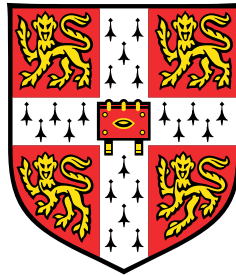


Accelerating the Design-Make-Test cycle of Drug Discovery with Machine Learning



William McCorkindale

Cavendish Laboratory, Department of Physics
University of Cambridge

Supervisor: Dr. Alpha Lee

Declaration

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

William McCorkindale
January 2023

Acknowledgements

And I would like to acknowledge ...

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

Abstract

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

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

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

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

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

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

Table of contents

Preface

	1
Chapter 1 (add link) talks about ...	2
In Chapter 2 we discuss ... This work resulted in the publication of the following article:	3
William McCorkindale, Ivan Ahel, Haim Barr, Galen J. Correy, James S. Fraser,	4
Nir London, Marion Schuller, Khriesto Shurrush, Alpha A. Lee. Fragment-Based	5
Hit Discovery via Unsupervised Learning of Fragment-Protein Complexes.	6
Specify who did what	7
Chapter 3:	8
Aaron Morris, William McCorkindale, The COVID Moonshot Consortium, Nir	9
Drayman, John D. Chodera, Savaş Tay, Nir London and Alpha A. Lee. Discovery	10
of SARS-CoV-2 main protease inhibitors using a synthesis-directed de novo design	11
model, <i>Chem. Commun.</i> , 2021,57, 5909-5912	12
Chapter 4	13
Dávid Péter Kovács, William McCorkindale and Alpha A. Lee. Quantitative	14
interpretation explains machine learning models for chemical reaction prediction	15
and uncovers bias. <i>Nature Communications</i> volume 12, Article number: 1695	16
(2021)	17
D.P.K. and W.M. implemented the algorithms, D.P.K. trained the models, designed the	18
experiments and analysed the model attributions for the various reaction classes. W.M. imple-	19
mented Tanimoto splitting and applied reaction templates for counting statistics and artificial	20
dataset generation. A.A.L. supervised and directed the project. All authors discussed the results	21
and approved the manuscript. D.P.K. and W.M. contributed equally to this study.	22
Chapter 5:	23
The final chapter ...	24

Chapter 1

25

Introduction

26

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

27

28

29

30

31

32

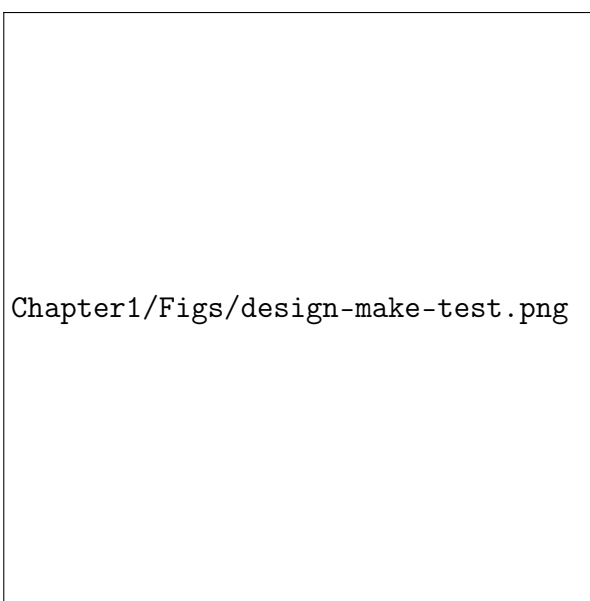


Fig. 1.1 An overview of the design-make-test cycle in drug discovery.

While computational methods have long been used in various stages of the cycle, there has been a recent surge in applying artificial intelligence to drug discovery following its success in various other fields, most notably computer vision and natural language processing. Since

33

34

35

molecular assaying is largely an automated process the application focus has been on ‘design’ and ‘make’ [?], for example in modelling quantitative structure-activity relationships (QSAR), designing generative models for proposing drug candidates, and planning retrosynthesis routes (Fig ??).

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

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

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

Chapter 2

65

Make - Understanding the Molecular Transformer

66

67

2.1 Introduction

68

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

69

70

71

72

73

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

74

75

76

77

78

79

80

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

81

82

83

84

85

86

87

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

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

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

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

The work in this chapter was done collaboratively with Dávid Péter Kovács. We did the code development together and discussed all of the results of the work. He executed the code, analysed the model attributions, and designed the majority of the experiments and all of the adversarial examples. I created the SMARTS templates for counting statistics from the patent datasets as well as for dataset generation in the synthetic experiments. Preliminary results from this work were presented at the ICML 2020 ‘ML Interpretability for Scientific Discovery’

Workshop [?]. All code including a README with the usage can be found in the GitHub repo
MTEExplainer [?].

2.2 Methods

2.2.1 Molecular Transformer

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

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

$$P(\text{tgt} \mid \text{src}) = \prod_{i=1}^N P(\text{tok}_i \mid \text{tok}_1, \dots, \text{tok}_{i-1}, \text{src}) \quad (2.1)$$

where tok_i is the i -th predicted token and N is the length of the prediction.

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

2.2.2 Data

We trained the model on a publicly available dataset of organic reactions mined from the US patent office [?] which has been filtered [?]. The data contains reactants, reagents, and products represented as SMILES (without including stereochemical information) which is a text based representation of molecules[? ?]. The training set was made up of 377 419 reactions

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

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

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

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

2.2.3 Integrated Gradients

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

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

$$IG_i(x) = (x_i - x'_i) \int_{\alpha=0}^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \quad (2.2)$$

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

It has been discussed before that the choice of baseline can have a large effect on the values of the attributions [?]. While we could have chosen unreactive molecules as our baseline, it is important to select baselines which are completely non-informative to avoid any ambiguity.

This would traditionally be the black image in the case of image recognition. In this work we use the embedding vector of the SMILES ‘. ’ token which is used to separate different molecules and hence does not contain any chemical information on its own.

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

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

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

2.2.4 Data Attribution

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

To successfully attribute to data, we must understand how ‘similar’ two input datapoints are according to the model by defining a similarity metric. For the Molecular Transformer which has an encoder-decoder architecture we use the output of the encoder layers as a basis for comparing data points. The challenge lies in the fact that these encoder hidden states have a

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

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

$$score = \frac{1}{1 + D} \quad (2.3)$$

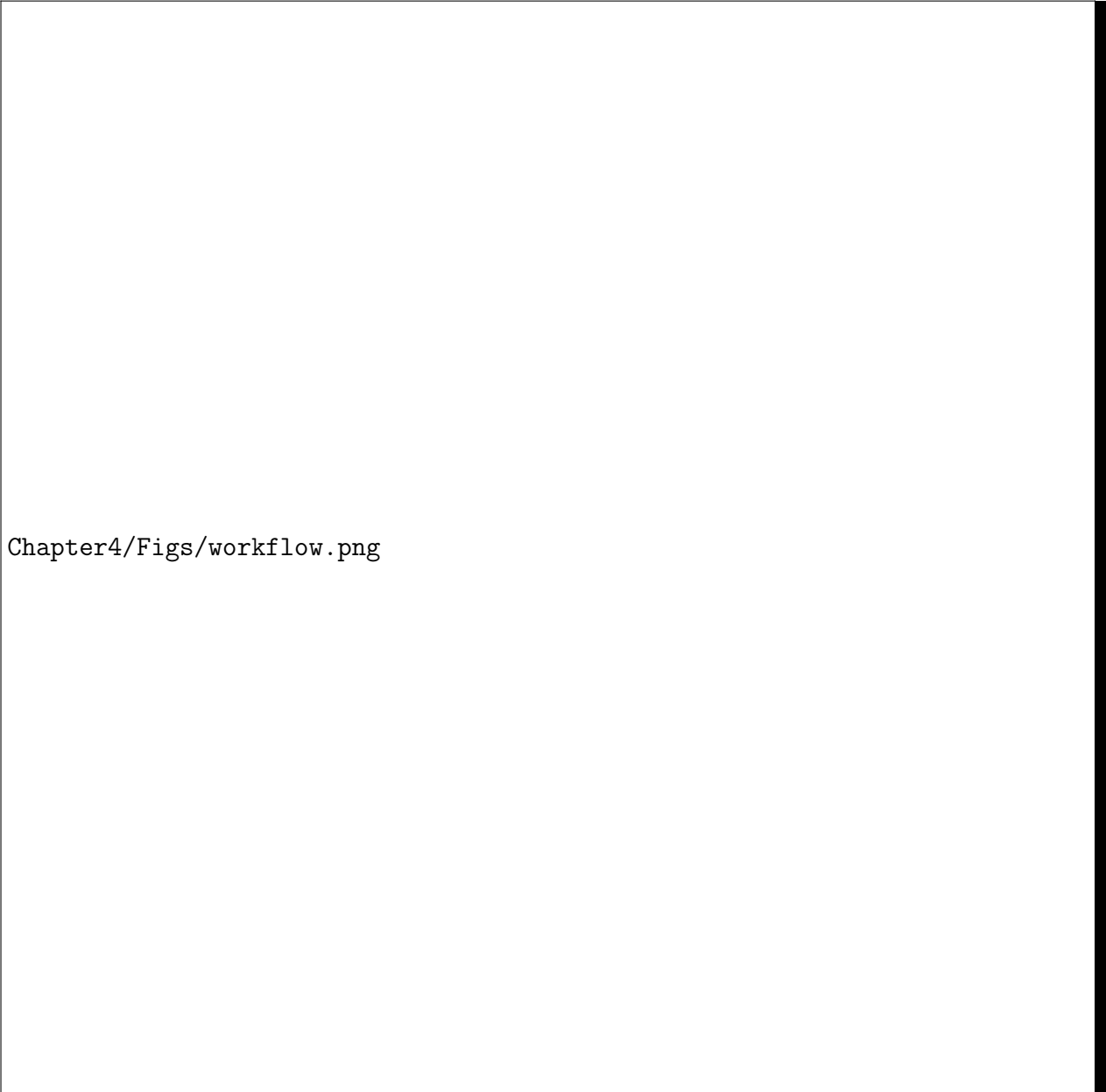
where D is the Euclidean distance between the vectors. This can be implemented in a vectorized way resulting in very quick computation even in the case of dataset sizes like Pistachio made up of over 2 million training examples. The top- n most similar reactions are returned where n is defined by the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

2.3 Results

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

2.3.1 Diels-Alder reactions

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



Chapter4/Figs/workflow.png

Fig. 2.1 An overview of the workflow for interpreting the Molecular Transformer.

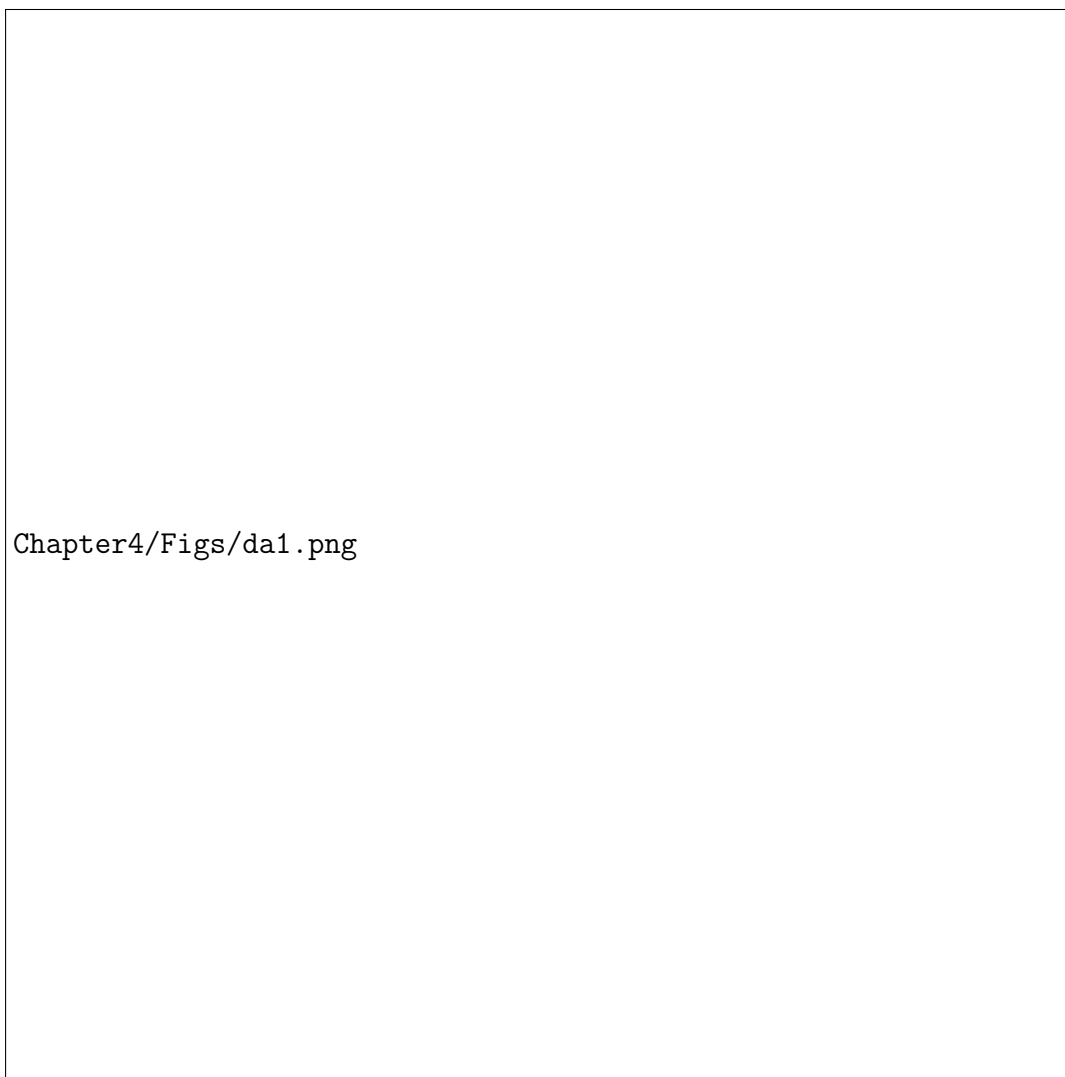


Fig. 2.2 A typical example of a Diels-Alder reaction with challenging selectivity.



