion of virtual screening libraries. *Nature Chemical Biology*.

[] 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. (2019). Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229.

[] 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.

[] Maggiora, G., Vogt, M., Stumpfe, D., and Bajorath, J. (2014). Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, 57(8):3186–3204.

[] Maggiora, G. M. (2006). On outliers and activity cliffswhy qsar often disappoints. *Journal of Chemical Information and Modeling*, 46(4):1535–1535.

[] Martin, Y. C., Kofron, J. L., and Traphagen, L. M. (2002). Do structurally similar molecules have similar biological activity? *Journal of Medicinal Chemistry*, 45(19):4350–4358.

[] Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., Clevert, D.-A., and Hochreiter, S. (2018). Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chem. Sci.*, 9:5441–5451.

[] McCloskey, K., Sigel, E. A., Kearnes, S., Xue, L., Tian, X., Moccia, D., Gikunju, D., Bazzaz, S., Chan, B., Clark, M. A., Cuozzo, J. W., Guié, M.-A., Guilinger, J. P., Huguet, C., Hupp, C. D., Keefe, A. D., Mulhern, C. J., Zhang, Y., and Riley, P. (2020). Machine learning on dna-encoded libraries: A new paradigm for hit finding. *Journal of Medicinal Chemistry*, 63(16):8857–8866.

[] 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.

[] McCorkindale, W., Poelking, C., and Lee, A. A. (2020). Investigating 3d atomic environments for enhanced qsar.

[] McGann, M. (2012). Fred and hybrid docking performance on standardized datasets. *Journal of Computer-Aided Molecular Design*, 26(8):897–906.

[] Meng, X.-Y., Zhang, H.-X., Mezei, M., and Cui, M. (2011). Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, 7(2):146–157.

[] Merget, B., Turk, S., Eid, S., Rippmann, F., and Fulle, S. (2017). Profiling prediction of kinase inhibitors: Toward the virtual assay. *Journal of Medicinal Chemistry*, 60(1):474–485.

[] Montavon, G., Samek, W., and Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal*, 73:1–15.

[] Morgan, H. L. (1965). The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113.

[] 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.

[] 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.

[] Mudrakarta, P. K., Taly, A., Sundararajan, M., and Dhamdhere, K. (2018). Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906, Melbourne, Australia. Association for Computational Linguistics.

[] 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.

[] Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594.

[] Niu, Z., Zhong, G., and Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62.

[] O'Boyle, N. M. and Sayle, R. A. (2016). Comparing structural fingerprints using a literature-based similarity benchmark. *Journal of Cheminformatics*, 8(1):36.

[] Obrezanova, O., Csányi, G., Gola, J. M. R., and Segall, M. D. (2007). Gaussian processes: A method for automatic qsar modeling of adme properties. *Journal of Chemical Information and Modeling*, 47(5):1847–1857.

[] 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^{pro} inhibitor clinical candidate for the treatment of covid-19. *Science*, 374(6575):1586–1593.

[] Pal, S., Kumar, V., Kundu, B., Bhattacharya, D., Preethy, N., Reddy, M. P., and Talukdar, A. (2019). Ligand-based pharmacophore modeling, virtual screening and molecular docking studies for discovery of potential topoisomerase i inhibitors. *Computational and Structural Biotechnology Journal*, 17:291–310.

[] Palmer, D. S., O'Boyle, N. M., Glen, R. C., and Mitchell, J. B. O. (2007). Random forest models to predict aqueous solubility. *Journal of Chemical Information and Modeling*, 47(1):150–158.

[] Parks, C. D., Gaieb, Z., Chiu, M., Yang, H., Shao, C., Walters, W. P., Jansen, J. M., McGaughey, G., Lewis, R. A., Bembenek, S. D., Ameriks, M. K., Mirzadegan, T., Burley, S. K., Amaro, R. E., and Gilson, M. K. (2020). D3r grand challenge 4: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *Journal of Computer-Aided Molecular Design*, 34(2):99–119.

