Chemical Space Analysis and Property Prediction for Carbon Capture Amine molecules

James L. McDonagh[a, b, c, 1,✉], **Stamatia Zavitsanou**[d, 2], **Alexander Harrison**[a], **Dimitry Zubarev**[e], **Flaviu Cipcigan**[a], **Theordore van Kessel**[f], **and Benjamin H. Wunsch**[f, 1, ✉]

[a]IBM Research Europe, Hartree Centre, SciTech Daresbury, Warrington, Cheshire WA4 4AD, U.K.

[b]University of Edinburgh, School of Mathematics, Bayes Centre, 47 Potterrow, Edinburgh EH8 9BT, U.K.
[c]Current address: Ladder Therapeutics doing business as Serna Bio, Lab F37, Stevenage Bioscience Catalyst, Gunnels Wood Road, Stevenage, Hertfordshire, SG1 2FX, UK
[d]University of Oxford, Physical and Theoretical Chemistry Laboratory, Oxford, UK
[e]IBM Research, IBM Almaden Research Center, San Jose, CA 95120, USA
[f]IBM Research, IBM T.J. Watson Research Center, Yorktown Heights, New York 10598, USA

Carbon capture and storage is part of the roadmap towards net zero for many countries around the world, since emissions from existing infrastructure are close to estimated carbon budgets. To address this problem, currently 87 carbon capture projects are proposed worldwide in the next 10 years. A major class of commercial carbon capture technology involves capture systems using solvents. Commonly carbon capture solvents feature blends of amines and water. Whilst these blends have proved valuable there is an increasing need to identify new candidate molecules which are more efficient and improve performance. Systematic approaches to improve on the current technology are now needed with increasing urgency to expedite the introduction of cutting edge carbon capture methods. Here, we present a chemical space analysis of carbon capture amines and proceed to show a framework for computational screening relevant to carbon capture solvents. The screening approach demonstrates the use of cloud computing, novel molecular representations and machine learning to screen potential candidates. Our show the utility of machine learning in this field for high throughput virtual screening with an exemplar application to absorption capacity classification. Additionally, this work discusses the opportunity for improved data awareness and accessibility in this field to advance at a pace of its e development.

carbon capture | data | machine learning | chemical representations
Correspondence: ***bhwunsch@us.ibm.com, james.mcdonagh@serna.bio***

Introduction

Climate change driven by emissions from human activities now poses the greatest environmental concern of this century.(1) Emissions of greenhouse gases such as CO_2, methane and nitrous oxides (NO_x) are the primary drivers of global warming. CO_2 is the largest fraction of greenhouse gases emitted. (2) Electricity generation from fossil fuel burning is the largest point source of CO_2 emissions around the world. Yet, fossil fuel burning infrastructure is still being built.(1) Due to this trend, committed emissions from existing energy generation infrastructure jeopardise climate targets.(3)

Modelling suggested that Carbon Capture, Utilization and Storage (CCUS) for CO_2 emissions is a necessary part of the technological solutions required to meet the Paris climate accord.(1, 4) CCUS is the only technology that can be used to help decarbonise existing energy infrastructure without decommissioning. CCUS is also important for hard-to-abate emissions, such as those in heavy industries.(5) There are approximately 87 planned CCUS plants between 2020–2030 according to the map of global CCUS projects by the International Association of Oil and Gas producers(6).

Of the currently available CCUS technologies, absorption using carbon capture solvents is the most mature, seeing commercial usage with further plans for new developments.(7, 8) The technology is dominated by the use of amine based solvents such as Monoethanolamine (MEA) or proprietary formulations of blends of amines. MEA has become a defacto standard as it has shown good performance in terms of capture capability as well as being relatively cheap. However, it has several drawbacks: high-energy penalty on regeneration, thermal degradation and corrosion.(7) As a result, new solvent candidates and new solvent mixtures are being investigated in both academic and industrial research laboratories.(9)

In this context, computational techniques can be used to screen, rank and predict new carbon capture solvents.(10–14) These computational techniques hold promise to improve the speed of discovery and innovation if paired with suitable data sets of solvent performance. In particular, the field of Chemical Informatics has developed a multitude of methods and practices, which can be used to address problems in the field of carbon capture(15). Access to good quality research data and methods is critical to the fast progress of a field, as demonstrated by examples such as those in solid state materials design (16) that have benefited from open innovation.

To inform this study and demonstrate the usefulness of computational approaches to this field, we have identi-

fied 167 unique amine molecules which have been reported in the literature(17–25) in relation to a range of carbon capture performance metrics. We have extracted string representations for these molecules from PubChem(26) and ChemSpider(27) in order to perform an analysis of the chemical space of carbon capture amines. In addition to this, we have created a new set of data for 98 amine molecules based on the absorption capacity of the amine molecules as an aqueous solution of 30% w/w. We have used a consistent set of experimental measures, making this dataset highly valuable for training Machine Learning (ML) models upon.

In this work, we use the dataset of 167 molecules to consider the chemical space of carbon capture amines. Additionally, we build high throughput virtual screening (HTVS) methods to predict the absorption capacity over our own dataset of 98 amine molecules. Our results are our first step towards accelerating the discovery of CCUS solvents.

Results

A Data Collection and Curation

Initially, we reviewed the literature searching for experimental absorption capacity measurements. It became clear that there were potentially issues comparing data over multiple experimental techniques and conditions and that the field lacked common data standards for carbon capture solvents research. Unlike counterparts in the solid state, such as Metal Organic Frameworks (MOFs), for which extensive crystal structure databases have been provided, carbon capture solvents is a relatively data poor field. This in many ways is likely related to the field's success in being one of the first commercially applied carbon capture technologies. As a result, data may often be considered too sensitive to be released. This is especially true of formulations and blended solvents.

This situation is historically reminiscent of fields such as pharmaceuticals, which, in some cases, have seen benefits from opening up some of the larger internal data sets from commercial organizations in recent years. (28) These benefits are both scientific (faster development of new ideas) (29) and also economic (30). Woelfe *et al* (31) provides an example case study on how a community accelerated the development of a route to enanitopure Praziquantel. The authors of this manuscript have demonstrated the use of open data sets towards predicting molecular and material properties such as water solubility and partition coefficients previously.(32–34)

Opening data in this field could enable a proliferation of data driven modelling and the establishment of common standards upon which to fairly compare methods. These comparisons can drive rapid advancement of computational screening in this area. This will help to bring research in this area in line with solid state carbon capture which sees wide spread modelling.(35, 36) Similar arguments have been proposed and discussed in other related fields for formulation chemistry.(37)

For these reasons, we have gathered our own data on a consistent experimental basis. We gathered 98 data points in total. These molecules were chosen as they represent a sub-set of previously explored molecules and unexplored molecules to the best of our knowledge. The unexplored molecules were chosen based upon expert input and computational similarity screening. The similarity screening was carried out against 11,000 purchasable amines from the ZINC database. We extracted from PubChem(26) and Chemspider(27) the SMILES representations of molecules previously tested for carbon capture from the literature.(17–19, 21, 22) We used this set as a comparison set for the similarity screening. The similarity was determined using Extended Murko hashes and Tanimoto similarity scores. A final set of 98 purchasable molecules was then selected from the screening and expert input.

We have performed in-house measurements on the 98 amine molecules using our testing laboratory, which measures the absorption capacity at $40^{\circ}C$ using a $200\mu L$ sample solution. The CO_2 capture measurements are made using a Non-Dispersive Infrared Sensor (NDIR) with a $4.3\mu m$ absorption band and a $3.9\mu m$ reference band. The experiments take approximately 60 minutes to complete and are run in duplicate. We have chosen to focus our HTVS efforts on binary classification for this initial work. The aim therefore, is to provide HTVS models which can be applied to prioritise molecules for more expensive exploration. The classes which we used as target data in this work are provided in the Table 1. Below we describe our data curation and classifcation process in detail.

For each of the 98 molecules we extracted the identifiers and 2D structures of the molecules. We proceeded to search the PubChem(26) and ChemSpider(27) databases for entries of these molecules and extracting further identifiers such that all molecules were specified by: IUPAC name, InChI, InChIKey and SMILES. In some cases an entry could not be found and we manually determined the name and generated the SMILES string, from which, we generated the InChI and InChIKey using RDKit(38) (version 2022.03.2). These representations are the most commonly used and are easily parsed by standard chemical informatics tool kits such as RDKit and OpenBabel(39). This information is provided in the .