Fig. 2.3 The USPTO transformer makes a completely incorrect prediction, while the Pistachio model correctly predicts the product and recognises the importance of the nitrile group. For the Pistachio model the IG attributions are shown together with the corresponding uniform attribution (ua) values.

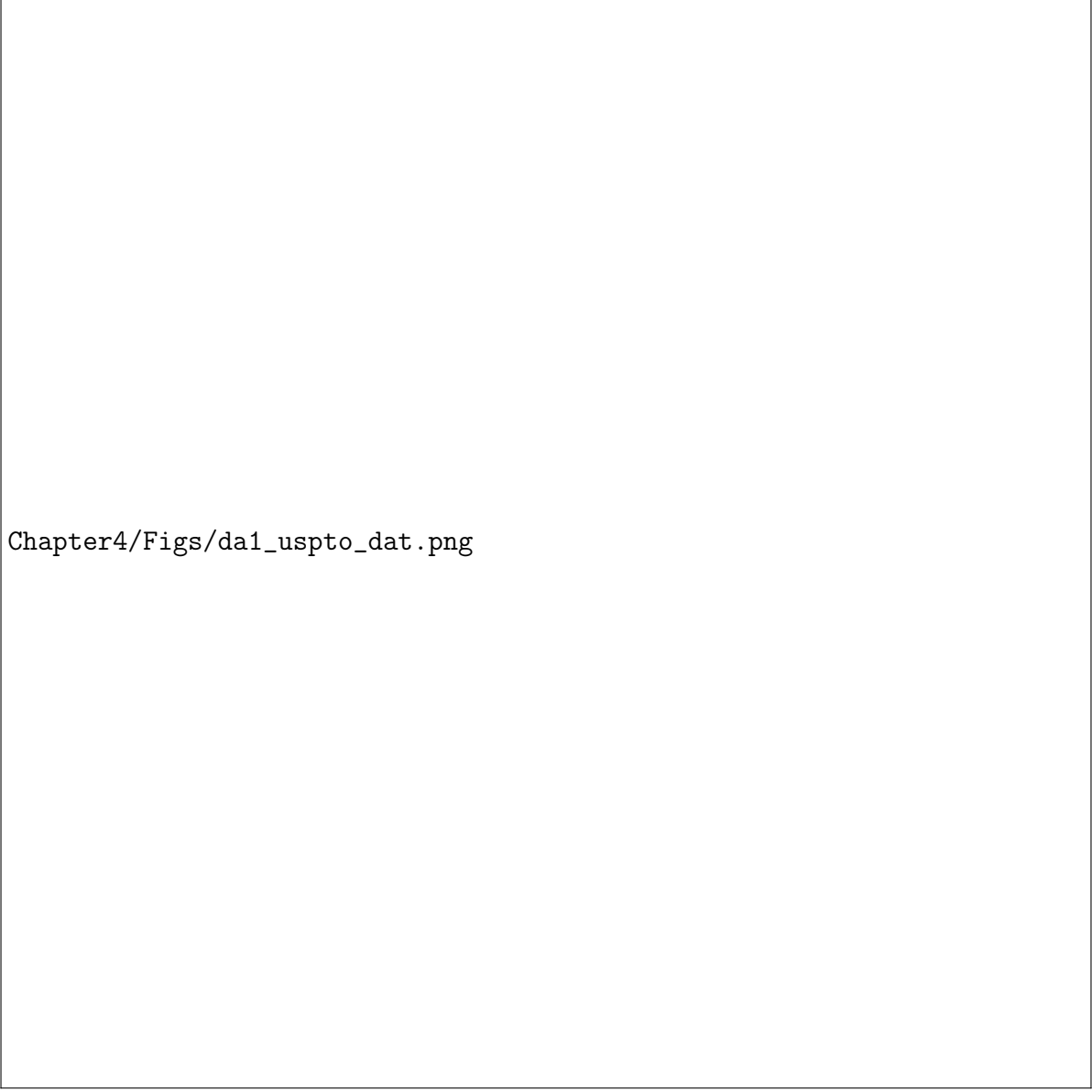
Fig ?? shows the Top-1 prediction of the USPTO and Pistachio models. The USPTO model does not seem to recognize the Diels-Alder reaction and gets the prediction wrong, indicating its own uncertainty by assigning a very low score of 0.300 to the prediction. To find the reason for the wrong prediction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

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

For the Pistachio model the Top-1 prediction is correct, as shown in Fig ??, and it has a confidence score of 0.819 indicating that it is fairly certain in the prediction. We also generated the IGs for this reaction to see if the selectivity is caused by the relevant nitrile and methoxy groups. The probability difference between the correct major and the minor products was 0.77 and it was distributed on the compounds as shown in Fig ?? alongside the uniform attribution values. It can be seen that the nitrile group received a higher than uniform attribution indicating that the model recognises its importance. The same cannot be said unambiguously about the methoxy group whose attribution is only slightly more than the corresponding uniform value. Based on this example we can conclude that the model has learnt to recognize Diels-Alder reactions, and the IGs point towards the fact that it has learnt the regioselectivity causes too. To confirm this a couple of reactions taken from publications were tested (Fig ??) and the Pistachio transformer is able to predict the correct products.

2.3.2 Friedel-Crafts acylation reactions

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



Chapter4/Figs/da1_uspto_dat.png

Fig. 2.4 Attribution to the USPTO training data shows that the USPTO transformer either completely fails to recognize Diels-Alder reactions or that no Diels-Alder reactions exist in the dataset.



Chapter4/Figs/da1_adv.png

Fig. 2.5 Further test reactions correctly predicted by the Pistachio model, validate that Pistachio correctly understands Diels-Alder reactions. [? ?]

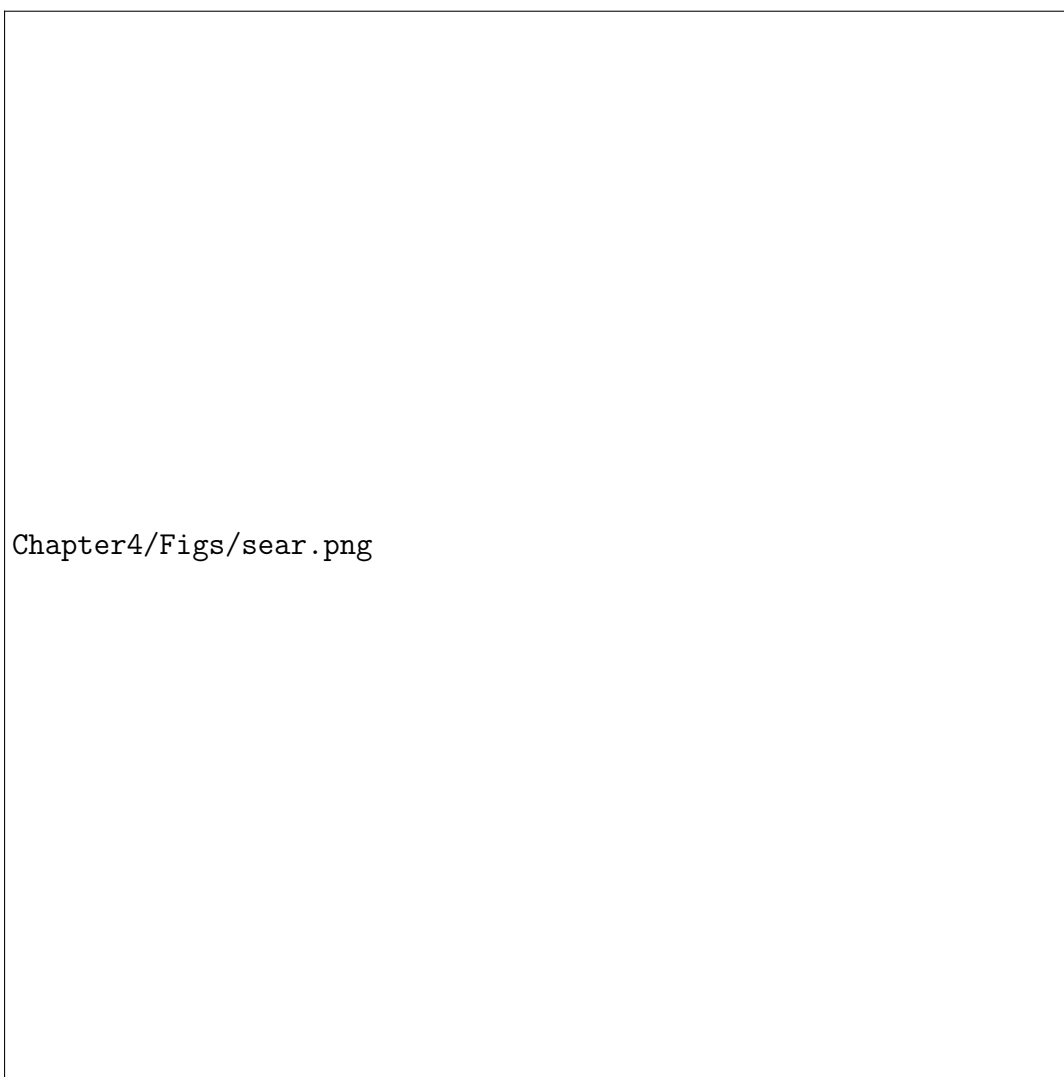


Fig. 2.6 Friedel-Crafts acylation reaction taken from the USPTO training set showing para selectivity.



Fig. 2.7 Both models predict the correct Friedel-Crafts acylation product but only the Pistachio model recognizes the importance of the -F atom in determining selectivity. The Integrated Gradients attributions are also shown along with the uniform attribution (ua) values.

The predictions of the two models are shown in Fig ???. Both models predict the para selectivity correctly with confidence scores close to 1.0. When inspecting the IG attributions it can be seen that the USPTO model puts a very small weight on the Fluorine, only a fifth of the uniform attribution value. There is a large attribution given to the reagent though which does not affect the selectivity of this reaction at all. The attributions indicate that the USPTO transformer has not learnt the importance of F as the cause of para selectivity in these reactions. On the other hand the Pistachio transformer assigns a very high attribution value to F, suggesting that it has recognized the reason for the selectivity.