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

[] Perera, D., Tucker, J. W., Brahmbhatt, S., Helal, C. J., Chong, A., Farrell, W., Richardson, P., and Sach, N. W. (2018). A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science*, 359(6374):429–434.

[] 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.

[] Polishchuk, P. (2020). Crem: chemically reasonable mutations framework for structure generation. *Journal of Cheminformatics*, 12(1):28.

[] Polishchuk, P. G., Muratov, E. N., Artemenko, A. G., Kolumbin, O. G., Muratov, N. N., and Kuz'min, V. E. (2009). Application of random forest approach to qsar prediction of aquatic toxicity. *Journal of Chemical Information and Modeling*, 49(11):2481–2488.

[] PostEra Inc. (2022). COVID moonshot. https://postera.ai/covid.

[] Pradeepkiran, J. A., Reddy, A. P., and Reddy, P. H. (2019). Pharmacophore-based models for therapeutic drugs against phosphorylated tau in alzheimer's disease. *Drug Discovery Today*, 24(2):616–623.

[] Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., and Koes, D. R. (2017). Protein–ligand scoring with convolutional neural networks. *Journal of Chemical Information and Modeling*, 57(4):942–957.

[] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents.

[] Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

[] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Augu:1135–1144.

[] Riniker, S. and Landrum, G. A. (2013). Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics*, 5(1):26.

[] Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754.

[] 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*.

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

C. A., Zhang, I., Pulido, I., Tomlinson, C., Taylor, J. C., Croll, T. I., and Brwewitz, L. (2023). Turning high-throughput structural biology into predictive inhibitor design. *Proceedings of the National Academy of Sciences*, 120(11):e2214168120.

[] Sacha, M., Błaż, M., Byrski, P., Włodarczyk-Pruszyński, P., and Jastrzębski, S. (2020). Molecule edit graph attention network: Modeling chemical reactions as sequences of graph edits.

[] Saldívar-González, F. I., Huerta-García, C. S., and Medina-Franco, J. (2020). Chemoinformatics-based enumeration of chemical libraries: a tutorial. *Journal of Cheminformatics*, 12(1):64.

[] Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C., and Glorius, F. (2020). A structure-based platform for predicting chemical reactivity. *Chem*, 6(6):1379–1390.

[] Santanilla, A. B., Regalado, E. L., Pereira, T., Shevlin, M., Bateman, K., Campeau, L.-C., Schneeweis, J., Berritt, S., Shi, Z.-C., Nantermet, P., Liu, Y., Helmy, R., Welch, C. J., Vachal, P., Davies, I. W., Cernak, T., and Dreher, S. D. (2015). Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science*, 347(6217):49–53.

[] Sapoval, N., Aghazadeh, A., Nute, M. G., Antunes, D. A., Balaji, A., Baraniuk, R., Barberan, C. J., Dannenfelser, R., Dun, C., Edrisi, M., Elworth, R. A. L., Kille, B., Kyrillidis, A., Nakhleh, L., Wolfe, C. R., Yan, Z., Yao, V., and Treangen, T. J. (2022). Current progress and open challenges for applying deep learning across the biosciences. *Nature Communications*, 13(1):1728.

[] 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.

[] Schneider, G. (2018). Automating drug discovery. *Nature Reviews Drug Discovery*, 17(2):97–113.

[] Schneider, N., Lowe, D. M., Sayle, R. A., and Landrum, G. A. (2015). Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *Journal of Chemical Information and Modeling*, 55(1):39–53. PMID: 25541888.

[] Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition.

[] Schreck, J. S., Coley, C. W., and Bishop, K. J. (2019). Learning retrosynthetic planning through simulated experience. *ACS Central Science*, 5(6):970.

[] 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.

[] 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.

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

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

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

[] Segler, M. H. S. (2019). World programs for model-based learning and planning in compositional state and action spaces.

[] Segler, M. H. S. and Waller, M. P. (2017). Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry – A European Journal*, 23(25):5966–5971.

