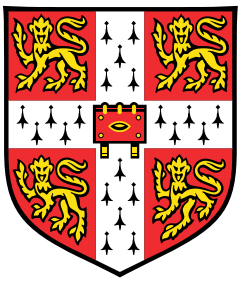
Accelerating the Design-Make-Test cycle of Drug Discovery with Machine Learning



William McCorkindale

Cavendish Laboratory, Department of Physics
University of Cambridge

Supervisor: Dr. Alpha Lee

St. John's College

April 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
April 2023

Acknowledgements

First and foremost, I would like to sincerely thank my supervisor Dr Alpha A. Lee for his invaluable guidance, support, and encouragement throughout my PhD. I am deeply grateful for his commitment to my success and for his role in shaping my personal and professional development.

I would also like to thank my colleagues in the Lee Group: Dávid, Emma, Felix, Janosh, Kadi, Penny, Rokas, Rhys. Their knowledge, expertise, and friendship have been invaluable in shaping my research and personal growth. I feel very fortunate to have had the opportunity to work with and learn from you all, and I will always cherish the memories and lessons learned from our time together.

I would like to thank the AlphaFold team at DeepMind, in particular Dr Jonas Adler, for generously hosting me as a research scientist intern. My time there was inspiring and eye-opening, and I am grateful for the opportunity to work with such a talented and passionate team.

I was generously funded by the Gates Cambridge Scholarship to undertake this research, and I am very thankful for their support, as well as for the many friendships I have made in the Gates scholar community. I would like to give special thanks in particular to David and Michelle, for their friendship and support.

During my PhD there was great upheaval in my birthplace, Hong Kong. I am thankful for the camaraderie and solidarity from my fellow Cambridge postgraduate students from Hong Kong, as well as friends with links to Hong Kong, who have helped me to navigate this difficult time.

Much of the research in this thesis was conducted during the COVID-19 pandemic, when I spent a considerable amount of time indoors with my dear housemates Dávid and Eszti. I will be forever grateful for their unwavering friendship, humor, and kindness during some of the most challenging times of my PhD journey.

I would like to thank Marika for always making me smile and reminding me what is truly important in life.

Lastly, I would like to thank my cats past and present as well as my parents, John and Maureen, for their unconditional love and continual support.

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 lead 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 in 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 1 billion molecules.

After identifying hit compounds, we enter the hit-to-lead stage where we wish to optimise their molecular structures to improve bioactivity. Framing bioactivity modelling as active/inactive classification 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.

Throughout the entire course of drug discovery, one needs to find a synthesis route to actually make the molecule. 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 the 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. COVID Moonshot, a drug discovery inititative that much of this thesis is a part of, is also introduced.

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 bioactivity 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.

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

Chapter 1

Introduction

The design of a new pharmaceutical treatment for a disease is a complex process involving many stages. Breakthroughs in medical research have enabled a wide variety of different treatment modalities ranging from vaccines to gene therapy, tailored for different diseases - in this thesis, we will narrow our scope to small-molecule drugs, which remain the most common form of treatment for a wide variety of diseases. The development of a new small molecule drug begins with the identification of a target, which is a protein or other biomolecule that plays a role in the disease's biochemical mechanism of action. Afterwards, a 'hit' compound, which is a small molecule that binds to the target with relatively weak potency, is found as the starting point for a drug design campaign. This brings us to the 'hit-to-lead' stage where modifications to the hit are performed to identify a lead compound, which is a small molecule that is potent against the target at very low concentrations. The lead is then optimized to improve its other molecular properties such as target selectivity, solubility, and toxicity. Finally, the development candidate is taken into preclinical studies and clinical trials to determine its safety and efficacy in humans.

The stages between target identification and development candidate nomination traditionally follow 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 bioactivity/molecular properties, 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 bioactivity/selectivity/toxicity criteria.

Fully completing a drug design campaign from target identification to preclinical studies can be very time-consuming and costly, and there is a pressing need to reduce the time and cost of the drug discovery process to enable faster development of new treatments.

Computational methods offer a promising avenue to accelerate the drug discovery process by automating the design-make-test cycle. The dream of replacing slow and expensive experiments

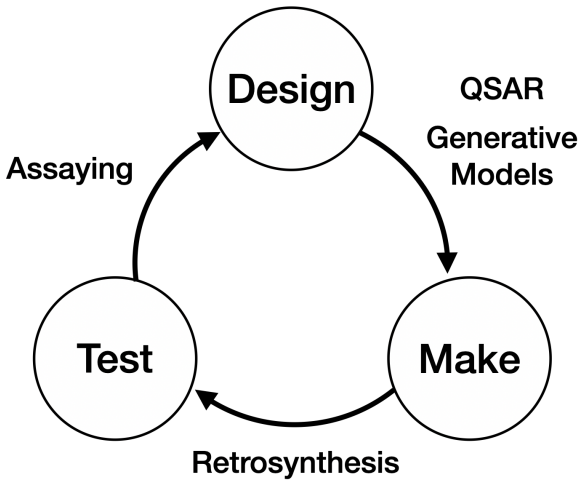


Fig. 1.1 An overview of the design-make-test cycle in drug discovery.

with fast and cheap predictive algorithms is not a new one, and continuous progress has been made in this direction over the past few decades. Recently, there has been a huge surge in applying machine learning (ML) methods to drug design following its success in various other fields, most notably computer vision and natural language processing.

The application focus for these ML methods has been on the 'design' and 'make' parts of the cycle, where they have outperformed traditional approaches on a variety of tasks from property prediction to synthesis route planning, and efforts are underway to use ML models to replace aspects of decision-making conventionally done by humans, such as the design of drug candidates.

Much of the progress in this area comes from adapting the latest state-of-the-art algorithms from the ML literature, and while this is a good starting point, it is not sufficient to fully leverage the potential of ML in drug discovery. The quantity and structure of data that is available in drug discovery are very different from that conventionally found in ML applications, and it is necessary to confront this fact and tailor ML models specifically for the unique problems and situations faced in pharmaceutical chemistry to drive the development of computational tools that truly suit our needs for accelerating the drug discovery process.

This is the key challenge that we shall explore in this thesis - how we can leverage data-driven approaches based on machine learning in the design-make-test cycle, grappling with the practical difficulties of a drug discovery campaign where we must make full use of the limited data available.

Design

The ‘Design’ stage can be broken down into two challenges: (i) predicting the property of a compound from its structure, and (ii) using that predictive model to propose a set of new compounds that are likely to have the desired property. The first challenge has historically been very well-studied: As stated by Alexander Crum Brown in 1868, the “physiological response of a compound is merely a function of its chemical constitution” [1] - chemists have been endeavoring to define this function, and predict the properties of compounds without synthesizing them since.

A milestone in the field was made by Hansch and Fujita in 1962 in defining Quantitative Structure-Activity Relationship (QSAR) equations. These were manually defined mathematical models that could be used to predict molecular properties such as the octanol/water partition coefficient (logP) and biological activity, from descriptor coefficients derived from molecular structure [2, 3].

Coincidentally, the field of artificial intelligence (AI) was also born in the late 1950s/1960s and the potential of applying computational methods to chemistry was recognized early on with the Dendral system in 1965, which assisted chemists in interpreting mass spectrometry data [4]. Dendral is not only recognised as the first formal application of AI to chemistry but also as the first expert system since it was capable of automating some tasks of organic chemists and thereby improving decision-making.

Further research into applying AI in chemistry greatly decreased following the general collapse in AI research from the 1970s onwards, and afterwards the workhorse for computational methods in drug design has been the use of simulations. Advances in computational processing power and the availability of protein-ligand crystal structures led to the development and use of techniques such as molecular docking, molecular dynamics, and free energy perturbation (FEP) [5]. The success of these methods has cemented the role of computational tools in drug discovery, and they remain in extensive use today.

The breakthroughs of machine learning in other fields such as image recognition and natural language processing in the 21st century have led to a resurgence of interest in applying these methods to drug discovery. The victory of a neural network model in the Merck molecular activity challenge [6] set in motion active research into the application of deep learning for molecular property prediction which continues to this day.

Much of this research consists of translating the latest state-of-the-art algorithms from the ML literature and applying them to the same retrospective QSAR model benchmarks repeatedly to compare performance between different models. While this has undeniable value in quantitatively driving the development of improved models, by focusing only on retrospective benchmarks and avoiding prospective deployment, model developers overlook

the unique challenges faced in real-world drug discovery and the practical needs of the end user [7].

This thesis addresses this in two places. In Chapter 3 we look at how to leverage fragment-protein structures from a crystallographic fragment screen to perform hit discovery in the absence of any bioactivity data via an unsupervised learning approach. Moving onwards to the early hit-to-lead stage, where bioactivity data is limited, noisy, and dominated by inactive molecules, in Chapter 4 we use a learning-to-rank framework to make use of inactive data and overcome experimental noise.

Once a predictive model is obtained, the next step of 'Design' is to propose a set of new compounds for synthesis and testing. Traditionally this is done by applying the predictive model to a library of compounds (either commercially available or constructed by human experts) and selecting the top-scoring compounds. However, this approach is limited by the coverage of the library, and vast chemical space remains unexplored. Generative deep learning models have been proposed as an alternative approach, where a neural network that can generate novel compounds is combined with a predictive model for exploring chemical space. This approach has shown to be successful for several toy problems and offers an interesting alternative to traditional library screening, particularly in the context of intellectual property [8]. However, even for toy examples, it has been shown that imperfections in the predictive model are exploited by the generative model to generate flawed compounds [9]. This is a problem that will be exacerbated in a real-world use case, and it is clear that improving predictive modelling of molecular properties is the fundamental bottleneck to successful 'Design'.

Make

After proposing a set of compounds, the next step is to synthesize them - the central focus of the field of synthetic organic chemistry. Designing a set of chemical reactions that transform a set of starting materials into a target molecule is approached via retrosynthetic analysis, a concept introduced by E.J. Corey [10]. In this approach, the target molecule is progressively simplified into precursor molecules until commercially available compounds are reached. This logical approach paved the way for the successful synthesis of many complex molecules, for which Corey was awarded the Nobel Prize in Chemistry in 1990 [11].

In addition to introducing the concept of retrosynthetic analysis, Corey was also a pioneer in applying computational methods to support the design of synthetic routes [12]. By utilising computers to search databases of available compounds as well as previously recorded chemical reactions instead of relying on the necessarily limited experience and domain knowledge of

individual chemists, computational tools allowed the design of shorter, more efficient synthesis routes.

Modern-day approaches build on this work, employing deep learning models instead. By taking advantage of vast repositories of reaction data held in public and proprietary databases as well as ever-growing libraries of commercially available building blocks, ML models for synthesis planning have found significant success in determining synthetic tractability as well as synthesis planning [13]. Indeed, recent state-of-the-art models have demonstrated their effectiveness through a variation of the Turing test, where computationally designed synthesis routes compete well against those from expert human organic chemists [14].

While there remain limitations on predicting subtle selectivities and reaction yields, it is generally accepted that ML models can predict the major products of robust reactions that are well-represented in public datasets. This has led to their increasing deployment and use in industry [15].

However, deep learning-based reaction prediction models critically suffer from a lack of interpretability. Their black-box nature means it is neither clear if the models are making correct predictions because of correct reasoning, nor is it clear what training data and biases they are relying on to reach a prediction [16]. As these models are made more and more available to non-expert end users, it is increasingly important to understand the decision-making of these models in order to avoid incorrect predictions and drive the development of more robust models. To address this, in Chapter 5 we showcase a workflow for explaining the reasoning of reaction prediction models.

Test

The major focus of machine learning approaches to property prediction has been on improving performance on retrospective datasets, typically involving curated biological activity data obtained from biochemical assays. A great deal of effort is spent on improving molecular representation and the predictive performance of these models, but relatively little thought is given to how this experimental data is generated, and the Test step of the drug discovery process is largely treated as a black box.

In reality, different biochemical assay techniques have different accuracies and idiosyncracies and the data generated from different assays should not be modelled equally [17]. Innovative experimental techniques such as phenotype screening [18], DNA-/Peptide-encoded libraries [19, 20], and high-throughput mass spectrometry [21] are being developed and there are exciting opportunities to apply machine learning methods for the interpretation of these new forms of experimental data.

Beyond just developing ML models to learn this kind of data, there is great potential in engaging with experimentalists to more tightly integrate ML with experimental techniques for the design of ML-centric design-make-test workflows. We explore this in Chapter 6 where we apply machine learning to crude bioactivity data from nanomolar-scale high-throughput chemistry.

Chapter 2

Background

2.1 Molecular Representation

2.1.1 SMILES

The simplified molecular-input line-entry system (SMILES) [22, 23] is a widely-used text-based description of molecular structure, developed to store chemical data in a computer-readable format. In SMILES strings, atoms are represented with their chemical symbols, and aromatic atoms are denoted in lowercase. Special characters are utilised to fully specify the chemical structure of a molecule following specific rules. For example, the characters = and # are used to represent double and triple bonds, @ characters are used to denote chirality, and \ and / characters specify local double bond configurations (see Table 2.1 for examples). Parentheses are used to denote 'branches' in the chemical structure, and cyclic substructures are encoded by breaking a single bond in the ring and labelling the matching atoms with numbers (Fig. 2.1).

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 are constructed by placing a > character between the SMILES strings of reactants, reagents, and products. If multiple molecules participate in the reaction, 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

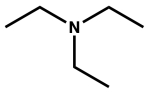
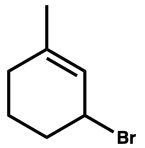
SMILES	Structure
C	CH_4
[Fe2+]	Fe^{2+}
C=O	CH_2O
C#N	HCN
CCN(CC)CC	
CC1=CC(CCC1)Br	

Table 2.1 Examples of SMILES strings.

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 [24], while USPTO consists of 480k reaction SMILES strings [25]. For text-based ML models such as the Molecular Transformer (see chapter 5), 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.

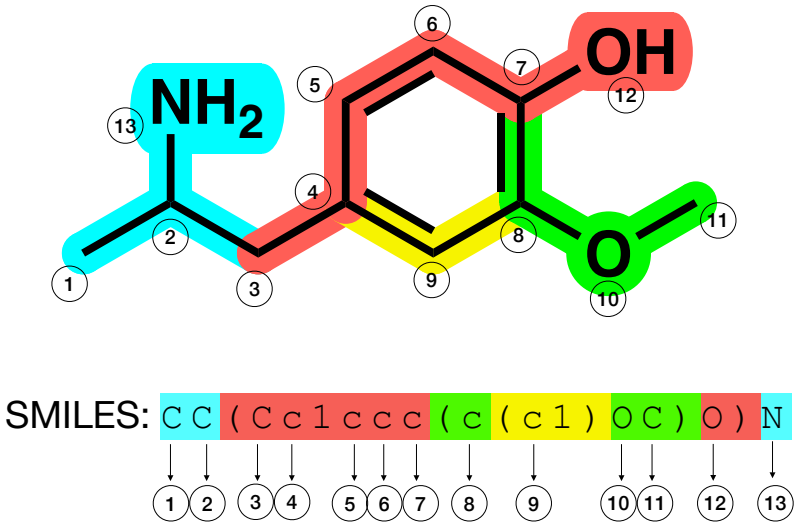


Fig. 2.1 **Illustration of the mapping from chemical structure to SMILES.** The chemical structure of a molecule is divided into several 'branches', which are shown in different colors. In SMILES, each branch is enclosed by parentheses. Adapted from [26].

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) [27] 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) [28] 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 outputs 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 [29]. 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, SMARTS also allows the specification of wildcard atoms and bonds, which allows expressive and precise definitions of substructures and atomic environments for searching chemical databases. 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

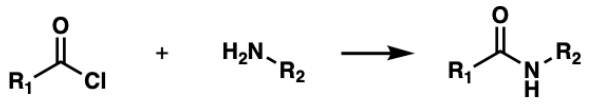
Table 2.2 Examples of SMARTS patterns.

The precise substructure specification of SMARTS is useful in many aspects of 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) [30] and PAINS (Pan Assay Interference Compounds) Filters [31]. Similarly, SMARTS queries are used to design 'structural alerts' that flag molecules containing reaction chemical substructures which may lead to undesirable toxicity in the compound itself or its metabolites [32].

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/proprietary sets of SMARTS strings tailored for their particular use cases. (see section 2.1.3 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 2.2). Beyond querying 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 [33, 34] *in silico* both for more specific use cases such as probing the SAR of a compound as well as the construction of ultra-large commercially available compound libraries eg ZINC [35], EnamineREAL.




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

Fig. 2.2 **An example reaction template.** This template describes 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 to the input to obtain the predicted outcome of the reaction. This approach was originally proposed for the reverse problem of retrosynthesis [12] and has had success in forward reaction prediction for the design of synthetic pathways to drug-like molecules [36].

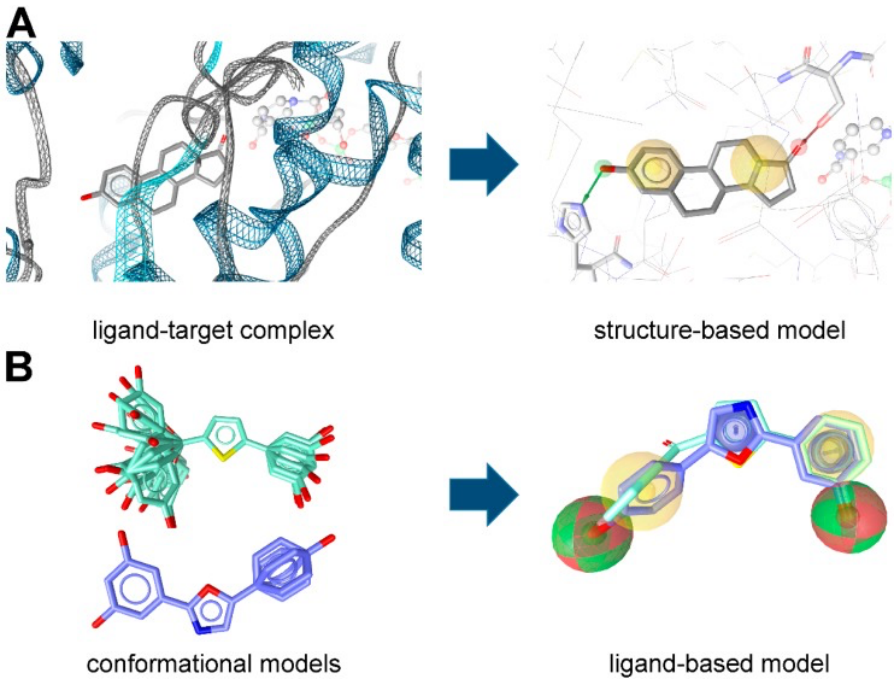


Fig. 2.3 **An illustration of pharmacophore model construction.** (**a**) Structure-based models are constructed from observed protein-ligand complexes while (**b**) ligand-based models are constructed by aligning conformers and identifying overlapping pharmacophores. Reproduced from [37].

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 [37]. 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 models can explain how 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.

Pharmacophore models can either be constructed using structural data or purely ligand-based data from a set of active compounds [38]. 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 2.3A).

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 2.3B). Although this approach circumvents the need for structural data which might not be possible to obtain, a downside is that all of the extracted pharmacophores 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 [39–42].

2.1.4 Fingerprints

Most classical machine learning methods require a (typically fixed-length) vector representation of molecules known in the literature as 'molecular fingerprints'. The most popular molecular fingerprint is the Morgan fingerprint [43], also known as the Extended-Connectivity FingerPrint (ECFP) [44]. 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 its 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 usefulness in calculating molecular similarity [45]. 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 [46]:

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

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 fingerprints. Two structures are usually considered dissimilar if the Tanimoto similarity is < 0.4 [47]. Alternate similarity measures exist but several comparison studies [48, 49] 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 [50] datasets of similar compounds as well as performing similarity-based virtual screening. Similarity-based virtual screening relies on the similarity property principle (SPP) [51], 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 [52], empirically it has been shown that structurally similar compounds are much more likely to be active compared to dissimilar ones [53]. Performing similarity-based virtual screening in practice involves calculating molecular similarities between known active compounds and unknown molecules from a database, then selecting those with the highest similarities; ECFP4 is a consistently well-performing fingerprint for this task [54, 55].

In addition, Morgan fingerprints have also been shown to be versatile as molecular descriptors for machine learning (ML). Machine learning models learn statistical patterns from data and can be used to make predictions on new data (see section 2.2.2). In the context of drug discovery, ML can be used to 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 [56], biological activities [57] as well as yields and stereoselectivities for chemical reactions [58]. 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 is one of the most frequently used methods in structure-based drug design. [59, 60] The binding mode is the relative location and 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.

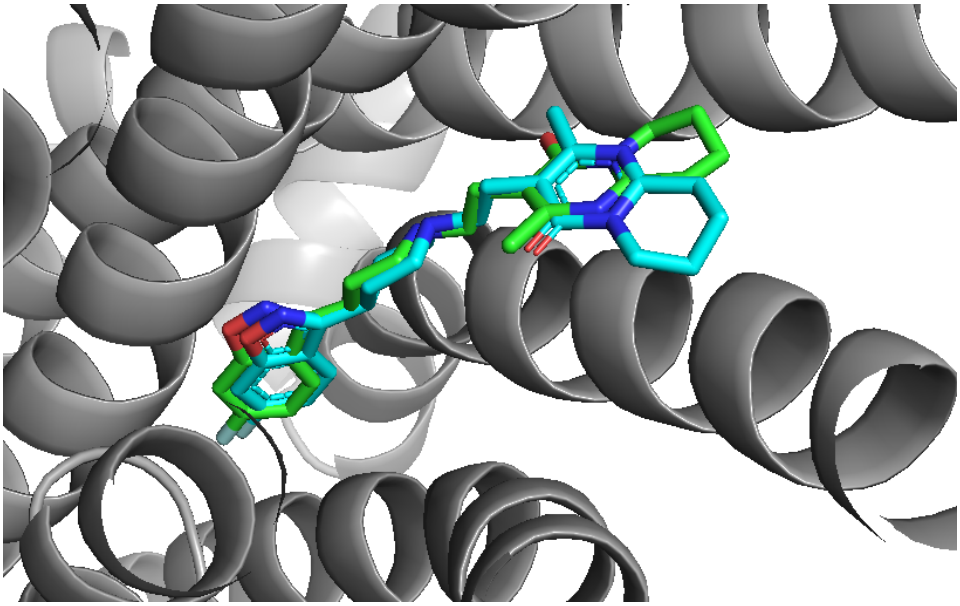


Fig. 2.4 **Example of a docked molecule.** The experimental structure of the ligand risperidone bound to the serotonin 2A receptor is shown in green, with the protein in grey (PDB: 6A93). The structure of the same ligand docked using GOLD via the CSDDiscovery suite is shown in cyan.

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 optimisation problem, where the

coordinates of the ligand and/or protein atoms are adjusted until the 'best fit' is achieved. 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 [61, 62] 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 is typically a physics-based molecular mechanics force field that estimates 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 [63]. 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 impossible 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 [64], Glide [65], AutoDock Vina [66], GOLD [67], and FRED [68]. The relative performance between these docking algorithms is 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 [69, 70]. There are also community assessments on the relative prospective performance of different docking approaches [71] and scoring functions [72].

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 [73]. 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) [74, 75].

Prospectively, large-scale virtual screening with molecular docking has had notable successes. A review specifically looking at G protein-coupled receptors (GPCRs) [76], 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 the 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 [77–79]. At the same time, limitations in the accuracy of scoring functions and the modelling of protein flexibility [80, 81] restrict the ability of docking to reliably distinguish active molecules from inactive ones [82, 83], leading to false positives which are exacerbated when screening large libraries [84].

In the absence of existing ligand bioactivity measurements for a protein target, virtual screening with molecular docking remains the only computational method of choice and is a starting point for beginning drug discovery against a brand new target. However, there have been recent results claiming that deep learning models which use neural networks to directly generate ligand binding poses [85, 86] 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 described in greater detail in refs [87, 88].

Very broadly, machine learning aims 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 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 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 (Fig. 2.5). 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.

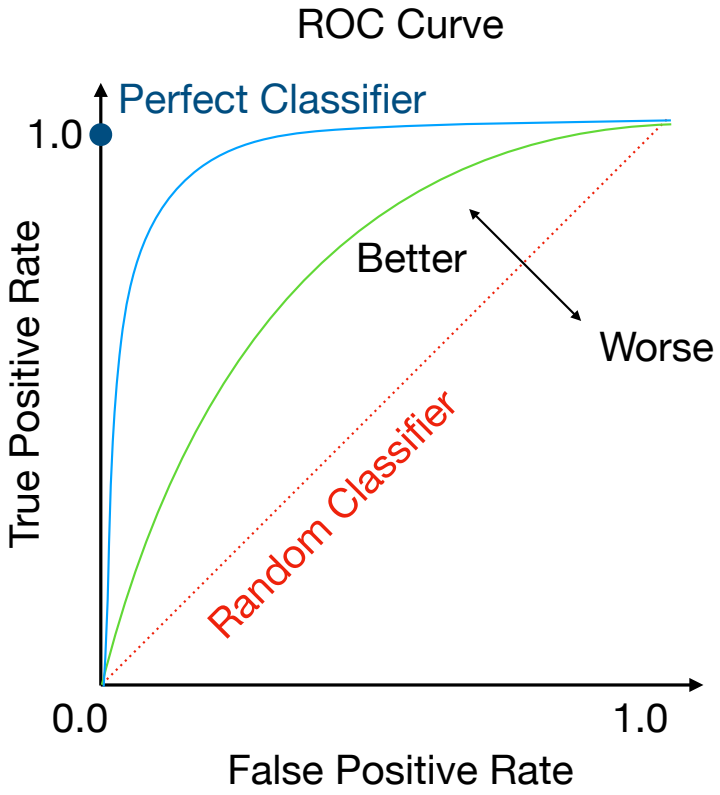


Fig. 2.5 **Example Receiver Operating Characteristic (ROC) curves.** The diagonal shows the performance of a random classifier. Two example classifiers (blue, green) are shown. Adapted from [89].