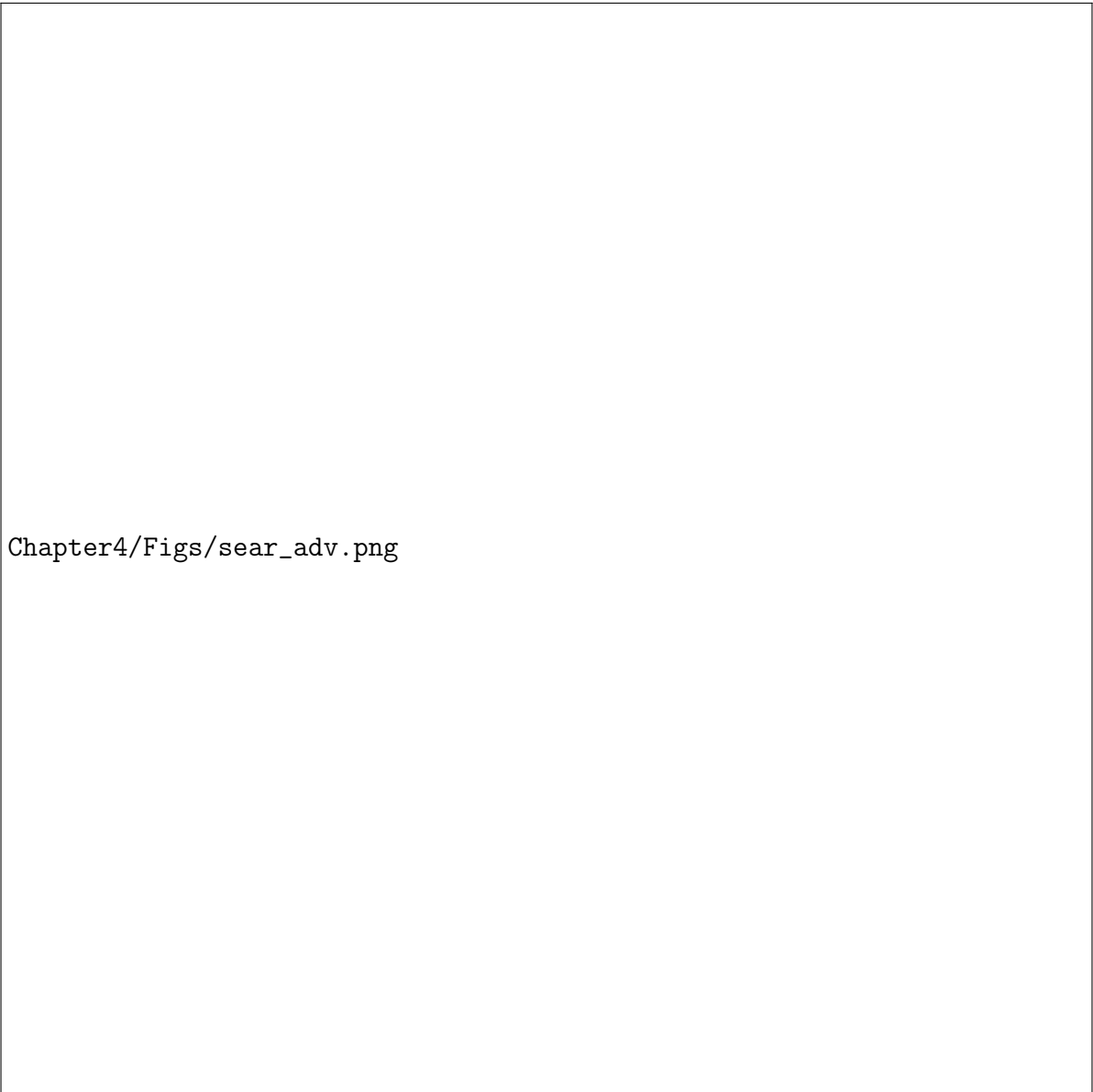
Guided by the attributions we designed a number of adversarial examples where we have changed only the Fluorine part of the reactant-reagent input. This choice was motivated by the fact that according to the USPTO model the selectivity was driven by the reagents instead of the substituent on the benzene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

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

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

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

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



Chapter4/Figs/sear_adv.png

Fig. 2.8 Adversarial examples designed using Fig ?? reveal that both models can easily fail to predict the correct meta product. The uniform attribution (ua) values for the IGs are also shown.

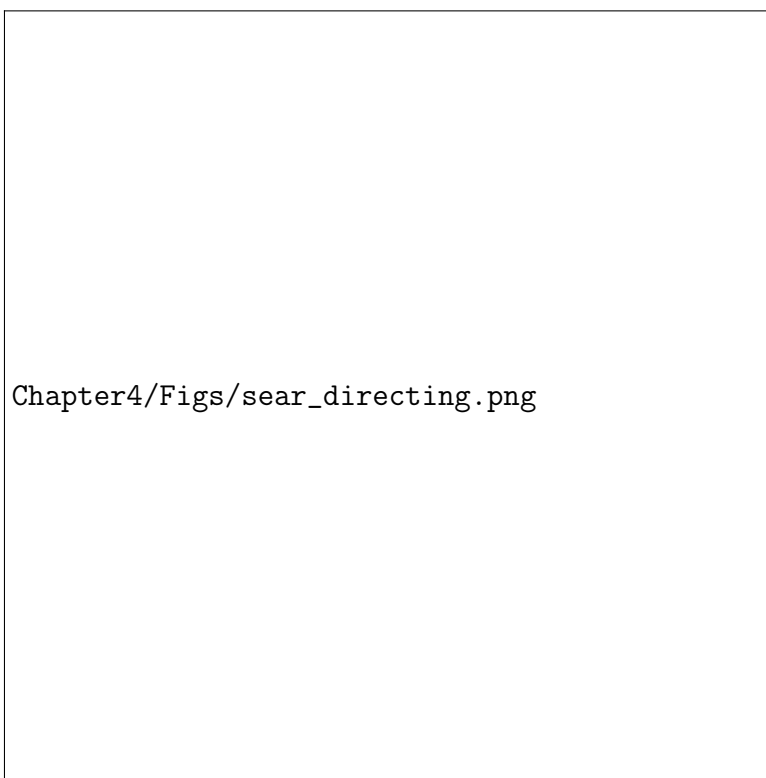


Fig. 2.9 Counting the number Friedel-Crafts acylation reactions in the training sets with ortho, meta and para selectivities reveals an alarming bias – the number of para reactions far outweigh those of meta or ortho reactions.

ring is only 7 and 23 for the two datasets, which are extremely small numbers compared to those for a single para directing substituent which are 347 and 609, a ratio of 25-50 times.

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

2.3.3 Selective reduction of aldehydes and ketones

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

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

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

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

To prove that one model was able to identify the importance of the cation and the other was not the IG attributions were generated and are shown in Fig ??.

Here positive attributions favour the selective reduction, and the negative attributions favour the correct product. It is

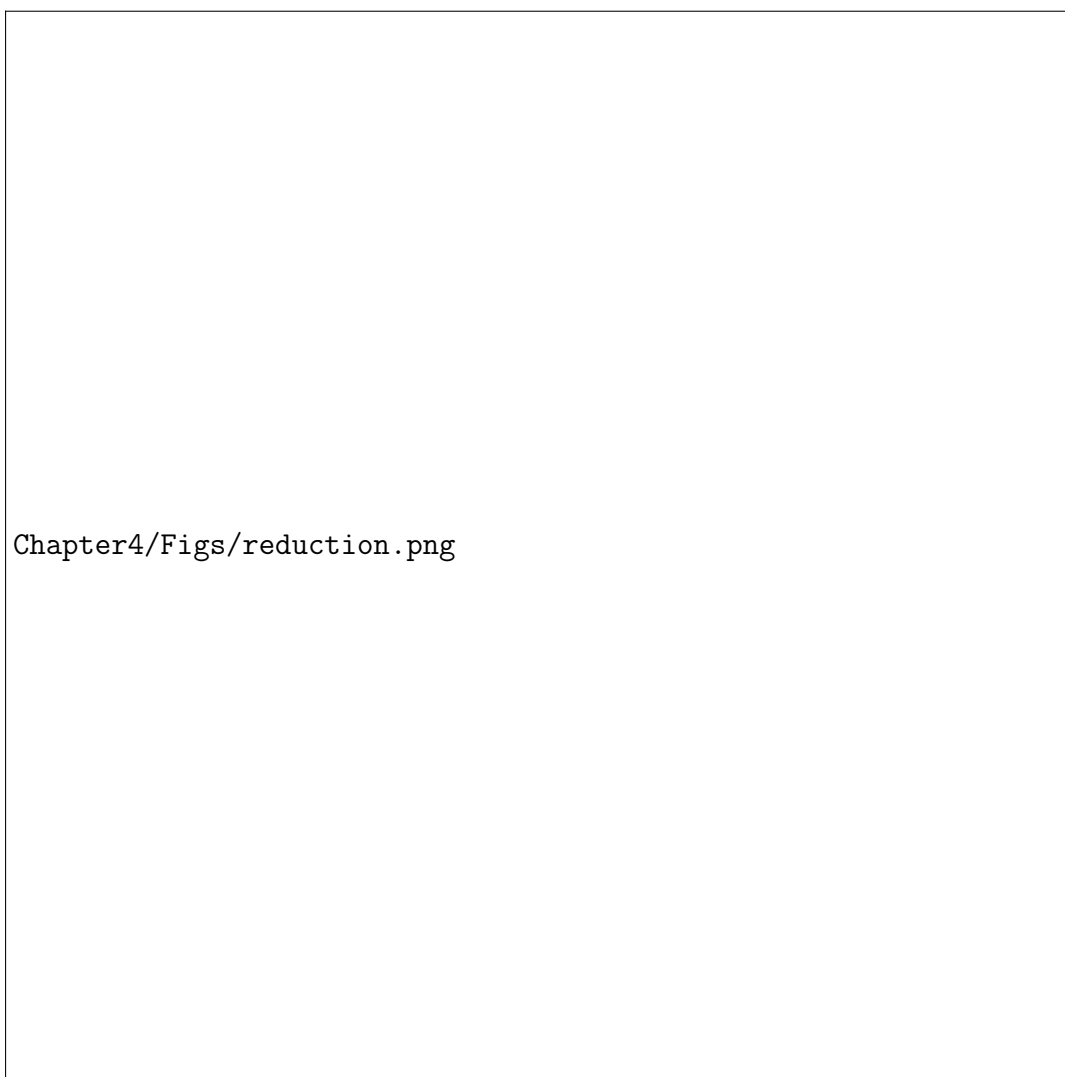


Fig. 2.10 An example of a reaction showing how NaBH_4 reduces the aldehydes selectively in the presence of an ester.



Fig. 2.11 Data attribution to the reaction in Fig ?? suggests that both models have correctly learnt the selectivity of reduction.

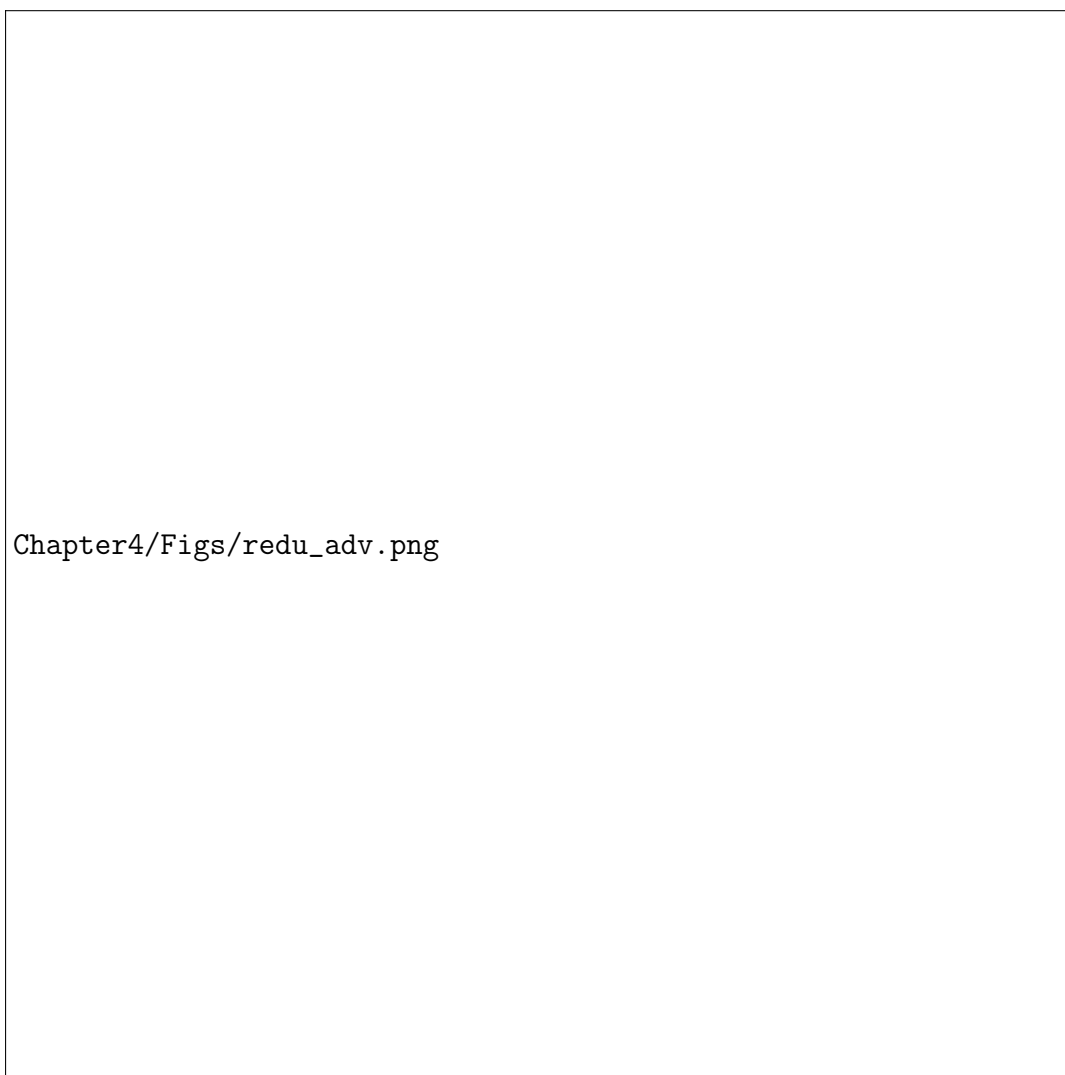


Fig. 2.12 Adversarial example for the borohydride reduction in Fig ?? where the Na⁺ ion was replaced by Li⁺ ion. The USPTO model continues predicting the selective reduction of the aldehyde wrongly whereas the Pistachio model predicts the correct product – the understanding of the models is reflecting in the IG attribution on the Li⁺ ion.

immediately obvious from the attributions that the USPTO model did not take into account the Li^+ ion as it was given an attribution score that is an order of magnitude smaller than the uniform attribution value.