A range of capacity units are used in the literature. The most common appear to be: $\frac{moles(CO_2)}{moles(\text{N atoms})}$, $\frac{moles(CO_2)}{moles(\text{amine molecules})}$ and $\frac{g(CO_2)}{g(\text{amine molecule})}$. Another unit which we encountered several times was $\frac{g(CO_2)}{L(solution)}$. This unit requires knowledge of density to accurately convert, as the solution includes the solvent volume as well as the amine volume. We have used the unit

$\frac{moles(CO_2)}{moles(\text{N atoms})}$ for our absorption capacities and provide conversion factors in the equation 1.

B Infrastructure

In this work we used cloud based computing as this offers us flexibility to scale the resources to our needs. This cluster consisted of eight nodes, each with 8 virtual CPUs and 32GB of RAM. This allowed us to quickly provision infrastructure to run our modeling.(37, 40, 41)

C Computational Modelling

In this work we have applied a range of methods to explore the properties of the proposed solvents. These methods broadly fall into the category of data driven chemical informatics, including chemical graph analysis, sub-structure searching and machine learning.(15) To our knowledge, the application of chemical space analysis and the subsequent bespoke fingerprinting is a novel contribution to this field and present a new analysis of the molecules most commonly used for carbon capture solvents.

C.1 Substructure searching and Topological Data Analysis

In the first part of this work, we analyze the structures of the molecules which have been considered as possible carbon capture solvents. We then compared these molecules with a set of $20,938$ commercially available amines taken from the ZINC database (42).

The purpose of this analysis is to identify chemical functionality strongly associated with carbon capture performance and to highlight potentially under-explored, yet synthetically accessible, regions of the amine chemical space. To achieve this, we used sub-structure searching over 3D molecular graphs which were generated from SMILES strings using RDKit. We extend this analysis with Topological Data Analysis (TDA) applied on the chemical space to produce a skeletonized representation of the high-dimensional molecular data set via Mapper TDA (43–54).

Mapper TDA is a technique to visualise the topology of high-dimensional data, such as point clouds. The construction is related to the concepts of a Reeb graph and pullback covers (45, 52). Mapper TDA tracks the evolution of the level sets of a real-valued function associated with the data points, known as the filter function. The filter function can be selected to reflect some geometric properties of the points in the dataset, such as eccentricity (position relative to the center of the data) or local density. The range of filter function values is split into overlapping intervals, also referred to as level sets. Mapper TDA tracks evolution of these level sets. For each interval, the corresponding subset of the data points is clustered. Finally, a graph is constructed where each node represents a cluster and two nodes are linked if the corresponding clusters overlap. Two Mapper TDA clusters can overlap because the filter function intervals are allowed to overlap. Further, it is customary to associate some attributes, such as filter function values or some scalar properties, with the nodes and visualize them as colors. The number of data points in the cluster is often visualized as the node size. The output of Mapper TDA is highly dependent on the choice of hyper-parameters. A comprehensive analysis of Mapper TDA parameters can be quite involved and equivalent to a standalone computational task (48).

C.2 Machine learning and Model Evaluation

In the second part of this work we describe a workflow for the classification of carbon capture molecules using several learning algorithms. The machine learning models include the Logistic Regression Classifier (55–57), Ada Boost Classifier (58, 59) and Gaussian Process classifier (60) as implemented in Scikit-learn (61) (version 1.0.2). We envision the classifiers as a first step towards high throughput virtual screening of carbon capture molecules. In many cases classification may be sufficient in order to prioritise and decide upon whether a molecule will go on to further more elaborate screening. The classification methods have been widely used for chemical property predictions previously.(62–64)

Gaussian Processes have been used in chemical modelling in many instances.(65–68) These are a stochastic process, which perform Bayesian inference over a space of functions that map a representation to a probability space, for the class of a molecule. A prior is used to define a probability distribution over functions. As data is provided to train the model, the functions which most suitably represent the data are selected leading to the posterior probability distribution. For classification, a logit function is used to output class probabilities. More details are give in chapter 3 of Williams *et al* (60).

Ada Boost, as implied by the name, is a boosting algorithm that combines multiple weak classifiers to increase the accuracy. In our case we use decision trees as our weak learners. The Ada Boost method works by initializing all training data with equal weights. After the first classifier is trained, examples which are incorrectly classified by the first classifier are given a higher weighting. The process is repeated for N weak learners.

Finally, Logistic Regression in its basic form uses a logistic function to model a binary dependent variable. This is done using a standard linear regression model which is mapped through a logistic function to give probabilities. Each molecule is assigned a probability for class 0 and 1 with a sum of one.

All models are assessed in terms of multiple performance metrics: accuracy, sensitivity, specificity, Receiver Operating Characteristics (ROC) curves and (69) Matthews Correlation Coefficient (MCC) (70, 71).

These metrics can all be formulated mathematically from a confusion matrix, which identifies the correct predictions, True Positives (TP) and True Negatives (TN), along its main diagonal and the two types of error associated with binary classification, False Positive (FP) and False Negative (FN), in the off diagonal elements. The equations used for these metrics are given in the equations 2 - 7.

Briefly, these metrics comprise the most commonly applied metrics for classification problems and well characterise the performance of our methods. Accuracy is likely the most common classification metric.(71) It is a ratio of the number of correct predictions over the total number of predictions. This leads to a ratio describing the fraction of predictions which are correctly classified in the set. This simple metric is a valuable high level overview of the performance of a classifier. The sensitivity and specificity each focus on the models ability to correctly predict the positive or negative class respectively. These metrics provide a greater insight into the potential errors and biases of the models. The ROC curves describe the model performance over decision thresholds with a FN rate on the x axis and TP rate on the y axis. These thresholds can be considered as balancing the positive an negative predictions, i.e. lowering the threshold will increase the number of positive predictions, which is the sum of true positive and false positive predictions. The Area Under the Curve (AUC) for a ROC curve is the integral of the area under the ROC curve and provides a single value metric for this trade off. The MCC metric is a powerful summary metric which ranges from -1 to +1 describing the skill of the classifier to predict positive cases as positive and negative cases as negative even when the classes are imbalanced.(71).

C.3 Computational workflow

The workflow to generate these models is given in Figure 1. The workflow contains two K-fold Cross Validations (CV) one nested within the other. The external CV holds a portion of the data set out as a validation set whilst providing all other points as training data. The internal CV uses the training points from the external CV to optimize the hyper-parameters and train a classifier for each external k-fold.(72) This means that the predictions are made for all 98 molecules over our external K-fold without biasing the models. Additionally, we can make an assessment of the models robustness to training set changes. We have chosen this method as it enables us to optimally use the small data set we have been able to gather from the literature.(72)

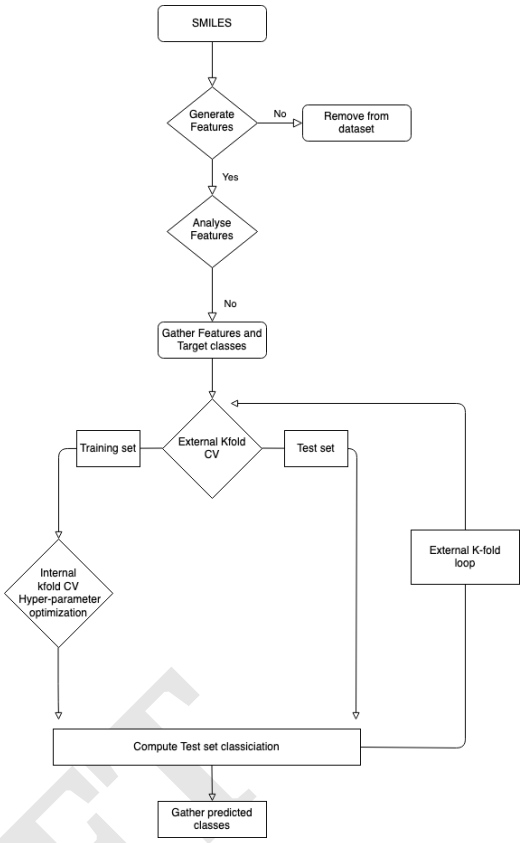

Fig. 1. Workflow to make classification predictions of each molecules in our data set.

To describe these molecules, we used three methods. The first are standard chemical informatics descriptors, generated through the Mordred descriptor calculator, (73) which produces over 1800 features of molecular characteristics. From the 1800 descriptors calculated, we identified the ones that correlate significantly with the properties of interest using the Spearman correlation coefficient between each Mordred descriptor and the respective property of interest.

Another way to describe molecules is *via* molecular fingerprints. Molecular fingerprints are vectors that encode structural information about a molecule. Commonly, this information is stored as binary digits representing presence and absence of a structural feature. There are different types of fingerprints available such as Morgan fingerprints (74), MACCS fingerprints (75) or MinHashed Atom Pair (MAP) fingerprints (76). In this work we have used the commonly applied MACCS fingerprints.