A large number of machine learning techniques have been described and applied for drug discovery, and an overview of them can be found in [90–92]. 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 [93]. 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 successively 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 achieve a lower variance and hence improved performance. To further reduce the 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 the prediction of properties such as solubility [94], biological activity [95, 96], and toxicity [97].

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 [98]. The kernel trick is the use of a kernel function to calculate the inner product between the feature 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 [99]. 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 has also led to their use for molecular property prediction [100–102], in particular incorporating the Tanimoto similarity in the kernel function [103, 104].

Deep Learning

In contrast to methods that 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 [105], natural language processing [106], speech recognition [107], and bioinformatics [61, 108]. 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 'backpropagation'. 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 [109]. Further improvements to model optimisation can be achieved by incorporating more sophisticated optimisation algorithms, such as Adam [110], as well as the use of regularisation techniques such as dropout [111] and batch normalisation [112], 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 [113], and ResNets [114]) while recurrent neural networks (RNNs) designed for learning temporal dependencies are applied to sequential data such as text [115] and speech [116] (e.g. Gated-Recurrent-Unit (GRU) [117] and Long-Short-Term-Memory (LSTM) networks [118]).

The Transformer architecture

The Transformer neural network architecture [119] developed for machine translation has had wide-ranging success in many natural language processing tasks, and has been naturally adapted to text-based SMILES representations of molecules and reactions. It has an encoder-decoder structure, where both the encoder and the decoder are made up of so-called transformer blocks which process the inputs by utilising a mechanism known as 'attention'. The Transformer performs text generation by autoregressively predicting the next word in a sentence given the previous words. The Transformer architecture is illustrated in Figure 2.6 (a).

Each string input (sentence) to the model is broken down into different individual 'tokens' (words) that are contained in a vocabulary. For each word in the vocabulary, we generate its own learnt fixed-length vector representation. Passing these vectors to the model would not be enough though as these vectors have no reference to where the given word appears inside the sequence. 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

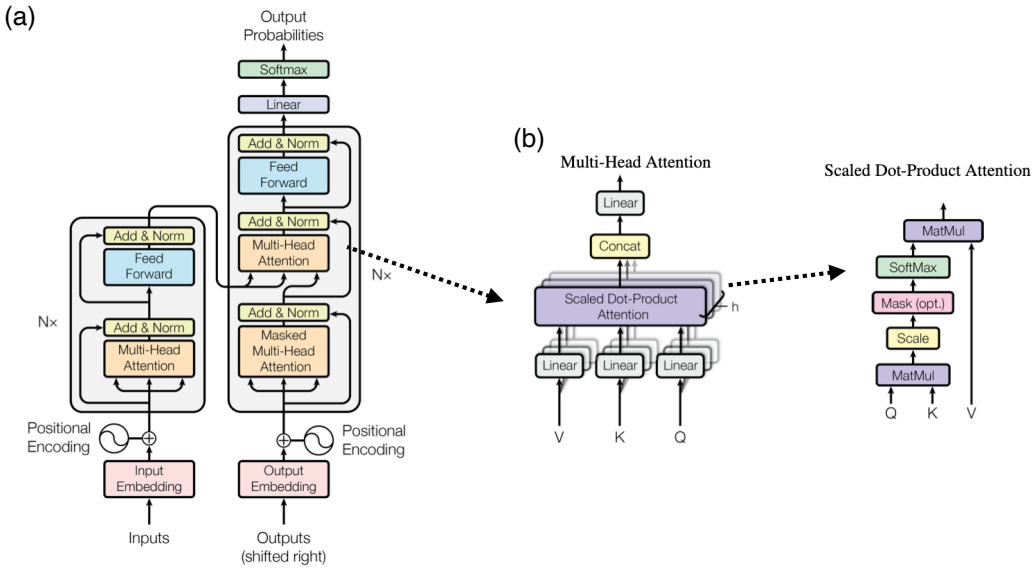


Fig. 2.6 **Graphical illustration of the Transformer model** (a) String inputs to the Transformer are converted into informative vector representations in the encoder (left) before they are fed into the decoder (right) to output probabilities for which token to generate next. (b) Multi-head self-attention is the key mechanism that enables Transformer models to learn complex contextual relationships between words. Adapted from [119].

positions (known as the positional encoding) and added to the token embeddings. An example positional encoding is simply the sine/cosine function of the token position, while more modern approaches use position embeddings that are learnt during model training [120].

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 feed into the decoder layer. The encoder layer is composed of multiple identical transformer blocks, each of which consists of a multi-head scaled attention layer followed by layer normalization and some fully connected feed-forward layers.

The multi-head scaled dot-product attention is the heart of the Transformer model. It is a specific version of a general deep learning technique called attention. Given a set of vector *values* and a vector *query* the attention mechanism computes a weighted sum of the *values* dependent on the *query*. The sum represents a selective summary of the *values* and the *query* determines the importance of each vector, i.e. determines how much each *value* vector is attended to.

The full attention mechanism operates by performing the following steps. Given some vector values $\boldsymbol{h_1}, \boldsymbol{h_2}, \ldots, \boldsymbol{h_N} \in \mathbb{R}^{d_1}$ and a query $\boldsymbol{q} \in \mathbb{R}^{d_2}$

1. First the attention scores $\boldsymbol{e} \in \mathbb{R}^N$ are computed. In the case of the scaled dot-product attention $d_1 = d_2$ and $\boldsymbol{e}$ is simply defined as the scaled vector of projections $e_i = \frac{\boldsymbol{q}^\intercal \boldsymbol{h}_i}{\sqrt{d_1}}$
2. To generate the attention distribution $\boldsymbol{\alpha}$ the softmax of $\boldsymbol{e}$ is taken:

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{N} \exp(e_j)} \tag{2.10}$$

3. Finally the attention distribution is used to take the weighted sum of the values to obtain the final output

$$\boldsymbol{a} = \sum_{i=1}^{N} \alpha_i \boldsymbol{h}_i \ \in \mathbb{R}^{d_1} \tag{2.11}$$

An in-depth review of the attention mechanism can be found in [121].

In the encoder of the Transformer, the attention mechanism is used to update the vector representation of each word in the input sequence by attending to the other words in the sequence, a technique known as self-attention. This allows the model to learn which words usually interact with each other.

Usually the input, be it a chemical formula or human langauge, contains a rich structure with the words affecting each others' meanings in multiple ways. The problem with the simple self-attention mechanism is that by defining a single attention distribution only one way of interaction can be learnt by the model. To give the model the ability to focus on different parts of the input sequence at different times, multi-head attention was introduced. In multi-head attention, the attention mechanism is applied multiple times in parallel in different lower dimensional latent vector spaces and the outputs of the different attention heads are concatenated and combined. The whole mechanism is illustrated on Figure 2.6 (b).

The decoder reads the output from the encoder layer to generate the output sequence. The attention mechanism in the decoder is slightly modified to allow the decoder to attend to outputs from the encoder. The first modification is the application of masking in the multi-head self-attention layers to make sure that every token in the output translation can only attend to the preceding ones. The second modification of the attention mechanism is that the encoder outputs are used as keys whilst the queries are the outputs of the previous decoder layer.

The generation of the translation happens in a sequential manner. First the embedding of the special `<start>` token is fed into the decoder and passed through the layers. The result is projected to a vector that has the same dimensionality as the size of the vocabulary. Finally the softmax function is applied to obtain the probability of the first word. The second word is generated by passing in the first word (and the `start` token) to the decoder, the third is generated by passing in the first and the second etc., in a process known as autoregressive translation. The

prediction is considered final when an <end> token is generated or a pre-specified maximum length is reached.

The Transformer architecture forms the basis of current state-of-the-art models for natural language processing, and active research is ongoing to improve the attention mechanism on long sequences [122] as well as memory efficiency [123] particularly since increases in model size have been shown to improve performance [124].

In the context of drug discovery, the obvious use case of the Transformer model is to apply them on text-based SMILES representations of molecules and reactions. Recent work has shown that Transformer models pre-trained on SMILES can learn useful vector representations for molecular property prediction [125], while the application of the Transformer models to reaction SMILES have shown promising results from reaction prediction [126] to atom-mapping [127]. In particular, The Molecular Transformer [126] uses this approach to perform reaction prediction by 'translating' the SMILES of reaction reagents and reactants to the SMILES of the reaction product, a model that we examine in further detail in Chapter 5.

Deep Learning for Drug Discovery

Beyond the application of Transformer models on SMILES, a wide range of other deep learning approaches has been used in drug discovery. The challenge of modelling molecular inputs has led to the development 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 [56, 128–130], and predicting reaction templates for input reactants [14]. More standard architectures such as three-dimensional voxel-based CNNs have also been trained on protein-ligand complexes to predict binding affinity [131–133].

Neural networks can also be used with pre-computed features such as molecular fingerprints. Example applications include the use of bioactivity prediction [57], reaction prediction [134, 135], and the prediction of docking scores [136].

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 [137].

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 does not

accurately reflect the true generalization performance of the model [138]. This has led to a growing movement towards measuring model performance using 'scaffold split' [56, 130] or 'time split' [138] 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 makes it difficult to understand the underlying reasons for a model's predictions. Without explanations of model predictions, it becomes difficult to avoid correct predictions for the wrong reasons (the so-called clever Hans effect) [139], 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 [16].

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

2.3 COVID Moonshot

This thesis was written during the COVID-19 pandemic, with more than 689 million confirmed cases & 6.8 million deaths worldwide up to Apr 2023 [140]. In parallel to the rapid development of vaccines against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) which is the cause of COVID-19, was the development of antiviral therapeutics to treat those who are infected.

The COVID Moonshot consortium was launched in March 2020, aiming to find oral antivirals against COVID-19 in an open-science, patent-free manner [141, 142]. The consortium focussed on developing inhibitors against the SARS-CoV-2 main viral protease (Mpro), an attractive target for antivirals for several reasons:

Firstly, Mpro is essential in the viral life cycle of SARS-CoV-2. The virus replicates itself after infecting a host cell by generating a long polyprotein that gets cleaved into individual proteins by multiple viral proteases. Mpro is one of these proteases, and it is responsible for cutting the viral polyprotein into smaller non-structural proteins (nsps) that are essential for replicating the viral RNA genome. By developing drugs to inhibit Mpro, the cleavage of the polyprotein into nsps and hence viral replication can be blocked [143].

Secondly, the binding site of Mpro is distinct from known human proteases, thus inhibitors of SARS-CoV-2 Mpro are unlikely to be toxic by interacting with human proteases and causing side effects [144, 145]. In addition, Mpro is highly conserved across different coronaviruses

so a potent inhibitor of SARS-CoV-2 Mpro could be fruitful for the future development of pan-coronavirus antivirals [146].

In vitro and in vivo studies clearly show that Mpro inhibition leads to antiviral activity in cell culture [147, 144], which is corroborated by the recent clinical success of nirmatrelvir (the Mpro inhibitor component of Paxlovid) [148] and ensitrevir [149, 150].

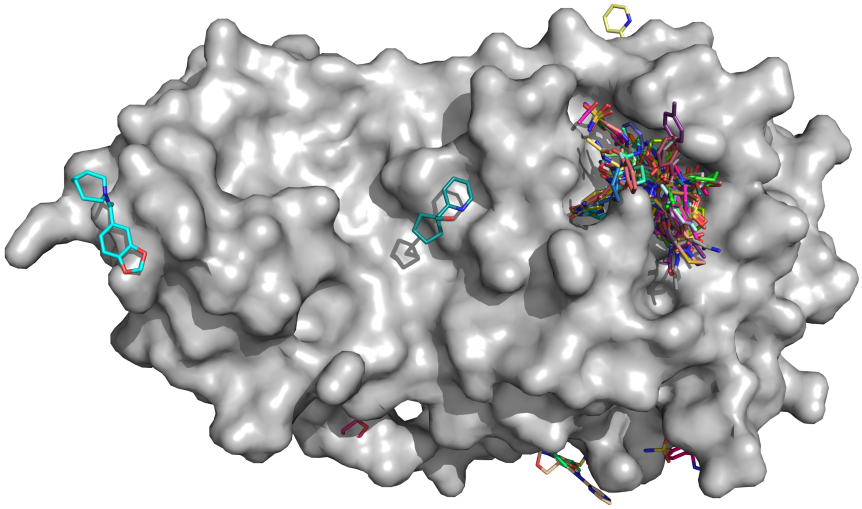


Fig. 2.7 **Crystallographic fragment screen against Mpro.** Overlaid X-ray crystal structures from a fragment screen against Mpro that formed the starting point of COVID Moonshot (PDB ID: 5R8T). Each fragment is shown in a different colour, with the majority of them bound in the active site of Mpro on the right side of the figure. Reproduced from [151].

COVID Moonshot embarked on a structure-based drug discovery campaign against Mpro, building on a crystallographic fragment screen that revealed potential binding interactions in the active site of Mpro (Figure 2.7). Leveraging crowdsourcing, high-throughput structural biology, machine learning, and exascale molecular simulations, the COVID Moonshot campaign was able to discover many potent non-covalent Mpro inhibitors, and the lead candidate from the campaign is currently undergoing preclinical studies.

Throughout the campaign from hit discovery to lead optimisation, all compound designs (>18,000 designs), crystallographic data (>840 ligand-bound X-ray structures), assay data (>10,000 measurements), and synthesized molecules (>2,400 compounds) were shared publicly [152].

Participation in COVID Moonshot forms the backdrop of many chapters of this thesis, for performing computational validation against retrospective data as well as for the prospective design of new compounds, exploring ways to apply the computational tools described above to accelerate the design-make-test cycle for Mpro inhibitor discovery and by extension drug discovery more broadly.

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 [17]. 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, protein crystals are soaked with high concencration of very low molecular weight compounds ('fragments' with typically less than 18 non-hydrogen atoms [153]) and the resulting protein-fragment complexes are resolved 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 [154]. 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 different spatially proximal fragments [155, 156]. 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 [157, 158]. 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 [142]. 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 virtual screen of 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 3.1a).

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 3.2). These are 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 [37]. 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.

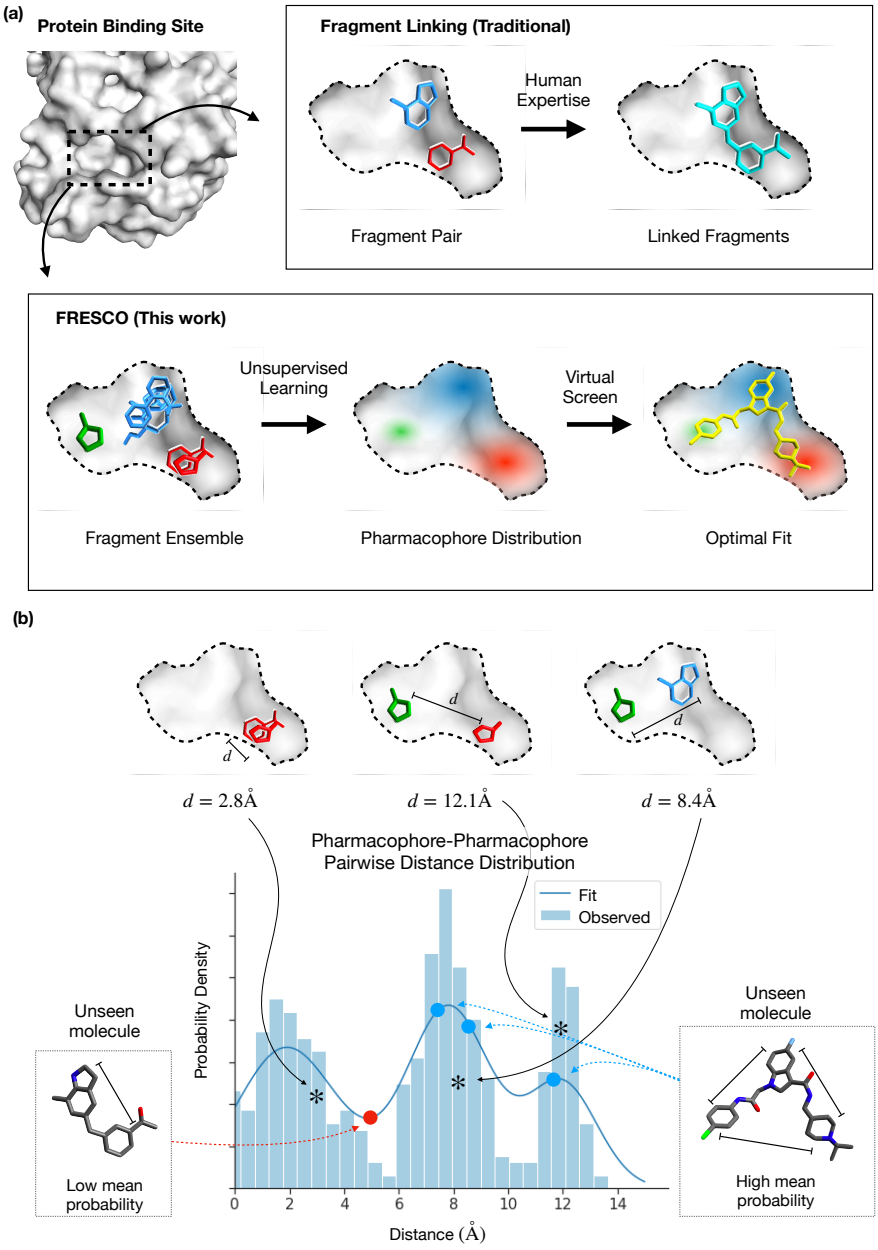


Fig. 3.1 **Illustrations of the FRESCO method.** (**a**) FRESCO conceptually differs from traditional fragment merging/linking by taking a distributional approach. (**b**) Unsupervised learning of the distribution of pairwise pharmacophore distances from fragment ensembles allows the virtual screening of unseen molecules.

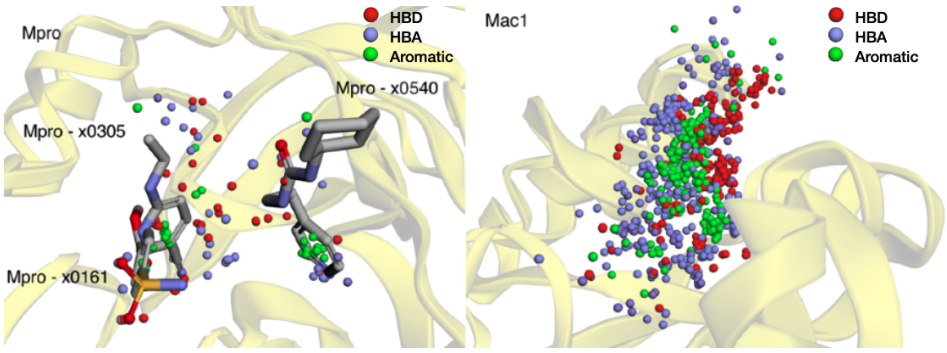


Fig. 3.2 **Examples of pharmacophore distributions from fragment ensembles.** The red, blue, and green spheres depict hydrogen bond donors, acceptors, and aromatic pharmacophores respectively from fragment ensembles in the 3D binding sites of (**a**) Mpro, and (**b**) nsp3-Mac1. Several of the Mpro fragments 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.

We utilise kernel density estimation (KDE) [159] to estimate this spatial distribution of pairwise pharmacophore distances. We then score unseen molecules by evaluating pharmacophore distances within that molecule against the probability distribution of pharmacophore distances derived from the fragment ensemble (Figure 3.1b). 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 [160], SyntaLinker [161], and Develop [162]), as well as data-mining methods such as Fragment Network [163], 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 be 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 [164]. 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 models of ligand-protein binding such as docking [77–79] 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 prospective 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 [7]. 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 [158]. For Mpro, non-covalent fragments from the XChem fragment screen [158] were used while for Mac1 both XChem and UCSF fragment data were used [165].

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` [166] 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 matrix of pairwise distances between all possible pharmacophore pairs (eg Donor-Donor & Aromatic-Acceptor) between different fragment molecules. The flattened distance matrix for each pharmacophore pair is the distribution that we choose to model.

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 pairwise distribution using the Improved Sheather-Jones algorithm [167]. Each pharmacophore pair is associated with its

own KDE, and the collection of KDE models for all of the pharmacophore pairs comprises the FRESCO model.

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 probabilities of each distribution 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. The 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 [142] which is targeting the SARS-CoV-2 main protease (Mpro). 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 (see Section 2.3 for more details on COVID Moonshot).

The Moonshot activity data for the retrospective study was accessed on 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 molecules during the fragment-to-lead stage of the Moonshot campaign, we only selected molecules that 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 [168, 169]. 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.

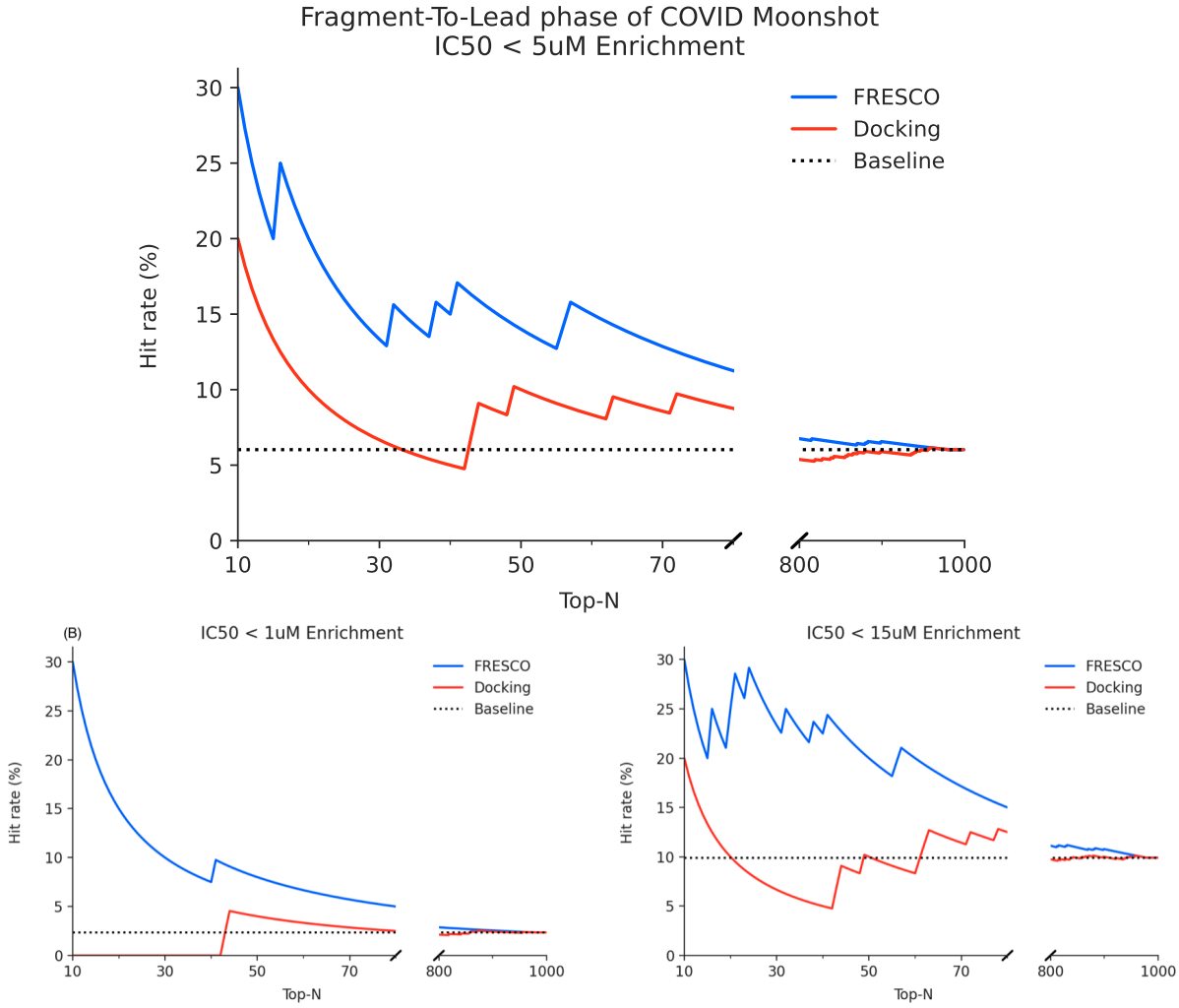


Fig. 3.3 **FRESCO is able to retrospectively perform hit detection.** High hit rates are achieved relative to docking and the human expert baseline when ranking molecules from the fragment-to-lead phase of COVID Moonshot.

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".

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

Figure 3.3 shows that FRESCO achieves higher hit rates compared to both computational docking and 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. The threshold for defining a hit compound is relatively arbitrary and repeated analysis for both lower and higher IC50 thresholds (1-15μM) show similar results (Figure 3.3B). 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 re-score 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.