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

2.3.4 Exploring the model with artificial data

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

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

Table 2.1 Number of reactions in the synthetic reduction datasets

Subset	Aldehyde	Ester	Both
NaBH_4	20 000	0	10 000
DIBAL	0	20 000	10 000
LiAlH_4	10 000	10 000	10 000

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



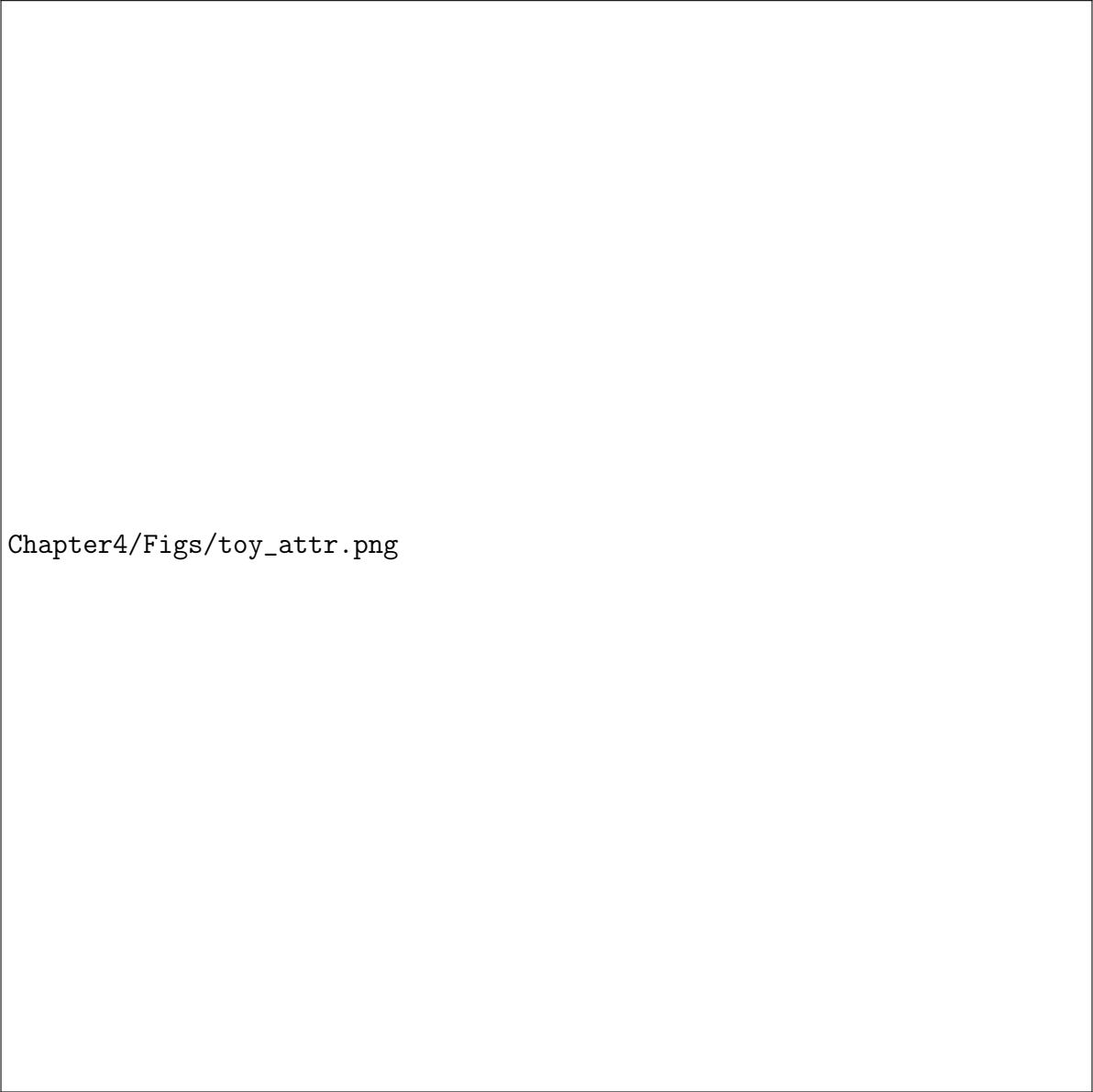
Fig. 2.13 Three uniquely selective reduction templates are chosen to challenge the transformer’s ability to learn selectivities if given enough data.

to learn selective transformations. To investigate how these findings are reflected in the IG 385
attributions and verify that they are able to capture these empirical observations we generated 386
the attributions for the reaction shown in Fig ??.

From the attributions we see that the model rightly gives high attribution to the Al^{3+} ion and 388
small attribution to Li^+ . This suggests that the model has learnt that it should only attend to one 389
of the two tokens because they always appear together in the reactions. The H^- ions get a low 390
attribution as expected. The attributions also verify that the model is finding the backbone part 391
of the input very important, as reconstruction of the backbone is the most challenging aspect of 392
making a correct prediction for this dataset of diverse backbones but restricted chemistry. 393

In the second more challenging synthetic experiment we investigated the models ability to 394
learn the selectivity of aromatic electrophilic substitution reactions. By constructing balanced 395
and biased datasets of Friedel-Crafts acylation reactions we hope to recover the observed 396
behaviour in Sec. ??.

In the balanced and biased datasets we chose 10 para and 10 meta 397
directing substituents which were placed on a benzene ring. For the last ‘super-biased’ dataset, 398
we only included three of the 10 meta directing substituents. These made the initial set of 20 399
reactant molecules which were reacted with a set of acyl chlorides generated by enumerating 400
all possible straight carbon chains up to 8 carbons with a maximum of one double bond. This 401



Chapter4/Figs/toy_attr.png

Fig. 2.14 The model trained on artificial reduction data correctly attributes importance to the Al^{3+} ion.

way we obtained 310 acyl chlorides that were reacted with the substituted benzenes to yield the meta or the para product. We use small carbon chains rather than the ZINC scaffolds to facilitate the learning of the backbones compared to learning the chemistry, which is partly why the dataset size was reduced as well and SMILES augmentation was applied. A summary of the training sets is shown in Table ??.

Table 2.2 Number of reactions in the synthetic Friedel-Crafts training datasets

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

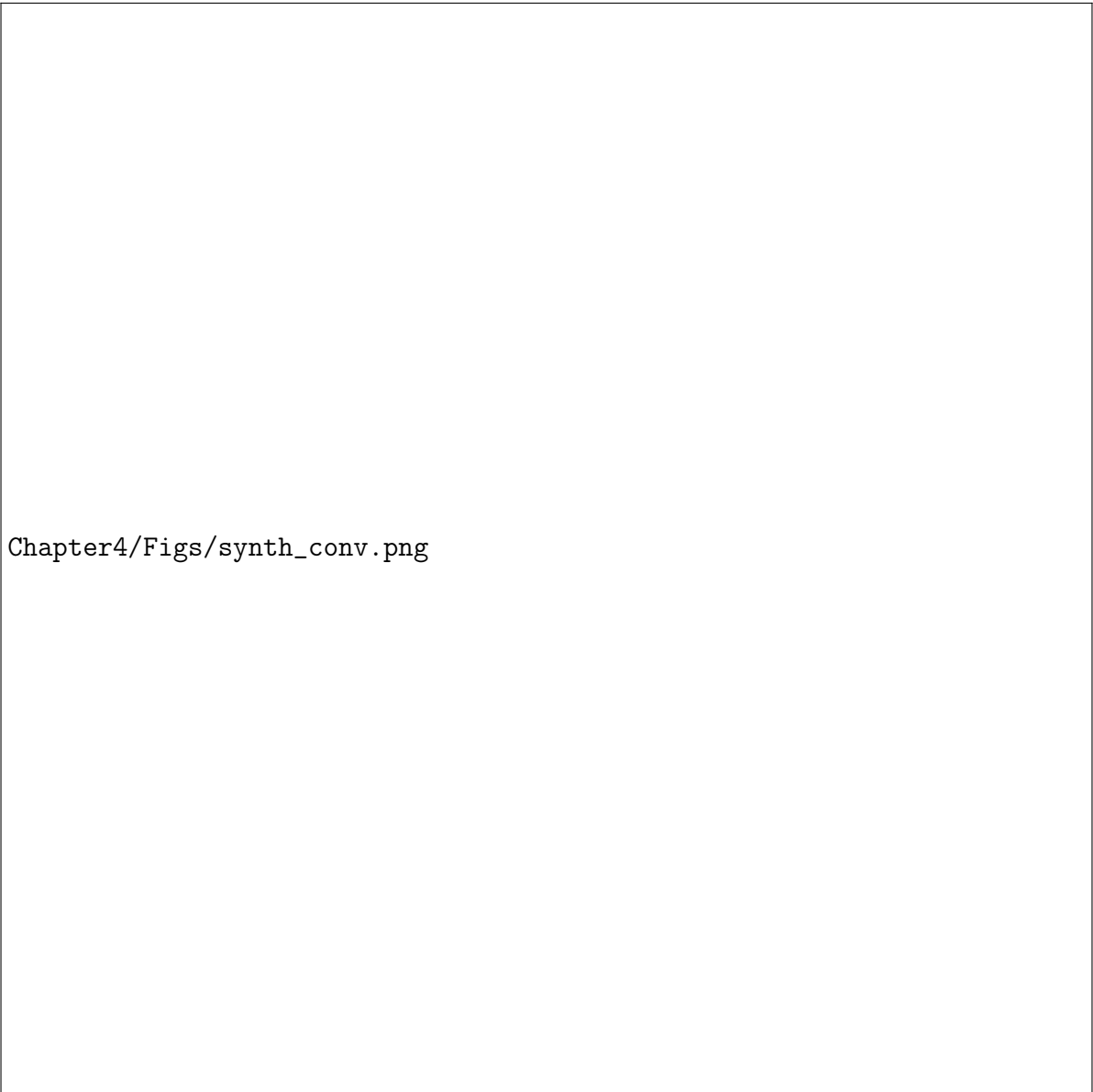
The USPTO and Pistachio transformers were trained for ~ 300 and ~ 100 epochs respectively, so this is the regime we wanted to investigate. We trained 10 transformer models on each of the datasets and saved checkpoints regularly. We created a test set using three meta directing and three para directing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

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

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

2.3.5 Outlook

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



Chapter4/Figs/synth_conv.png

Fig. 2.15 Dataset bias is reflected in the model predictions. The figure shows the proportion of para (solid line) and meta (dashed line) predictions on a balanced test set as a function of the number of training epochs for different biased training sets.

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

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

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

Using evidence from these attribution methods, we hypothesized that many of the erroneous predictions of the Molecular Transformer model stem from data biases. We have validated this hypothesis by designing an artificial dataset of Friedel-Crafts acylation reactions where we could show how biases in the dataset manifest in the predictions of the model. In addition, we observe that the model trained on the Pistachio dataset had in general better predictions and much better calibrated uncertainty scores, in spite of the fact that this model only achieved 76% test-set accuracy. This suggests that Pistachio is not as biased as USPTO, and hints that the addition of training data can substantially improve model performance.

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

Overall, from this work we believe that improvements to the training data can be just as 466
impactful as improving the machine learning models themselves. By demonstrating the power 467
of interpretability methods when rigorously applied to scientific questions, we have shown that 468
these methods can be useful beyond just giving explanations of predictions by exposing dataset 469
biases. Applying our approach for data and input interpretation beyond chemical reaction 470
prediction to other fields will likely be equally constructive, illuminating the path to improved 471
training data and hence improved artificial intelligence models. 472

Chapter 3

473

Design - QSAR with SOAP-GP

474

3.1 Introduction

475

Challenging though synthesis planning may be, at least it is fairly well-established and straight-forward – the difficulty in designing drug candidates, on the other hand, continues to be the scientific bottleneck in drug discovery. A key question is how to predict physical properties or bioactivity from molecular structure — quantitative structure–activity relationships (QSAR) modelling. While statistical and machine learning (ML) approaches have been developed for QSAR since the 1970s, significant amount of innovation has occurred in the space of models – from classical ML methods such as random forest and support vector machine, to the latest technologies based on deep neural networks. Central to any machine learning methodology is the way molecules are described within the model. Most methodologies to date treat a molecule as a 2D object – a graph where the atoms are nodes and the bonds are edges. This leads to the extended connectivity fingerprint (ECFP) [?] and more recent advances that extract the best possible representation of a 2D molecule graph using an end-to-end differentiable framework [?].