[] Sheridan, R. P. (2013). Time-split cross-validation as a method for estimating the goodness of prospective prediction. *Journal of Chemical Information and Modeling*, 53(4):783–790.

[] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

[] Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

[] Stärk, H., Ganea, O., Pattanaik, L., Barzilay, R., and Jaakkola, T. (2022). Equibind: Geometric deep learning for drug binding structure prediction. In *International Conference on Machine Learning*, pages 20503–20521. PMLR.

[] Struble, T. J., Alvarez, J. C., Brown, S. P., Chytil, M., Cisar, J., DesJarlais, R. L., Engkvist, O., Frank, S. A., Greve, D. R., Griffin, D. J., Hou, X., Johannes, J. W., Kreatsoulas, C., Lahue, B., Mathea, M., Mogk, G., Nicolaou, C. A., Palmer, A. D., Price, D. J., Robinson, R. I., Salentin, S., Xing, L., Jaakkola, T., Green, W. H., Barzilay, R., Coley, C. W., and Jensen, K. F. (2020). Current and future roles of artificial intelligence in medicinal chemistry synthesis. *Journal of Medicinal Chemistry*, 63(16):8667–8682. PMID: 32243158.

[] Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., and Wang, R. (2019). Comparative assessment of scoring functions: The casf-2016 update. *Journal of Chemical Information and Modeling*, 59(2):895–913.

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

[] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: A classification and regression tool for compound classification and qsar modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958.

[] Swamidass, S. J., Chen, J., Bruand, J., Phung, P., Ralaivola, L., and Baldi, P. (2005). Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics*, 21(suppl_1):359–368.

[] Systems, D. C. I. (2022). Smarts - a language for describing molecular patterns.

[] Temml, V., Voss, C. V., Dirsch, V. M., and Schuster, D. (2014). Discovery of new liver x receptor agonists by pharmacophore modeling and shape-based virtual screening. *Journal of Chemical Information and Modeling*, 54(2):367–371.

[] Tetko, I. V. (2002). Neural network studies. 4. introduction to associative neural networks. *Journal of Chemical Information and Computer Sciences*, 42(3):717–728. PMID: 12086534.

[] Tetko, I. V., Karpov, P., Van Deursen, R., and Godin, G. (2020). State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature communications*, 11(1):1–11.

[] Thakkar, A., Kogej, T., Reymond, J.-L., Engkvist, O., and Bjerrum, E. J. (2020). Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chem. Sci.*, 11:154–168.

[] 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*.

[] 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*.

[] Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M., and Willett, P. (2012). Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets. *Journal of Chemical Information and Modeling*, 52(11):2884–2901. PMID: 23078167.

[] Trnka, T. M. and Grubbs, R. H. (2001). The development of l2x2ruchr olefin metathesis catalysts: An organometallic success story. *Accounts of Chemical Research*, 34(1):18–29. PMID: 11170353.

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

[] 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.

[] Vandenberk, J., Kennis, L. E. J., Van Heertum, A. H. M. T., and Van der Aa, M. J. M. C. (1981). 1,3-dihydro-1-[(1-piperidinyl)alkyl]-2h-benzimidazol-2-one derivatives.

[] 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.

[] Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., and Taylor, R. D. (2003). Improved protein–ligand docking using gold. *Proteins: Structure, Function, and Bioinformatics*, 52(4):609–623.

[] Volkamer, A., Riniker, S., Nittinger, E., Lanini, J., Grisoni, F., Evertsson, E., Rodríguez-Pérez, R., and Schneider, N. (2023). Machine learning for small molecule drug discovery in academia and industry. *Artificial Intelligence in the Life Sciences*, 3:100056.

[] Vuorinen, A. and Schuster, D. (2015). Methods for generating and applying pharmacophore models as virtual screening filters and for bioactivity profiling. *Methods*, 71:113–134.

[] Walters, W., Stahl, M. T., and Murcko, M. A. (1998). Virtual screening—an overview. *Drug Discovery Today*, 3(4):160–178.