Additionally, we have defined our own structure based fingerprint (CCS fingerprint) following consideration of the literature and our own chemical space analysis. The latest version of the source code for generating these fingerprints and the data set we generated can be found `https://github.com/Jammyzx1/Carbon-capture-fingerprint-generation` (as of submission) and archived under DOI 10.5281/zenodo.7828285. This fingerprint is a fixed length (72 elements) with each element representing a chemical

group or groups. These chemical groups comprise those commonly seen in carbon capture solvents and those found more broadly across amine chemical space. Each bit is defined by a SMARTS string and substructure searching is carried out in parallel using DASK(77) (version 2022.02.0) and RDKit(38) (version 2022.03.2) to generate the fingerprint vector.

In this section we outline our chemical space analysis, models and property predictions. We begin exploring the molecular data set, seeking trends across the data in terms of the chemical structures. We proceed with using the learning models discussed previously to predict absorption capacity. We complete this work by evaluating our models and considering the impact of our predictions.

1 Results and Discussion

A Chemical Space Analysis of Carbon Capture Amine Molecules

First, we explore and compare the structures of the amine molecules we extracted from carbon capture literature with those we extracted from the ZINC database(42) which are commercially available.

Several authors have reported chemical sub-structures which influence carbon capture capabilities.(10, 17–19) In particular, Singh *et al*(18, 19) developed structure activity relationships based on chemical functionalities. Their work studies the effects of many chemical functionalities on carbon capture loading and develops design considerations for carbon capture amines. These included alkyl chain length and functional group separation measured in number of carbon atoms. Additionally, consideration of ring substituent and their positions was provided in a later publication.(19) Work by Papadopoulos *et al*(10) provided a computational design system. This work also identified a small number of chemical structures which were useful as descriptors for their models. Work by Puxty *et al*(17) reports the position of OH moieties relative to the amine nitrogen to be important. Steric hindrance around the amine nitrogen is another chemical feature reported to be of importance. It has been shown that steric hindrance can change the reaction route of primary and secondary amines towards that of tertiary amines. This is an important observation owing to the differing atom efficiency between the two routes. Primary and secondary amines have been shown to react with CO_2 through a pathway requiring a second molecule to complete the reaction, see figure 2. The second molecule may be water in some cases or a second primary or secondary amine. Tertiary amines have been shown to react in a one to one fashion with CO_2 effectively acting as a catalyst see figure 3.(12, 17, 78, 79)

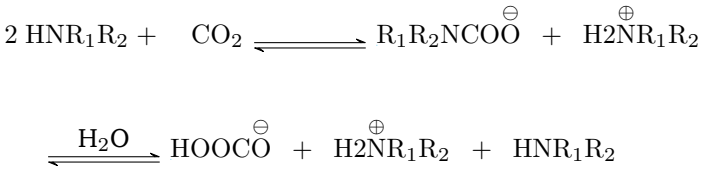

Fig. 2. Primary and secondary amine general reaction scheme.

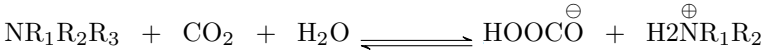

Fig. 3. Tertiary amine general reaction scheme.

We have taken these considerations a step further, defining the CCS chemical fingerprint based upon these observations and our own analysis of commercial amines. Our analysis identified common functionalities in commercial amines such as benzene rings, five member carbon rings, nitrogen containing hetrocycles and halogen groups some of which are not commonly found among amines tested for carbon capture. The CCS fingerprint we define combines the SMARTS definitions for common chemical sub-structures in molecules tested for carbon capture and wider commercial amines. We apply this tool here in consideration of the relative abundance of these sub-structures in carbon capture and commercial amines.

The inclusion of both chemical functionalities common in carbon capture amines and those more broadly in synthetic amines was done to enable the fingerprint to capture the differentiation between the two groups. We use sub-structure searching over a fixed order of chemical sub-structures, defined by SMARTS, in order to produce the CCS fingerprint. The fingerprint definition in terms of the order and SMARTS patterns used for substructure matching are included in the . Each of the SMARTS patterns defines one bit in our fingerprint. In total there are 72 elements and hence 72 sub-structure searches per molecule. In order to make this computationally reasonable we apply sub-structure searching through RDKit and parallelize over batches of 1000 molecules using DASK(77, 80). With this implementation we are able to produce the fingerprint in approximately 5 minutes on a laptop for the set of $20,938$ molecules compared to several hours when run in serial.

Considering these points Figure 4 displays a fingerprint based comparison of the 167 amines trialled for carbon capture compared to the $20,938$ amines collected from ZINC which are commercially available.

The list of carbon capture molecules collected in this work is not exhaustive, but is a representative sample of the published amine solvent molecules which have been openly reported. As a result the aim here is to provide an analysis which highlights the most explored regions of the amine chemical space and point out synthetically accessible areas of amine chemical space which may be under explored in terms of carbon capture. Figure 4 displays a histogram with the normalized count of occurrences of the given sub-structures across molecules in both sets (blue is the carbon capture trialled data set of

167 molecules, red is the commercial amines data set of 20,938 amines). Clearly there is a substantial difference in the size of these data sets, hence the normalization allows one to consider relative abundance rather than absolute counts.

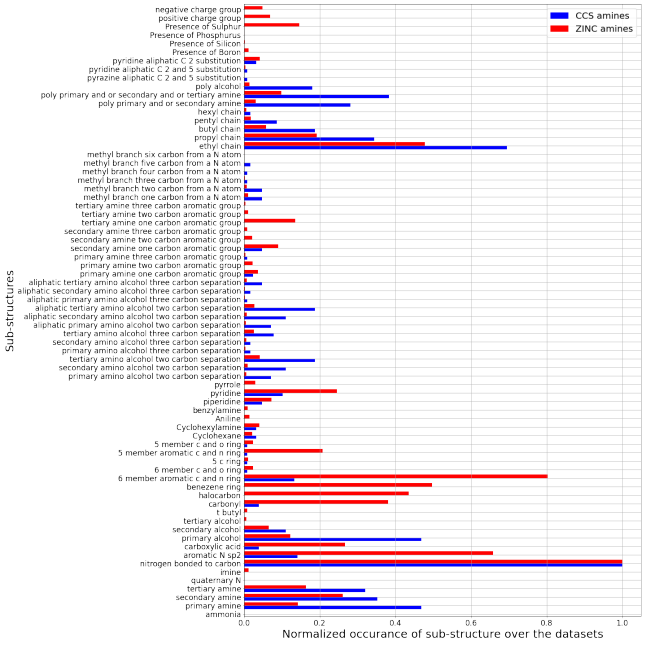

Fig. 4. Fingerprint comparison over two data sets of amines, 20,938 commercially available amines and 167 amines tested for carbon capture abilities. All bits are found in the larger data set at least once except ammonia, however their occurrence may be rare enough that it is not clearly visible on the normalized x-axis. Where this occurs we have decided to include the bit as it has been noted in other literature sources as potentially important.

From Figure 4 it is clear that the carbon capture data set includes molecules which contain a sub-set of chemical moieties more commonly compared to the commercial amine data set. For example, in the alkanolamines sub-structures in the centre of the y-axis. This subset may be somewhat expected given the wide spread use of MEA and related molecules. It is also clear that structures such as carbonyls, halo-carbons and aromatic groups are relatively less common in the carbon capture data set compared with commercially available amines. We note that substances such as benzylamine have been used as promoters within formulated blends rather than capture solvents themselves. Such molecules are not captured in this analysis.(81, 82) This analysis suggests there is likely a defined sub-space of the amine chemical space, which is more likely to be associated with amines suitable for carbon capture.

Figure 5 displays the chemical space graphically and follows the protocol described in some of the author's previous work.(83) In this figure each molecule is represented as a node in the graph and the most similar (Tanimoto similarity scores of ≥ 0.7 using Morgan fingerprints with a radius of 2 and 2048 bits) are connected. The graph topology is generated through the Fruchterman-Reingold force-directed algorithm(84) using Python's NetworkX package (v.2.6.3). This algorithm treats the nodes as a set of spring connected particles and simulates the graphs topology to a quasi-equilibrium state. In this case the springs were weighted by the Tanimoto similarity score, making those more similar node relatively more attractive to one another. The highlighted nodes are molecules which have been reported in the literature as trialled for carbon capture capability previously.