Nonetheless, the physical mechanism that underlies biological activity is favourable interactions between local regions on the 3D surface of a molecule (pharmacophores) and residues in the receptor binding site. As such, one would expect that the 3D shape of the molecule would be a more appropriate input. Approaches that attempt to capture this such as Comparative Molecular Field Analysis [?] and Rapid Overlay of Chemical Structures [?] have been developed in the literature. However, those methods either require manual alignment [?] – introducing bias – or consider the similarity between the shapes of entire molecules [?], overlooking the fact that it is often specific regions of the molecule that drive binding or determine physicochemical properties. Overcoming this limitation, one can coarse-grain the molecule into sites of salient interactions [?], but this requires prior insights on what are the important

molecule-receptor interactions. Focusing on locality, Axen et al. [?] developed an approach inspired by extended connectivity fingerprints, where the local 3D environments around each atom are mapped into a fixed length vector via hashing. However, this representation is lossy, and the process of folding into a fix length representation leads to different 3D environments being indistinguishable. (need to deal with this)

The problem of representing and comparing local atomic environments has received significant attention in the field of interpolating potential energy surfaces. Recent pioneering studies proposed a powerful mathematical formalism – Smooth Overlap of Atomic Positions (SOAP) [?]. The key idea is to first represent each atomic environment using a sum of Gaussian densities, then ensure rotational invariance by integrating over all rotations (analytically tractable using the mathematics of spherical harmonics), and finally compute molecular similarity between two molecules by the similarity between the atomic environments that are the most similar. SOAP has found success in cracking challenging problems in materials science such as the phase behavior and defect structure of carbon [?], boron [?] and silicon [?]. Although SOAP has become the workhorse in computational physics, it has not been extensively tested and deployed in cheminformatics. Some examples of ligand-based virtual screening using SOAP [?] have been reported, but a systematic exploration of the utility of SOAP as a general purpose 3D QSAR method has thus far not been done.

In this chapter, we will first describe an ML model utilising SOAP descriptors (SOAP-GP) and show that it can comfortably compete and outperform traditional fingerprint-based approaches as well as state-of-the-art graph neural networks on predicting binding affinity, even in challenging scaffold splits which address dataset bias [? ?]. That being said, we emphasize that we are not advocating for SOAP-GP to become the new paradigm in cheminformatics QSAR. We instead argue that one should exploit the richness of chemical featurisation by combining ML models using different molecular representations rather than relying on any one particular method. We justify this claim by demonstrating that an ensemble of models with different representations outperform equivalently-sized ensembles of the same model/representation.

The work in this chapter is a continuation of my MSci research project from the year before. A paper based on prior results had been submitted to the Journal of Medicinal Chemistry, and will shortly be resubmitted after taking into account comments from reviewers. All code used in this chapter can be found in the GitHub repo soapgp [?].

3.2 Methods

3.2.1 SOAP-GP

In the SOAP framework [?], the local atomic environment of an atom \mathbf{x} is represented by the sum of element-specific Gaussian densities centered on the positions of neighbourhood atoms. The “similarity” between two atomic environments is given by $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \cdot \mathbf{x}_j$, where the similarity function k between environments represents the overlap of these neighbourhood densities integrated over all rotations (normalized so that self-similarity is unity). Using spherical harmonics and radial basis functions, this integral can be analytically computed as a truncated sum of coefficients – the vector of these coefficients forms the descriptor \mathbf{x} . This construction is invariant to rotations, translations, and permutations, and thus alignment-free. Please provide more details about the loss of resolution due to integrating over all rotations. This means that the per-atom-environment comparisons could be nonphysical in the sense that one maximal “alignment” is not possible in the context of another.

To find the geometric similarity between two molecules A and B (Fig ??), the local similarities of the best possible pairing of the atomic environments in A and B are used:

$$K(A, B) = \sum_{\substack{i \in A \\ j \in B}} P_{ij} k(\mathbf{x}_i, \mathbf{x}_j) \quad (3.1)$$

where P_{ij} is $i - j$ th element of the permutation matrix P that maximises K . This can be expressed as an optimal assignment problem and computed efficiently using an entropy-regularization approach, and is known as the “REMatch” similarity kernel [?]. The ‘distance’ between A and B , which can be understood to be a measure of the geometric difference between these two molecules, can then be easily evaluated as

$$d(A, B) = \sqrt{2 - 2K(A, B)} \quad (3.2)$$

The SOAP framework thus allows us to represent differences in three-dimensional molecular shape between individual molecules in a principled, alignment-free manner as a singular metric. Since the binding of ligands to proteins strongly depends on the three-dimensional interactions between the ligand and the receptor binding site, there is reason to expect that such a precise measure of molecular shape could act as an informative descriptor for predicting bioactivity.

This method differs in approach from conventional QSAR methods in that no explicit chemistry descriptors (eg bonding, hybridization, aromaticity) are used at all in the featurisation of a molecule. Instead, chemical information is implicitly learned from the conformational shape of the molecule, from the coordinates of the atoms relative to one another, and completely encoded in the form of a numerical distance metric.

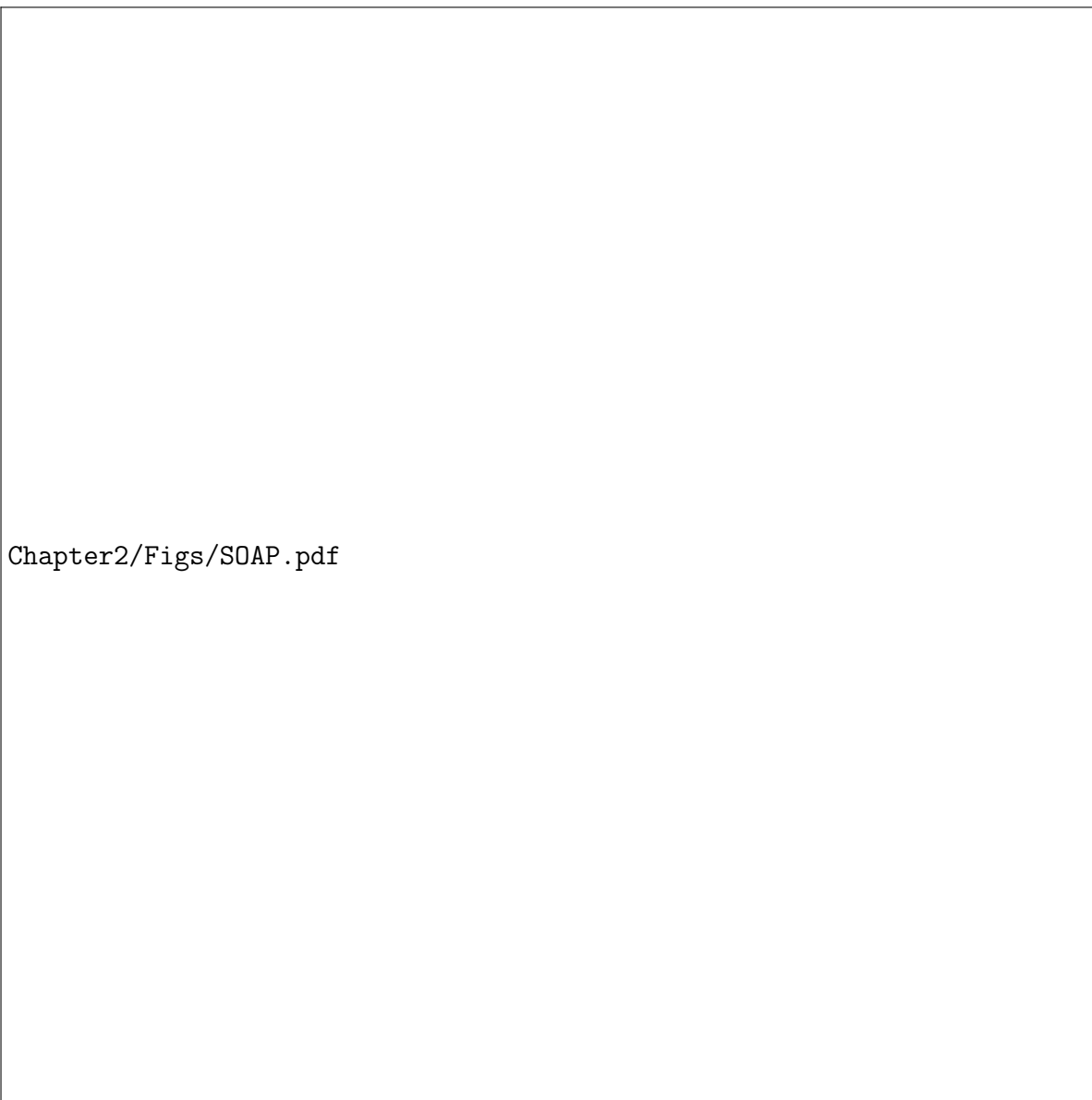


Fig. 3.1 An illustration of how molecular similarity is defined by permuting and maximising the similarity of atomic environments between molecules.

Such an approach naturally lends itself into the framework of kernel-based machine learning methods. Well-known example of such methods are Gaussian Process Regression [?] (GP) models, which we implement using the Matérn kernel. We refer to GP models utilizing SOAP features as SOAP-GP (Fig ??).

GP Regression is a Bayesian ML method which searches over a probabilistic distribution of functions of functions which could model the data. The kernel $K(X_i, X_j)$ between data points is used as the covariance of the prior distribution over functions, and the training data is

used to construct a likelihood. With Bayes’ theorem this defines a posterior distribution for prediction. The model is trained by optimizing the kernel hyperparameters in order to maximise the marginal likelihood of the distribution of functions which model the data.

To incorporate smoothness and differentiability into the GP kernel in order assist in the learning of the model, we augment the REMatch distance $d(A,B)$ (Eq. ??) with the $\nu = \frac{3}{2}$ Matérn kernel $K^\dagger(A,B)$:

$$K^\dagger(A,B) = \sigma^2 \left(1 + \frac{\sqrt{3}d}{\rho} \right) \exp \left(- \frac{\sqrt{3}d}{\rho} \right) \quad (3.3)$$

where σ and ρ are the kernel hyperparameters (both initialised at unity) to be optimized.

Computationally, SMILES strings of molecules are converted into (x,y,z) atomic coordinates using the ETKDG conformer generation method (cite) implemented in RDKit [?]. Only one conformer is generated – how one would incorporate multiple conformers and the fact that only one conformer is needed to achieve good predictive performance is discussed later. SOAP descriptors and kernels were computed from the resultant atomic coordinates using the soapxx and dscribe packages [? ?], with the basis function parameters $n_{max} = 12, l_{max} = 8$. Two sets of SOAP descriptors with $r_{cut} = 3.0\text{\AA}$, $\sigma = 0.2\text{\AA}$ and $r_{cut} = 6.0\text{\AA}$, $\sigma = 0.4\text{\AA}$ were evaluated and concatenated for each molecule. For the REMatch kernel, the entropy regularization parameter α was set to 0.5 with a convergence threshold of 10^{-6} . The GP model itself was implemented using GPFLOW.

3.2.2 Comparative models

The performance of our model was compared directly with that of several others which use representations of differing dimensionality and complexity. The intention of this exercise is not to establish an authoritative benchmark of QSAR model architectures by any means, merely as an illustration of how SOAP-GP compares to representative example models which utilise particular molecular featurisations. Indeed, SOAP itself is merely one example of the many ways in which those in the field of materials science seek to precisely describe atomic and crystal structures (eg atom-centered symmetry functions [? ?], many-body tensor representations [?]).

The industry standard approach for representing molecules is to use the extended connectivity fingerprint (ECFP), which considers molecules as 2D graphs and encodes the topological structural features of a molecule into a fixed-length binary bit string. ECFPs are a popular similarity search tool in drug discovery as the distance between two molecules can be simply



Chapter2/Figs/SOAPGP-horizontal.pdf

Fig. 3.2 An overview of the SOAP-GP model implementation.

defined as the Tanimoto distance between the bit strings [?]. We implement ECFPs in a random forest model (ECFP-RF), which is an established benchmark model for QSAR tasks.

An extension of ECFPs is to consider molecules as 3D structures instead of molecular graphs, which leads to the extended three-dimensional fingerprint (E3FP) [?]. The logic behind such an approach is that the 3D fingerprints are better able to encode stereochemistry and include information on the relationship between atoms close in space but distant in bond connectivity. The E3FP approach only considers radial distances between atoms in its featurisation, while

SOAP features also encode angular information. Just like ECFPs, the similarity between two molecules can be calculated using the Tanimoto distance between their E3FP fingerprints.

For E3FPs there is no standard model implementation - both random forest and Gaussian Process models were attempted and the GP models on average performed better so we from hereon utilise E3FPs in a GP framework (E3FP-GP) with the Matérn kernel in an identical fashion as the SOAP-GP except that the Tanimoto distances between molecular fingerprints are used in place of the SOAP REmatch distance. The difference performance can be isolated to the quality of the molecular distance measures – this would illustrate the importance of including angular information in featurising molecular shape.

Last but not least, we also consider the Directed Message Passing Neural Network (DMPNN) model [?], a state-of-the-art graph neural network which uses 2D molecular graphs explicitly encoding atomic and bond properties such as formal charge and conjugation as input features, usually as one-hot vectors. In graph neural networks, atom and bond features are combined with those of their neighbours via message-passing and convolutional embedding to construct a learnt global descriptor of a molecule, which is then passed through a neural network for property prediction. Graph neural networks have been gaining popularity in the cheminformatics community for property prediction [? ?], and most recently the DMPNN model was utilised in a successful landmark deep learning search for novel antibiotics [?].