[] Walters, W. P. (2019). Virtual chemical libraries. *Journal of Medicinal Chemistry*, 62(3):1116–1124.

[] Wang, R., Fang, X., Lu, Y., and Wang, S. (2004). The pdbbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry*, 47(12):2977–2980.

[] Wei, J. N., Duvenaud, D., and Aspuru-Guzik, A. (2016). Neural networks for the prediction of organic chemistry reactions. *ACS Central Science*, 2(10):725–732.

[] 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.

[] 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.

[] Willett, P., Barnard, J. M., and Downs, G. M. (1998). Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 38(6):983–996.

[] Wong, F., Krishnan, A., Zheng, E. J., Stärk, H., Manson, A. L., Earl, A. M., Jaakkola, T., and Collins, J. J. (2022). Benchmarking alphafold-enabled molecular docking predictions for antibiotic discovery. *Molecular Systems Biology*, 18(9):e11081.

[] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. (2018). Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.*, 9:513–530.

[] Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., and Barzilay, R. (2019). Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388. PMID: 31361484.

[] 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.

[] 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.

Appendix A

Computational Details

A.1 Docking against SARS-CoV-2

All molecules synthesised by the COVID Moonshot Consortium were docked against structure x2908 reported by Diamond XChem [**?**]. 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.

A.2 Machine learning

A.2.1 FRESCO

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.

A.2.2 Ranking model

Our training set, de novo design method and generated molecules are available on https://github.com/wjm41/mpro-rank-gen.

A.2.3 Transformer model

The Molecular Transformer architecture used throughout this work is based on the model described in Schwaller et. al. [**?**].

The model uses a 256 dimensional learnt embedding for each SMILES token. The encoder and the decoder are both made up of 4 standard transformer layers and dropout is applied with probability 0.1 [**?**]. For weight optimization the Adam optimizer is used and the model is trained for 500 000 steps. Checkpoints are saved every 10 000 steps and the final model is obtained by averaging the weights of the last 20 checkpoints for USPTO.

The model was implemented with OpenNMT-py package [**?**] which makes use of the PyTorch framework [**?**].

All code used for implementing the attribution tools for the Molecular Transformer, generating the artificial Friedel-Crafts dataset, and Tanimoto-splitting USPTO can be accessed in the GitHub repo MTExplainer[**?**]. The USPTO dataset used to train the model [**? ?**], as well as the Tanimoto similarity-based train/test splits of USPTO can also be found in the GitHub repo.

Appendix B

Experimental Details

B.1 Mpro assay

The experimental procedure for measuring Mpro inhibition via Homogeneous Time Resolved Fluorescence (HTRF) assay is the same as that previously reported by COVID Moonshot[**?**], which is repeated below.

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

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%.

B.2 OC43 antiviral assay

A549 expressing H2B-mRuby were seeded in 384 well plates (4,000 cells per well) in DMEM+2% FCS in a total volume of 30ul. One day later, 20ul of OC43 were added to the wells for a final MOI of 0.3. one hour after viral addition, the drug (or DMSO as control)

was added to the cells. Drugs were added at a volume of 50nl, in a final dose range of 0.3-20mM. Cells were incubated at 33C, 5% CO2 for 2 days, fixed with paraformaldehyde and stained for the presence of the viral nucleoprotein. Images were captured and quantified using the Incucyte machine and software. 3 biological repeated were performed.

B.3 Model generated reaction schemes

B.4 Mac1 assay

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 [**?**]. Compounds were dispensed into ProxiPlate-384 Plus (PerkinElmer) assay plates using an Echo 525 liquid handler (Labcyte). Binding assays were conducted in a final volume of 16 μl 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).