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 3.4. 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 [170]. 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 [171] were kept. Then, we remove molecules that match known filters for pan-assay interference compounds (PAINS) [31] as well as filters for moieties that are undesirable for medicinal chemistry (eg furan, thiophene, nitro groups). Duplicate tautomers for each molecule are also

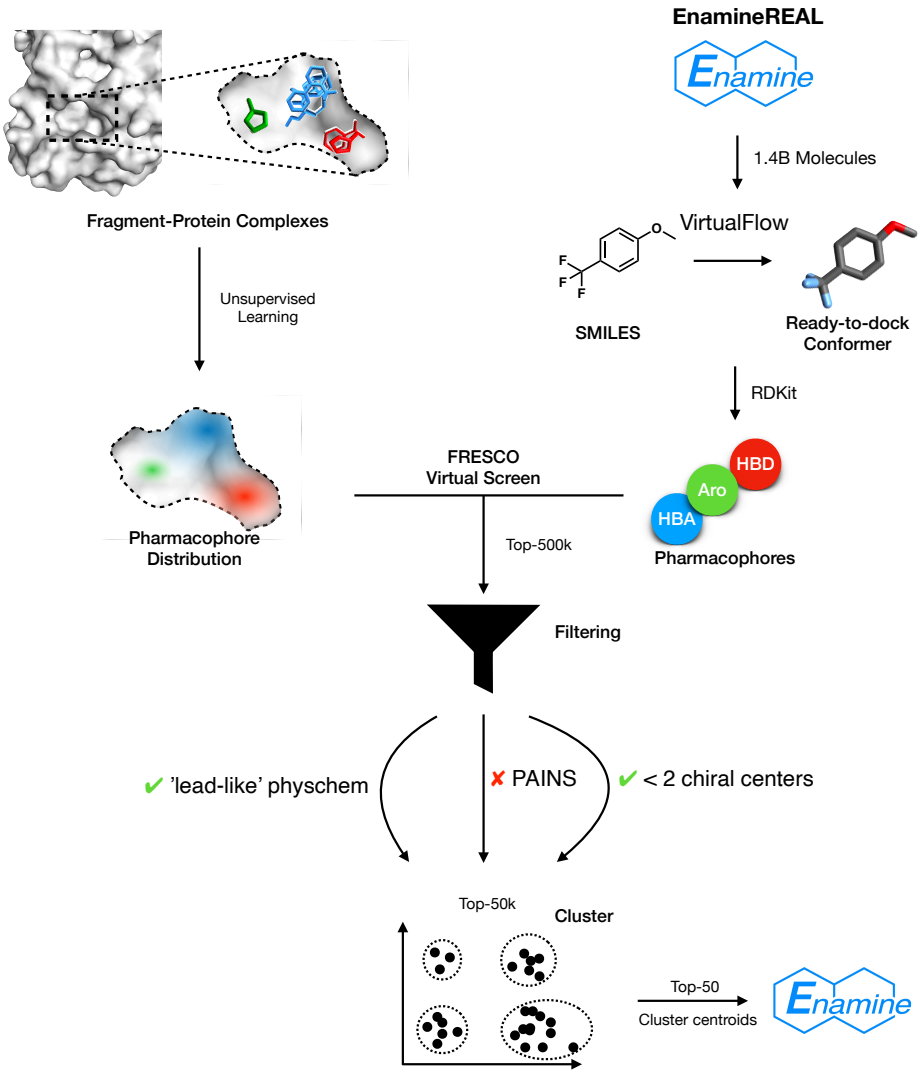


Fig. 3.4 **A schematic of the FRESCO screening workflow.** Target-specific FRESCO models are applied on pre-processed conformers of compounds from EnamineREAL to score them. The top-ranked compounds are then filtered by their physical properties, and clustered by structural similarity. Diverse compounds are chosen for synthesis by picking centroids of the most populous clusters.

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 [50] 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.

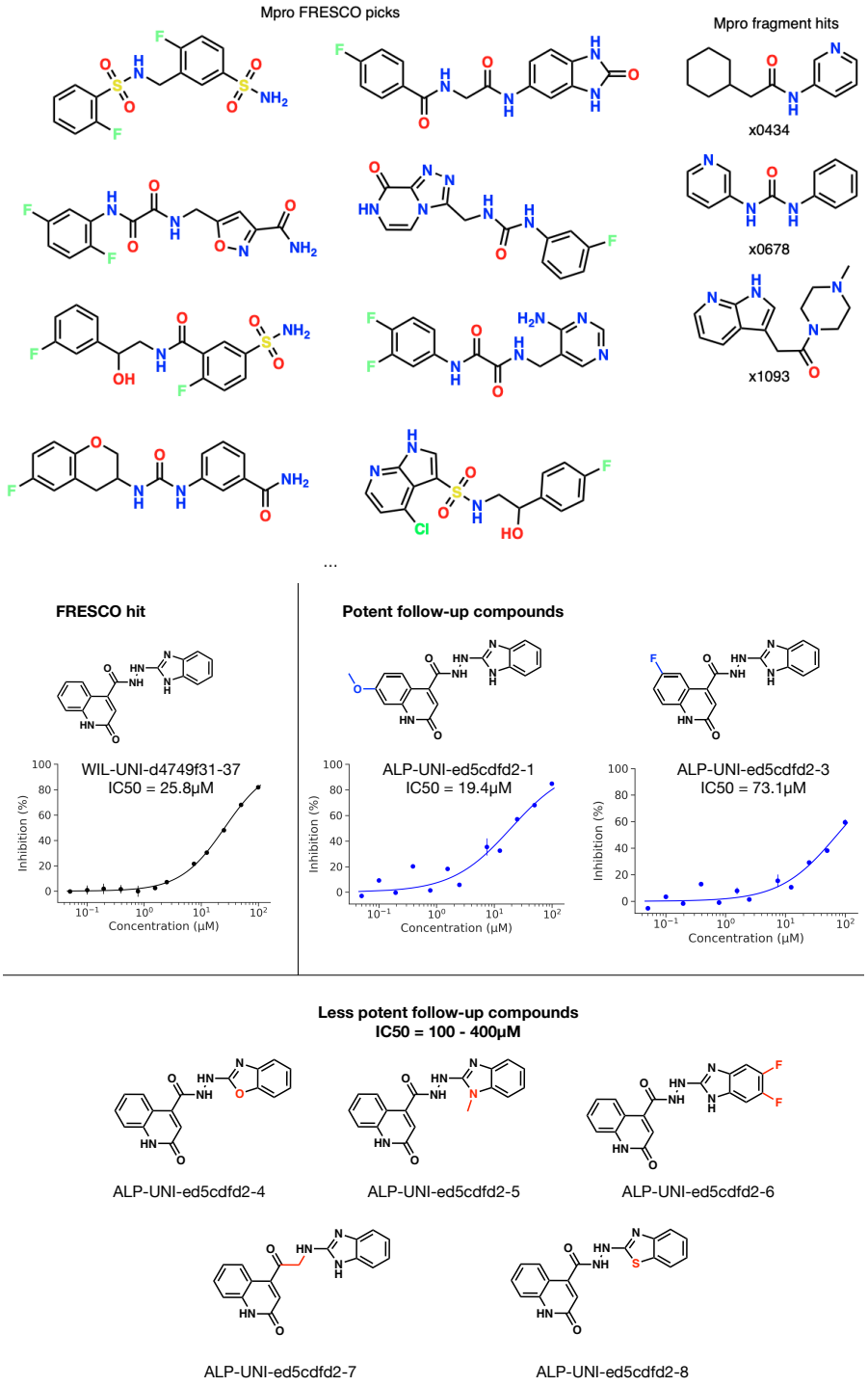


Fig. 3.5 **FRESCO prospectively identifies a hit against Mpro.** (**a**) The molecules favoured by FRESCO tend to have 2 aromatic moieties connected via an amide or an amide isostere, similarly exhibited by three of the initial fragment hits whose structures are also shown. (**b**) Compound WIL-UNI-d4749f31-37 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.

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 most of the other fragment hits possessing an aromatic group bound at similar locations (Figure 3.5a). 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, the COVID Moonshot Consoritium synthesized 8 close analogues to demonstrate the existence of responsive Structure-Activity Relationship [172, 173] (Figure 3.5b). 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 the generalisability of FRESCO in performing hit detection. nsp3-Mac1 is a viral ADP-ribosylhydrolase that counteracts host immune response by cleaving ADP-ribose that is

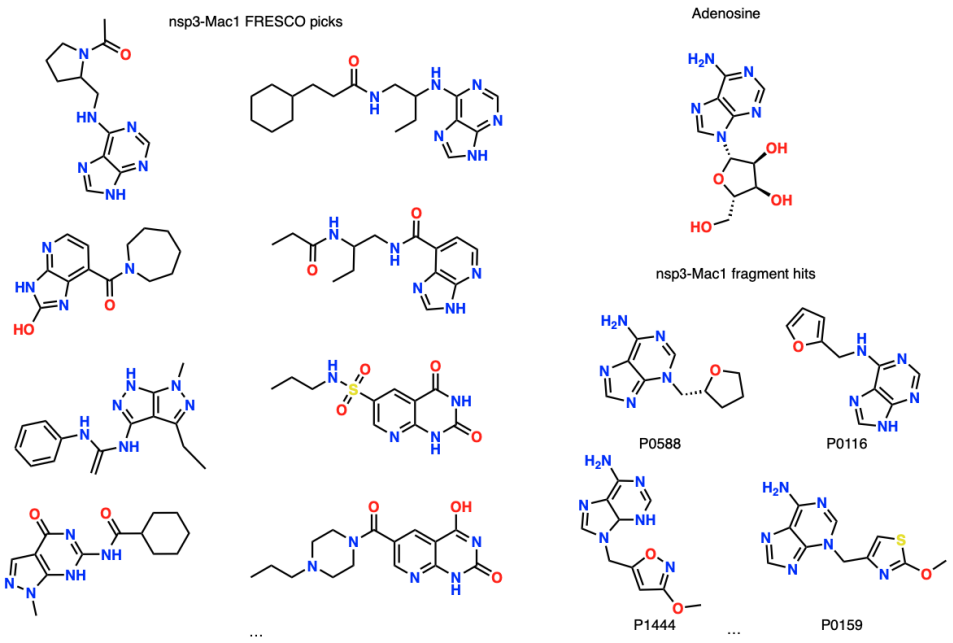


Fig. 3.6 **Top-ranked compounds from FRESCO mimic the natural substrate of nsp3-Mac1.** The molecules favoured by FRESCO tend to contain an acceptor-donor pair spatially proximal to a heterocyclic motif. This mimics adenosine, a core in the natural substrate. This motif is also shared in many of the initial fragment hits, with example structures shown in the figure.

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 [165], 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 heterocyclic motif. This mimics adenosine, a core in the natural substrate, and this

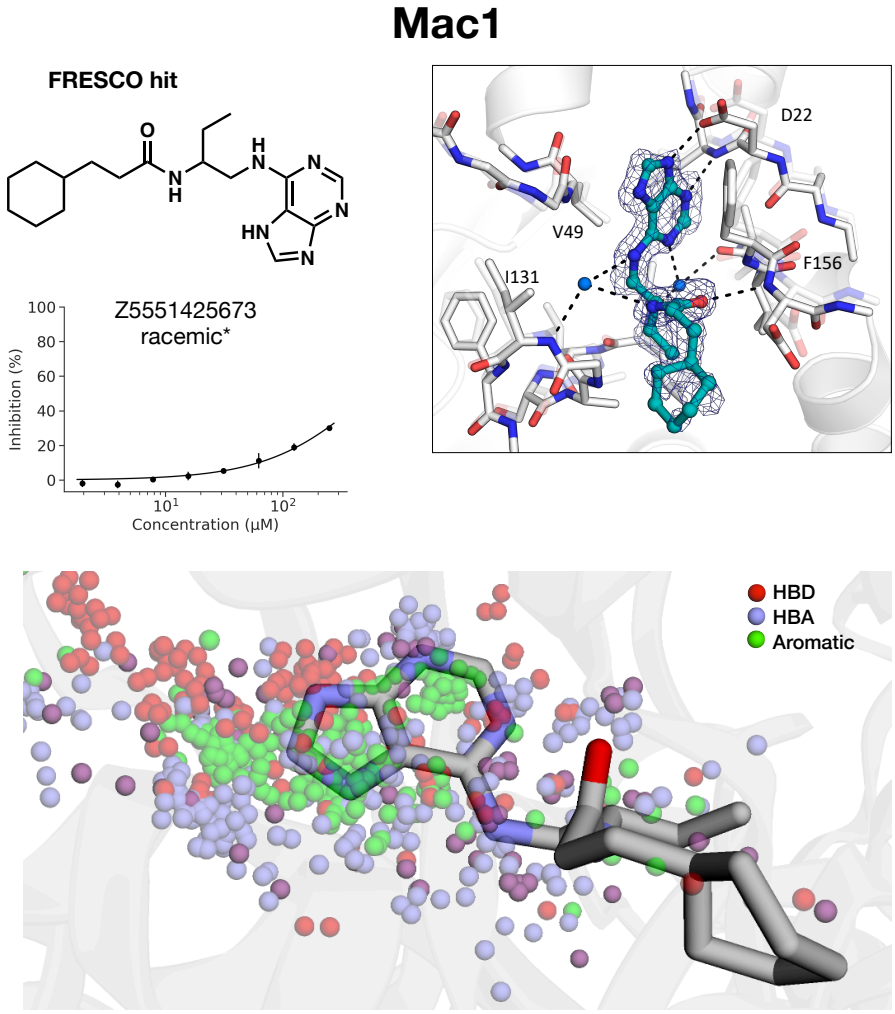


Fig. 3.7 **FRESCO identifies a hit against nsp3-Mac1 with structural confirmation.** (**a**) Compound Z5551425673 is identified as a hit against Mac1 via HTRF assay, with (**b**) hit confirmation via resolution of a crystal structure of Z5551425673 (colored in cyan) bound to the Mac1 active site. (**c**) 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.

motif is shared in many of the initial fragment hits (Figure 3.6). 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 3.7). Crystal structures of 9 other compounds chosen via the FRESCO workflow were also obtained though 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 3.8). These sets 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 multiple fragments are 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

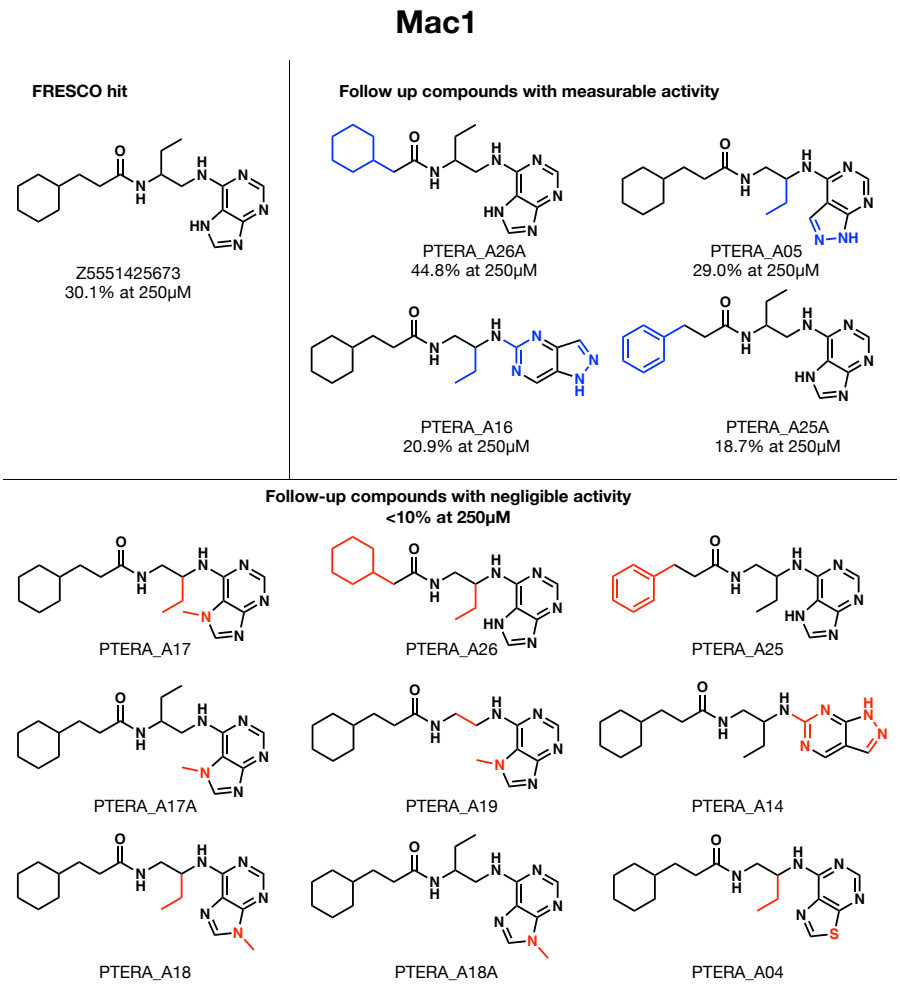


Fig. 3.8 **nsp3-Mac1 hit confirmation via determination of structural-activity relationship.** Close analogues around the hit compound identified by FRESCO, Z5551425673, reveals structure-activity relationship which derisks singleton artefacts.

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 re-scoring "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 constraints 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

In the hit expansion stage of drug discovery, bioactivity modelling and hence compound design is hindered by data insufficiency as the majority of tested compounds are inactive, as well as data noise because 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 make a regression approach challenging. In this chapter, we discuss the prospective use of algorithmic *de novo* design to rapidly expand hits against SARS-CoV-2 Mpro utilising machine learning (ML) models for ranking compounds by bioactivity as well as synthesis route prediction. We demonstrate that our learning-to-rank model outperforms docking on retrospective data, and shows enrichment in a time-split. Prospectively, we algorithmically design 5 new compounds predicted to have higher activity, together with predicted synthetic routes. All designs 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. Instead of predicting IC50 values directly, we use a learn-to-rank approach [174, 175] 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.

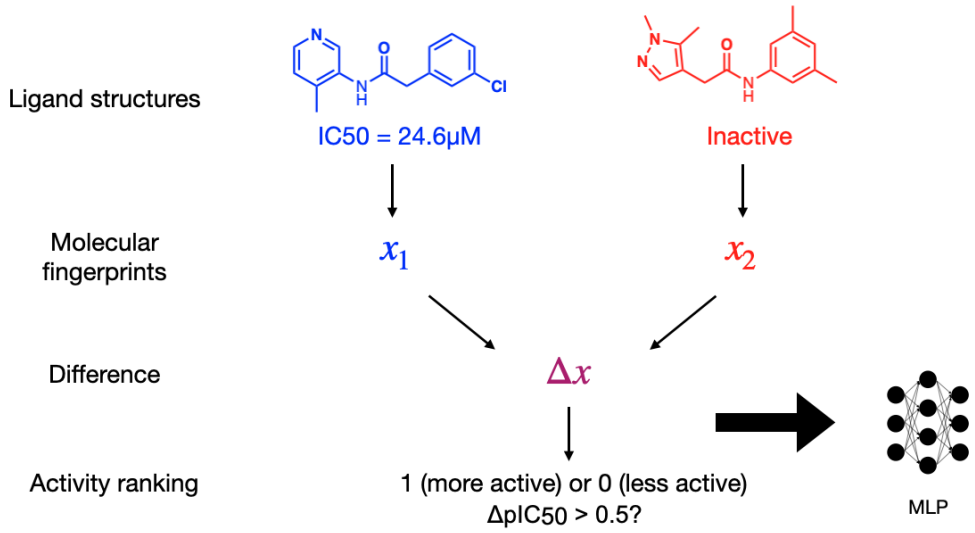


Fig. 4.1 **A schematic of the model setup.** A classifier takes the difference in molecular fingerprint between two molecules and predicts where one molecule is more or less active than the other.

To implement this ranking model (Figure 4.1), 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 the FastAI Tabular framework [176], 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 A.3).

The activity data available from COVID Moonshot at the time of this study consisted of 42 compounds with IC_{50} within assay dynamic range ($< 100\mu$M) and 515 inactives. To construct

a suitable dataset for training the model, the data must be reformated into pairs of molecules. This is done by pairing all actives with all inactives, as well as pairing up actives where there is a significant difference 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 with each other as the ranking of inactivity is not relevant to the task at hand and potentially noisy/misleading.

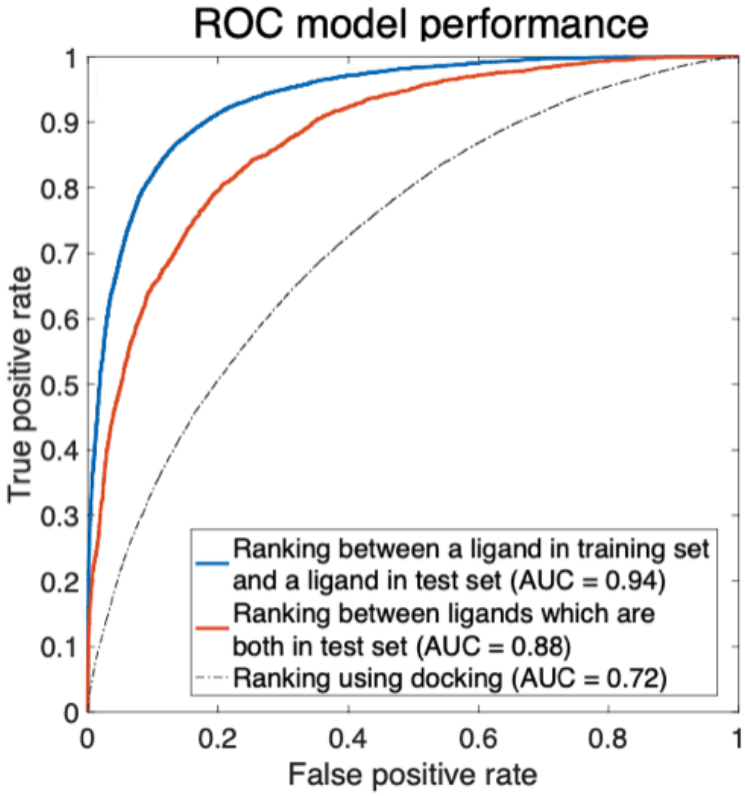


Fig. 4.2 **Relative ranking of ligands can be predicted by our learning-to-rank machine learning model.** The Receiver Operating Characteristic curve of classifying whether a molecule is more/less active than the other.

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 4.2 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 an AUC of 0.94 (95% CI: [0.91,0.98]) where we compare a ligand in the training set against another ligand in the test set; the latter is more relevant as our goal is finding ligands more active than the best incumbent. The 95% confidence interval is computed using bootstrapping. We also compare our model against 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 A.1). Note that docking does not require ligand bioactivity as training data, thus is not a direct 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 this work. During that time, the COVID Moonshot Consortium (a team of expert medicinal chemists) has independently designed, synthesised, and tested 356 compounds [177], out of which 15% were better than the top 2 compounds (having IC_{50} comparable within error) in our dataset. Table 4.1 shows that our model has an enrichment factor of $\sim$2, i.e. if we re-score the 356 compounds synthesized by the medicinal chemistry team using our model, and pick the top 1%-10% percentile, the proportion of molecules that would be better than the top 2 compounds would be $\sim$2x higher than human selection.

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 compounds 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 neural network-based ML model developed in this work.

Instead, we consider a more targeted approach by exploring the smaller, local chemical space of chemical substructures contained within our initial dataset. Building on rule-based fragmentation methods such as BRICS [178] and CReM [179], 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.

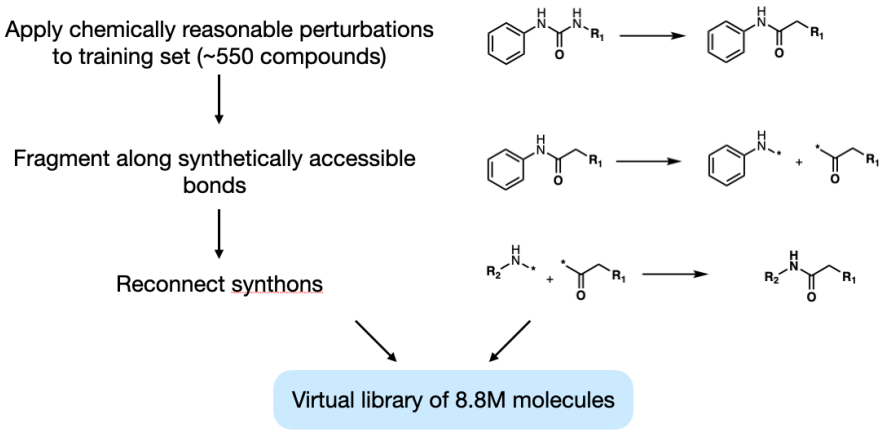


Fig. 4.3 **A schematic of the methodology for the library generation process.**

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 4.3). 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.

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 [180, 181]. The top 5 molecules from the screening library with <4 steps in their predicted routes were synthesised and tested (Figure 4.4A). For comparison, the most potent molecules from the training set are shown in Figure 4.4B. 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 is exploring novel scaffolds.

Figure 4.5 shows that for Compounds **1**, **2**, **4** and **5** our retrosynthesis algorithm generates successful routes, thus provides a reasonable estimate of synthetic complexity. The syntheses were carried out at the Wuxi AppTec and compounds were assayed as received. Minor variations in building blocks were employed depending on what was readily available. We note that our algorithm failed to estimate the synthetic complexity of Compound **3**. The final amide formation step was unexpectedly challenging, and no desired product was seen despite significant efforts in condition screening. Compound **3** was furnished via an alternative strategy, employing an Ullmann coupling to arylate the amide, which was not predicted by our approach.