ECFP fingerprints were generated with a radius of 3 and 1024 bits using RDKit, while E3FP fingerprints also with 1024 bits were generated using the e3fp package [?]. The DMPNN model was implemented using the chemprop package. The training procedure regarding the molecular features used as well as the initial hyperparameter optimization was done following the guidelines from [?].

3.2.3 Datasets

We used IC50 datasets for 24 diverse protein targets extracted from ChEMBL which have been previously investigated in several screening and modelling studies [? ?]. IC50 measures the concentration of a compound required for the inhibition of a target to drop by 50% - the IC50 (or pIC50) values are a direct metric of ligand-protein binding affinity, and modelling these values are thus a suitable challenge for comparing QSAR models. We further filter the datasets as we found that they contained many large compounds such as glycans which are beyond the scope of small molecule drug discovery. We only keep molecules with molecular weight below 500 daltons (as per Lipinski's rules) which reduces the datasets by 19% on average (Table ??). experimental error

Table 3.1 ChEMBL IC50 bioactivity data used in this study

Target	Initial Length	Post-Filter Length
A2a	203	166
ABL1	773	536
Acetylcholinesterase	3159	2491
Aurora-A	2125	1612
B-raf	1730	824
Cannabinoid	1116	820
Carbonic	603	556
Caspase	1606	1362
Coagulation	1700	862
COX-1	1343	1278
COX-2	2855	2704
Dihydrofolate	584	548
Dopamine	479	405
Ephrin	1740	1716
erbB1	4868	3598
Estrogen	1705	1546
Glucocorticoid	1447	1077
Glycogen	1757	1655
HERG	5207	4042
JAK2	2655	2252
LCK	1352	954
Monoamine	1379	1344
opioid	840	611
Vanilloid	1923	1656

For SOAP-GP, ECFP-RF, and E3FP-GP, the datasets are split 80/20 into train/test sets and
for the DMPNN models the split is 70/10/20 for train/validation/test sets. The random split
results are given as the mean results from 3 runs each with 5-fold cross validation.

Besides random splitting, we also evaluate on these datasets using scaffold split, where
molecules are binned by Murcko scaffold (evaluated using RDKit). This method of splitting
better simulates the real-life drug discovery cycle where prior activity data only exists for a
class of chemical compounds that are different from those that are being evaluated, in other
words posing a greater extrapolation challenge. Bins larger than half of the required test set
size are placed in the training/validation set and all remaining bins are distributed randomly
such that the required train/test split sizes are met. The scaffold split results are given as the
mean results from 15 runs using different random seeds for the distribution of scaffolds.

Table 3.2 pIC50 Random Split RMSE Results - the lowest RMSE for each dataset are bolded.

Dataset	Random Split			
	ECFP-RF	E3FP-GP	DMPNN	SOAP-GP
A2a	0.839 ± 0.030	0.793 ± 0.034	0.993 ± 0.062	0.924 ± 0.064
ABL1	0.848 ± 0.018	0.843 ± 0.019	0.965 ± 0.030	0.798 ± 0.017
AChE	0.784 ± 0.006	0.868 ± 0.007	0.783 ± 0.011	0.761 ± 0.009
Aurora-A	0.830 ± 0.010	0.900 ± 0.008	0.842 ± 0.008	0.844 ± 0.009
B-raf	0.712 ± 0.008	0.786 ± 0.008	0.778 ± 0.010	0.720 ± 0.008
Cannabinoid	0.747 ± 0.015	0.800 ± 0.011	0.845 ± 0.019	0.716 ± 0.011
Carbonic	0.659 ± 0.016	0.670 ± 0.013	0.702 ± 0.023	0.839 ± 0.095
Caspase	0.587 ± 0.008	0.662 ± 0.009	0.597 ± 0.012	1.096 ± 0.061
Coagulation	0.909 ± 0.010	1.010 ± 0.009	1.019 ± 0.022	0.984 ± 0.037
COX-1	0.729 ± 0.013	0.744 ± 0.013	0.732 ± 0.011	0.706 ± 0.013
COX-2	0.790 ± 0.007	0.826 ± 0.007	0.804 ± 0.012	0.762 ± 0.007
Dihydrofolate	0.799 ± 0.025	0.849 ± 0.019	0.890 ± 0.023	0.811 ± 0.021
Dopamine	0.747 ± 0.013	0.816 ± 0.014	0.921 ± 0.020	0.777 ± 0.017
Ephrin	0.722 ± 0.011	0.749 ± 0.007	0.719 ± 0.009	0.701 ± 0.008
erbB1	0.756 ± 0.003	0.818 ± 0.005	0.748 ± 0.010	0.772 ± 0.003
Estrogen	0.691 ± 0.005	0.697 ± 0.005	0.670 ± 0.007	0.633 ± 0.007
Glucocorticoid	0.612 ± 0.010	0.663 ± 0.008	0.691 ± 0.008	0.613 ± 0.007
Glycogen	0.743 ± 0.007	0.788 ± 0.008	0.806 ± 0.009	0.769 ± 0.006
HERG	0.610 ± 0.006	0.679 ± 0.005	0.615 ± 0.007	0.569 ± 0.005
JAK2	0.672 ± 0.007	0.737 ± 0.007	0.719 ± 0.007	0.683 ± 0.009
LCK	0.829 ± 0.010	0.867 ± 0.012	0.918 ± 0.021	0.827 ± 0.010
Monoamine	0.676 ± 0.007	0.680 ± 0.008	0.724 ± 0.012	0.680 ± 0.009
opioid	0.729 ± 0.011	0.781 ± 0.018	0.748 ± 0.020	0.692 ± 0.015
Vanilloid	0.724 ± 0.006	0.774 ± 0.006	0.744 ± 0.008	0.720 ± 0.006

Dataset	Scaffold Split			
	ECFP-RF	E3FP-GP	DMPNN	SOAP-GP
A2a	1.113 ± 0.087	1.134 ± 0.128	1.434 ± 0.194	1.028 ± 0.065
ABL1	0.933 ± 0.036	0.971 ± 0.047	1.069 ± 0.043	0.951 ± 0.050
AChE	0.990 ± 0.023	1.045 ± 0.025	0.994 ± 0.030	0.952 ± 0.022
Aurora-A	0.928 ± 0.025	1.011 ± 0.021	0.953 ± 0.029	0.942 ± 0.017
B-raf	0.866 ± 0.038	0.916 ± 0.038	0.959 ± 0.032	0.841 ± 0.035
Cannabinoid	0.874 ± 0.026	0.943 ± 0.028	0.967 ± 0.027	0.827 ± 0.022
Carbonic	0.682 ± 0.032	0.816 ± 0.044	0.809 ± 0.060	0.689 ± 0.049
Caspase	0.721 ± 0.040	0.764 ± 0.027	0.770 ± 0.025	0.922 ± 0.063
Coagulation	0.996 ± 0.014	1.076 ± 0.023	1.100 ± 0.025	0.989 ± 0.025
COX-1	0.793 ± 0.017	0.789 ± 0.014	0.768 ± 0.008	0.781 ± 0.009
COX-2	1.008 ± 0.039	1.009 ± 0.031	1.010 ± 0.037	0.960 ± 0.033
Dihydrofolate	0.914 ± 0.058	0.938 ± 0.051	1.012 ± 0.044	0.967 ± 0.057
Dopamine	0.869 ± 0.020	0.882 ± 0.018	0.940 ± 0.020	0.894 ± 0.020
Ephrin	0.881 ± 0.018	0.908 ± 0.028	0.904 ± 0.025	0.882 ± 0.021
erbB1	0.888 ± 0.013	0.947 ± 0.012	0.864 ± 0.010	0.891 ± 0.007
Estrogen	0.795 ± 0.018	0.786 ± 0.015	0.744 ± 0.014	0.708 ± 0.011
Glucocorticoid	0.742 ± 0.023	0.790 ± 0.024	0.859 ± 0.022	0.738 ± 0.014
Glycogen	0.869 ± 0.021	0.910 ± 0.022	0.963 ± 0.022	0.906 ± 0.020
HERG	0.690 ± 0.018	0.747 ± 0.021	0.706 ± 0.023	0.656 ± 0.018
JAK2	0.746 ± 0.010	0.803 ± 0.013	0.783 ± 0.019	0.738 ± 0.021
LCK	0.909 ± 0.014	0.954 ± 0.018	1.056 ± 0.030	0.918 ± 0.012
Monoamine	0.818 ± 0.022	0.813 ± 0.023	0.927 ± 0.030	0.817 ± 0.023
opioid	0.781 ± 0.032	0.797 ± 0.031	0.811 ± 0.028	0.747 ± 0.021
Vanilloid	0.770 ± 0.018	0.814 ± 0.018	0.826 ± 0.026	0.762 ± 0.015

3.3 Results

3.3.1 Performance Comparison

The above models are compared by evaluating the root-mean-square errors (RMSE) of their predictions on the same train/test splits of the datasets.

With random splitting (Table ??), both ECFP-RF and SOAP-GP each outperform the others on 10 out of the 24 tasks, while DMPNN only does so on 3, E3FP-GP on one. A similar picture is seen under scaffold splitting where SOAP-GP does best on 12 of the 24 tasks, 9 for ECFP-RF, and only two for DMPNN and one for E3FP-GP. Overall both the RMSEs and standard deviations are higher for the scaffold split results because of the increased challenge from scaffold splitting.

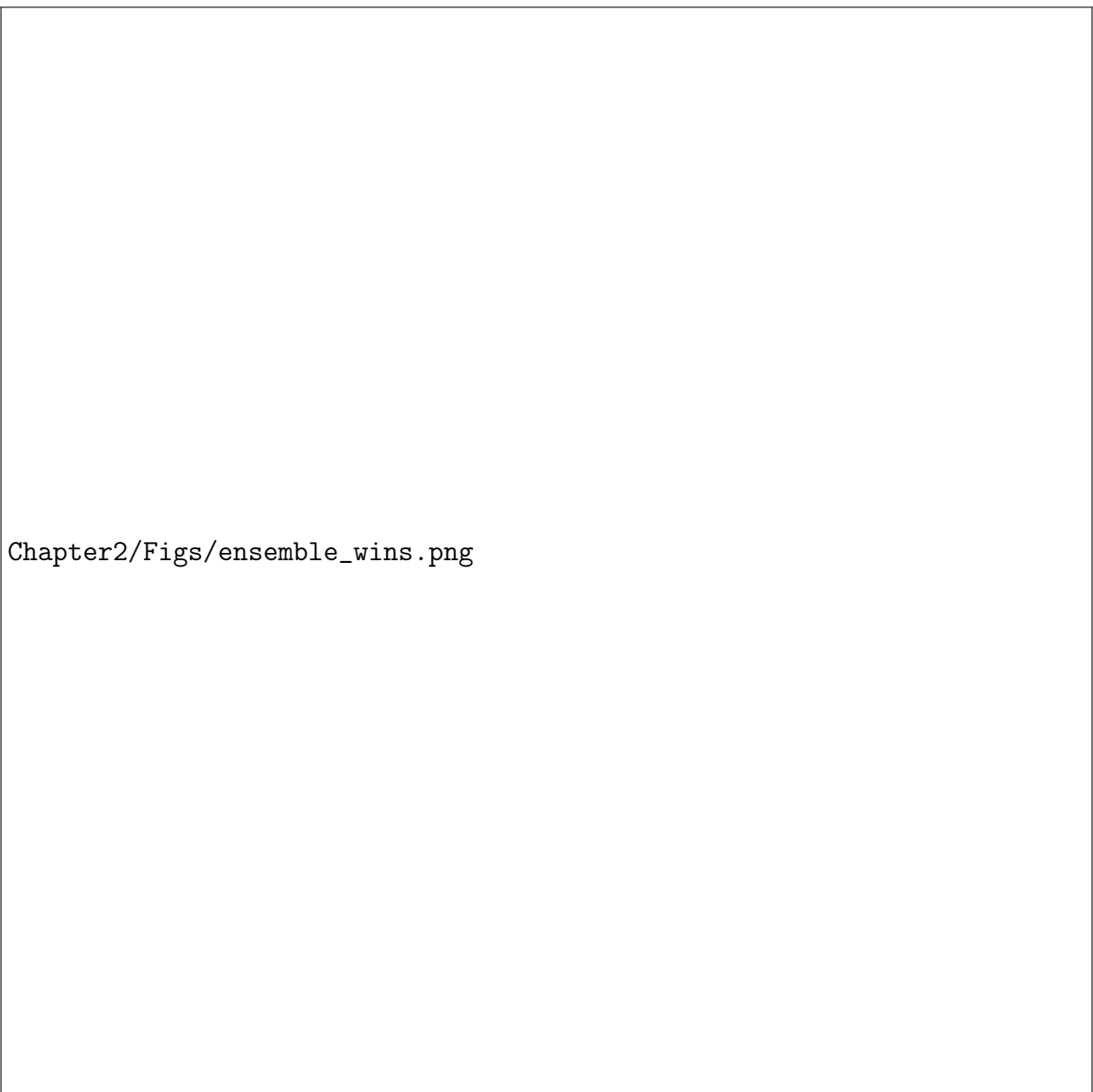
These results show that SOAP-GP, utilising out of the box open-source molecular descriptors of three-dimensional molecular shape, is competitive with both conventional and current state-of-the-art ML QSAR models. In particular, comparing SOAP-GP against E3FP-GP which is often the worst-performing out of the four models suggests that merely accounting for radial distances is an insufficiently informative description of shape. The additional angular information contained within the SOAP descriptors allows the SOAP-GP to far better model binding affinity.