Fig. 5. Force directed graph of the amine chemical space. The highlighted nodes are molecules which have been reported in the literature as trialled for carbon capture capability previously. The cyan nodes are commercially available amines which to the best of our knowledge have not been tested for carbon capture capability.

We can see that molecules which have reported carbon capture properties are not evenly distributed. The nodes tend to be away from the centre and distributed throughout the shell of the graph. The graph is generated based upon molecular similarity such that those with more connections remain closer to the centre of the graph. As the carbon capture molecules tend to exist in the shell they can be considered relatively dissimilar to the commercial amines which remain in the centre. Still most of the carbon capture amines posses at least one connection, suggesting they are not special isolated cases. Generally the carbon capture amines appear to inhabit sub-sections of the amine chemical space based upon molecular similarity.

To elucidate this sub-space more clearly we have applied TDA. A skeletonized representation of the set of the topological data associated with both sets of amines described above is shown in Figure 6. Mapper TDA is applied to the molecular point cloud in the space of the CCS structural fingerprints equipped with pair-wise dice distances. During Mapper construction, we chose eccentricity of the molecules in the point cloud as the filter function. Here, eccentricity refers to the position of the molecule relative to the "center" of the point cloud; it increases further from the center towards the outskirts. The range of the eccentricity values was split into 40 intervals with 50% overlap between intervals. This produced 40 level sets of amines which were clustered with Agglomerative Clustering on the pre-computed matrix of dice distances.

Figure 6A shows the produced Mapper graph where nodes represent clusters within level sets, nodes are linked if respective clusters have common members, color encodes the filter function (eccentricity), and the node size encodes the number of amines in the respective cluster. Figures 6B and C maintain the layout of the graph in Figure 6A and the encoding of the number of amines in a cluster by the node size. Figure 6B shows the anomaly scores of the molecules in the dataset evaluated using the Isolation Forest algorithm, averaged over clusters, and encoded as the node color. High positive values of the anomaly score indicate inliers, decreasing values indicate higher level of abnormality, and negative values indicate outliers. Figure 6C uses color to encode

the fraction of the carbon capture amines in each cluster. We note that the highest content of carbon capture amines in the Mapper clusters does not exceed 20%.
Comparison of Figures 6A and C suggests that carbon capture amines live primarily on the outskirts of the data set. This finding can be interpreted as a sign of under-utilization of the space of amines in the studies of utility for carbon capture. One possible reason could be a bias of the majority of amines towards biochemical/medicinal applications leading to unnecessarily complex and/or expensive structures. Comparison of Figures 6B and C shows that carbon capture amines are not outliers, as the only cluster with the average anomaly score characteristic of outliers has zero fraction of carbon capture amines. Carbon capture amines are not the most “normal" amines either, the average anomaly scores of the clusters rich in carbon capture amines are shifted towards zero.

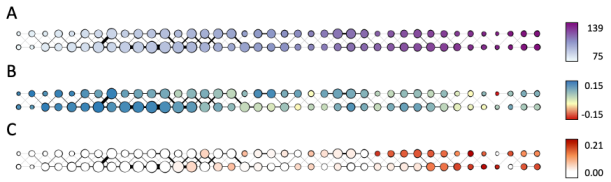

Fig. 6. Mapper graph of the combined dataset of amines. Eccentricity of amines in the combined dataset is used as the filter during Mapper construction. Node size is proportional to the number of amines associated with the node. Thickness of a link between two nodes is proportional to the number of amines that are associated with both nodes. Panel A: color encodes mean eccentricity of the molecules associated with the node. Panel B: color encodes mean anomaly score (Isolation Forest) of the molecules associated with the node. Panel C: fraction of amines from CCS dataset among molecules associated with the node.

Considering all aspects of this analysis it appears that the carbon capture amines considered here are representatives of a sub-space in amine chemistry. Many of the commerical amines are likely to have been developed for diverse industrial applications and as such many will be unsuitable (too costly, over complex or only available in small quantities) for carbon capture. The analysis does suggest though that there is considerable unexplored, or at least unreported, areas of amine chemical space which may hold novel candidates for carbon capture.

B Carbon Capture Absorption Capacity Classification

In this section we outline our absorption capacity classifications. We begin generating QSAR models for the classification of molecules based on absorption capacity. We complete this work by evaluating our models and considering the impact of our predictions.
Here, we report the results for the classification models generated with MACCS fingerprints, CCS fingerprints and Mordred descriptors against absorption capacity in units of ($\mathrm{mol}_{CO_2}/\mathrm{mol}_N$).

There are 98 molecules in our absorption capacity data set, classified to binary classes. Class 1 represents higher values and class 0 represents lower values of absorption capacity. The molecules are classified based upon the amine functionalities they contain. Both primary and secondary amines are thought to react with CO_2 through a mechanism requiring two amine molecules to complete the reaction. Therefore, a primary or secondary amine has a theoretical absorption capacity of 0.5 per primary or secondary amine group. Tertiary amines are thought to react in a one to one mechanism therefore have a theoretical absorption capacity of 1.0 per tertiary amine group. We classify molecules by summing up these expected contributions per amine group. Where mixtures of primary or secondary with tertiary amines arise we applying a weighting based upon the number of tertiary amine groups, as both of the proposed amine reaction routes are possible and can be competitive in terms of the kinetics. We therefore down scale the tertiary contributions to 0.5. If this value is below the experimental absorption capacity then class 0 is assigned to the molecule; if the experimental absorption capacity is greater than or equal to the value then class 1 is assigned to the molecule. From this dataset, 71 molecules are class 0, and 27 molecules are class 1.

The two classes are highly imbalanced. To achieve better performance in the models, we generate additional sampling points for the minority class using the Synthetic Minority Over-Sampling Technique (SMOTE) (85) for non-categorical features and Synthetic Minority Over-sampling Technique for Nominal (SMOTEN)(85) for categorical features. This is implemented in the imbalenced learn Python package (version 0.9.0). In both cases, these methods select the five nearest minority class neighbours in feature space to the kth example minority point, choose at random one of the five and generate a synthetic sample point along the connecting line between the example point and the random neighbour. Note that the methods have no information about the majority class.

These techniques provide a better balance between the classes and hence improve the learning of a decision boundary. We apply the SMOTE algorithms to each training set in the K-fold cross validation independently to avoid data leakage from the test sets. We note that pre-computing the SMOTE synthetic points prior to train test splits in the K-fold cross validation can lead to notable data leakage and over optimistic metrics for the model performance. We explored the impact of this in our work and found that on the headline accuracy metrics data leakage could provide approximately an 7-8% over estimate in a models predictive accuracy. Here we present how Gaussian Process, Logistic Regression and Ada Boost methods perform on the balanced data sets.

B.1 Mordred descriptors as features

For each molecule, we generate over 1500 descriptors using Mordred.(73) The list of Mordred descriptors can be found at reference (86). From these descriptors, we are only interested in those that have a notable correlation with absorption capacity. We thus set a Spearman correlation cutoff of 0.5 and further analysed these features for significance using a two-tailed p-test(87) over 5000 random sample permutations using the Spearman correlation coefficient as the test statistic, which leaves us with 35 features which have a significant p-value at 95%. The list of features which correlate are given in the . Following feature generation, we apply one-hot encoding for categorical features and min-max scaling for continuous features. There were 6 features considered as categorical out of the 35 (nBondsM, nBondsKD, C1SP2, HybRatio, FCSP3, ETA_beta_ns). Categorical in this case includes features with specific increments such as counts. Following one hot encoding the feature set extends to 84 as every unique value of the categorical features becomes a binary feature array. Scikit-learn(88) was employed to perform one hot encoding and min-max scaling.(61)

B.2 Molecular fingerprints as features

As discussed above we have developed a new fingerprint, CCS fingerprint, for carbon capture solvents based upon the chemical space analysis and the presence or absence of substructure searches using SMARTS strings. They are composed of 72 binary features. The features are not pre-processed in any other way. The SMARTS definitions are provided in . The use of such fingerprints can enhance the interpretability of models in terms of the chemical structures and their correlation with the properties of interest.

Additionally, we compared our CCUS fingerprint with the well established MACCS keys (89, 90). The MACCS keys are composed of 166 binary bits which also represent the presence and absence of chemical features. MACCS keys have been widely used, especially in the pharmaceutical industry. The bits represent a wide subset of chemical space.

B.3 Results for Mordred Descriptors

We begin our modelling of absorption capacity using the Mordred descriptors as features to represent the molecules. Figure 7 and table 1 provide a summary of the performance of the three models generated from Logistic Regression, Ada Boost and Gaussian Process classification methods.

Table 1. Classifier metrics for balanced data for absorption capacity with models built from Mordred features. MCC is the Matthew's correlation coefficient.