Compounds **1**-**5** were tested for Mpro activity using a fluorescence assay. Figure 4.6 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

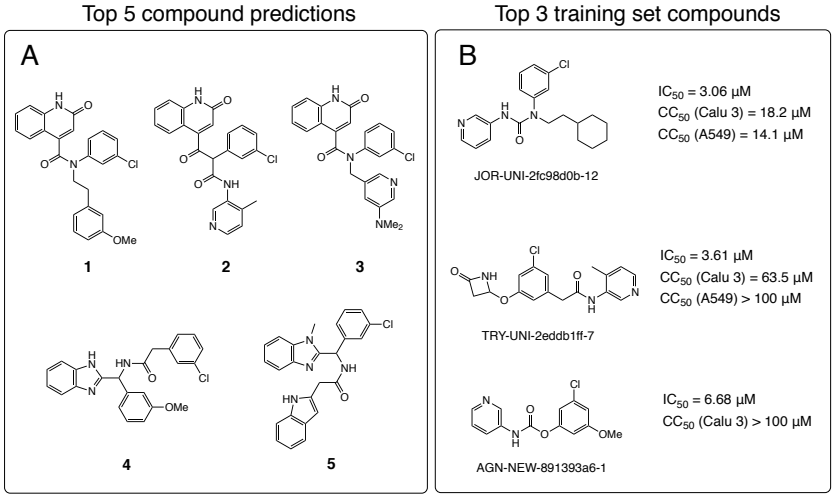


Fig. 4.4 **Our synthesis-driven design model prioritises molecular scaffolds that are not in the top hits.** (A) The 5 compounds selected by our methodology for synthesis and testing. (B) The top 3 compounds from the training set, with potency and cytotoxicity measurements.

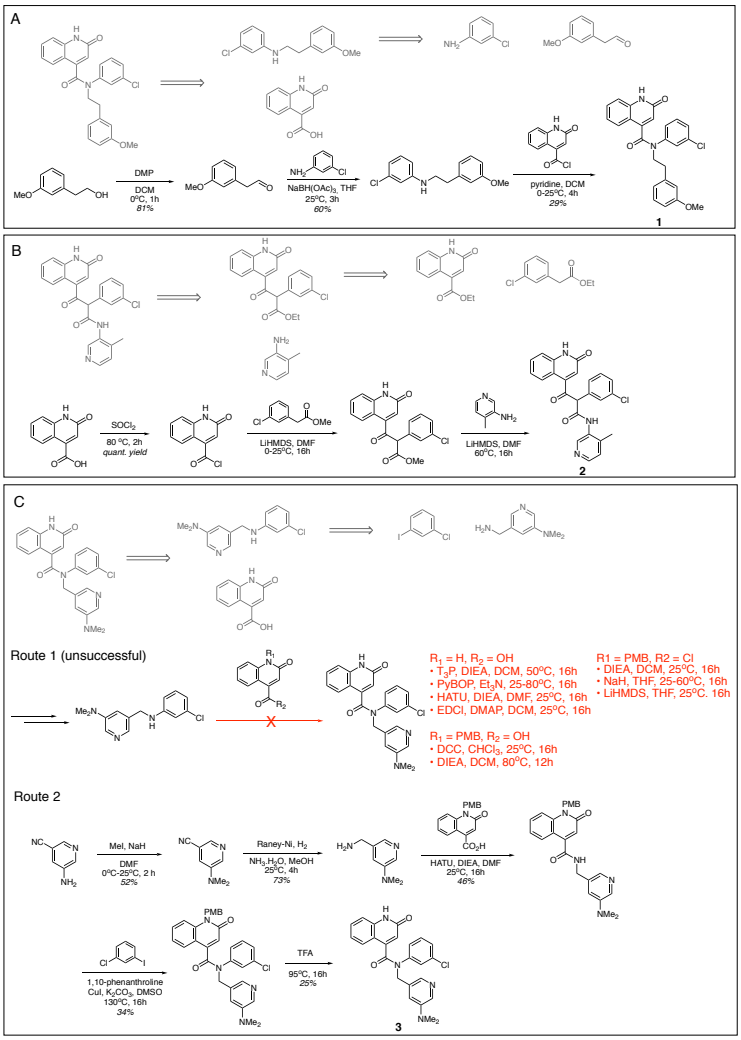


Fig. 4.5 **Model generated synthetic schemes that are experimentally validated.** Schemes (A)-(C) show the synthesis schemes generated by our model (grey) and experimental schemes (black) for Compounds **1**-**3**. The schemes for compounds **4** and **5** can be found in Appendix B.3.

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 B for assay details).

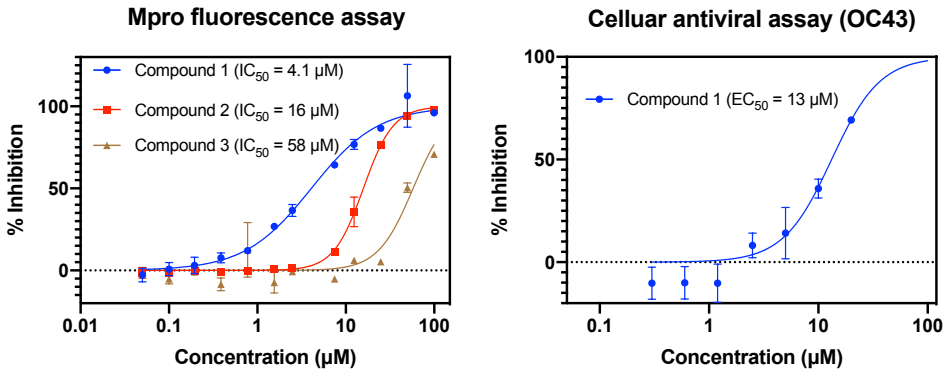


Fig. 4.6 **Three compounds generated using our synthesis-directed model exhibit Mpro activity.** Our most active compound has measurable antiviral activity against the OC43 coronavirus and no measurable cytotoxic effect.

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 [182]. 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 to 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 [183].

Beyond the capability to triage libraries by synthesizability 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 [184, 185]. 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 [186] or reinforcement learning feedback loop [187] will be a challenging but exciting area of future work.

Chapter 5

Quantitative Interpretation of Reaction Prediction Models

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 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 counter-intuitive 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 [188, 189]. The challenge of synthesis planning lies in searching through a 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 [135, 190–193]. However, accurately predicting the outcome of chemical reactions remains a hurdle [13, 194, 195].

The current state-of-the-art in reaction prediction is the Molecular Transformer [196], which employs the transformer neural network architecture that was first introduced for neural machine translation [119]. 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 the 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 were text mined from US patents [25] and filtered [197]. Recent work shows that thorough dataset augmentation improves model performance by allowing it to consider different equivalent SMILES representations [198].

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 the 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 mask model shortcomings, which we resolved by introducing a new unbiased train/test split.

5.2 Molecular Transformer

The Molecular Transformer [126] is a reaction prediction model based on the Transformer architecture [119] designed for natural language processing tasks (see Chapter 2.2.2 for details). By 'translating' reactant/reagent SMILES to reaction product SMILES (Fig 5.1), the Molecular Transformer has shown state-of-the-art performance on reaction prediction benchmarks, outperforming human organic chemists and is the basis for several computational retrosynthesis platforms.

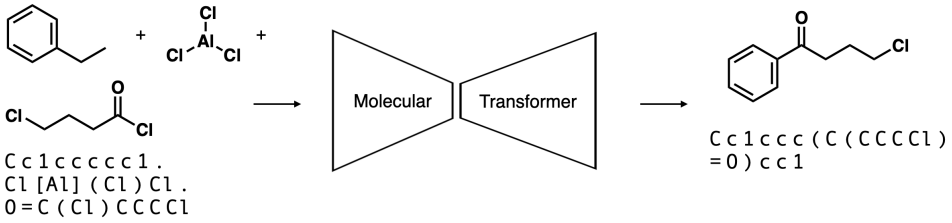


Fig. 5.1 **Schematic illustration of the Molecular Transformer.** The inputs to the model are tokenized SMILES of the reactants and reagents, and the model performs machine translation to predict the most likely product molecule with a probability score.

In Transformer models, strings of text are broken down ('tokenized') into individual 'tokens' which in the case of the Molecular Transformer are the individual characters or subsets of characters of SMILES strings [199]. 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 [200] compared to only 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.

The predictions of the model are generated in an autoregressive way meaning that the output tokens are predicted one at a time conditoned on the previously generated tokens. 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 src is the input to the model, tgt is the predicted output, tok_i is the i-th predicted token, and N is the length of the prediction. The probability of a token is influenced by both grammar (the output should be a valid SMILES string) and semantics (the output SMILES should

correspond to the chemical reaction taking place), which we must disentangle to understand the model's reasoning (see Section 5.3.1).

5.2.1 Training data

The Molecular Transformer is typically trained on datasets of organic reactions text-mined from patent or industry databases. A commonly used one is the USPTO dataset [25, 197] which contains ∼480,000 organic reactions from US patents filed between 1976 and 2016, filtered to remove duplicates and erroneous reactions.

We found that this dataset still contained a large number of erroneous reactions whose sole products 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 in the original paper [196]. We trained a Molecular Transformer model on this dataset and achieved 88.8% Top-1 accuracy on the test set, similar to that reported on the standard USPTO test set (90.4 %). 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 model's 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 [201] but it also hinders generalization because popular reactions are not necessarily better as has recently been shown in the case of inorganic chemical reactions [202].

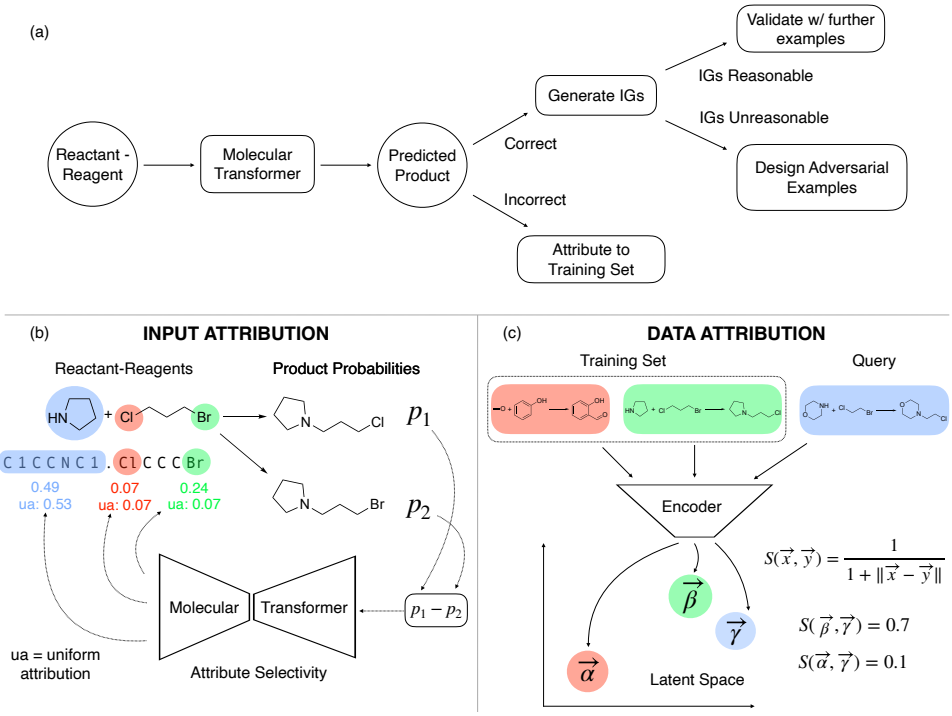


Fig. 5.2 **Schematic illustration of the attribution workflow.** (**a**) Overview of our workflow to interpret the Molecular Transformer. (**b**) Schematic of how the predicted probability difference between two products is attributed back to the reactant-reagent string in order to interpret the model's 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 is used to infer the learnt similarity between query reactants and those from the training set.

5.3 Quantitative Interpretation methods

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 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 the absence of an outcome. Furthermore, the investigator needs to show pieces of evidence that support the theory.

We employ Integrated Gradients [203] 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 in Figure 5.2b. 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 5.3.1.

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 to the model were then identified using the Euclidean distance of these latent space vectors. A schematic overview of our method is shown in Figure 5.2c. Details of the methods can be found in Section 5.3.2.

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 mean that they have multiple plausible outcomes. These reactions are most fit for identifying if the model is making the predictions on a true chemical basis because the underlying chemical causes are well established. Our general framework for interpreting chemical reactions is shown in Figure 5.2a.

Once a suitable chemical reaction with two possible target molecules is chosen the Molecular Transformer probability scores of the products are generated (Eq.5.1). The difference in probability score between the true and the incorrect but plausible products is then 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 [204–206, 203]. 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. Several methods have their roots in cooperative game theory and are proven to yield fair attributions as defined by the axioms of fairness [203]. For machine learning models where

the gradients are not readily available, there are so-called Shapley-values and the closely related SHAP method [205]. For models such as the Transformer where the gradients are easy to evaluate the Integrated Gradients (IGs) method is a more natural choice [203] though other methods such as LRP have also been applied successfully [207]. 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 [208]. A graphical illustration of IGs is shown in Figure 5.2b. Our approach builds on the work of McCloskey et. al. [209] 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 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 two possible products 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 between 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 an understanding of the model's inner workings [210]. 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 network's layer outputs (activations) concatenated together. This similarity measure has been shown to be useful for judging the reliability of toxicity model predictions by comparing molecules not in the training set [211].

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 the 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 5.2b). The similarity score of the input reaction vector $\mathbf{u}$ to a training set vector $\mathbf{v}$ is calculated by

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

this function returns a score of 1 if the two vectors $\mathbf{u}$ and $\mathbf{v}$ are identical, and 0 if the reactions are very different i.e. $\|\mathbf{u}-\mathbf{v}\|$ is very large. In this work, we use the similarity score to find the top-5 most similar reactions in 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 5.2(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 5.2a, 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 several adversarial examples.

5.4.1 Epoxidation

The oxidation of alkenes to form epoxides is an important intermediate reaction in many synthesis plans [213]. The common oxidants in these reactions are peroxy compounds. The most widely used example of them is mCPBA, which is a versatile reagent appearing 2052 times in the USPTO dataset. This is in the high data regime where we would expect the model to do well as there are a large number of different training examples available.

Epoxidation reactions can be regioselective, with more substituted alkenes reacting faster because they are more electron-rich [213]. A typical example reaction showing this type of selectivity is shown in Figure 5.3a.

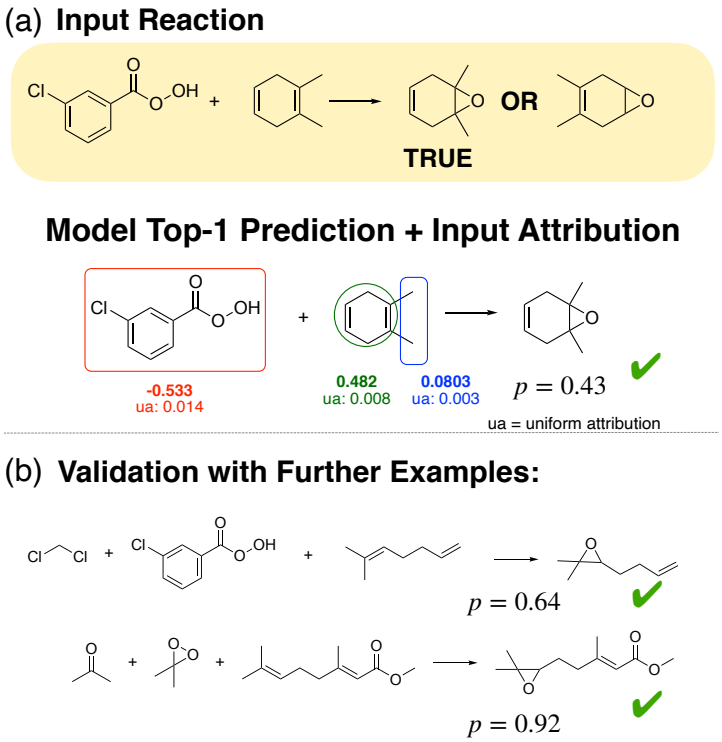


Fig. 5.3 **IG attributions highlight correct model reasoning.** (**a**) The model correctly predicts the product of a typical epoxidation reaction and shows significant positive attributions to the two methyl groups that are responsible for the selectivity. (**b**) We validate the model's knowledge on two unseen epoxidation reactions from chemical literature [212]

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 5.3a 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 in

Figure 5.3b. 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 more subtle chemistry, the ester group being electron withdrawing. The model can predict the correct outcome here as well confirming that the predictions are correct for the right reason.

5.4.2 Diels-Alder

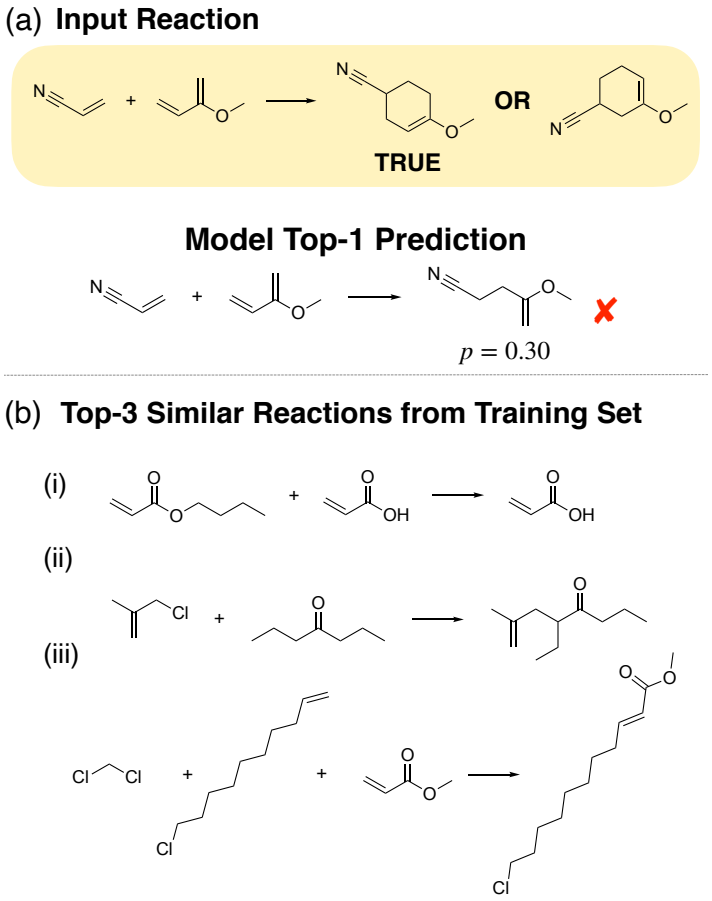


Fig. 5.4 **Data attribution explains erroneous predictions.** (**a**) The model makes an incorrect prediction on a typical example of a Diels-Alder reaction with challenging selectivity. (**b**) Attribution to the USPTO training data shows that the model either completely fails to recognize Diels-Alder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

The Diels-Alder reaction transforms a conjugated diene and an alkene (called dienophile) to a six-membered ring with a double bond [213]. 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 in Figure 5.4a.

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 5.2a and generated the most similar training reactions to see what causes this erroneous prediction.

Figure 5.4b 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 [214] 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.

5.4.3 Friedel-Crafts Acylation

Friedel-Crafts acylation reactions are an example of electrophilic aromatic substitution [215]. 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 determines the selectivity of these reactions. An example of a selective Friedel-Crafts reaction is shown in Figure 5.5(a) where according to the patent the para product is formed with a yield of 90% [216]. In this reaction, the 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.

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 with 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

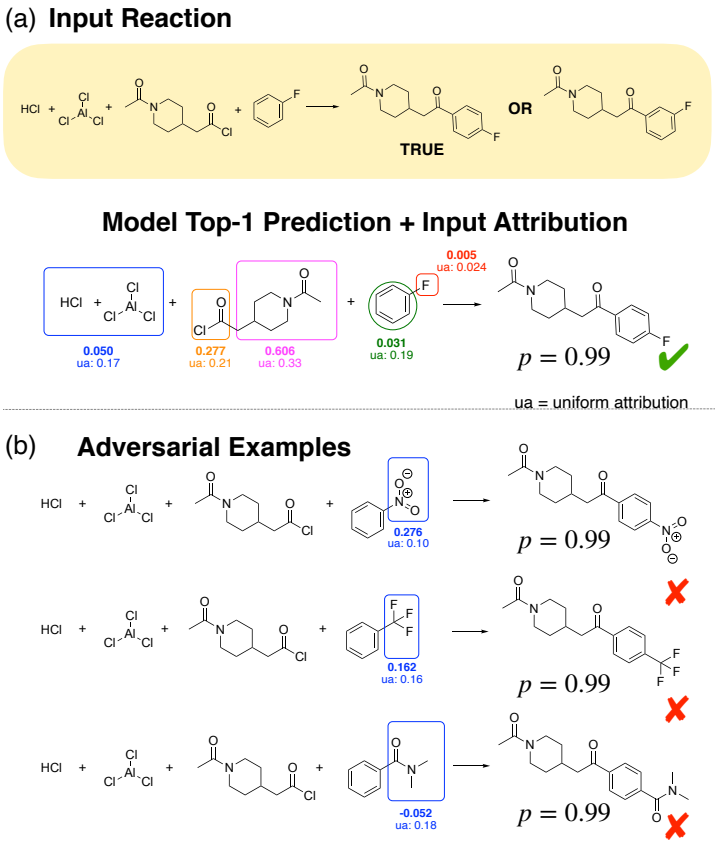


Fig. 5.5 **IG attributions reveal incorrect reasoning and guide the design of adversarial examples.** (**a**) The model correctly predicts the major para product of a typical Friedel-Crafts acylation, but low attribution is given to the para-directing -F group. (**b**) The model is fooled into incorrectly 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.

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 5.6), we find that there are many more para substitution reactions than meta in the training dataset. The origin of this bias is unclear, but a possible explanation is that medicinal chemists are encouraged to perform para substitutions when investigating the structure-activity relationship of a bioactive compound [217].

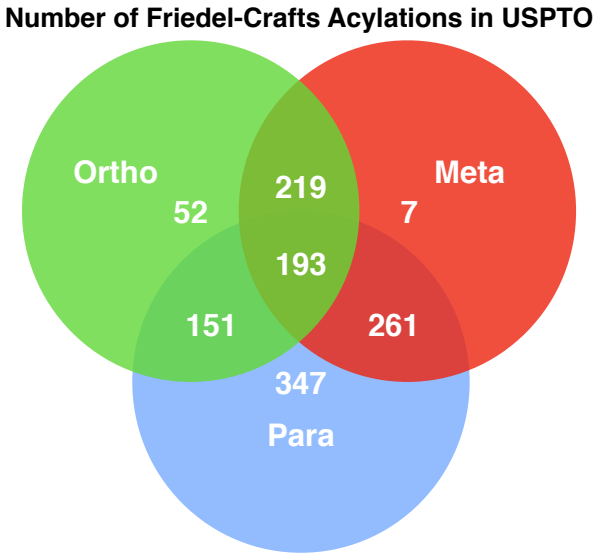


Fig. 5.6 **The number of para Friedel-Crafts acylation reactions in USPTO far outweigh those of meta or ortho reactions.** Overlaps in the Venn diagram denote cases where the benzene has more than 1 substituent.

This data bias may 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. 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 a 1:1 proportion of para:meta reactions (Table 5.1). 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 fewer meta reactions are present. Finally, we generate a 'Severely Biased' dataset with a 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 a fair comparison, this is the regime we wanted to investigate with the artificial datasets. We trained 10 transformer models on each of the three training sets and saved checkpoints from the beginning of model training up to 256 epochs.

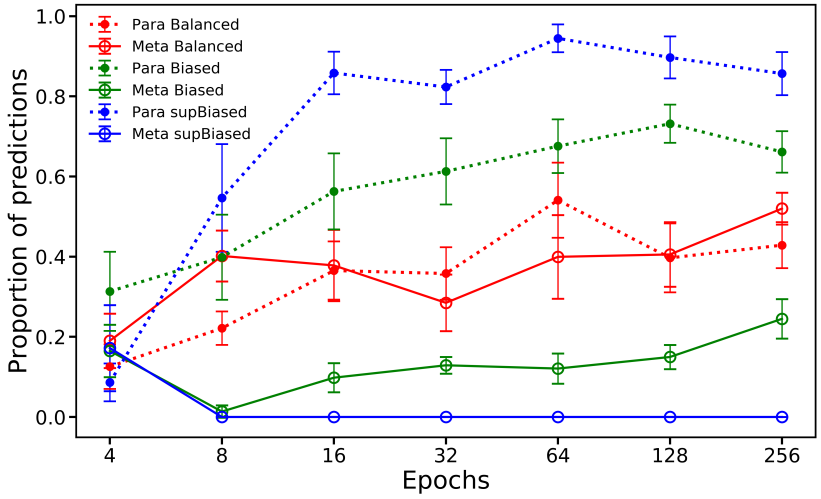


Fig. 5.7 **Biased training data leads to biased predictions from the Molecular Transformer.** The figure shows the proportion of para (solid line) and meta (dashed line) predictions on a balanced test set as a function of the number of training epochs for different biased training sets. The error bars 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 some 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 (Fig 5.7). 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 is 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 translating due to the presence of gender stereotypes in the training data [218].

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 [129]. This is the phenomenon 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 manifests 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 [49]. Figure 5.8 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 [219].

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.