3.3.2 Ensembling Representations

Although we have showcased the competitive capabilities of SOAP-GP, we do not suggest that SOAP-GP should become the new paradigm in cheminformatics QSAR, nor indeed that any sole representation/model should be. From our dataset of 24 targets alone we can already observe that model performance can vary substantially and that it is hard to know a priori which model would do best. We argue that this is the case generally in bioactivity prediction and it is unhelpful to solely rely on any particular ‘state of the art’ model or molecular featurisation for QSAR modelling.

Our argument is specifically for the learning of accurate ligand-binding affinity. When conducting a vast virtual screen of a molecular library or attempting to classify activity from a large high-throughput-screen, practical considerations such as time constraints and the memory scaling behaviour of models should dictate the choice of model. However, exact IC50-accuracy activity data essentially never exceeds $N = 10^4$ in size [?], and in this regime computational constraints are typically minor compared to demands on predictive performance.

In this scenario, a straightforward way to achieve improved performance is to combine QSAR models in an ensemble learning approach where the predictions from several models



Chapter2/Figs/ensemble_wins.png

Fig. 3.3 Ensembling diverse representations is superior to ensembling similar representations regardless of model architecture. Green/red colour intensity indicates the number of tasks out of 24 in which the ensemble on the y-axis has a better/worse RMSE than the ensemble on the x-axis.

are averaged to give better results [?]. Such an approach is only successful if there is sufficient 685
diversity such that each model captures trends in the dataset that are neglected by the others. 686
The power of model ensembling lies not merely via the principle of ‘strength in numbers’, but 687
‘strength in diversity’. 688

Model ensembling in QSAR has not been well-explored, and we propose that as a general principle it is important to include a diversity of representations in the ensemble, as opposed to ensembling different model architectures on the same representation. Unlike the conventional applications of machine learning, chemistry possesses an vast richness of featurisations and we should take advantage of this fact. Models trained on hybridization states and stereochemistry will capture distinct effects from those trained using conformational shapes, and these differences are much more significant than the particular architecture that is employed.

We demonstrate this by comparing the performance of ensembles with diverse representations pairwise against equivalently-sized ensembles that utilise the same features (Fig ??), as well as only single non-ensembled models. It can be seen that ensembling diverse representations almost always outperforms ensembling the same representation, which in turn tends to be better than the single models on their own. These differences are most accentuated in the scaffold split scenario. The best performing ensemble out of the possible combinations is an ensemble of SOAP-GP and ECFP-RF. This is not entirely surprising given that these were the two best-performing single models on their own, but it demonstrates that the trends learnt by the two models complement one another, that combining 2D topological information with precise 3D atomic features can push the predictive frontier of QSAR modelling.

3.4 Outlook

We described SOAP, a novel alignment-free 3D QSAR method which employs a GP model on the intermolecular similarity between local atomic environments featurized using SOAP descriptors. This model was compared with a 2D fingerprint-based random forest model, a 3D fingerprint-based GP, as well as a state-of-the-art graph convolutional neural network, DMPNN, on 24 IC50 regression tasks from ChEMBL. We showed that SOAP-GP is competitive with these on both random and challenging scaffold splits.

The success of SOAP-GP in modelling ligand-protein binding affinity suggests that many other atomic/structural descriptors from the field of materials science, as well as the particular models that utilise those descriptors for the purpose of predicting quantum energies, have the potential to also be useful in QSAR modelling.

That being said, we believe that it is unrealistic to expect any single descriptor or model to be the best for any particular QSAR task. Thus, ensembling different models is the most principled way of conducting QSAR modelling in general. We further argue that ensembling models with diverse representations is fundamentally more important than ensembling diverse model architectures, and demonstrate this by comparing the pairwise performance of equivalently sized ensembles with/without diverse representations. We find that ensembles with diverse

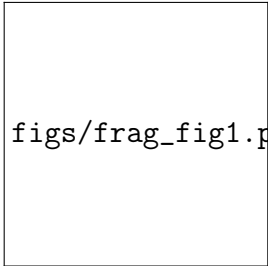
representations outperform those with the same representation, and that SOAP-GP combined with ECFP-RF is by far the best performing model, showcasing the power of combining 2D and 3D features in capturing information relevant to predicting binding affinity.

Although much research is done on continually developing novel and improved QSAR models in a competitive basis, by instead taking a collaborative approach and complementing models with one another to exploit feature-rich chemistry, sizeable performance gains can also be achieved.

3.4 Outlook

47

[journal=jacsat,manuscript=article]achemso graphicx 730
[version=3]mhchem [breaklinks]hyperref 731
aal44@cam.ac.uk [Cavendish]Cavendish Laboratory, University of Cambridge, Cambridge 732
CB3 0HE, United Kingdom [PostEra] PostEra Inc., 2 Embarcadero Center, San Francisco, CA 733
94111, USA 734
FRESCO]Bioactive Ligand Discovery via Unsupervised Learning of Fragment-Protein 735
Complexes 736



figs/frag_fig1.pdf

737

Abstract

738

In this work, we describe an end-to-end hit detection approach that bridges the the paradigms of both fragment-based drug design and virtual screening. Our method, named FRESCO, utilises unsupervised learning to learn pharmacophore distributions directly from experimental 3D fragment-protein structures. The trained model evaluates whether or not a particular compound possesses pharmacophores matching the distribution exhibited by the bound fragments, replicating the intuition of a medicinal chemist deducing spatial correlations between pharmacophores from different fragments. Our approach is computationally validated with a retrospective study on SARS-CoV-2 main protease (Mpro) ligands using data from COVID Moonshot [?], showing high enrichment. Then, we conduct an experimental search for novel hits for Mpro and the Mac1 domain of SARS-CoV-2 non-structural protein 3 (nsp3) by scoring a library of 1.4 billion purchasable compounds from EnamineREAL, resulting in 1 (novel?) hit for MPro and 2 (novel?) hits for Mac-1. (Scaffold exploration of the detected hits hopefully lead to novel potent ligands!) Our results are the first experimentally validated demonstration of hit detection via a fully computational workflow starting directly from an experimental fragment screen.

3.5 Introduction

753

TODO - shorten introduction, put some of it in discussion?

754

Developing a new drug from original idea to the launch of a finished product is a complex process which can take 12–15 years and cost in excess of \$1 billion [?]. A key step in the early stages of the drug discovery process following the identification of a biological target is hit detection. Broadly speaking, a ‘hit’ is a compound that interacts with the identified target sufficiently strong enough to act as a starting point for optimisation of the compound structure towards a candidate drug.

755
756
757
758
759
760

Approaches towards hit detection generally involve the screening of libraries of compounds. For example, in high throughput screening (HTS) often hundreds of thousands of chemical compounds are synthesised and tested, requiring substantial resources as well as complex logistics. While experimental techniques such as DNA-Encoded libraries are being developed

761

762

763

764

to increase the efficiency of large-scale compound screening [?], there has been a growing push towards conducting hit detection computationally instead to decrease the cost and accelerate this step of the drug discovery process [?].

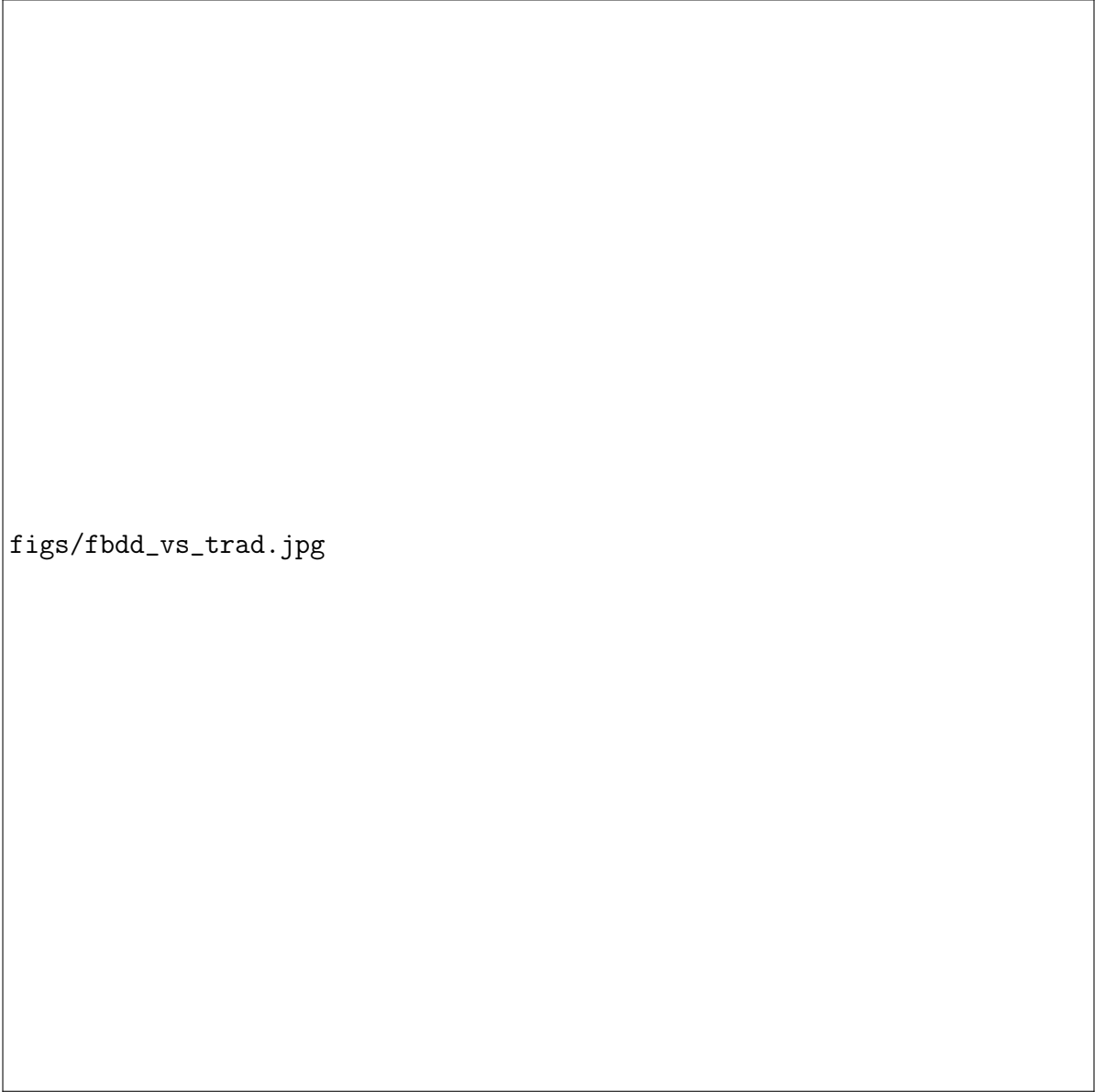
In this approach, known as virtual screening, a computational scoring function is used to estimate the potency of a compound. After computing the scores for all of the compounds in a library, only those ranked highly by the scoring function are chosen for synthesis and experimental validation. Currently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

An alternative to these methodologies is fragment-based drug discovery (FBDD). In this approach, a library of very low molecular weight compounds (‘fragments’ typically less than 18 nonhydrogen atoms [?]) are screened at high concentrations alongside the generation of bound fragment-protein structures via X-ray crystallography or cryo-EM. By obtaining these experimental structures and examining the binding interactions between individual fragments with protein residues, fragments can be used as building blocks for larger molecules by linking or merging together disparate fragments in order to increase potency. Conceptually, FBDD is based on a coarse-graining of fragments to specific moieties or groups that are associated with interactions to the target, with the goal of maintaining and optimising these interactions in larger molecules.

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

In this work, we describe an end-to-end hit detection approach that bridges the the paradigms of both fragment-based drug design and virtual screening. Our method, named FRESCO, utilises unsupervised machine learning to learn pharmacophore distributions directly from experimental 3D fragment-protein structures. The trained model acts as a scoring function that can be used to perform virtual screening, evaluating whether or not a particular compound possesses pharmacophores matching the distribution exhibited by the bound fragments.