Algorithm	Accuracy	Sensitivity	Specificity	MCC
Gaussian Process	0.73	0.30	0.90	0.25
Logistic Regression	0.81	0.63	0.87	0.51
Adaboost	0.74	0.48	0.85	0.34

From the results in figure 7 and table 1 overall all models have a fair predictive accuracy's between 0.73 and 0.81. The Gaussian Process and Ada Boost methods have broadly performed similarly in terms of accuracy, but the Logistic Regression method has a notable improvement with an accuracy over 0.80. However, for all three model there are notable differences in the sensitivity and specificity. The Gaussian Process and Ada Boost models both struggle similarly in terms of sensitivity. This is demonstrated clearly in figure 7 A and C. Plot A shows roughly the same number of and CCUS fingerprint predictions coupled with a larger number of predictions whilst plot C shows a near even spread over , CCUS fingerprint and . This suggests the models are very poor in terms of predicting the positive class. The Logistic Regression model shows improvement beyond Gaussian Process and Ada Boost with respect to sensitivity, with noatbely higher prediction proportion. All models show much better performance in terms of predicting . The MCC values highlight this imbalanced predictive accuracy with fairly low values; noting that values of 0.0 for MCC correspond to random, these predictions are showing limited improvement above this.

B.4 Results for MACCS fingerprints

Turning to the MACCS fingerprint representation, figure 8 and table 2 provide a summary of the models performance.

Table 2. Classifier metrics for balanced data for absorption capacity with models built from MACCS fingerprint features. MCC is the Matthew's correlation coefficient.

Algorithm	Accuracy	Sensitivity	Specificity	MCC
Gaussian Process	0.78	0.48	0.89	0.40
Logistic Regression	0.83	0.63	0.90	0.55
Adaboost	0.78	0.56	0.86	0.43

Using the MACCS fingerprints, and considering the metrics in figure 8 and table 2 all three models again make a reasonable prediction of the molecules class considering the accuracy metric that ranges between 0.78 and 0.83. As for the Mordred descriptors, delving a bit deeper using the sensitivity and specificity metrics we find that predictions of the positive class are poorer that for the negative class. Again we the Logistic Regression model out performing the other two, however, there is a notable improvement in the prediction of the positive class for the Gaussian Process and Ada Boost models. The specificity has remained at a similar level of accuracy compared to the Mordred models. We note that the MCC scores have improved overall representing the

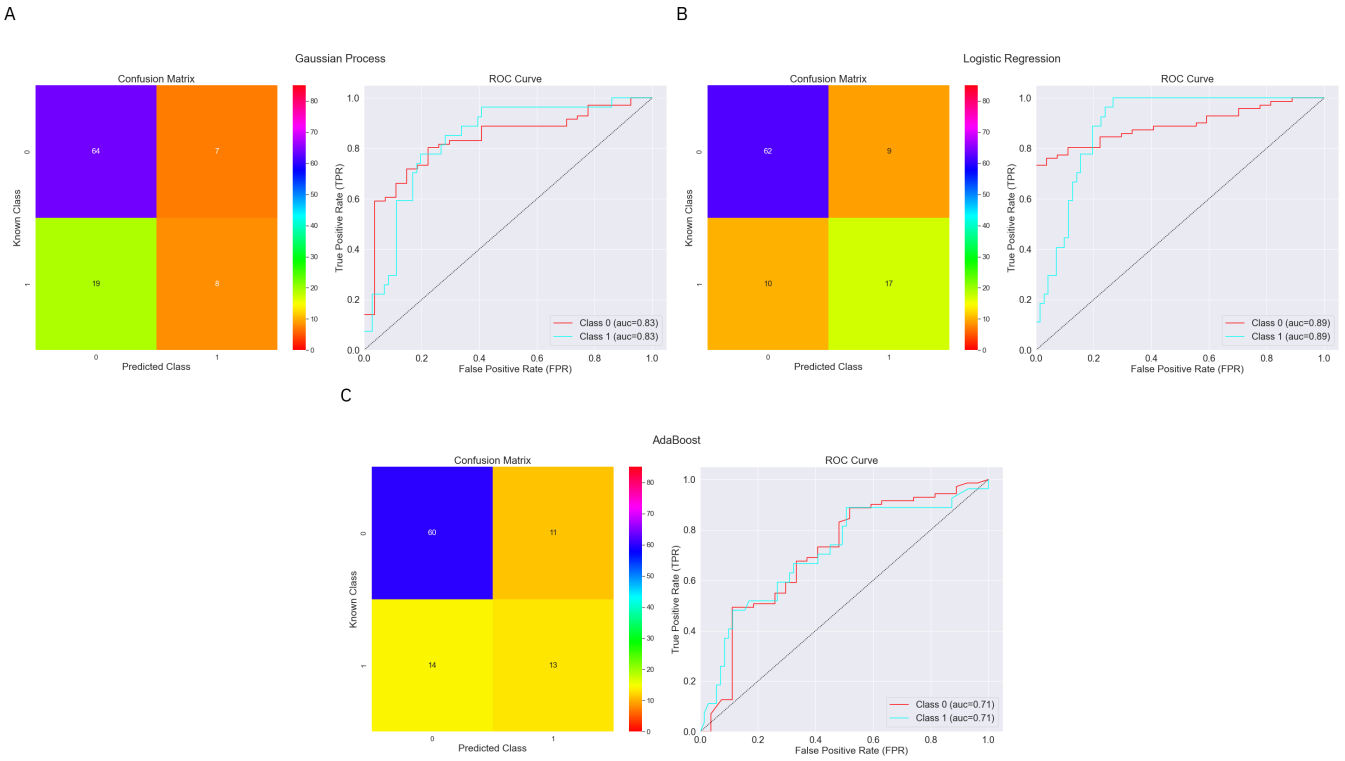

Fig. 7. Confusion matrices and ROC Curves for the balanced data against absorption capacity classification using the Mordred chemical features.

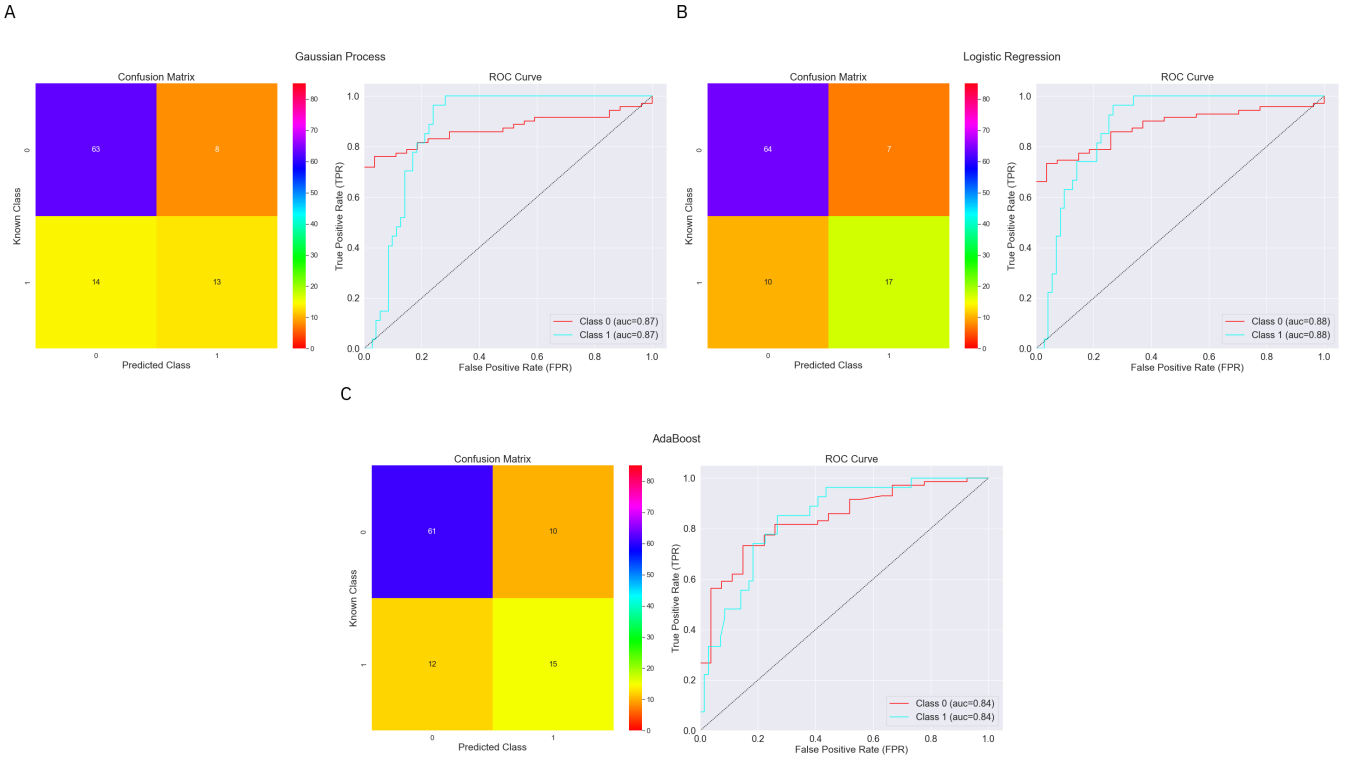

Fig. 8. Confusion matrices and ROC Curves for the balanced data against absorption capacity classification using the MACCS keys as features.

better balance over the three model in predicting both classes.