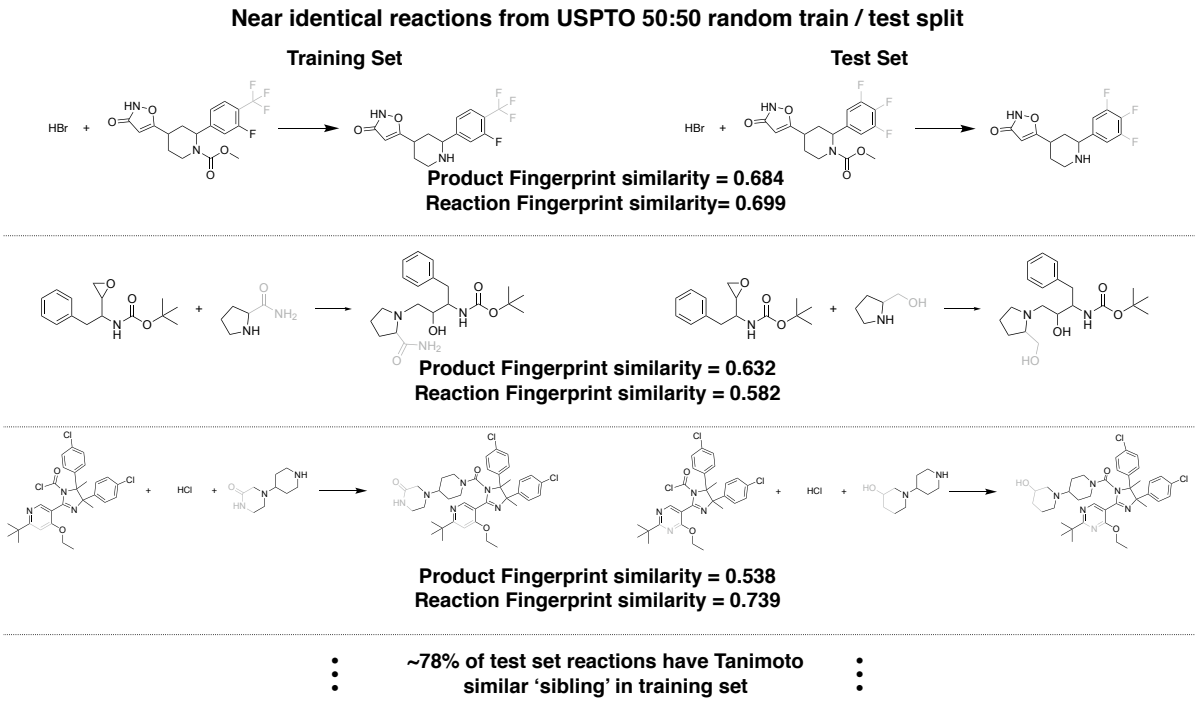


Fig. 5.8 **Randomly splitting USPTO results in a large number of near-identical reactions shared between train/test sets.** 78% of reactions in the test set have products that are within Tanimoto similarity 0.5 of a product in the training set following a 50:50 random split. By eye, it can be seen that 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%.

Such a dataset split intends to remove structural bias but we must also make sure that the distribution of different reaction types in the train and test sets are 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 5.9).

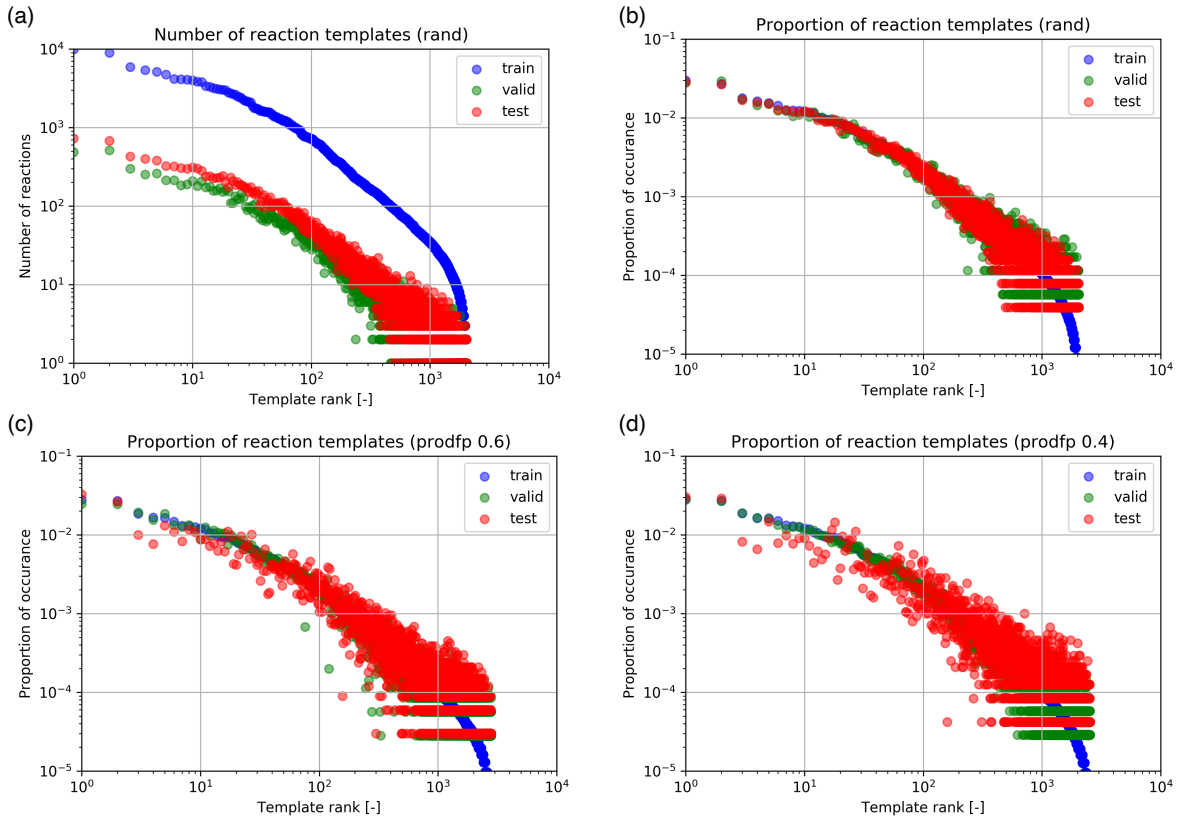


Fig. 5.9 **Tanimoto splitting minimally affects reaction template distribution.** (**a** - **b**) The absolute occurrence (**a**) and fractional occurrence (**b**) of reaction templates in train/valid/test sets of USPTO from a random split. The distribution of test set reactions closely resembles that 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 [14] was applied to the training, validation, and test sets from different dataset splitting methods (Fig 5.9 (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 qualitative trends observed.

These graphs show that the distribution of templates in the test set closely follows that of the training set in all cases (Fig 5.9 (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$4.9k, suggesting that this splitting method can produce test sets that better represent the distribution of reaction types from the full dataset ($\sim$26k templates) compared to a random split. This is similar to an importance sampling scheme that helps sample 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 for 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. [14] which is a widely-used graph-based machine learning reaction prediction model. This model explicitly represents molecules as graphs and considers reactions as a series of graph edits instead of the Molecular Transformer's text-based translation of SMILES strings.

Table 5.2 shows that the model performance of both the graph-based model and the Molecular Transformer significantly decreases 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 improvement 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. The inclusion of additional physical insight into models, as done in recent work incorporating explicit reaction mechanisms for reaction prediction [220] and machine-learning regio-selectivity prediction [221], may 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 [222].

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 the loss of important metadata, like reaction conditions.

Finally, we have shown that scaffold bias is a phenomenon 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 for continued refinement of predictive models.

Chapter 6

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

Accurately determining the biological activity of a molecule requires compound synthesis and purification followed by the preparation of a solution of known concentration to measure compound activity via an assay. However, 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 [223–225]. 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 a 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 advanced intermediates in the synthesis route.

The drawback, however, is that it is rarely possible to perform purification for reactions conducted at the nanomole scale. The output of these reactions is therefore a mixture of chemical reactants, reagents, and products known as a crude reaction mixture. These reaction mixtures can still be assayed for measuring biological activity, but there will necessarily be additional noise as detected potency will be influenced by variations in reaction yield or interference from the reactants. Nevertheless, nanomolar-scale synthesis combined with biochemical screening of crude reaction mixtures has been exploited for the discovery of

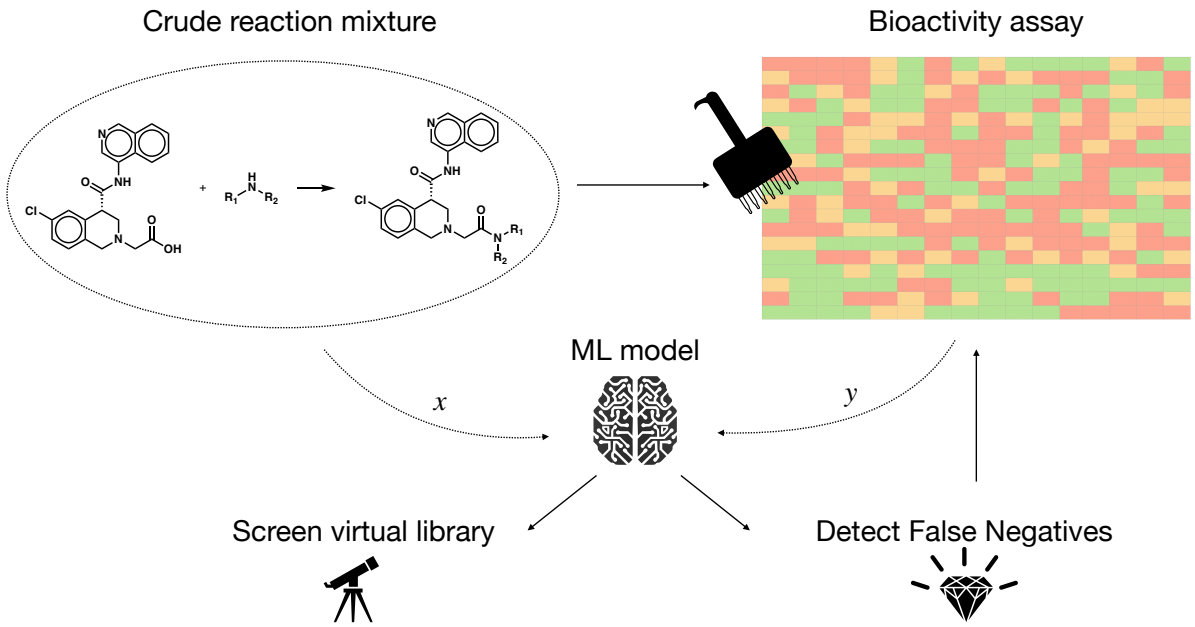


Fig. 6.1 **A schematic of the screening workflow.** Nanomolar-scale crude reaction mixtures are assayed for activity against Mpro, and a machine learning model is trained to predict the bioactivity measurements from the chemical structure of the compounds. The trained model is used to identify false negatives in the dataset and to identify novel hits in a prospective virtual screen.

potent inhibitors for several kinases [226, 225], illustrating the potential of this approach for accelerating hit-to-lead drug design.

A related method for high-throughput synthesis is DNA-encoded libraries (DEL) [19]. By linking small molecule building blocks with DNA fragments before chemically enumerating the building blocks with one another, large and diverse combinatorial libraries can be generated. These libraries can be biochemically screened in crude because potent compounds which bind to the protein target can be identified by their DNA tags using high-throughput sequencing methods. There has been recent success in training ML models on DEL bioactivity screen for hit-finding [227, 228] as well as toxicology screening [229], and the success of this approach suggests that applying ML to nanomolar screening data may also be possible.

In this chapter, we investigate the application of ML models on bioactivity data from nanomolar-scale high-throughput screening of crude reaction mixtures against SARS-CoV-2 Mpro. We show that ML models can be trained on this data to identify false negatives within the dataset missed due to experimental noise and that these models can be used to identify novel hits in a prospective virtual screen.

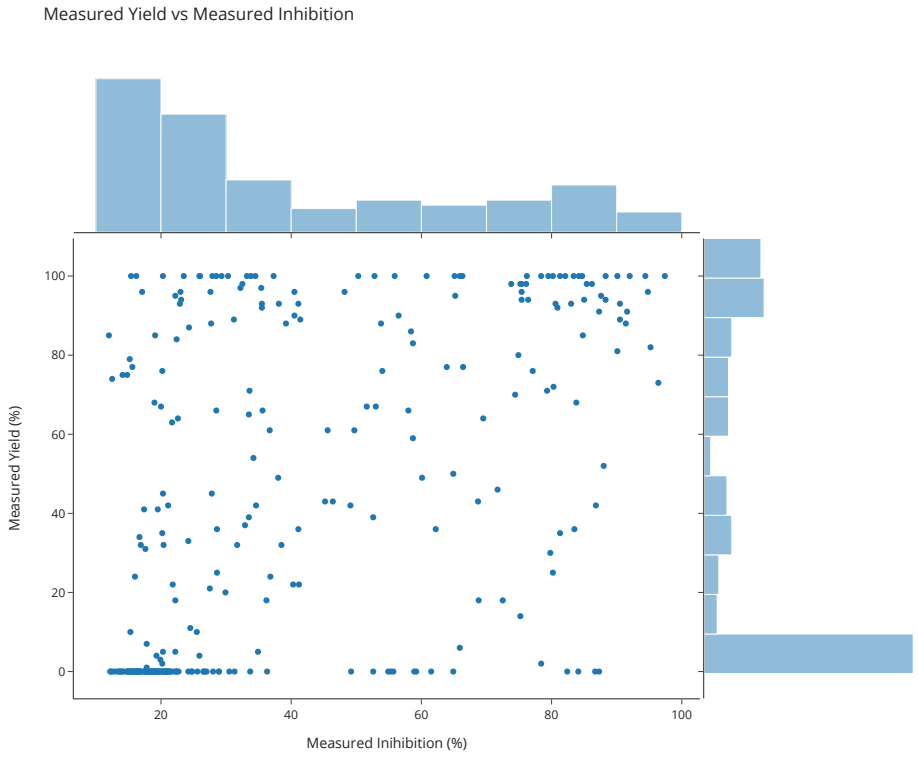


Fig. 6.2 **The bioactivity of the crude reaction mixtures is influenced by reaction yield.** High-yielding compounds are likelier to exhibit high activity (strongly inhibiting Mpro) while low-yielding compounds tend to have low activity, though the overall correlation is not strong (Spearman correlation $\rho = 0.54$).

6.1 Modelling high-throughput crude screening data

The COVID Moonshot campaign [142] against SARS-CoV-2 main protease (Mpro) investigated multiple chemically diverse molecular series simultaneously in order to minimise the risk of failure. One of the series explored was MAT-POS-4223bc15-21, whose molecular structure had the potential for optimisation towards targeting the P4 pocket of Mpro.

To rapidly evaluate the SAR of this series, nanomolar high-throughput chemistry was used to enumerate a library of 300 amine building blocks via amide coupling. Yield estimation via integration of UV spectra showed 151 of the library yielded >30% of the desired product, and the crude reaction mixtures were then assayed against Mpro (Experimental details can be found in Appendix B.1 & B.6).

Although the bioactivity measured in crude is a strong indicator of the potency of compounds, it is also influenced by the yield of the crude reaction mixture. Inspection of the crude activity data reveals a clear relationship (Spearman correlation $\rho = 0.54$) between the

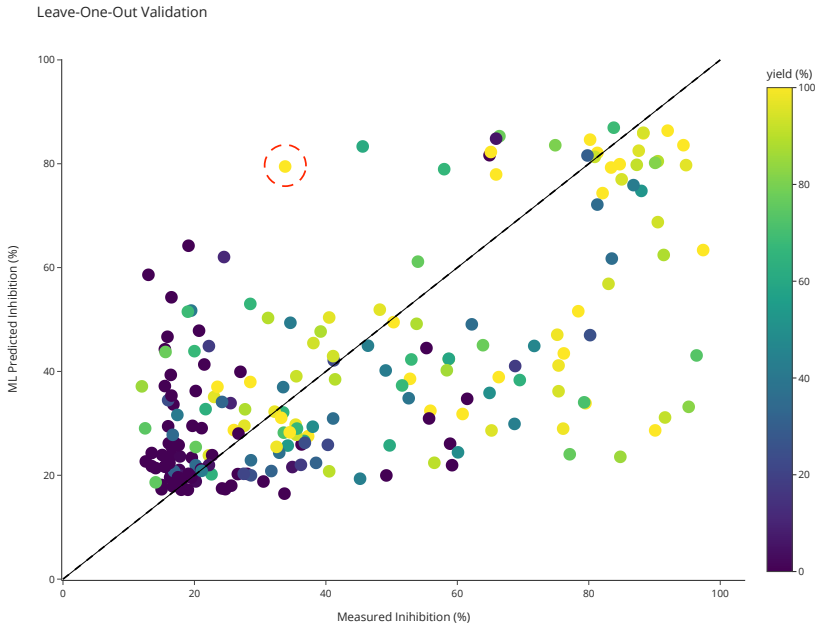


Fig. 6.3 **Machine learning is able to model crude bioactivity data.** The model predictions (y-axis) are generated in a leave-one-out fashion, where the model is trained on all but one datapoint, and then evaluated on the remaining datapoint which is then repeated for each datapoint in the dataset. The datapoint circled in red is a false negative identified by the model and subsequently experimentally validated (Figure 6.4).

measured activity and the yield of the crude reaction mixture (Figure 6.2). Compounds with high yields tend to have strong inhibition, while many compounds with low measured yields have no activity. This means that potentially potent inhibitors may be missed due to low yield from the amide coupling reaction, and this confounding variable adds additional difficulty to the already challenging task of modelling bioactivity with machine learning.

We model the bioactivity data using two different ML models, taking the mean of their predictions as the final output, to avoid overfitting to the small number of datapoints in this dataset. The two models are: (A) a random forest (RF) model that takes the Morgan fingerprint representation of the amide product in the crude reaction mixture, and (B) a gaussian process (GP) model that takes the Morgan fingerprint representation of the *amine* building block as that is the only varying component in the reactants of the crude reaction mixture. Both models aim to predict the measured activity of the crude reaction mixture.

Due to the small size of this dataset, we also choose to use a leave-one-out cross-validation approach to train the models. This means that the model is trained on all but one datapoint, and then evaluated on the remaining datapoint. This is repeated for each datapoint in the

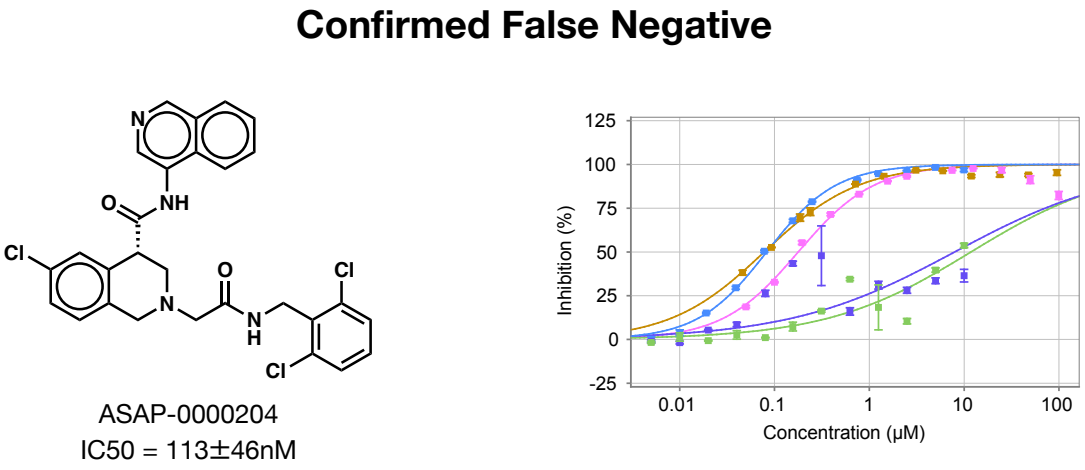


Fig. 6.4 **Experimental validation of ML identified false negative.** A compound with a large positive difference between predicted and measured activity (Figure 6.3) was re-synthesized and assayed to obtain full dose-response curves. Two sets of curves were obtained with one set demonstrating significantly higher potency than the other, suggesting that solubility issues were interfering with the assay measurements. Examining the higher potency dose-response curves show that the compound has IC50 = $0.113 \pm 0.046 \mu$M , which would place it among the top 20 potent compounds in the original dataset.

dataset (Figure 6.3). Overall, the model predictions moderately correlate with the experimental measurements (Spearman correlation $\rho = 0.64$) and identify distinct clusters of high and low-activity compounds.

While the model is generally quite conservative in its predictions, it does identify several compounds with a high predicted activity that are not active in the crude assay. We suspected that these were false negatives due to either the low yield of the crude reaction mixture or some other form of experimental noise. We identified 5 compounds with the largest positive difference between predicted and measured activity 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 indeed potent with IC50 = $0.113 \pm 0.046 \mu$M (ASAP-0000204) (Figure 6.4).

6.2 Prospective Virtual Screening

Having demonstrated that our ML model can meaningfully learn insights from crude activity data and identify false negatives, we further investigate whether it can extrapolate to novel compounds and be used for virtual screening. The ultimate aim of the SAR exploration of MAT-POS-4223bc15-21 is to optimise potency, and by virtually screening commercially

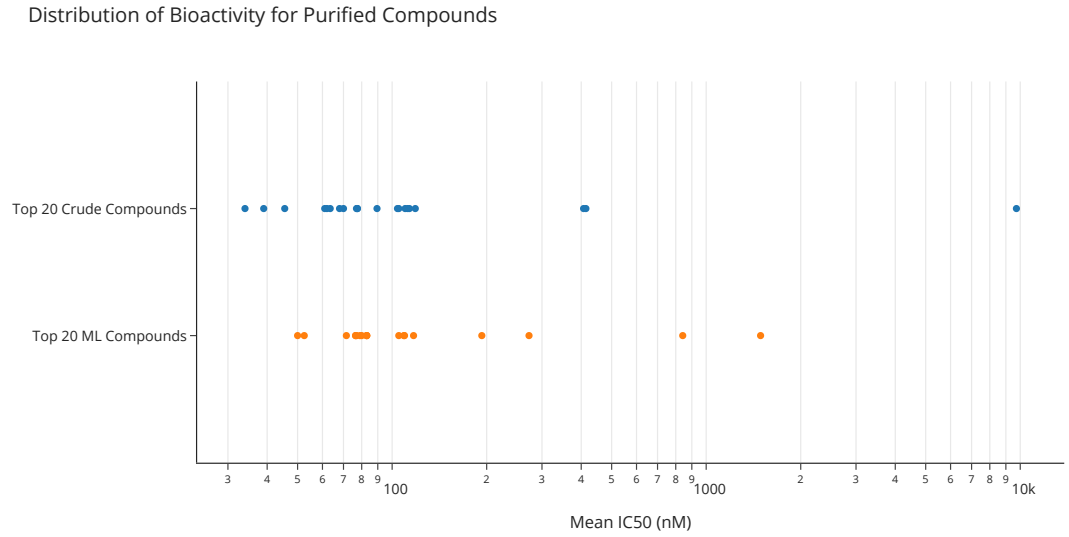


Fig. 6.5 **The bioactivity distribution of ML-identified compounds from virtual screening is similar to those from the most potent crude compounds.** The top 20 crude compounds are shown in blue, and the top 20 ML compounds are shown in orange.

available amine building blocks for the amide coupling reaction we may be able to identify novel compounds with even higher potency.

We construct the screening library by virtually enumerating primary and secondary amine building blocks from Enamine with the same carboxylic acid substructure from MAT-POS-4223bc15-21, resulting in a library of 62,814 amides. These compounds were then filtered for structural alerts and molecules with more than one chiral center were removed to avoid complications in assessing potency. This leaves 58,082 compounds which were scored by the trained GP and RF models, and the top 20 ("ML") compounds with a high mean predicted activity were selected for synthesis, purification, and assaying to obtain IC50 values.

In parallel, the top 20 ("crude") compounds with the highest measured crude activity from the initial 300 amides were also selected for resynthesis. Crystal structures of the resynthesised compounds bound to Mpro were obtained, verifying that 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.

Both sets of compounds exhibit similar distributions in bioactivity (Figure 6.5), with several compounds improving IC50 by up to 300-fold on MAT-POS-4223bc15-21 (up to 3-fold on the corresponding methyl-amide). The top 2 ML compounds showed promising average IC50 values of 53nM (ASAP-0000169) and 75nM (ASAP-0000211), respectively (Figure 6.6). Although the top 2 most potent crude compounds were more potent with IC50 = 28nM

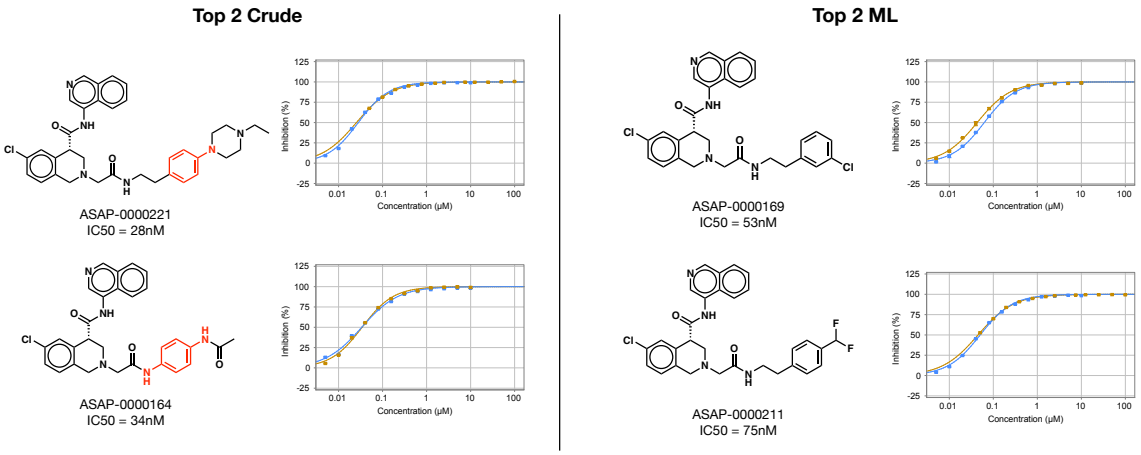


Fig. 6.6 **Comparison of top ML compounds vs top crude compounds.** The top compounds from the crude screening are more potent but contain aniline motifs (highlighted in red) that are undesirable due to toxicity concerns.

(ASAP-0000221) and IC50 = 34nM (ASAP-0000164), they contain aniline motifs that are generally avoided due to their propensity for the formation of reactive metabolites [230]. The top 2 crude compounds without the aforementioned motif had IC50s of 46nM (ASAP-0000155) and 64nM (ASAP-0000225), which are similar to the top 2 ML compounds.