This methodology aims to replicate the intuition of a medicinal chemist performing fragment-based drug discovery, abstracting fragments to pharmacophores and deducing spatial correlations between pharmacophores from different fragments. As a matter of fact, we go beyond the typical strategem of growing one individual fragment independently of the others,



figs/fbdd_vs_trad.jpg

Fig. 3.4 An illustration comparing fragment-based drug discovery to traditional approaches.

or merging two particular fragments - by training our model on all of the fragment-protein structures we leverage information from all existing fragments, ensuring no pharmacophore correlations are overlooked in the hit detection process.

We first computationally validate our approach with a retrospective study on bioactivity prediction for SARS-CoV-2 main protease (Mpro) ligands using data from COVID Moonshot [?], showing high enrichment. Then, we conduct an experimental search for novel hits for Mpro and the Mac1 domain of SARS-CoV-2 non-structural protein 3 (nsp3) by performing a virtual screen with FRESKO on a library of 1.4 billion purchasable compounds from EnamineREAL. This resulted in 1 (novel?) hit for MPro and 2 (novel?) hits for Mac-1, with crystallographic poses for the Mac-1 hits. Follow-up compounds for performing scaffold exploration of the detected hits were synthesised and assayed demonstrating credible structure-activity relationships, confirming that the detected hits were genuine.

Our method is unique not only in its philosophy, but also its use of unsupervised learning in the form of kernel density estimation. Although there has been a rapid growth in machine learning methods applied to drug discovery in recent years particularly in QSAR/molecular property prediction, the vast majority of them are supervised learning techniques requiring not merely the existence of experimental assay data, but its existence in sufficient quantity and quality to train a useful model. The nature of the problem of hit detection in early-stage drug discovery is one where such data is nonexistent, which underlies the necessity of having methods for directly performing hit detection from fragment screen data.

The only comparable methods that the authors are aware of is recent work in using deep generative models for proposing merges between two fragments (DeLinker [?], SyntaLinker [?], and Develop [?]). These approaches differ from ours in several ways: Firstly, these models all require human intervention from an expert in choosing which fragments to merge, or what pharmacophoric constraints need to be obeyed, whereas our model is fully end-to-end. Secondly, these methods utilise neural network-based generative models which are very sensitive to training hyperparameter choice in general (citation about mode collapse [?]), and for molecular generation in particular known to propose invalid and/or unsynthesizable molecules [?]. In contrast, our method relies on kernel density estimation which is simple, robust and free from hyperparameter tuning, and we explicitly only screen purchasable, easily-synthesised molecules. Lastly, the proposed models have only been studied computationally and lack validation in the real world.

Our results are the first experimentally validated demonstration of hit detection via a fully computational workflow starting directly from an experimental fragment screen. This work opens the door for bridging fragment-based drug discovery with virtual screening.

3.6 Results

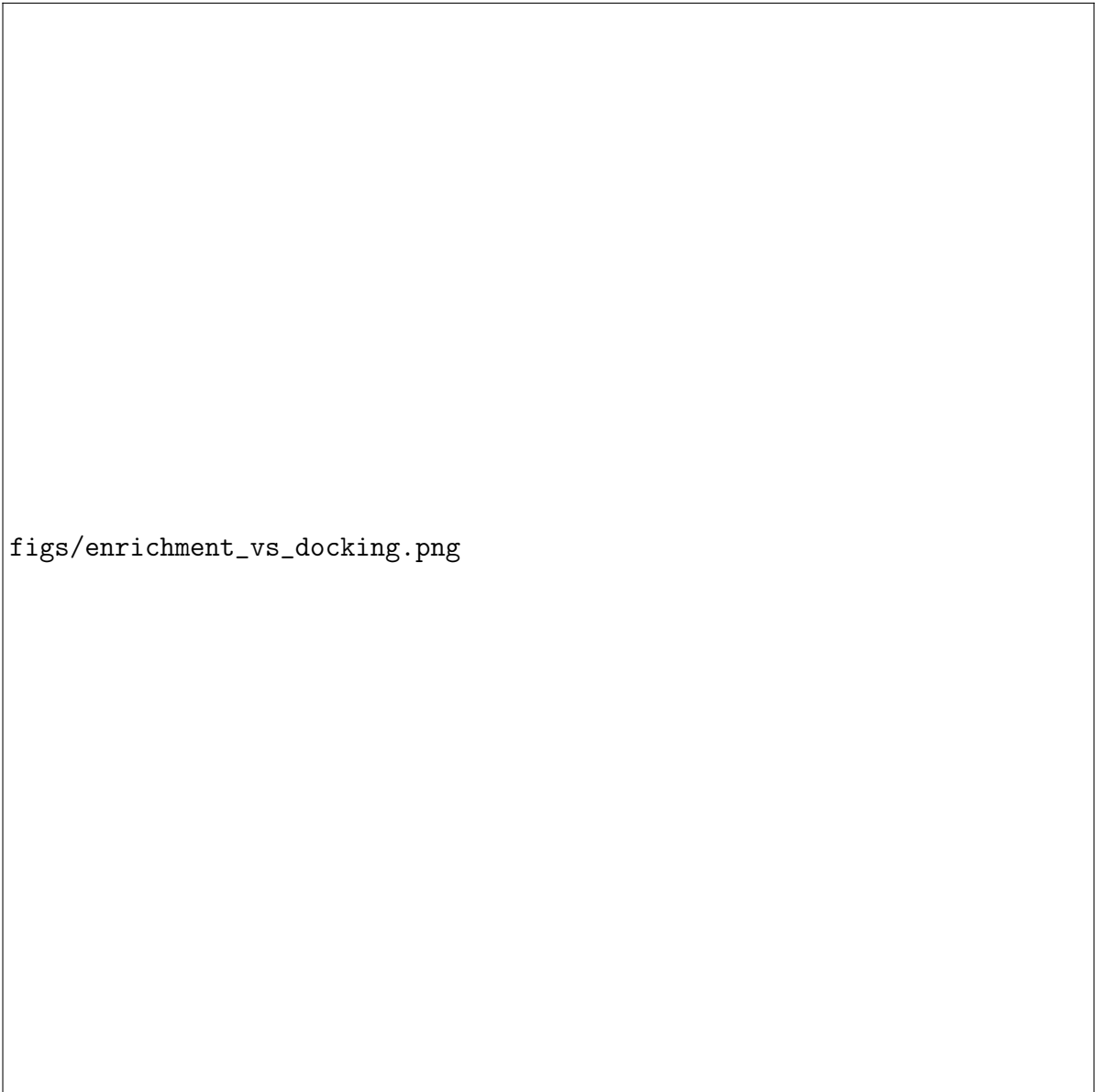
3.6.1 Computational Retrospective Study

To validate that our hypothesised methodology has merit, a computational study was conducted on the SARS-CoV-2 main protease (Mpro). The COVID Moonshot campaign [?] was established as an open science effort towards developing a patent-free antiviral drug for the SARS-CoV-2, specifically targeting the inhibition of Mpro as that would prevent the virus from further replication. Throughout the campaign, activity data consisting of the structures of molecules that were synthesised and assayed was continually released. As a proof-of-concept, the model predictions for the assayed molecules are compared against the measured activity and analysed.

Firstly, the FRESCO model was fit on publicly reported crystallographic structures of non-covalent fragments bound to the SARS-CoV-2 Mpro protein [?]. Next, conformer coordinates for Moonshot compounds reported before March 22nd 2021 were obtained from parallel work within the Moonshot consortium performing docking studies on Mpro (TODO - citation/explanation?). Pharmacophore features were generated from these conformers and the molecules were scored using the model. The enrichment factor for picking compounds above an IC₅₀ of 10 μ M is computed as we are most interested in the ability of this model to sample potent compounds. We also compute the enrichment from ranking the compounds with the Chemgauss4 scoring function.

The results are shown in Fig. ?? illustrating high enrichment, validating the hypothesis that it is possible to correlate bioactivity with unsupervised learning of fragment pharmacophore distributions. The date 21st July 2021 marks when additional experimental data was released and the direction of the COVID Moonshot campaign shifted from hit detection to optimisation of existing lead compounds [?]. The optimisation process relies on the improvement of energetic interactions between the ligand and residues in the active site, likely growing beyond the volume covered by the initial fragment hits and so cannot in principle be captured by FRESCO. This is consistent with the observed decrease in enrichment relative to docking when using data from latter stages of Moonshot where more of the submitted compounds are designed for reaching nanomolar affinity. This shows the value in using FRESCO for hit detection in the early stages of a drug discovery campaign.

TODO - Which enrichment curves to show? Put some in the SI? Should I include ZINC graph?



figs/enrichment_vs_docking.png

Fig. 3.5 FRESCO is able to detect potent compounds purely based on unsupervised learning. High enrichment relative to docking is achieved when performing a retrospective study on activity data COVID Moonshot.

3.6.2 Experimental Prospective Study

After confirming the merit of this approach via a retrospective computational study, a prospective experimental search for novel hits was performed to demonstrate the capability of this methodology. We study two targets: the main protease (Mpro) of SARS-CoV-2 and the Mac1 domain of SARS-CoV-2 non-structural protein 3 (nsp3).

For both targets we followed the same workflow to discover novel hits. We first fit FRESKO models on experimental fragment-protein crystal complexes, and used the models to screen the VirtualFlow [?] library of commercially available compounds. The top predicted compounds were filtered by their physical properties and were clustered by structural similarity. The centroids of the 50 most populous clusters were selected as hit candidates and ordered for synthesis and testing. Details on the methodology can be found in Sec. ??.

Mpro

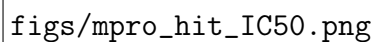
For Mpro, 38 compounds were successfully synthesised and assayed. One of the compounds, WIL-UNI-d4749f31-37, was recorded with an IC₅₀ of 25.8 μM measured via fluorescence assay while the remaining compounds were found to be inactive. To confirm that the compound activity was not a false positive (eg measured potency due to assay interference) and that genuine ligand-protein interactions existed, a follow-up series of 8 compounds (ALP-UNI-ed5cdfd2) consisting of structural perturbations to the molecular scaffold was also synthesised and assayed. All 8 compounds exhibited inhibition at high concentrations and one compound (ALP-UNI-ed5cdfd2-1) had a lower IC₅₀ of 19.4 μM, demonstrating a genuine structure–activity relationship SAR for this hit compound.

Mpro order 1st batch = WIL-UNI-d4749f31, order 2nd batch = WIL-UNI-2a57d06c (no actives).

Mac-1

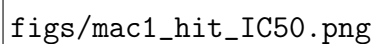
For Mac-1, 52 compounds were successfully synthesised and assayed. Two of the compounds show non-negligible activity at high concentration - at 250 μM, compound Z5551425673 (as a racemic mixture) has an inhibition of 30.1%, while compound Z1102995175 has 24.8%. In addition, crystallographic structures of Z5551425673 (modelled as the S-stereoisomer) bound to the active site was also found alongside that of 3 other compounds (Z2890189003, Z2890182452, Z1423250928), confirming that Z5551425673 is a true hit.

As with Mpro, a follow-up series of compounds were designed to perturb the chemical structure of Z5551425673 in order to confirm the existence of SAR for this compound against Mac-1. 26 compounds were ordered and ... TODO.



figs/mpro_hit_IC50.png

Fig. 3.6 Dose-Response curve for WIL-UNI-d4749f31-37 and follow-up compounds demonstrating SAR relationship.



figs/mac1_hit_IC50.png

Fig. 3.7 Dose-Response curve for Z5551425673 and follow-up compounds demonstrating SAR relationship. Also showing crystal structure TODO.

TODO - table and structures in SI

901

3.7 Discussion

902

TODO - improvements that could be made? Future extensions? Transfer learning between targets? Docked fragments? Ensemble dock scores with FRESCO scores?

903

904

3.8 Conclusion

905

3.9 Methods

906

3.9.1 Datasets

907

The crystal structures were downloaded from [Fragalysis](#). For Mpro, non-covalent fragments from the XChem fragment screen [?] were used while for Mac-1 both XChem and UCSF fragment data were used.

908

909

910

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

911

912

913

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

914

915

3.9.2 Model Construction

916

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

917

918

919

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

920

921

922

923

924

925

926

927

928

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 have nothing to do with binding to the binding site - strong correlations in pharmacophore distribution between multiple independent fragments are indicative of useful binding interactions and these are what we hope to capture with this methodology.

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

Virtual screening of molecular libraries is done by evaluating the probability of the pharmacophore distribution of each molecule using the KDEs. For each pharmacophore combination, the mean log-probability of the distribution is calculated. The overall score for the molecule is returned as the mean log-probability over all of the pharmacophore combinations.

TODO - description of runtime? Emphasise no GPUs needed?

3.9.3 Compound Selection

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

The top-50k molecules remaining from the filtering were then clustered via Butina Clustering [?] with a Tanimoto distance threshold of 0.2. This resulted in 24748 and 22358 clusters for Mpro and Mac-1, respectively. For both targets the centroids of the 50 most populous clusters (or the closest purchasable analogue if it wasn’t available) were chosen as the candidate

compounds. These compounds were ordered for synthesis from Enamine which resulted in 38
and 52 successfully made molecules for Mpro and Mac-1, respectively.

TODO - Confirm made/assayed molecule numbers.

3.10 Author Contributions

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

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

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