B.5 Results for CCS fingerprints

The last representation is that of our CCS fingerprint; figure 9 and table 3 provide the summary results for the three models trained on this representation.

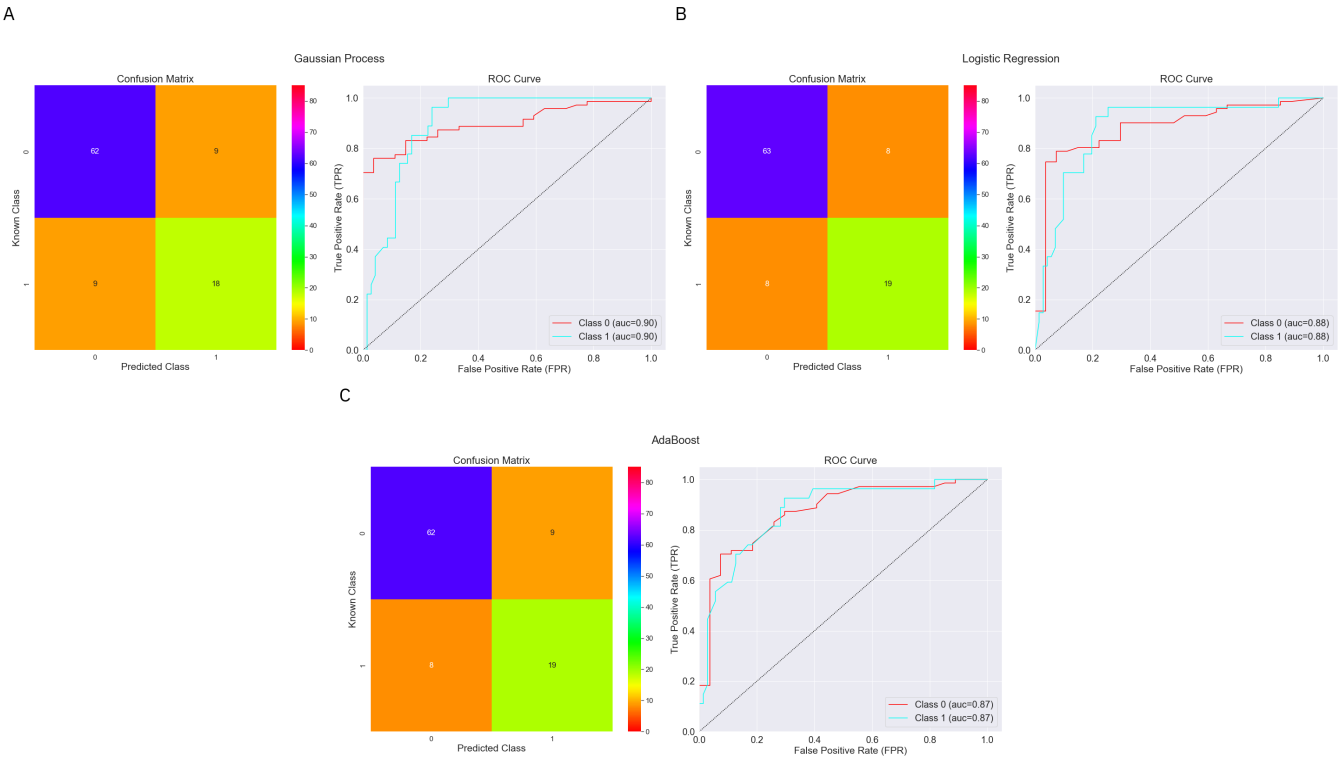

Fig. 9. Confusion matrices and ROC Curves for the balanced data against absorption capacity using CCS fingerprints as features.

Table 3. Classifier metrics for balanced data for absorption capacity with models built from CCS fingerprint features. MCC is the Matthew's correlation coefficient.

Algorithm	Accuracy	Sensitivity	Specificity	MCC
Gaussian Process	0.82	0.67	0.87	0.54
Logistic Regression	0.84	0.70	0.89	0.59
Adaboost	0.83	0.70	0.87	0.57

From figure 9 and table 3 it appears that all three models make good predictions of the molecules classes. The accuracy of all models is greater than 0.8, with the accuracy range of 0.82 - 0.84. In the Logistic Regression and Ada Boost models we note a much improved sensitivity of 0.70 shown diagrammatically in figure 9 where we can now see the majority of positive class molecules are predicted correctly by all three models.There is a slight improvement in the specificity also overt he three models compared to the models using Mordred or MACCS representations. Overall the MCC scores are now all over 0.5 showing the more balanced predictive accuracy.

Comparing the models on their summary metrics we see that in general figures 7 - 9 and tables 1 - 3 suggest that classification of molecules using shallow learning algorithms for absorption capacity is a difficult task. Across the models presented we have used several molecular representations. The Mordred descriptors are composed of a range of well known 2D molecular descriptors encoding information of electronic state, graph topologies and molecular properties. We found 35 had a notable correlation with absorption capacity but this vector extended to 84 when one-hot encoding was applied. This means a notable part of the representation contains a null representation. It is possible that with a larger data set the most explanatory features could be more readily identified and the models improved. The current models struggle particularly to correctly separate molecules into the promising class, with a fairly balanced error rate across CCUS fingerprint and predictions.

The MACCS fingerprints are a standard fingerprint representation which has been employed many times in materials modelling. To our knowledge, it has not been applied previously to predicting absorption capacity. In this work we see that the MACCS fingerprint performs reasonably as a representation but struggles with the classification of molecules in the promising class. This is clearly shown in the sensitivity and specificity values. The MACCS fingerprints are the largest representation used in this work at 164 elements each, with every element requiring a sub-structure match to build the representation. This can be a relatively computationally expensive task.

Having considered these two standard representation methods, we developed our own fingerprint, inspired by the MACCS scheme, which encoded the sub-structures noted by the carbon capture community to correlate with carbon capture performance. We also wished to generate a more condensed representation which with equivalent software implementation could reasonably be expected to be generated with fewer sub-structure matches. From this we developed the CCS fingerprint. The models generated above show the result is promising. All of the models built using the CCS fingerprint perform with an accuracy higher than the standard fea-

tures together with much improved predictive accuracy for the positive class, of approximately 70%. The models using the CCS fingerprint maintain high predictive accuracy for the negative class inline with the values seen from the standard features of 0.85 - 0.90. Owing to the improved predictive performance of the positive class these models also achieve the highest MCC scores demonstrating a more balanced predictive capability over the classes.

The best overall positive class predictor comes from the use of the CCS fingerprint features using the Logistic Regression classifier with 0.89 promising class correctly predicted 0.89 negative class correctly predicted and an overall accuracy of 0.84. The Logistic Regression models across all feature sets have tended to provide the most promising predictive accuracy over the classes. All models show a reasonable capability to predict the molecules which are unlikely to be promising in terms of capacity, which for HTVS may still be a useful and computationally inexpensive filter. The Use of the CCS fingerprint provides improved predictions of the positive class suggesting it could be useful in HTVS in terms of prioritisation of laboratory testing.

C Feature Importance

We have performed feature importance analysis using the Logistic Regression classifiers over the difference feature sets. The importance of a feature is reflected by the magnitude of the linear regression coefficients in the models. We show in figure 10 the mean feature importance over the cross validation.

Whilst being careful not to over interpret these figures, as they are based on no underlying fundamental physics or chemical theory, we can see some trends in the feature which are important. Looking at subfigure A, using Mordred descriptors we note number of auto-correlation feature have large magnitude coefficients. These auto-correlation coefficients relate to valence electrons and charges suggesting the model is largely relying on fairly simplistic representations of the electronic structure of the molecule. These models my be improved with a better description of the electronic structure.