This result highlights ML's ability to identify promising yet overlooked scaffolds without compromising potency.

6.3 Discussion

In this work, we demonstrate the application of ML models on bioactivity data from nanomolar-scale high-throughput screening of crude reaction mixtures and illustrate the potential of this approach for accelerating compound optimisation in drug discovery. Focusing on the SAR exploration of an inhibitor against SARS-CoV-2 Mpro, we show that ML models trained on nanomolar-scale crude screening data can be used to identify false negatives within the dataset missed due to experimental noise. Furthermore, we show that these models can be used for virtual screening of compound libraries and identify novel potent inhibitors.

The combination of nanomolar-scale synthesis with cellular/biochemical screening is a powerful experimental technique that can greatly speed up compound optimisation and has already demonstrated success in the discovery of potent kinase inhibitors [226, 225]. Despite the relatively high throughput of this technique, the number of compounds that can be screened is still dwarfed by the size of chemical space, intrinsically limiting the degree to which

compound potency can be optimised. Not only can machine learning augment this technique by deconvolution of the experimental noise, they fundamentally extend its reach by utilising the data to virtually screen compound libraries and explore chemical space.

While only a single iteration of model training and prospective testing was investigated in this work, it is clear that the combination of ML with nanomolar high-throughput screening could be very effective in an automated decision-making workflow. As discussed in a previous chapter, an automated workflow for the generation of molecular designs could potentially reduce iteration cycle time, require fewer compounds and iterations to produce a candidate, and scale to more programs [184, 185, 183]. Nanomolar high-throughput screening could be used as the testing step in such a workflow for performing Bayesian optimisation of compound potency [186].

More broadly, the concept of using crude reaction mixtures in place of pure compounds in drug discovery is not limited to biochemical assays. Recent work in utilising crude reaction mixtures for crystallographic fragment screening demonstrated approximately 7x saving of reagents and solvents and up to a 4x reduction in time [231]. The higher throughput of these approaches necessarily results in additional noise, and perhaps the application of machine learning can be used to detect outliers in such data and perform extrapolation also with the ultimate aim of accelerating drug discovery.

Chapter 7

Outlook

The research presented in this thesis explores ways in which data-driven approaches based on machine learning can be leveraged in the design-make-test cycle of drug discovery. In each of the three steps of 'design', 'make', and 'test' we encounter the same underlying challenge, of grappling with the practical difficulties of a drug discovery campaign where we must make full use of the limited data available.

In Chapter 3 we looked at how to leverage fragment-protein structures from a crystallographic fragment screen to hit discovery in the absence of any bioactivity data. Using an unsupervised learning approach, we learn the geometric distribution of pharmacophores from the fragment-protein complexes and use these to screen potential molecules for bioactivity. We showed that this approach outperforms docking in distinguishing active compounds from inactive ones on retrospective data. Further, we prospectively found novel hits for SARS-CoV-2 Mpro and the Mac1 domain of SARS-CoV-2 non-structural protein 3 by virtually screening a library of 1 billion molecules.

Chapter 4 takes us to the early stages of hit-to-lead molecular optimisation where bioactivity data is limited, noisy, and dominated by inactive molecules. We overcame this challenge with a learning-to-rank framework via an ML model that predicts whether a compound is more or less active than another. This approach allowed us to make use of inactive data and threshold the bioactivity differences above measurement noise, and validation on retrospective data for SARS-CoV-2 Mpro showed that we can outperform docking on ranking ligands. Combining this model with AI-based synthesis tools, we prospectively screened a library of 8.8M molecules to arrive at a potent compound with a novel scaffold.

While AI-based synthesis tools have already shown demonstrable success in accelerating the synthesis of new molecules, they are still prone to failure and suffer from a lack of transparency in their decision-making due to their black-box nature. To address this, in Chapter 5 we showcased a workflow for quantitatively interpreting a state-of-the-art deep learning model

for reaction prediction. By analysing chemically selective reactions, we showed 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.

In Chapter 6 we explored how to accelerate testing procedures by applying machine learning to bioactivity data from nanomolar-scale high-throughput chemistry. While this experimental technique greatly increases the number of molecules that can be tested, there is additional noise resulting from having to assay crude reaction mixtures instead of pure samples. Nevertheless, we showed that machine learning models trained on this data can cut through this noise and identify a false negative assay measurement, as well as prospectively screen a library of 62K molecules to discover new SARS-CoV-2 Mpro inhibitors just as potent as those from the original assay.

7.1 Directions for Future Research

While we have explored some methods to accelerate the design-make-test cycle in this thesis, there is still much more work to be done to bring us closer to the dream of efficient and automated drug discovery. Below, we outline some of the most prominent directions for future research.

7.1.1 Deploying and Extending ML-based Synthesis Tools

Machine learning-based models for synthesis prediction have made significant strides in recent years, and are already sufficiently accurate to be deployed in industry for synthesis route design [232]. However, translating these models into a practical tool that can be used by non-specialists is non-trivial. Connecting these models with commercial compound databases, updating models in response to changes in stock availability, and developing more user-friendly interfaces are fruitful engineering challenges for enabling more widespread adoption of these tools.

An intriguing extension of this idea is allowing a large language model (LLM) to access ML-based synthesis tools. Large language models, using the same architectures on which many ML synthesis tools are based, have demonstrated incredible success in natural language processing as well as in understanding programming languages [106, 233, 234]. An exciting use-case is to allow LLMs to make API calls to other software, meaning that one can abstract away complex programming tasks by just giving a natural language description of the task to the LLM. A recent example demonstrated how GPT-4, a powerful LLM, can perform molecular queries in the PubChem database, determine whether the molecule is purchasable, find a

supplier that sells it, and either purchase the compound or draft an email to a synthesis CRO to order [235]. Integrating LLMs with ML-based synthesis tools, or augmenting LLMs by training them to understand SMILES and/or SMARTS, could extend such capabilities and facilitate complex user requests in synthesis planning.

In tandem with optimising model deployment, continued improvement in the accuracy of synthesis predictions can be achieved in several ways. Novel model architectures that better reflect the underlying structure of chemical reaction data, both the graph nature of molecules as well as the correlations between chemical reactions from the same patent, could potentially improve the accuracy of synthesis predictions. The incorporation of proprietary reaction data from industry should also lead to improved performance, although there are difficulties in standardising data from electronic laboratory notebooks (ELNs) [236] and there has been conflicting evidence regarding its usefulness [237, 238]. Finally, the creation of better benchmarks, likely incorporating disclosed industrial data [239], can help ensure that models are evaluated under realistic conditions, providing a more accurate and robust representation of their capabilities.

Last but not least, there is much room for improvement in the prediction of optimal reaction conditions, catalysts, and reaction yields, [240–242] which are not currently handled by most ML-synthesis tools. While existing models are effective at predicting the outcomes of small-scale chemical reactions, they may not be as accurate when it comes to predicting the behavior of reactions that must be scaled up for large-scale production. Including the modelling of reaction yields and reaction conditions, for example via training on high-throughput experimentation data [243, 244] combined with Bayesian optimisation [245], would allow for greater precision overall, and expand the scope of these models beyond small-scale reactions to bridge the gap between research and production.

7.1.2 Addressing Data Scarcity

A common challenge in drug discovery is the limited availability of data. While it is true that data scarcity is a persistent issue, it is also true that we are not making the most of the data that we do have [91]. Existing datasets can be enhanced by utilizing them in more effective ways, and by designing models that take into account the unique structure inherent in the data. For example, imputation-based methods explicitly consider data sparsity, filling in missing data points with estimates derived from other data points which allow researchers to make use of incomplete datasets and exploit correlations between different endpoint measurements [246–248]. Transfer learning approaches involve using knowledge gained from one dataset to inform the learning process in another domain, and there is a large scope for designing models that can utilise abundant noisy data to improve the prediction of low-data properties [249, 250].

By developing and utilizing models that are specifically designed for low-data regimes instead of merely adapting models performant in data-rich scenarios, we can make more accurate predictions and gain insights that would not be possible with traditional models.

Federated learning is another promising approach that can help overcome data scarcity by building models that can learn from proprietary data while maintaining data privacy [251]. The pharmaceutical industry possesses orders of magnitude more data on experimental measurements than publicly available but that data is often proprietary and cannot be shared. Federated learning approaches allow us to pool data from multiple industrial partners while maintaining the privacy and security of the underlying data, allowing us to build more accurate models than possible by any individual partner. A landmark example of this is the MELLODDY project [252, 253], which brought together ten pharmaceutical companies and aggregated over 2.6+ billion confidential experimental data points for 21+ million different small molecules and 40+ thousand on-target and pharmacokinetics assays. The massive increase in collective training data boosted predictive performance, particularly for pharmacokinetics and safety panel assays. The methodology of this project is a strong foundation on which to build more powerful federated learning platforms in the future.

In parallel to maximising the use of existing data, there are also exciting opportunities to use machine learning methods to leverage noisy but high-throughput experimental techniques in drug discovery. One example of such a technique is DNA-encoded libraries, which enable the synthesis and screening of large numbers of compounds in a relatively short amount of time. While the resulting data may be noisy, the sheer volume of data generated makes it a valuable resource for building ML models [227, 229, 228]. There is a large scope for developing models that can leverage data from other under-utilised high-throughput techniques, from microfluidics platforms [254, 255] to high-throughput crystallography [256, 157], to deliver novel insights in drug discovery.

7.1.3 Integration of Protein Bioinformatics

Although bioactivity involves the complementary interactions between a small molecule and its protein target, the focus of machine learning methods in drug discovery has largely been on molecules alone while the protein component has been largely neglected. Recent stunning advances in machine learning for protein bioinformatics, such as AlphaFold [61] and RFDiffusion [257], suggest that ML bioinformatics models have a rich understanding of protein structure and function. It is clear that by integrating these models and workflows with those for ML in chemistry and cheminformatics, we may be able to make significant breakthroughs in bioactivity prediction which remains a fundamental challenge in drug discovery.

Expanding beyond bioactivity prediction, these models and workflows should improve the modelling of small molecule interactions with proteins in general. This could enable better understanding of metabolic pathways and the dynamics of biochemical processes which are relevant for the prediction of pharmacokinetics and toxicity properties, areas in which data scarcity and assay variability pose an even greater challenge than bioactivity prediction [258, 249, 259, 250]. Greater accuracy in predicting in-vivo measurements from in-vitro data would allow us to design safer drugs and reduce clinical-stage attrition in the latter stages of drug development.

Demonstrable achievements in predicting protein function [260] as well as enzyme design [261] also open the door for designing efficient biocatalysed synthesis routes for otherwise unsynthesisable drug candidates. This could have a particular impact in the latter stages of drug design where functionalisation of a complex drug candidate is required to optimise its pharmacokinetics while retaining bioactivity. The inclusion of basic enzyme information into ML-based synthesis tools has already shown predictive capabilities on retrospective data [262], and successful prospective validation experiments will likely follow.

Extrapolating further, the ability to predict molecular interactions with a diverse range of proteins could also be used to determine molecular mechanisms of action, as well as identify new drug targets [263]. By combining insights from both chemical and biological data, we can gain a more comprehensive understanding of how small molecules interact with the body at a molecular level, opening up new opportunities for drug discovery and design.

7.1.4 Full Automation of Drug Discovery

Full automation of the drug discovery process is an ambitious goal that is the ultimate destination of research in computational drug design. Developing systems that can perform multiple iterations of the design-make-test process autonomously, with no human intervention in the decision-making process, is a long-term goal that is still far from being realised [264, 265, 184].

At the moment the most advanced systems still require significant human intervention in the decision-making process, such as constraining the search space of potential drug designs, ordering/performing the synthesis/assaying of selected candidates, and interpreting experimental results [183]. Achieving full automation of drug discovery will require integrating all of the models and workflows mentioned above, and developing new approaches to optimize decision-making and resource allocation. Incorporating generative models with reinforcement learning [266, 267, 187, 268], in particular, could be the ultimate embodiment of this vision, enabling systems to learn from experience and continually improve their performance.

That said, the computational and material resources required for training such an approach may be prohibitive in the near/medium term, and significant technological breakthroughs will

be required to make full automation of drug discovery a practical reality. Nonetheless, the potential benefits of such an approach are enormous, with the potential to transform the field of drug discovery and accelerate the pace of new drug development.

References

[1] A. Crum Brown and Thomas R. Fraser. V.—on the connection between chemical constitution and physiological action. part. i.—on the physiological action of the salts of the ammonium bases, derived from strychnia, brucia, thebaia, codeia, morphia, and nicotia. *Earth and Environmental Science Transactions of The Royal Society of Edinburgh*, 25(1):151–203, 1868.

[2] Corwin Hansch, Peyton P Maloney, Toshio Fujita, and Robert M Muir. Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. *Nature*, 194(4824):178–180, 1962.

[3] Corwin Hansch and Toshio Fujita. p-σ-π analysis. a method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, 86(8):1616–1626, 1964.

[4] Robert K. Lindsay, Bruce G. Buchanan, Edward A. Feigenbaum, and Joshua Lederberg. Dendral: A case study of the first expert system for scientific hypothesis formation. *Artificial Intelligence*, 61(2):209–261, 1993.

[5] Wenbo Yu and Alexander D. MacKerell. *Computer-Aided Drug Design Methods*, pages 85–106. Springer New York, New York, NY, 2017.

[6] Junshui Ma, Robert P. Sheridan, Andy Liaw, George E. Dahl, and Vladimir Svetnik. Deep neural nets as a method for quantitative structure–activity relationships. *Journal of Chemical Information and Modeling*, 55(2):263–274, 2015.

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

[8] Joshua Meyers, Benedek Fabian, and Nathan Brown. De novo molecular design and generative models. *Drug Discovery Today*, 26(11):2707–2715, 2021.

[9] Philipp Renz, Dries Van Rompaey, Jörg Kurt Wegner, Sepp Hochreiter, and Günter Klambauer. On failure modes in molecule generation and optimization. *Drug Discovery Today: Technologies*, 32-33:55–63, 2019.

[10] Elias James Corey. *The logic of chemical synthesis*. Wiley, 1991.

[11] The Nobel Foundation. The nobel prize in chemistry 1990, 190.

[12] E J Corey, A K Long, and S D Rubenstein. Computer-assisted analysis in organic synthesis. *Science*, 228(4698):408, 4 1985.

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

[14] Connor W. Coley, Wengong Jin, Luke Rogers, Timothy F. Jamison, Tommi S. Jaakkola, William H. Green, Regina Barzilay, and Klavs F. Jensen. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.*, 10:370–377, 2019.

[15] Thomas J. Struble, Juan C. Alvarez, Scott P. Brown, Milan Chytil, Justin Cisar, Renee L. DesJarlais, Ola Engkvist, Scott A. Frank, Daniel R. Greve, Daniel J. Griffin, Xinjun Hou, Jeffrey W. Johannes, Constantine Kreatsoulas, Brian Lahue, Miriam Mathea, Georg Mogk, Christos A. Nicolaou, Andrew D. Palmer, Daniel J. Price, Richard I. Robinson, Sebastian Salentin, Li Xing, Tommi Jaakkola, William. H. Green, Regina Barzilay, Connor W. Coley, and Klavs F. Jensen. Current and future roles of artificial intelligence in medicinal chemistry synthesis. *Journal of Medicinal Chemistry*, 63(16):8667–8682, 2020.

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

[17] J P Hughes, S Rees, S B Kalindjian, and K L Philpott. Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249, 03 2011.

[18] Srinivas Niranj Chandrasekaran, Hugo Ceulemans, Justin D. Boyd, and Anne E. Carpenter. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nature Reviews Drug Discovery*, 20(2):145–159, 2021.

[19] Adrián Gironda-Martínez, Etienne J. Donckele, Florent Samain, and Dario Neri. Dna-encoded chemical libraries: A comprehensive review with succesful stories and future challenges. *ACS Pharmacology & Translational Science*, 4(4):1265–1279, 08 2021.

[20] Simon L. Rössler, Nathalie M. Grob, Stephen L. Buchwald, and Bradley L. Pentelute. Abiotic peptides as carriers of information for the encoding of small-molecule library synthesis. *Science*, 379(6635):939–945, 2023.

[21] Maria Emilia Dueñas, Rachel E Peltier-Heap, Melanie Leveridge, Roland S Annan, Frank H Büttner, and Matthias Trost. Advances in high-throughput mass spectrometry in drug discovery. *EMBO Molecular Medicine*, 15(1):e14850, 2023.

[22] David Weininger. 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, 1988.

[23] David Weininger, Arthur Weininger, and Joseph L. Weininger. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *Journal of Chemical Information and Computer Sciences*, 29(2):97–101, 1989.

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

[25] Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. Phd, University of Cambridge, 2012.

[26] Hyunseung Kim, Jonggeol Na, and Won Bo Lee. 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, 2021.

[27] Stephen Heller, Alan McNaught, Stephen Stein, Dmitrii Tchekhovskoi, and Igor Pletnev. Inchi - the worldwide chemical structure identifier standard. *Journal of Cheminformatics*, 5(1):7, 2013.

[28] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100string representation. *Machine Learning: Science and Technology*, 1(4):045024, oct 2020.

[29] Daylight Chemical Information Systems. Smarts - a language for describing molecular patterns, 2022.

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

[31] Jonathan B. Baell and Georgina A. Holloway. 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, 2010. PMID: 20131845.

[32] Carmen Limban, Diana C. Nuţă, Cornel Chiriţă, Simona Negreș, Andreea L. Arsene, Marina Goumenou, Spyros P. Karakitsios, Aristidis M. Tsatsakis, and Dimosthenis A. Sarigiannis. The use of structural alerts to avoid the toxicity of pharmaceuticals. *Toxicology Reports*, 5:943–953, 2018.

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

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

[35] John J. Irwin, Khanh G. Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R. Wong, Munkhzul Khurelbaatar, Yurii S. Moroz, John Mayfield, and Roger A. Sayle. Zinc20—a free ultralarge-scale chemical database for ligand discovery. *Journal of Chemical Information and Modeling*, 60(12):6065–6073, 2020.

[36] Tomasz Klucznik, Barbara Mikulak-Klucznik, Michael P. McCormack, Heather Lima, Sara Szymkuć, Manishabrata Bhowmick, Karol Molga, Yubai Zhou, Lindsey Rickershauser, Ewa P. Gajewska, Alexei Toutchkine, Piotr Dittwald, Michał P. Startek, Gregory J. Kirkovits, Rafał Roszak, Ariel Adamski, Bianka Sieredzińska, Milan Mrksich, Sarah L.J. Trice, and Bartosz A. Grzybowski. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem*, 4(3):522–532, 2018.

[37] Teresa Kaserer, Katharina R. Beck, Muhammad Akram, Alex Odermatt, and Daniela Schuster. Pharmacophore models and pharmacophore-based virtual screening: Concepts and applications exemplified on hydroxysteroid dehydrogenases. *Molecules*, 20(12):22799–22832, 2015.

[38] Anna Vuorinen and Daniela Schuster. Methods for generating and applying pharmacophore models as virtual screening filters and for bioactivity profiling. *Methods*, 71:113–134, 2015.

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

[40] Veronika Temml, Constance V. Voss, Verena M. Dirsch, and Daniela Schuster. 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, 02 2014.

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

[42] Sourav Pal, Vinay Kumar, Biswajit Kundu, Debomita Bhattacharya, Nagothy Preethy, Mamindla Prashanth Reddy, and Arindam Talukdar. 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, 2019.

[43] H. L. Morgan. 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, 1965.

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

[45] Gerald Maggiora, Martin Vogt, Dagmar Stumpfe, and Jürgen Bajorath. Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, 57(8):3186–3204, 2014.

[46] Peter Willett, John M. Barnard, and Geoffrey M. Downs. Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 38(6):983–996, 1998.

[47] Pierre Baldi and Ramzi Nasr. 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, 2010.

[48] Roberto Todeschini, Viviana Consonni, Hua Xiang, John Holliday, Massimo Buscema, and Peter Willett. 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, 2012. PMID: 23078167.

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

[50] Darko Butina. 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, 1999.

[51] Mark A Johnson and Gerald M Maggiora. *Concepts and applications of molecular similarity*. Wiley, 1990.

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

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

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

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

[56] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.*, 9:513–530, 2018.

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

[58] Frederik Sandfort, Felix Strieth-Kalthoff, Marius Kühnemund, Christian Beecks, and Frank Glorius. A structure-based platform for predicting chemical reactivity. *Chem*, 6(6):1379–1390, 2020.

[59] Xuan-Yu Meng, Hong-Xing Zhang, Mihaly Mezei, and Meng Cui. Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, 7(2):146–157, 2011.

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

[61] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[62] Felix Wong, Aarti Krishnan, Erica J Zheng, Hannes Stärk, Abigail L Manson, Ashlee M Earl, Tommi Jaakkola, and James J Collins. Benchmarking alphafold-enabled molecular docking predictions for antibiotic discovery. *Molecular Systems Biology*, 18(9):e11081, 2022.

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

[64] Ryan G. Coleman, Michael Carchia, Teague Sterling, John J. Irwin, and Brian K. Shoichet. Ligand pose and orientational sampling in molecular docking. *PLOS ONE*, 8(10):1–19, 10 2013.

[65] Richard A. Friesner, Jay L. Banks, Robert B. Murphy, Thomas A. Halgren, Jasna J. Klicic, Daniel T. Mainz, Matthew P. Repasky, Eric H. Knoll, Mee Shelley, Jason K. Perry, David E. Shaw, Perry Francis, and Peter S. Shenkin. 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, 2004.

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

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

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

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

[70] Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, 31(3):405–412, 10 2014.

[71] Conor D. Parks, Zied Gaieb, Michael Chiu, Huanwang Yang, Chenghua Shao, W. Patrick Walters, Johanna M. Jansen, Georgia McGaughey, Richard A. Lewis, Scott D. Bembenek, Michael K. Ameriks, Tara Mirzadegan, Stephen K. Burley, Rommie E. Amaro, and Michael K. Gilson. 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, 2020.

[72] Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: The casf-2016 update. *Journal of Chemical Information and Modeling*, 59(2):895–913, 2019.

[73] Brian J. Bender, Stefan Gahbauer, Andreas Luttens, Jiankun Lyu, Chase M. Webb, Reed M. Stein, Elissa A. Fink, Trent E. Balius, Jens Carlsson, John J. Irwin, and Brian K. Shoichet. A practical guide to large-scale docking. *Nature Protocols*, 16(10):4799–4832, 2021.

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

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

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

[77] Jiankun Lyu, Sheng Wang, Trent E. Balius, Isha Singh, Anat Levit, Yurii S. Moroz, Matthew J. O'Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, Andrey A. Tolmachev, Brian K. Shoichet, Bryan L. Roth, and John J. Irwin. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, 2019.

[78] Assaf Alon, Jiankun Lyu, Joao M. Braz, Tia A. Tummino, Veronica Craik, Matthew J. O'Meara, Chase M. Webb, Dmytro S. Radchenko, Yurii S. Moroz, Xi-Ping Huang, Yongfeng Liu, Bryan L. Roth, John J. Irwin, Allan I. Basbaum, Brian K. Shoichet, and Andrew C. Kruse. Structures of the

σ

2 receptor enable docking for bioactive ligand discovery. *Nature*, 600(7890):759–764, 2021.

[79] Elissa A. Fink, Jun Xu, Harald Hübner, Joao M. Braz, Philipp Seemann, Charlotte Avet, Veronica Craik, Dorothee Weikert, Maximilian F. Schmidt, Chase M. Webb, Nataliya A. Tolmachova, Yurii S. Moroz, Xi-Ping Huang, Chakrapani Kalyanaraman, Stefan Gahbauer, Geng Chen, Zheng Liu, Matthew P. Jacobson, John J. Irwin, Michel Bouvier, Yang Du, Brian K. Shoichet, Allan I. Basbaum, and Peter Gmeiner. Structure-based discovery of nonopioid analgesics acting through the α_{2A}-adrenergic receptor. *Science*, 377(6614):eabn7065, 2022.

[80] Jon A. Erickson, Mehran Jalaie, Daniel H. Robertson, Richard A. Lewis, and Michal Vieth. Lessons in molecular recognition: The effects of ligand and protein flexibility on molecular docking accuracy. *Journal of Medicinal Chemistry*, 47(1):45–55, 2004.

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

[82] Manuel A. Llanos, Melisa E. Gantner, Santiago Rodriguez, Lucas N. Alberca, Carolina L. Bellera, Alan Talevi, and Luciana Gavernet. 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, 2021. PMID: 34313128.

[83] Guillem Macip, Pol Garcia-Segura, Júlia Mestres-Truyol, Bryan Saldivar-Espinoza, María José Ojeda-Montes, Aleix Gimeno, Adrià Cereto-Massagué, Santiago Garcia-Vallvé, and Gerard Pujadas. 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, 03 2022.

[84] Jiankun Lyu, John J. Irwin, and Brian K. Shoichet. Modeling the expansion of virtual screening libraries. *Nature Chemical Biology*, 2023.

[85] Hannes Stärk, Octavian Ganea, Lagnajit Pattanaik, Regina Barzilay, and Tommi Jaakkola. Equibind: Geometric deep learning for drug binding structure prediction. In *International Conference on Machine Learning*, pages 20503–20521. PMLR, 2022.

[86] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi S. Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. In *The Eleventh International Conference on Learning Representations*, 2023.

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

[88] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

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

[90] Andreas Bender and Isidro Cortés-Ciriano. 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, 2021.

[91] Andreas Bender and Isidro Cortes-Ciriano. 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, 2021.

[92] Andrea Volkamer, Sereina Riniker, Eva Nittinger, Jessica Lanini, Francesca Grisoni, Emma Evertsson, Raquel Rodríguez-Pérez, and Nadine Schneider. Machine learning for small molecule drug discovery in academia and industry. *Artificial Intelligence in the Life Sciences*, 3:100056, 2023.