B.5 Crystallographic screening on SARS-CoV-2 nsp3-Mac1

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

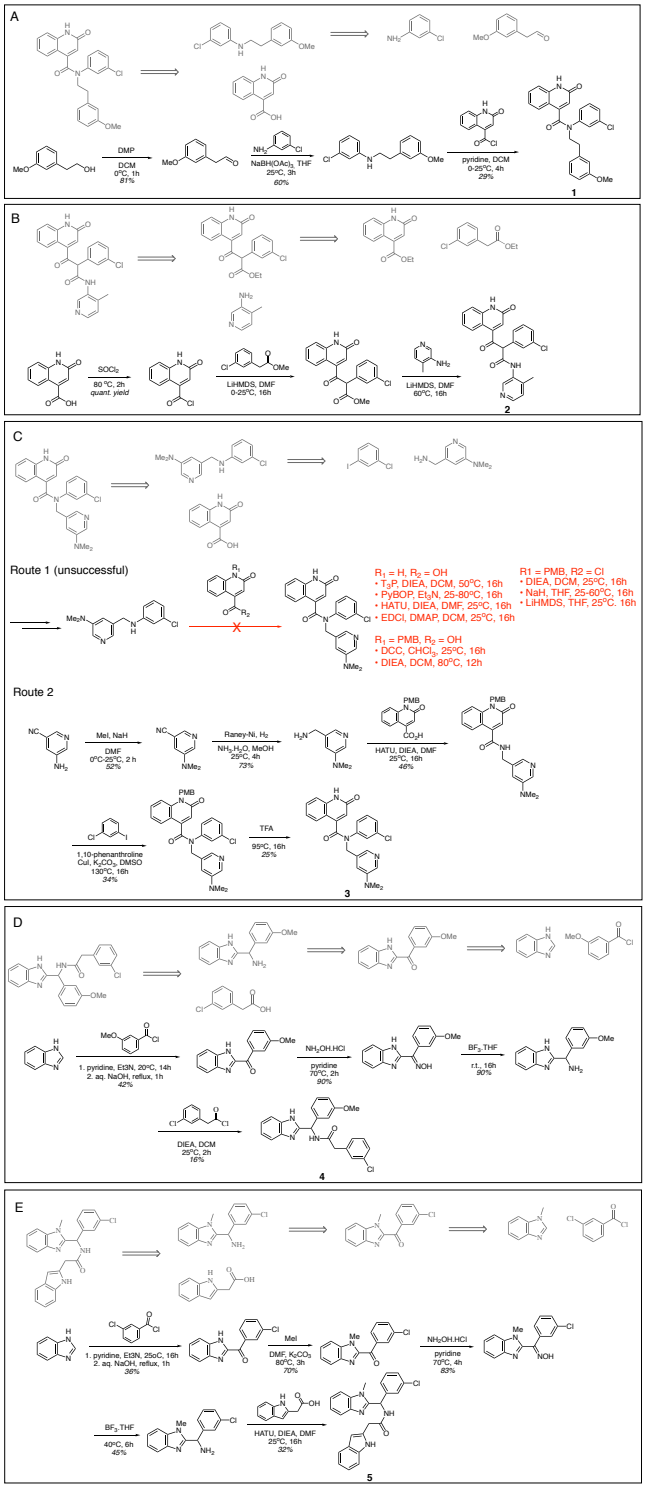


Fig. B.1 Model generated synthetic schemes for compounds **1**-**5**. The synthesis schemes generated by our model (grey) and the experimental schemes (black).

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/.

B.6 High-Throuhput Amide Coupling

The amide library was made by reacting the carboxylic acid under the optimized reaction conditions (2 eq. amine; 2 eq. EDC; 2 eq. HOAt; 5 eq. DIPEA; DMSO; RT; 24h) with 300 amines (202 aromatics, 49 primary, and 49 secondary aliphatic amines). For library production, we used Echo LDV plates and an Echo 555 acoustic dispenser for liquid handling. Plate copies were made after diluting the reaction mixture with 4 L DMSO. For yield estimation, 1 L of the diluted library was transferred to an LC/MS-ready 384-well plate, followed by dilution with 20% ACN in water to the final volume of 50 L. The desired product was identified in 60% of wells.