We see large magnitudes in the linear coefficients for the CCS fingerprint for features related to the nitrogen environment, separating distances between amine and alcohol groups and chain lengths together with whether a molecule contain multiple amine functionalities. These are structural features which have been highlight by others as correlating with absorption capacity. As for the MACCS keys this is a reassuring set of feature importance's. We provide in the a SHAP (91) analysis of each of these models over cross validation for the 20 most important features as determine by SHAP. This analysis was performed on a subset of the each folds test data. This analysis shows similar trends to the feature importance.

2 Conclusions

This work displays an analysis of the chemical space of carbon capture amines against a background of commercially available amines. This analysis shows that carbon capture amines inhabit an edge region of the chemical space, but are not outliers in their structure compared to the wide set of commercially available amines. This is promising as it suggests that there may be other commercially available amines which will be suitable for carbon capture with out expensive new synthesis pathways being required. It also highlights chemical functional groups which are relatively less common in carbon capture amines. It remains unclear whether these are less common due to a lack of reporting on carbon capture capabilities for molecules containing these functionalities or due to these chemical functionalities having a consistent detrimental impact on carbon capture performance. This is an area for further exploration which could have a notable impact on the field by improving knowledge, data availability and thus modelling validation capabilities.

We used this chemical space analysis to define a novel fingerprint for the modelling of amine molecules used in carbon capture. This fingerprint has been shown to be an effective featurization method for QSAR modelling and a way to analyze the chemical space. We have also tested the use of commonly applied featurization methods through the Mordred engine and MACCS fingerprints. The models built here show that QSAR prediction for absorption capacity is challenging with the limited available data. Some of our model show promise for HTVS of carbon capture amines in the future. The use of the CCS fingerprint gave the most accurate classification models for each class. The CCS fingerprint also showed the most balanced model in terms of predictive accuracy for each class.

One of the biggest challenges to this work is relative lack of open available data in this field. This leads to small-data issues and limits the potential use of more complex modelling. We have used our own data for our HTVS models in this work. We will be publishing this data in due course. Opening data in machine readable formats (such as csv, json and HDF5 files for example) will enable computational scientist to better explore this area.

As policy shifts towards a net zero carbon world and carbon capture, usage and storage is deployed, the release of more data in the open literature related to these technologies will become more vital. This data can be enhanced with computation to help in the search for more efficient solvents, and carbon capture materials more generally, as we have demonstrated in this work. Further, the overlap of computational and experimental work is a powerful combination. Computation can

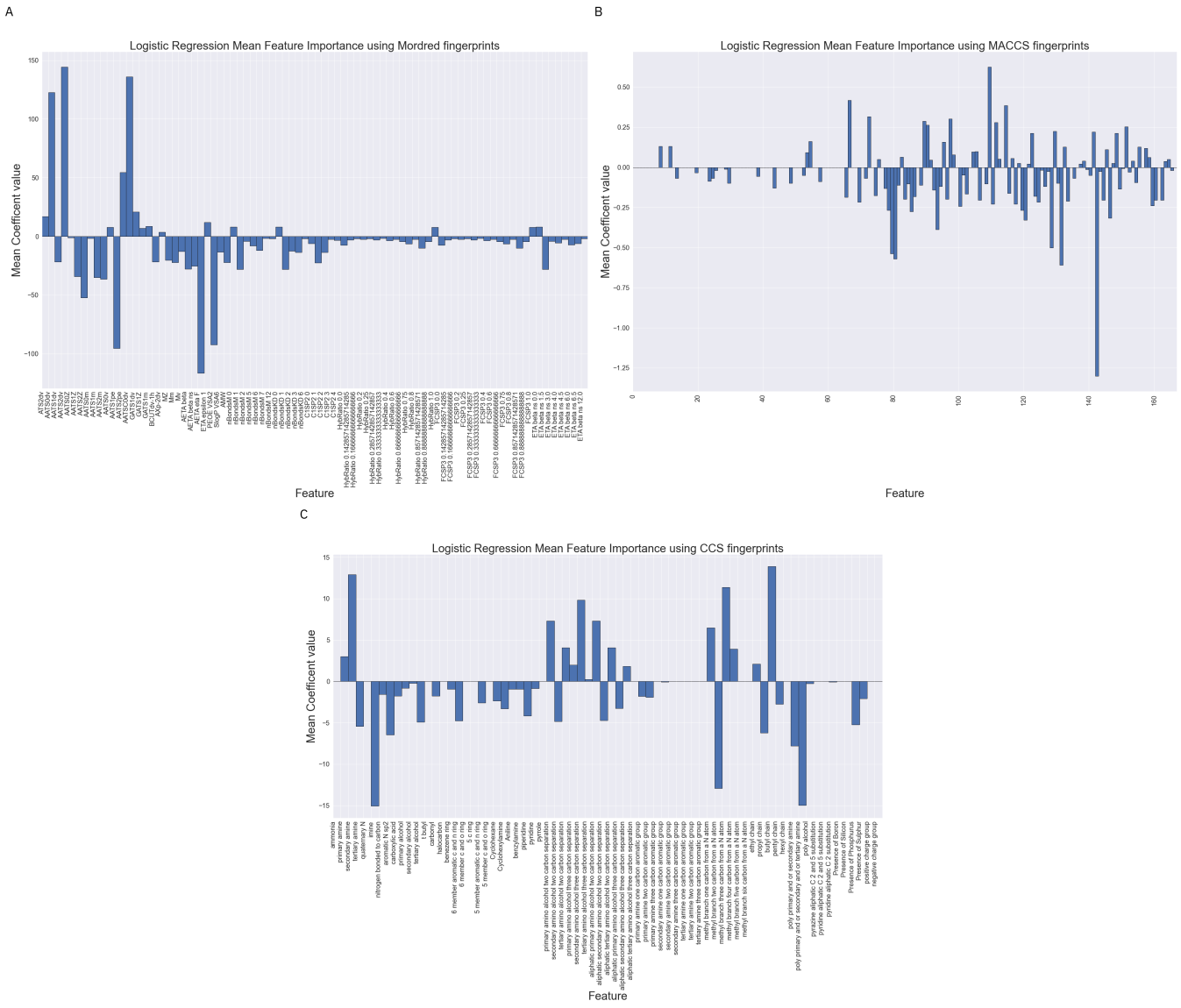

Fig. 10. Feature importance metrics using Logistic Regression over all feature sets. The mean regression coefficients are plotted as measures of importance. Sub-figure A is for Logistic Regression using the Mordred feature set, sub-figure B is for Logistic Regression using the MACCS fingerprints and sub-figure C is for Logistic Regression using the CCS fingerprints

rapidly screen and rank materials. Discovering more efficient materials for carbon capture is a goal that is required to avoid the more catastrophic effects of climate. Additionally, these tools can help to mitigate against potential future environmental threats from the use of carbon capture technology using predictive models for a wide range of properties. To mitigate the effects of climate change is likely to require great urgency in collaborating at scale across the world to accelerate the development and understanding of the most promising net zero technologies.

Author Contributions

James McDonagh: Conceptualization, Data curation, Formal analysis, Methodology, Software, Project administration, Supervision Writing – original draft. *Stamatia Zavitsanou*: Data curation, Formal analysis, Methodology, Software, Writing – original draft, Writing – review & editing. *Alexander Harrison*: Resources, Project administration, Software. *Dimitry Zubarev*: Formal analysis, Methodology, Software, Writing – original draft. *Flaviu Cipcigan*: Conceptualization, Project administration, Formal analysis, Writing – original draft, Writing – review editing.

Conflicts of interest

There are no conflicts to declare

Acknowledgements

The authors thank Bruce Elmegreen, Mathias Steiner, Stacey Gifford, Binquan Luan, James Hendrick and Nathaniel Park for insightful conversations. Data and required materials for this work can be found in the .

Code and Data Avaliability

The code and data set generated by the authors can be found at https://github.com/Jammyzx1/Carbon-capture-fingerprint-generation.