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

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

[95] Vladimir Svetnik, Andy Liaw, Christopher Tong, J. Christopher Culberson, Robert P. Sheridan, and Bradley P. Feuston. Random forest: A classification and regression tool for compound classification and qsar modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958, 2003.

[96] Benjamin Merget, Samo Turk, Sameh Eid, Friedrich Rippmann, and Simone Fulle. Profiling prediction of kinase inhibitors: Toward the virtual assay. *Journal of Medicinal Chemistry*, 60(1):474–485, 2017.

[97] Pavel G. Polishchuk, Eugene N. Muratov, Anatoly G. Artemenko, Oleg G. Kolumbin, Nail N. Muratov, and Victor E. Kuz'min. Application of random forest approach to qsar prediction of aquatic toxicity. *Journal of Chemical Information and Modeling*, 49(11):2481–2488, 2009.

[98] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

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

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

[101] William McCorkindale, Carl Poelking, and Alpha A. Lee. Investigating 3d atomic environments for enhanced qsar, 2020.

[102] Kjell Jorner, Tore Brinck, Per-Ola Norrby, and David Buttar. Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. *Chem. Sci.*, 12:1163–1175, 2021.

[103] S. Joshua Swamidass, Jonathan Chen, Jocelyne Bruand, Peter Phung, Liva Ralaivola, and Pierre Baldi. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics*, 21(suppl_1):359–368, 06 2005.

[104] 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. *Chem. Sci.*, 13:13541–13551, 2022.

[105] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[106] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[107] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition, 2019.

[108] Nicolae Sapoval, Amirali Aghazadeh, Michael G. Nute, Dinler A. Antunes, Advait Balaji, Richard Baraniuk, C. J. Barberan, Ruth Dannenfelser, Chen Dun, Mohammadamin Edrisi, R. A. Leo Elworth, Bryce Kille, Anastasios Kyrillidis, Luay Nakhleh, Cameron R. Wolfe, Zhi Yan, Vicky Yao, and Todd J. Treangen. Current progress and open challenges for applying deep learning across the biosciences. *Nature Communications*, 13(1):1728, 2022.

[109] Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–43, 2018.

[110] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

[111] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[112] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift, 2015.

[113] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[114] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[115] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 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, October 2014. Association for Computational Linguistics.

[116] Zachary C. Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning, 2015.

[117] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014.

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

[119] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):5999–6009, 2017.

[120] Philipp Dufter, Martin Schmitt, and Hinrich Schütze. Position information in transformers: An overview, 2021.

[121] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

[122] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers, 2022.

[123] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019.

[124] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020.

[125] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction, 2020.

[126] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Costas Bekas, and Alpha A. Lee. Molecular Transformer - A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science*, 5(9):1572–1583, 11 2019.

[127] Philippe Schwaller, Benjamin Hoover, Jean-Louis Reymond, Hendrik Strobelt, and Teodoro Laino. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances*, 7(15):eabe4166, 2021.

[128] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 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, 2017.

[129] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Jörg K. Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chem. Sci.*, 9:5441–5451, 2018.

[130] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019. PMID: 31361484.

[131] Matthew Ragoza, Joshua Hochuli, Elisa Idrobo, Jocelyn Sunseri, and David Ryan Koes. Protein–ligand scoring with convolutional neural networks. *Journal of Chemical Information and Modeling*, 57(4):942–957, 2017.

[132] Fergus Imrie, Anthony R. Bradley, Mihaela van der Schaar, and Charlotte M. Deane. 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, 2018.

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

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

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

[136] Francesco Gentile, Vibudh Agrawal, Michael Hsing, Anh-Tien Ton, Fuqiang Ban, Ulf Norinder, Martin E. Gleave, and Artem Cherkasov. Deep docking: A deep learning platform for augmentation of structure based drug discovery. *ACS Central Science*, 6(6):939–949, 2020.

[137] Dejun Jiang, Zhenxing Wu, Chang-Yu Hsieh, Guangyong Chen, Ben Liao, Zhe Wang, Chao Shen, Dongsheng Cao, Jian Wu, and Tingjun Hou. 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, 2021.

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

[139] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096, 2019.

[140] World Health Organization. Coronavirus (covid-19) dashboard, 2023. https://covid19.who.int.

[141] John Chodera, Alpha A Lee, Nir London, and Frank von Delft. Crowdsourcing drug discovery for pandemics. *Nature Chemistry*, 12(7):581–581, 2020.

[142] The COVID Moonshot Consortium, Hagit Achdout, Anthony Aimon, Elad Bar-David, Haim Barr, Amir Ben-Shmuel, James Bennett, Vitaliy A. Bilenko, Vitaliy A. Bilenko, Melissa L. Boby, Bruce Borden, Gregory R. Bowman, Juliane Brun, Sarma BVNBS, Mark Calmiano, Anna Carbery, Daniel Carney, Emma Cattermole, Edcon Chang, Eugene Chernyshenko, John D. Chodera, Austin Clyde, Joseph E. Coffland, Galit Cohen, Jason Cole, Alessandro Contini, Lisa Cox, Milan Cvitkovic, Alex Dias, Kim Donckers, David L. Dotson, Alice Douangamath, Shirly Duberstein, Tim Dudgeon, Louise Dunnett, Peter K. Eastman, Noam Erez, Charles J. Eyermann, Mike Fairhead, Gwen Fate, Daren Fearon, Oleg Fedorov, Matteo Ferla, Rafaela S. Fernandes, Lori Ferrins, Richard Foster, Holly Foster, Ronen Gabizon, Adolfo Garcia-Sastre, Victor O. Gawriljuk, Paul Gehrtz, Carina Gileadi, Charline Giroud, William G. Glass, Robert Glen, Itai Glinert, Andre S. Godoy, Marian Gorichko, Tyler Gorrie-Stone, Ed J. Griffen, Storm Hassell Hart, Jag Heer, Michael Henry, Michelle Hill, Sam Horrell, Victor D. Huliak, Matthew F.D. Hurley, Tomer Israely, Andrew Jajack, Jitske Jansen, Eric Jnoff, Dirk Jochmans, Tobias John, Steven De Jonghe, Anastassia L. Kantsadi, Peter W. Kenny, J. L. Kiappes, Serhii O. Kinakh, Lizbe Koekemoer, Boris Kovar, Tobias Krojer, Alpha Lee, Bruce A. Lefker, Haim Levy, Ivan G. Logvinenko, Nir London, Petra Lukacik, Hannah Bruce Macdonald, Beth MacLean, Tika R. Malla, Tatiana Matviiuk, Willam McCorkindale, Briana L. McGovern, Sharon Melamed, Kostiantyn P. Melnykov, Oleg Michurin, Halina Mikolajek, Bruce F. Milne, Aaron Morris, Garrett M. Morris, Melody Jane Morwitzer, Demetri Moustakas, Aline M. Nakamura, Jose Brandao Neto, Johan Neyts, Luong Nguyen, Gabriela D. Noske, Vladas Oleinikovas, Glaucius Oliva, Gijs J. Overheul, David Owen, Ruby Pai, Jin Pan, Nir Paran, Benjamin Perry, Maneesh Pingle, Jakir Pinjari, Boaz Politi, Ailsa Powell, Vladimir Psenak, Reut Puni, Victor L. Rangel, Rambabu N. Reddi, St Patrick Reid, Efrat Resnick, Emily Grace Ripka, Matthew C. Robinson, Ralph P. Robinson, Jaime Rodriguez-Guerra, Romel Rosales, Dominic Rufa, Kadi Saar, Kumar Singh Saikatendu, Chris Schofield, Mikhail Shafeev, Aarif Shaikh, Jiye Shi, Khriesto Shurrush, Sukrit Singh, Assa Sittner, Rachael Skyner, Adam Smalley, Bart Smeets, Mihaela D. Smilova, Leonardo J. Solmesky, John Spencer, Claire Strain-Damerell, Vishwanath Swamy, Hadas Tamir, Rachael Tennant, Warren Thompson, Andrew Thompson, Susana Tomasio, Igor S. Tsurupa, Anthony Tumber, Ioannis Vakonakis, Ronald P. van

Rij, Laura Vangeel, Finny S. Varghese, Mariana Vaschetto, Einat B. Vitner, Vincent Voelz, Andrea Volkamer, Frank von Delft, Annette von Delft, Martin Walsh, Walter Ward, Charlie Weatherall, Shay Weiss, Kris M. White, Conor Francis Wild, Matthew Wittmann, Nathan Wright, Yfat Yahalom-Ronen, Daniel Zaidmann, Hadeer Zidane, and Nicole Zitzmann. Open science discovery of oral non-covalent sars-cov-2 main protease inhibitor therapeutics. *bioRxiv*, 2022.

[143] Rolf Hilgenfeld. From sars to mers: crystallographic studies on coronaviral proteases enable antiviral drug design. *The FEBS journal*, 281(18):4085–4096, 2014.

[144] Zhenming Jin, Xiaoyu Du, Yechun Xu, Yongqiang Deng, Meiqin Liu, Yao Zhao, Bing Zhang, Xiaofeng Li, Leike Zhang, Chao Peng, et al. Structure of mpro from sars-cov-2 and discovery of its inhibitors. *Nature*, 582(7811):289–293, 2020.

[145] Yuzhi Liu, Chengyuan Liang, Liang Xin, Xiaodong Ren, Lei Tian, Xingke Ju, Han Li, Yongbo Wang, Qianqian Zhao, Hong Liu, et al. The development of coronavirus 3c-like protease (3clpro) inhibitors from 2010 to 2020. *European journal of medicinal chemistry*, page 112711, 2020.

[146] Sven Ullrich and Christoph Nitsche. The sars-cov-2 main protease as drug target. *Bioorganic & Medicinal Chemistry Letters*, page 127377, 2020.

[147] Linlin Zhang, Daizong Lin, Xinyuanyuan Sun, Ute Curth, Christian Drosten, Lucie Sauerhering, Stephan Becker, Katharina Rox, and Rolf Hilgenfeld. Crystal structure of sars-cov-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science*, 368(6489):409–412, 2020.

[148] Dafydd R. Owen, Charlotte M. N. Allerton, Annaliesa S. Anderson, Lisa Aschenbrenner, Melissa Avery, Simon Berritt, Britton Boras, Rhonda D. Cardin, Anthony Carlo, Karen J. Coffman, Alyssa Dantonio, Li Di, Heather Eng, RoseAnn Ferre, Ketan S. Gajiwala, Scott A. Gibson, Samantha E. Greasley, Brett L. Hurst, Eugene P. Kadar, Amit S. Kalgutkar, Jack C. Lee, Jisun Lee, Wei Liu, Stephen W. Mason, Stephen Noell, Jonathan J. Novak, R. Scott Obach, Kevin Ogilvie, Nandini C. Patel, Martin Pettersson, Devendra K. Rai, Matthew R. Reese, Matthew F. Sammons, Jean G. Sathish, Ravi Shankar P. Singh, Claire M. Steppan, Al E. Stewart, Jamison B. Tuttle, Lawrence Updyke, Patrick R. Verhoest, Liuqing Wei, Qingyi Yang, and Yuao Zhu. An oral sars-cov-2 m^{pro} inhibitor clinical candidate for the treatment of covid-19. *Science*, 374(6575):1586–1593, 2021.

[149] Hiroshi Mukae, Hiroshi Yotsuyanagi, Norio Ohmagari, Yohei Doi, Takumi Imamura, Takuhiro Sonoyama, Takahiro Fukuhara, Genki Ichihashi, Takao Sanaki, Keiko Baba, Yosuke Takeda, Yuko Tsuge, and Takeki Uehara. A randomized phase 2/3 study of ensitrelvir, a novel oral sars-cov-2 3c-like protease inhibitor, in japanese patients with mild-to-moderate covid-19 or asymptomatic sars-cov-2 infection: Results of the phase 2a part. *Antimicrobial Agents and Chemotherapy*, 66(10):e00697–22, 2022.

[150] Hiroshi Mukae, Hiroshi Yotsuyanagi, Norio Ohmagari, Yohei Doi, Hiroki Sakaguchi, Takuhiro Sonoyama, Genki Ichihashi, Takao Sanaki, Keiko Baba, Yuko Tsuge, and Takeki Uehara. Efficacy and Safety of Ensitrelvir in Patients With Mild-to-Moderate Coronavirus Disease 2019 (COVID-19): The Phase 2b Part of a Randomized, Placebo-Controlled, Phase 2/3 Study. *Clinical Infectious Diseases*, 12 2022.

[151] Diamond Light Source. Main protease structure and xchem fragment screen, 2020. https://www.diamond.ac.uk/covid-19/for-scientists/Main-protease-structure-and-XChem.html.

[152] PostEra Inc. COVID moonshot. https://postera.ai/covid, 2022.

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

[154] Michael M Hann, Andrew R Leach, and Gavin Harper. 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, 2001.

[155] Osamu Ichihara, John Barker, Richard J. Law, and Mark Whittaker. Compound design by fragment-linking. *Molecular Informatics*, 30(4):298–306, 2011.

[156] Haoyu S. Yu, Kalyan Modugula, Osamu Ichihara, Kimberly Kramschuster, Simon Keng, Robert Abel, and Lingle Wang. 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, 2021. PMID: 33372778.

[157] Johannes Schiebel, Stefan G Krimmer, Karine Röwer, Anna Knörlein, Xiaojie Wang, Ah Young Park, Martin Stieler, Frederik R Ehrmann, Kan Fu, Nedyalka Radeva, et al. High-throughput crystallography: reliable and efficient identification of fragment hits. *Structure*, 24(8):1398–1409, 2016.

[158] Alice Douangamath, Daren Fearon, Paul Gehrtz, Tobias Krojer, Petra Lukacik, C. David Owen, Efrat Resnick, Claire Strain-Damerell, Anthony Aimon, Péter Ábrányi-Balogh, José Brandão-Neto, Anna Carbery, Gemma Davison, Alexandre Dias, Thomas D. Downes, Louise Dunnett, Michael Fairhead, James D. Firth, S. Paul Jones, Aaron Keeley, György M. Keserü, Hanna F. Klein, Mathew P. Martin, Martin E. M. Noble, Peter O'Brien, Ailsa Powell, Rambabu N. Reddi, Rachael Skyner, Matthew Snee, Michael J. Waring, Conor Wild, Nir London, Frank von Delft, and Martin A. Walsh. Crystallographic and electrophilic fragment screening of the sars-cov-2 main protease. *Nature Communications*, 11(1):5047, 2020.

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

[160] Fergus Imrie, Anthony R. Bradley, Mihaela van der Schaar, and Charlotte M. Deane. Deep generative models for 3d linker design. *Journal of Chemical Information and Modeling*, 60(4):1983–1995, 04 2020.

[161] Yuyao Yang, Shuangjia Zheng, Shimin Su, Chao Zhao, Jun Xu, and Hongming Chen. Syntalinker: automatic fragment linking with deep conditional transformer neural networks. *Chem. Sci.*, 11:8312–8322, 2020.

[162] Fergus Imrie, Thomas E. Hadfield, Anthony R. Bradley, and Charlotte M. Deane. Deep generative design with 3d pharmacophoric constraints. *Chem. Sci.*, 12:14577–14589, 2021.

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

[164] Eugene N Muratov, Jürgen Bajorath, Robert P Sheridan, Igor V Tetko, Dmitry Filimonov, Vladimir Poroikov, Tudor I Oprea, Igor I Baskin, Alexandre Varnek, Adrian Roitberg, et al. Qsar without borders. *Chemical Society Reviews*, 49(11):3525–3564, 2020.

[165] Marion Schuller, Galen J. Correy, Stefan Gahbauer, Daren Fearon, Taiasean Wu, Roberto Efraín Díaz, Iris D. Young, Luan Carvalho Martins, Dominique H. Smith, Ursula Schulze-Gahmen, Tristan W. Owens, Ishan Deshpande, Gregory E. Merz, Aye C. Thwin, Justin T. Biel, Jessica K. Peters, Michelle Moritz, Nadia Herrera, Huong T. Kratochvil, null null, Anthony Aimon, James M. Bennett, Jose Brandao Neto, Aina E. Cohen, Alexandre Dias, Alice Douangamath, Louise Dunnett, Oleg Fedorov, Matteo P. Ferla, Martin R. Fuchs, Tyler J. Gorrie-Stone, James M. Holton, Michael G. Johnson, Tobias Krojer, George Meigs, Ailsa J. Powell, Johannes Gregor Matthias Rack, Victor L. Rangel, Silvia Russi, Rachael E. Skyner, Clyde A. Smith, Alexei S. Soares, Jennifer L. Wierman, Kang Zhu, Peter O'Brien, Natalia Jura, Alan Ashworth, John J. Irwin, Michael C. Thompson, Jason E. Gestwicki, Frank von Delft, Brian K. Shoichet, James S. Fraser, and Ivan Ahel. Fragment binding to the nsp3 macrodomain of sars-cov-2 identified through crystallographic screening and computational docking. *Science Advances*, 7(16):eabf8711, 2021.

[166] Greg Landrum. RDKit: Open-source cheminformatics. http://www.rdkit.org, 2023.

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

[168] 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.*, 57:5909–5912, 2021.

[169] Kadi L. Saar, Daren Fearon, The COVID Moonshot Consortium, Frank von Delft, John D. Chodera, and Alpha A. Lee. Turning high-throughput structural biology into predictive inhibitor design. *bioRxiv*, 2021.

[170] Christoph Gorgulla, Andras Boeszoermenyi, Zi-Fu Wang, Patrick D. Fischer, Paul W. Coote, Krishna M. Padmanabha Das, Yehor S. Malets, Dmytro S. Radchenko, Yurii S. Moroz, David A. Scott, Konstantin Fackeldey, Moritz Hoffmann, Iryna Iavniuk, Gerhard Wagner, and Haribabu Arthanari. An open-source drug discovery platform enables ultra-large virtual screens. *Nature*, 580(7805):663–668, 2020.

[171] Chemspace. Lead-like compounds, 2023.

[172] Johannes C. Hermann, Yingsi Chen, Charles Wartchow, John Menke, Lin Gao, Shelley K. Gleason, Nancy-Ellen Haynes, Nathan Scott, Ann Petersen, Stephen Gabriel, Binh Vu, Kelly M. George, Arjun Narayanan, Shirley H. Li, Hong Qian, Nanda Beatini, Linghao Niu, and Qing-Fen Gan. Metal impurities cause false positives in high-throughput screening campaigns. *ACS Medicinal Chemistry Letters*, 4(2):197–200, 02 2013.

[173] Francesca E. Morreale, Andrea Testa, Viduth K. Chaugule, Alessio Bortoluzzi, Alessio Ciulli, and Helen Walden. 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, 10 2017.

[174] Nigel P Duffy. Molecular property modeling using ranking, April 20 2010. US Patent 7,702,467.

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

[176] Jeremy Howard et al. fastai. https://github.com/fastai/fastai, 2018.

[177] The COVID Moonshot Consortium. 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*, 2020.

[178] Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, 3(10):1503–1507, 2008.

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

[180] Alpha A Lee, Qingyi Yang, Vishnu Sresht, Peter Bolgar, Xinjun Hou, Jacquelyn L Klug-McLeod, Christopher R Butler, et al. Molecular transformer unifies reaction prediction and retrosynthesis across pharma chemical space. *Chemical Communications*, 55(81):12152–12155, 2019.

[181] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.

[182] Kadi L. Saar, William McCorkindale, Daren Fearon, Melissa Boby, Haim Barr, Amir Ben-Shmuel, null null, Nir London, Frank von Delft, John D. Chodera, Alpha A. Lee, Matthew C. Robinson, Nir London, Efrat Resnick, Daniel Zaidmann, Paul Gehrtz, Rambabu N. Reddi, Ronen Gabizon, Haim Barr, Shirly Duberstein, Hadeer Zidane, Khriesto Shurrush, Galit Cohen, Leonardo J. Solmesky, Alpha Lee, Andrew Jajack, Milan Cvitkovic, Jin Pan, Ruby Pai, Emily Grace Ripka, Luong Nguyen, Mikhail Shafeev, Tatiana Matviiuk, Oleg Michurin, Eugene Chernyshenko, Vitaliy A. Bilenko, Serhii O. Kinakh, Ivan G. Logvinenko, Kostiantyn P. Melnykov, Victor D. Huliak, Igor S. Tsurupa, Marian Gorichko, Aarif Shaikh, Jakir Pinjari, Vishwanath Swamy, Maneesh Pingle, Sarma BVNBS, Anthony Aimon, Frank von Delft, Daren Fearon, Louise Dunnett, Alice Douangamath, Alex Dias, Ailsa Powell, Jose Brandao Neto, Rachael Skyner, Warren Thompson, Tyler Gorrie-Stone, Martin Walsh, David Owen, Petra Lukacik, Claire Strain-Damerell, Halina Mikolajek, Sam Horrell, Lizbé Koekemoer, Tobias Krojer, Mike Fairhead, Elizabeth M. MacLean, Andrew Thompson, Conor Francis Wild, Mihaela D. Smilova, Nathan Wright, Annette von Delft, Carina Gileadi, Victor L. Rangel,

Chris Schofield, Eidarus Salah, Tika R. Malla, Anthony Tumber, Tobias John, Ioannis Vakonakis, Anastassia L. Kantsadi, Nicole Zitzmann, Juliane Brun, J. L. Kiappes, Michelle Hill, Karolina D Witt, Dominic S Alonzi, Laetitia L Makower, Finny S. Varghese, Gijs J. Overheul, Pascal Miesen, Ronald P. van Rij, Jitske Jansen, Bart Smeets, Susana Tomésio, Charlie Weatherall, Mariana Vaschetto, Hannah Bruce Macdonald, John D. Chodera, Dominic Rufa, Matthew Wittmann, Melissa L. Boby, Michael Henry, William G. Glass, Peter K. Eastman, Joseph E. Coffland, David L. Dotson, Ed J. Griffen, Willam McCorkindale, Aaron Morris, Robert Glen, Jason Cole, Richard Foster, Holly Foster, Mark Calmiano, Rachael E. Tennant, Jag Heer, Jiye Shi, Eric Jnoff, Matthew F.D. Hurley, Bruce A. Lefker, Ralph P. Robinson, Charline Giroud, James Bennett, Oleg Fedorov, St Patrick Reid, Melody Jane Morwitzer, Lisa Cox, Garrett M. Morris, Matteo Ferla, Demetri Moustakas, Tim Dudgeon, Vladimír Pšenák, Boris Kovar, Vincent Voelz, Anna Carbery, Alessandro Contini, Austin Clyde, Amir Ben-Shmuel, Assa Sittner, Boaz Politi Einat B. Vitner, Elad Bar-David, Hadas Tamir, Hagit Achdout, Haim Levy, Itai Glinert, Nir Paran, Noam Erez, Reut Puni, Sharon Melamed, Shay Weiss, Tomer Israely, Yfat Yahalom-Ronen, Adam Smalley, Vladas Oleinikovas, John Spencer, Peter W. Kenny, Walter Ward, Emma Cattermole, Lori Ferrins, Charles J. Eyermann, Bruce F. Milne, Andre S. Godoy, Gabriela D. Noske, Glaucius Oliva, Rafaela S. Fernandes, Aline M. Nakamura, Victor O. Gawriljuk, Kris M. White, Briana L. McGovern, Romel Rosales, Adolfo Garcia-Sastre, Daniel Carney, Edcon Chang, Kumar Singh Saikatendu, Laura Vangeel Johan Neyts, Kim Donckers, Dirk Jochmans, Steven De Jonghe, Gregory R. Bowman, Bruce Borden, Sukrit Singh, Andrea Volkamer, Jaime Rodriguez-Guerra, Gwen Fate, Storm Hassell Hart, Vitaliy A. Bilenko, Serhii O. Kinakh, Ivan G. Logvinenko, Kostiantyn P. Melnykov, Victor D. Huliak, Igor S. Tsurupa, Kadi L Saar, Benjamin Perry, Laurent Fraisse, Peter Sjö, Pascale Boulet, Sophie Hahn, Charles Mowbray, Lauren Reid, Paul Rees, Qiu Yu Judy Huang, Sarah N Zvornicanin, Ala M. Shaqra, Nese Kurt Yilmaz, Celia A. Schiffer, Ivy Zhang, Iván Pulido, Charlie Tomlinson, Jenny C. Taylor, Tristan Ian Croll, and Lennart Brwewitz. Turning high-throughput structural biology into predictive inhibitor design. *Proceedings of the National Academy of Sciences*, 120(11):e2214168120, 2023.

[183] Brian Goldman, Steven Kearnes, Trevor Kramer, Patrick Riley, and W. Patrick Walters. Defining levels of automated chemical design. *Journal of Medicinal Chemistry*, 65(10):7073–7087, 2022.

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

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

[186] Ksenia Korovina, Sailun Xu, Kirthevasan Kandasamy, Willie Neiswanger, Barnabas Poczos, Jeff Schneider, and Eric P. Xing. Chembo: Bayesian optimization of small organic molecules with synthesizable recommendations, 2019.

[187] Jannis Born, Matteo Manica, Ali Oskooei, Joris Cadow, Karsten Borgwardt, and María Rodríguez Martínez. Paccmann^{RL}: Designing anticancer drugs from transcriptomic data via reinforcement learning, 2019.

[188] David C Blakemore, Luis Castro, Ian Churcher, David C Rees, Andrew W Thomas, David M Wilson, and Anthony Wood. Organic synthesis provides opportunities to transform drug discovery. *Nature chemistry*, 10(4):383, 2018.

[189] Jonas Boström, Dean G Brown, Robert J Young, and György M Keserü. Expanding the medicinal chemistry synthetic toolbox. *Nature Reviews Drug Discovery*, 2018.

[190] Marwin HS Segler, Mike Preuss, and Mark P Waller. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604, 2018.

[191] Akihiro Kishimoto, Beat Buesser, Bei Chen, and Adi Botea. 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, 2019.

[192] John S Schreck, Connor W Coley, and Kyle JM Bishop. Learning retrosynthetic planning through simulated experience. *ACS Central Science*, 5(6):970, 2019.

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

[194] Simon Johansson, Amol Thakkar, Thierry Kogej, Esben Bjerrum, Samuel Genheden, Tomas Bastys, Christos Kannas, Alexander Schliep, Hongming Chen, and Ola Engkvist. Ai-assisted synthesis prediction. *Drug Discovery Today: Technologies*, 2020.

[195] Thomas J. Struble, Juan C. Alvarez, Scott P. Brown, Milan Chytil, Justin Cisar, Renee L. DesJarlais, Ola Engkvist, Scott A. Frank, Daniel R. Greve, Daniel J. Griffin, Xinjun Hou, Jeffrey W. Johannes, Constantine Kreatsoulas, Brian Lahue, Miriam Mathea, Georg Mogk, Christos A. Nicolaou, Andrew D. Palmer, Daniel J. Price, Richard I. Robinson, Sebastian Salentin, Li Xing, Tommi Jaakkola, William. H. Green, Regina Barzilay, Connor W. Coley, and Klavs F. Jensen. Current and future roles of artificial intelligence in medicinal chemistry synthesis. *Journal of Medicinal Chemistry*, 63(16):8667–8682, 2020. PMID: 32243158.

[196] Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Costas Bekas, and Alpha A. Lee. Molecular Transformer - A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science*, 5(9):1572–1583, 11 2019.

[197] Wengong Jin, Connor W. Coley, Regina Barzilay, and Tommi Jaakkola. Predicting organic reaction outcomes with weisfeiler-lehman network. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):2608–2617, 2017.

[198] Igor V Tetko, Pavel Karpov, Ruud Van Deursen, and Guillaume Godin. State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature communications*, 11(1):1–11, 2020.

[199] Philippe Schwaller, Théophile Gaudin, Dávid Lányi, Costas Bekas, and Teodoro Laino. "found in translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.*, 9:6091–6098, 2018.

[200] Esben Jannik Bjerrum. Smiles enumeration as data augmentation for neural network modeling of molecules, 2017.

[201] Amol Thakkar, Thierry Kogej, Jean-Louis Reymond, Ola Engkvist, and Esben Jannik Bjerrum. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chem. Sci.*, 11:154–168, 2020.

[202] Xiwen Jia, Allyson Lynch, Yuheng Huang, Matthew Danielson, Immaculate Lang'at, Alexander Milder, Aaron E. Ruby, Hao Wang, Sorelle A. Friedler, Alexander J. Norquist, and Joshua Schrier. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature*, 573:251–255, 2019.

[203] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *34th International Conference on Machine Learning, ICML 2017*, 7:5109–5118, 2017.

[204] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "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, 2016.

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

[206] Grégoire Montavon, Wojciech Samek, and Klaus Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal*, 73:1–15, 2018.

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

[208] Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 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, 2018. Association for Computational Linguistics.

[209] Kevin McCloskey, Ankur Taly, Federico Monti, Michael P. Brenner, and Lucy J. Colwell. 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, 2019.

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

[211] Timothy E. H. Allen, Andrew J. Wedlake, Elena Gelžinytė, Charles Gong, Jonathan M. Goodman, Steve Gutsell, and Paul J. Russell. Neural network activation similarity: a new measure to assist decision making in chemical toxicology. *Chem. Sci.*, 11:7335–7348, 2020.

[212] Anna Maria Lluch, Francisco Sánchez-Baeza, Angel Messeguer, Caterina Fusco, and Ruggero Curci. Regio- and chemoselective epoxidation of fluorinated monoterpenes and sesquiterpenes by dioxiranes. *Tetrahedron*, 49(28):6299–6308, 1993.

[213] Jonathan Clayden, Nick Greeves, and Stuart Warren. *Organic Chemistry*. Oxford University Press, 2nd edition, 2012.

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

[215] Charles Friedel and James Crafts. Sur une nouvelle méthode générale de synthèse d'hydrocarbures, d'acétones, etc., 1877.

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

[217] John G. Topliss. A manual method for applying the hansch approach to drug design. *Journal of Medicinal Chemistry*, 20(4):463–469, 1977.

[218] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy, July 2019. Association for Computational Linguistics.

[219] Nadine Schneider, Daniel M. Lowe, Roger A. Sayle, and Gregory A. Landrum. 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, 2015. PMID: 25541888.

[220] John Bradshaw, Matt J. Kusner, Brooks Paige, Marwin H. S. Segler, and José Miguel Hernández-Lobato. A generative model for electron paths, 2019.

[221] Yanfei Guan, Connor W. Coley, Haoyang Wu, Duminda Ranasinghe, Esther Heid, Thomas J. Struble, Lagnajit Pattanaik, William H. Green, and Klavs F. Jensen. Regioselectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chem. Sci.*, 12:2198–2208, 2021.

[222] Mikołaj Sacha, Mikołaj Błaż, Piotr Byrski, Paweł Włodarczyk-Pruszyński, and Stanisław Jastrzębski. Molecule edit graph attention network: Modeling chemical reactions as sequences of graph edits, 2020.

[223] Alexander Buitrago Santanilla, Erik L. Regalado, Tony Pereira, Michael Shevlin, Kevin Bateman, Louis-Charles Campeau, Jonathan Schneeweis, Simon Berritt, Zhi-Cai Shi, Philippe Nantermet, Yong Liu, Roy Helmy, Christopher J. Welch, Petr Vachal, Ian W. Davies, Tim Cernak, and Spencer D. Dreher. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science*, 347(6217):49–53, 2015.

[224] Damith Perera, Joseph W. Tucker, Shalini Brahmbhatt, Christopher J. Helal, Ashley Chong, William Farrell, Paul Richardson, and Neal W. Sach. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science*, 359(6374):429–434, 2018.

[225] Paul Gehrtz, Shir Marom, Mike Bührmann, Julia Hardick, Silke Kleinbölting, Amit Shraga, Christian Dubiella, Ronen Gabizon, Jan N. Wiese, Matthias P. Müller, Galit Cohen, Ilana Babaev, Khriesto Shurrush, Liat Avram, Efrat Resnick, Haim Barr, Daniel Rauh, and Nir London. Optimization of covalent mkk7 inhibitors via crude nanomole-scale libraries. *Journal of Medicinal Chemistry*, 65(15):10341–10356, 2022.

[226] Nathan J. Gesmundo, Bérengère Sauvagnat, Patrick J. Curran, Matthew P. Richards, Christine L. Andrews, Peter J. Dandliker, and Tim Cernak. Nanoscale synthesis and affinity ranking. *Nature*, 557(7704):228–232, 2018.

[227] Kevin McCloskey, Eric A. Sigel, Steven Kearnes, Ling Xue, Xia Tian, Dennis Moccia, Diana Gikunju, Sana Bazzaz, Betty Chan, Matthew A. Clark, John W. Cuozzo, Marie-Aude Guié, John P. Guilinger, Christelle Huguet, Christopher D. Hupp, Anthony D. Keefe, Christopher J. Mulhern, Ying Zhang, and Patrick Riley. Machine learning on dna-encoded libraries: A new paradigm for hit finding. *Journal of Medicinal Chemistry*, 63(16):8857–8866, 2020.

[228] Katherine S. Lim, Andrew G. Reidenbach, Bruce K. Hua, Jeremy W. Mason, Christopher J. Gerry, Paul A. Clemons, and Connor W. Coley. Machine learning on dna-encoded library count data using an uncertainty-aware probabilistic loss function. *Journal of Chemical Information and Modeling*, 62(10):2316–2331, 2022.

[229] Vincent Blay, Xiaoyu Li, Jacob Gerlach, Fabio Urbina, and Sean Ekins. Combining dels and machine learning for toxicology prediction. *Drug Discovery Today*, 27(11):103351, 2022.

[230] Antonia F. Stepan, Daniel P. Walker, Jonathan Bauman, David A. Price, Thomas A. Baillie, Amit S. Kalgutkar, and Michael D. Aleo. Structural alert/reactive metabolite concept as applied in medicinal chemistry to mitigate the risk of idiosyncratic drug toxicity: A perspective based on the critical examination of trends in the top 200 drugs marketed in the united states. *Chemical Research in Toxicology*, 24(9):1345–1410, 2011.

[231] Lisa M. Baker, Anthony Aimon, James B. Murray, Allan E. Surgenor, Natalia Matassova, Stephen D. Roughley, Patrick M. Collins, Tobias Krojer, Frank von Delft, and Roderick E. Hubbard. Rapid optimisation of fragments and hits to lead compounds from screening of crude reaction mixtures. *Communications Chemistry*, 3(1):122, 2020.

[232] Zhengkai Tu, Thijs Stuyver, and Connor W. Coley. Predictive chemistry: machine learning for reaction deployment, reaction development, and reaction discovery. *Chem. Sci.*, 14:226–244, 2023.

[233] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M.

Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

[234] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[235] OpenAI. GPT-4 Technical Report. https://cdn.openai.com/papers/gpt-4.pdf, 2023.

[236] Kevin Maik Jablonka, Luc Patiny, and Berend Smit. Making the collective knowledge of chemistry open and machine actionable. *Nature Chemistry*, 14(4):365–376, 2022.

[237] Amol Thakkar, Thierry Kogej, Jean-Louis Reymond, Ola Engkvist, and Esben Jannik Bjerrum. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chem. Sci.*, 11:154–168, 2020.

[238] Olaf Wiest, Mandana Saebi, Bozhao Nan, John E Herr, Jessica Wahlers, Zhichun Guo, Andrzej Zuranski, Thierry Kogej, Per-Ola Norrby, Abigail G Doyle, and Nitesh V. Chawla. On the use of real-world datasets for reaction yield prediction. *Chem. Sci.*, pages –, 2023.

[239] Steven M. Kearnes, Michael R. Maser, Michael Wleklinski, Anton Kast, Abigail G. Doyle, Spencer D. Dreher, Joel M. Hawkins, Klavs F. Jensen, and Connor W. Coley. The open reaction database. *Journal of the American Chemical Society*, 143(45):18820–18826, 2021.

[240] Derek T. Ahneman, Jesús G. Estrada, Shishi Lin, Spencer D. Dreher, and Abigail G. Doyle. Predicting reaction performance in c-n cross-coupling using machine learning. *Science*, 360(6385):186–190, 2018.

[241] Philippe Schwaller, Alain C Vaucher, Teodoro Laino, and Jean-Louis Reymond. Prediction of chemical reaction yields using deep learning. *Machine Learning: Science and Technology*, 2(1):015016, mar 2021.

[242] Daniel Probst, Philippe Schwaller, and Jean-Louis Reymond. Reaction classification and yield prediction using the differential reaction fingerprint drfp. *Digital Discovery*, 1:91–97, 2022.

[243] Emma King-Smith, Simon Berritt, Louise Bernier, Xinjun Hou, Jacquelyn Klug-McLeod, Jason Mustakis, Neal Sach, Joseph Tucker, Qingyi Yang, Roger Howard, et al. Probing the chemical" reactome" with high throughput experimentation data. 10 2022.

[244] Jessica Xu, Dipannita Kalyani, Thomas Struble, Spencer Dreher, Shane Krska, Stephen L Buchwald, and Klavs F Jensen. Roadmap to pharmaceutically relevant reactivity models leveraging high-throughput experimentation. 9 2022.

[245] Benjamin J. Shields, Jason Stevens, Jun Li, Marvin Parasram, Farhan Damani, Jesus I. Martinez Alvarado, Jacob M. Janey, Ryan P. Adams, and Abigail G. Doyle. Bayesian reaction optimization as a tool for chemical synthesis. *Nature*, 590(7844):89–96, 2021.

[246] Benedict W. J. Irwin, Julian R. Levell, Thomas M. Whitehead, Matthew D. Segall, and Gareth J. Conduit. Practical applications of deep learning to impute heterogeneous drug discovery data. *Journal of Chemical Information and Modeling*, 60(6):2848–2857, 2020.

[247] Benedict W. J. Irwin, Thomas M. Whitehead, Scott Rowland, Samar Y. Mahmoud, Gareth J. Conduit, and Matthew D. Segall. Deep imputation on large-scale drug discovery data. *Applied AI Letters*, 2(3):e31, 2021.

[248] Edwin G. Tse, Laksh Aithani, Mark Anderson, Jonathan Cardoso-Silva, Giovanni Cincilla, Gareth J. Conduit, Mykola Galushka, Davy Guan, Irene Hallyburton, Benedict W. J. Irwin, Kiaran Kirk, Adele M. Lehane, Julia C. R. Lindblom, Raymond Lui, Slade Matthews, James McCulloch, Alice Motion, Ho Leung Ng, Mario Öeren, Murray N. Robertson, Vito Spadavecchio, Vasileios A. Tatsis, Willem P. van Hoorn, Alexander D. Wade, Thomas M. Whitehead, Paul Willis, and Matthew H. Todd. An open drug discovery competition: Experimental validation of predictive models in a series of novel antimalarials. *Journal of Medicinal Chemistry*, 64(22):16450–16463, 2021.

[249] Jan Wenzel, Hans Matter, and Friedemann Schmidt. Predictive multitask deep neural network models for adme-tox properties: Learning from large data sets. *Journal of Chemical Information and Modeling*, 59(3):1253–1268, 2019.

[250] Olga Obrezanova, Anton Martinsson, Tom Whitehead, Samar Mahmoud, Andreas Bender, Filip Miljković, Piotr Grabowski, Ben Irwin, Ioana Oprisiu, Gareth Conduit, Matthew Segall, Graham F. Smith, Beth Williamson, Susanne Winiwarter, and Nigel Greene. Prediction of in vivo pharmacokinetic parameters and time–exposure curves in rats using machine learning from the chemical structure. *Molecular Pharmaceutics*, 19(5):1488–1504, 2022.

[251] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[252] Wouter Heyndrickx, Lewis Mervin, Tobias Morawietz, Noe Sturm, Lukas Friedrich, Adam Zalewski, Anastasia Pentina, Lina Humbeck, Martijn Oldenhof, Ritsuya Niwayama, Peter Schmidtke, Nikolas Fechner, Jaak Simm, Adam Arany, Nicolas Drizard, Rama Jabal, Arina Afanasyeva, Regis Loeb, Shlok Verma, Simon Harnqvist, Matthew Holmes, Balasz Pejo, Maria Telenczuk, Nicholas Holway, Arne Dieckmann, Nicola Rieke, Djork-Arne Clevert, Michael Krug, Christopher Luscombe, Darren Green, Peter Ertl, Peter Antal, David Marcus, Nicolas Do Huu, Hideyoshi Fuji, Stephen Pickett, Gergely Acs, Eric Boniface, Bernd Beck, Yax Sun, Arnaud Gohier, Friedrich Rippmann, Ola Engkvist, Andreas H. Goller, Yves Moreau, Mathieu N. Galtier, Ansgar Schuffenhauer, and Hugo Ceulemans. MELLODDY: cross pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. 10 2022.

[253] Martijn Oldenhof, Gergely Acs, Balazs Pejo, Ansgar Schuffenhauer, Nicholas Holway, Noe Sturm, Arne Dieckmann, Oliver Fortmeier, Eric Boniface, Clement Mayer,

Arnaud Gohier, Peter Schmidtke, Ritsuya Niwayama, Dieter Kopecky, Lewis Mervin, Prakash Chandra Rathi, Lukas Friedrich, András Formanek, Peter Antal, Jordon Rahaman, Adam Zalewski, Wouter Heyndrickx, Ezron Oluoch, Manuel Stossel, Michal Vanco, David Endico, Fabien Gelus, Thaïs de Boisfosse, Adrien Darbier, Ashley Nicollet, Matthieu Blottiere, Maria Telenczuk, Van Tien Nguyen, Thibaud Martinez, Camille Boillet, Kelvin Moutet, Alexandre Picosson, Aurelien Gasser, Inal Djafar, Antoine Simon, Adam Arany, Jaak Simm, Yves Moreau, Ola Engkvist, Hugo Ceulemans, Camille Marini, and Mathieu Galtier. Industry-scale orchestrated federated learning for drug discovery, 2022.

[254] Petra S. Dittrich and Andreas Manz. Lab-on-a-chip: microfluidics in drug discovery. *Nature Reviews Drug Discovery*, 5(3):210–218, 2006.

[255] Aleksander Skardal, Thomas Shupe, and Anthony Atala. Organoid-on-a-chip and body-on-a-chip systems for drug screening and disease modeling. *Drug Discovery Today*, 21(9):1399–1411, 2016.

[256] Tom L. Blundell, Harren Jhoti, and Chris Abell. High-throughput crystallography for lead discovery in drug design. *Nature Reviews Drug Discovery*, 1(1):45–54, 2002.

[257] Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv*, 2022.

[258] Barun Bhhatarai, W. Patrick Walters, Cornelis E. C. A. Hop, Guido Lanza, and Sean Ekins. Opportunities and challenges using artificial intelligence in adme/tox. *Nature Materials*, 18(5):418–422, 2019.

[259] Andreas H. Göller, Lara Kuhnke, Floriane Montanari, Anne Bonin, Sebastian Schneckener, Antonius ter Laak, Jörg Wichard, Mario Lobell, and Alexander Hillisch. Bayer's in silico admet platform: a journey of machine learning over the past two decades. *Drug Discovery Today*, 25(9):1702–1709, 2020.

[260] Maxwell L. Bileschi, David Belanger, Drew H. Bryant, Theo Sanderson, Brandon Carter, D. Sculley, Alex Bateman, Mark A. DePristo, and Lucy J. Colwell. Using deep learning to annotate the protein universe. *Nature Biotechnology*, 40(6):932–937, 2022.

[261] Andy Hsien-Wei Yeh, Christoffer Norn, Yakov Kipnis, Doug Tischer, Samuel J. Pellock, Declan Evans, Pengchen Ma, Gyu Rie Lee, Jason Z. Zhang, Ivan Anishchenko, Brian Coventry, Longxing Cao, Justas Dauparas, Samer Halabiya, Michelle DeWitt, Lauren Carter, K. N. Houk, and David Baker. De novo design of luciferases using deep learning. *Nature*, 614(7949):774–780, 2023.

[262] Daniel Probst, Matteo Manica, Yves Gaetan Nana Teukam, Alessandro Castrogiovanni, Federico Paratore, and Teodoro Laino. Biocatalysed synthesis planning using data-driven learning. *Nature Communications*, 13(1):964, 2022.

[263] Monica Schenone, Vlado Dancik, Bridget K Wagner, and Paul A Clemons. Target identification and mechanism of action in chemical biology and drug discovery. *Nature Chemical Biology*, 9(4):232–240, 2013.

[264] Connor W. Coley, Natalie S. Eyke, and Klavs F. Jensen. Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angewandte Chemie - International Edition*, pages 2–38, 2019.

[265] Connor W. Coley, Natalie S. Eyke, and Klavs F. Jensen. Autonomous Discovery in the Chemical Sciences Part II: Outlook. *Angewandte Chemie - International Edition*, pages 2–25, 2019.

[266] Mariya Popova, Olexandr Isayev, and Alexander Tropsha. Deep reinforcement learning for de novo drug design. *Science Advances*, 4(7):eaap7885, 2018.

[267] Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N. Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific Reports*, 9(1):10752, 2019.

[268] Vijil Chenthamarakshan, Samuel C. Hoffman, C. David Owen, Petra Lukacik, Claire Strain-Damerell, Daren Fearon, Tika R. Malla, Anthony Tumber, Christopher J. Schofield, Helen M. E. Duyvesteyn, Wanwisa Dejnirattisai, Loic Carrique, Thomas S. Walter, Gavin R. Screaton, Tetiana Matviiuk, Aleksandra Mojsilovic, Jason Crain, Martin A. Walsh, David I. Stuart, and Payel Das. Accelerating inhibitor discovery for multiple sars-cov-2 targets with a single, sequence-guided deep generative framework, 2022.

[269] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[270] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[271] Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates, 2018.

[272] PostEra. Manifold. https://app.postera.ai.

[273] Guillaume Klein, Yoon Kim, Jean Senellart, and Alexander M. Rush. OpenNMT, 2017.

[274] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank

Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

Appendix A

Computational Details

A.1 Docking against SARS-CoV-2 Mpro

All molecules synthesised by the COVID Moonshot Consortium were docked against structure x2908 reported by Diamond XChem [158]. 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 FRESCO

The workflow for implementing efficient and accurate KDE fitting utilises several software packages. The Improved Sheather-Jones algorithm for KDE selection utilises the implementation in `KDEpy`. Each KDE is then constructed using the chosen bandwidths with `scikit-learn` [269] 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 [270].

The pharmacophore processing, substructure filtering, and butina clustering workflow was implemented using the `rdkit` [166] Python package.

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

A.3 Ranking model

The MLP model for ranking compounds was implemented using the FastAI Tabular framework [176]. We found that the default hyperparameters often resulted in overfitting, so we use a smaller network with non-negligible dropout for additional regularization. Two layers of size 10 and 5 were used, with a dropout of 0.3 applied to both layers. The network was trained for 5 epochs with the Adam optimizer [110] and the 1cycle cyclical learning rate scheduler [271].

The reaction routes were generated using Manifold [272], an ML-based synthesis tool and search engine.

The model training sets, screening library, and code implementation are available at https://github.com/wjm41/mpro-rank-gen.

A.4 Molecular Transformer interpretation

The Molecular Transformer architecture used is based on the model described in Schwaller et. al. [126].

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 [111]. 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 [273] which makes use of the PyTorch framework [274].

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 at https://github.com/davkovacs/MTExplainer. The USPTO dataset used to train the model [25, 197], as well as the Tanimoto similarity-based train/test splits of USPTO can also be found in the GitHub repo.

A.5 Crude bioactivity modelling

The RF model was implemented with default hyperparameters in scikit-learn [269]. The code for implementing the RF model, as well as the bioactivity data, screening library and chosen molecules are available on https://gitlab.com/wjm41/noisyamides.

The GP model was also implemented in scikit-learn [269] with default length scale hyperparameters of 1.0 for the radial basis function as well as Matern kernels. The code for implementing the GP model can be found at https://github.com/emmaking-smith/Moonshot.

Appendix B

Experimental Details

B.1 SARS-CoV-2 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[142], 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 ML generated reaction schemes

B.4 SARS-CoV-2 nsp3-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 [165]. 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

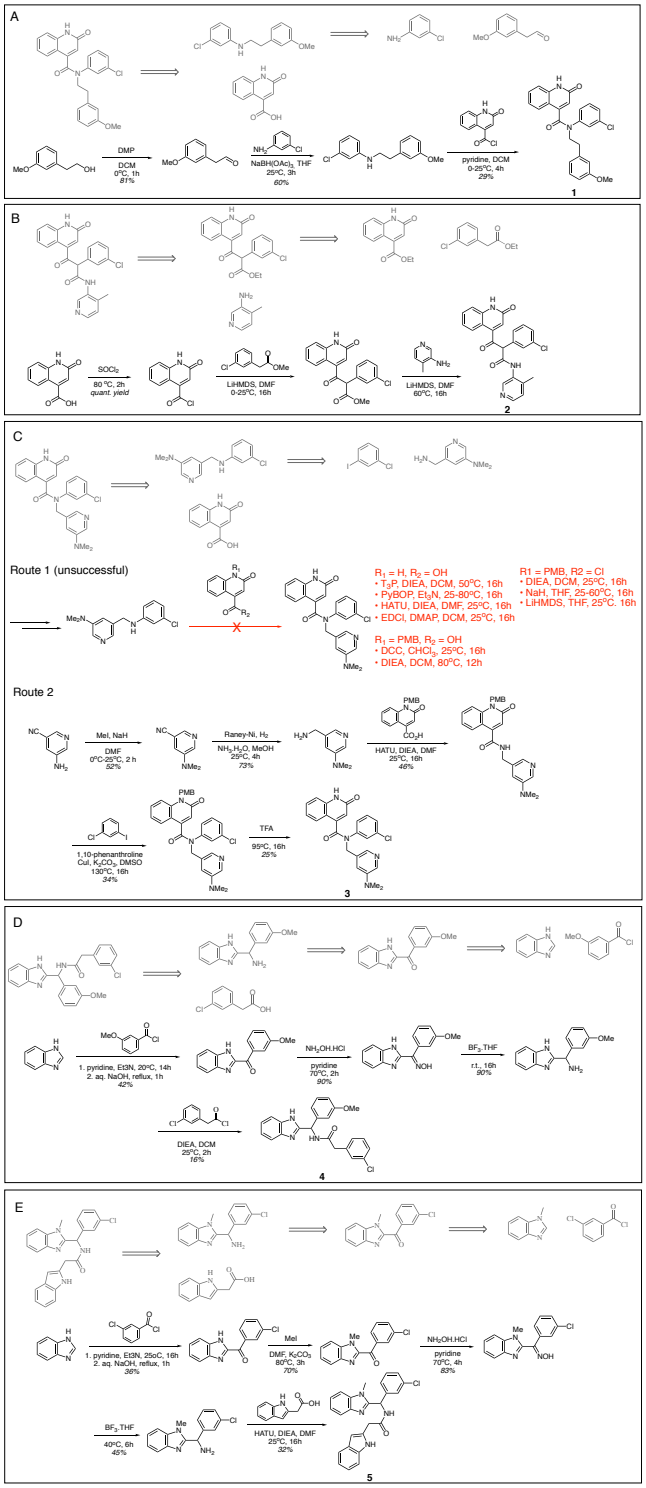


Fig. B.1 Model generated synthetic schemes for compounds **1**-**5** in chapter 4. The synthesis schemes generated by our model (grey) and the experimental schemes (black).

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-Throughput 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 μL DMSO. For yield estimation, 1 μL 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 μL. The desired product was identified in 60% of wells.