Bibliography

1. Xiao Wu, Meihong Wang, Peizhi Liao, Jiong Shen, and Yiguo Li. Solvent-based post-combustion co2 capture for power plants: A critical review and perspective on dynamic modelling, system identification, process control and flexible operation. *Applied Energy*, 257:113941, 2020.
2. Jos GJ Olivier, KM Schure, and JAHW Peters. Trends in global co2 and total greenhouse gas emissions. *PBL Netherlands Environmental Assessment Agency*, 5, 2017.
3. Dan Tong, Qiang Zhang, Yixuan Zheng, Ken Caldeira, Christine Shearer, Chaopeng Hong, Yue Qin, and Steven J. Davis. Committed emissions from existing energy infrastructure jeopardize 1.5 °c climate target. *Nature*, 572(7769):373–377, July 2019. doi: 10.1038/s41586-019-1364-3.
4. Thomas Bruhn, Henriette Naims, and Barbara Olfe-Kräutlein. Separating the debate on co2 utilisation from carbon capture and storage. *Environmental Science & Policy*, 60:38–43, 2016.
5. International Energy Association (IEA). Ccus in clean energy transitions. Technical report, International Energy Association, 2020.
6. International Association of Oil and Gas Producers. Map of global ccus projects, 2020.
7. Cong Chao, Yimin Deng, Raf Dewil, Jan Baeyens, and Xianfeng Fan. Post-combustion carbon capture. *Renewable and Sustainable Energy Reviews*, page 110490, 2020.
8. Mai Bui, Claire S Adjiman, André Bardow, Edward J Anthony, Andy Boston, Solomon Brown, Paul S Fennell, Sabine Fuss, Amparo Galindo, Leigh A Hackett, et al. Carbon capture and storage (ccs): the way forward. *Energy & Environmental Science*, 11(5):1062–1176, 2018.
9. Ida M. Bernhardsen and Hanna K. Knuutila. A review of potential amine solvents for CO 2 absorption process: Absorption capacity, cyclic capacity and pKa. *International Journal of Greenhouse Gas Control*, 61:27–48, June 2017. doi: 10.1016/j.ijggc.2017.03.021.
10. Athanasios I Papadopoulos, Sara Badr, Alexandros Chremos, Esther Forte, Theodoros Zarogiannis, Panos Seferlis, Stavros Papadokonstantakis, Amparo Galindo, George Jackson, and Claire S Adjiman. Computer-aided molecular design and selection of co 2 capture solvents based on thermodynamics, reactivity and sustainability. *Molecular Systems Design & Engineering*, 1(3):313–334, 2016.
11. Graeme Puxty and Marcel Maeder. Using chemometrics tools to gain detailed molecular information on chemical processes. *Journal of Chemometrics*, 34(7):e3207, 2020.
12. Xin Yang, Robert J Rees, William Conway, Graeme Puxty, Qi Yang, and David A Winkler. Computational modeling and simulation of co2 capture by aqueous amines. *Chemical reviews*, 117(14):9524–9593, 2017.
13. Hsueh-Chien Li, Jeng-Da Chai, and Ming-Kang Tsai. Assessment of dispersion-improved exchange-correlation functionals for the simulation of co2 binding by alcoholamines. *International Journal of Quantum Chemistry*, 114(12):805–812, 2014.
14. Alexey A Orlov, Alain Valtz, Christophe Coquelet, Xavier Rozanska, Erich Wimmer, Gilles Marcou, Dragos Horvath, Bénédicte Poulain, Alexandre Varnek, and Frédérick de Meyer. Computational screening methodology identifies effective solvents for co2 capture. *Communications Chemistry*, 5(1):1–7, 2022.
15. Eugene N. Muratov, Jürgen Bajorath, Robert P. Sheridan, Igor V. Tetko, Dmitry Filimonov, Vladimir Poroikov, Tudor I. Oprea, Igor I. Baskin, Alexandre Varnek, Adrian Roitberg, Olexandr Isayev, Stefano Curtalolo, Denis Fourches, Yoram Cohen, Alan Aspuru-Guzik, David A. Winkler, Dimitris Agrafiotis, Artem Cherkasov, and Alexander Tropsha. QSAR without borders. 49(11):3525–3564, 2020. doi: 10.1039/d0cs00098a.
16. Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin a. Persson. The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. ISSN 2166532X. doi: 10.1063/1.4812323.
17. Graeme Puxty, Robert Rowland, Andrew Allport, Qi Yang, Mark Bown, Robert Burns, Marcel Maeder, and Moetaz Attalla. Carbon dioxide postcombustion capture: a novel screening study of the carbon dioxide absorption performance of 76 amines. *Environmental science & technology*, 43(16):6427–6433, 2009.
18. Prachi Singh, John PM Niederer, and Geert F Versteeg. Structure and activity relationships for amine based co2 absorbents—i. *International Journal of Greenhouse Gas Control*, 1(1):5–10, 2007.
19. Prachi Singh, John PM Niederer, and Geert F Versteeg. Structure and activity relationships for amine-based co2 absorbents-ii. *Chemical Engineering Research and Design*, 87(2):135–144, 2009.
20. Young Eun Kim, Soung Hee Yun, Jeong Ho Choi, Sung Chan Nam, Sung Youl Park, Soon Kwan Jeong, and Yeo Il Yoon. Comparison of the co2 absorption characteristics of aqueous solutions of diamines: absorption capacity, specific heat capacity, and heat of absorption. *Energy & Fuels*, 29(4):2582–2590, 2015.
21. Firoz Alam Chowdhury, Hidetaka Yamada, Takayuki Higashii, Kazuya Goto, and Masami Onoda. Co2 capture by tertiary amine absorbents: a performance comparison study. *Industrial & engineering chemistry research*, 52(24):8323–8331, 2013.
22. Sigvart Evjen, Oda Siebke Løge, Anne Fiksdahl, and Hanna K Knuutila. Aminoalkyl-functionalized pyridines as high cyclic capacity co2 absorbents. *Energy & Fuels*, 33(10):10011–10015, 2019.
23. Ardi Hartono, Solrun Johanne Vevelstad, Arlinda Ciftja, and Hanna K Knuutila. Screening of strong bicarbonate forming solvents for co2 capture. *International Journal of Greenhouse Gas Control*, 58:201–211, 2017.
24. Bijan Rezaei, Siavash Riahi, and Ali Ebrahimpoor Gorji. Molecular investigation of amine performance in the carbon capture process: least squares support vector machine approach. *Korean Journal of Chemical Engineering*, 37(1):72–79, 2020.
25. Qi Yang, Graeme Puxty, Susan James, Mark Bown, Paul Feron, and William Conway. Toward intelligent co2 capture solvent design through experimental solvent development and amine synthesis. *Energy & Fuels*, 30(9):7503–7510, 2016.
26. Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem in 2021: new data content and improved web interfaces. *Nucleic acids research*, 49(D1):D1388–D1395, 2021.
27. Harry E Pence and Antony Williams. Chemspider: an online chemical information resource, 2010.
28. Martin Simonovsky and Joshua Meyers. DeeplyTough: Learning structural comparison of protein binding sites. *Journal of Chemical Information and Modeling*, 60(4):2356–2366, February 2020. doi: 10.1021/acs.jcim.9b00554.
29. Lucy Lu Wang and Kyle Lo. Text mining approaches for dealing with the rapidly expanding literature on COVID-19. *Briefings in Bioinformatics*, 22(2):781–799, 12 2020. ISSN 1477-4054. doi: 10.1093/bib/bbaa296.
30. Michael J. Fell. The economic impacts of open science: A rapid evidence assessment. *Publications*, 7(3):46, July 2019. doi: 10.3390/publications7030046.
31. Michael Woelfle, Piero Olliaro, and Matthew H. Todd. Open science is a research accelerator. 3(10):745–748, September 2011. doi: 10.1038/nchem.1149.
32. JL McDonagh, Tanja van Mourik, and John BO Mitchell. Predicting melting points of organic molecules: applications to aqueous solubility prediction using the general solubility equation. *Molecular informatics*, 34(11-12):715–724, 2015.
33. James L McDonagh, David S Palmer, Tanja van Mourik, and John BO Mitchell. Are the sublimation thermodynamics of organic molecules predictable? *Journal of chemical information and modeling*, 56(11):2162–2179, 2016.
34. Payel Das, Tom Sercu, Kahini Wadhawan, Inkit Padhi, Sebastian Gehrmann, Flaviu Cipcigan, Vijil Chenthamarakshan, Hendrik Strobelt, Cicero Dos Santos, Pin-Yu Chen, et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nature Biomedical Engineering*, pages 1–11, 2021.
35. GUOJING CONG, Anshul Gupta, Rodrigo Neumann Barros Ferreira, Breanndan O'Conchuir, and Maira De Bayser. Prediction of co2 adsorption in nano-pores with graph neural networks. In *AAAI Conference on Artificial Intelligence*, 2022.
36. Yi Luo, Saientan Bag, Orysia Zaremba, Adrian Cierpka, Jacopo Andreo, Stefan Wuttke, Pascal Friederich, and Manuel Tsotsalas. Mof synthesis prediction enabled by automatic data mining and machine learning. *Angewandte Chemie International Edition*, 61(19):e202200242, 2022.
37. James L McDonagh, William C Swope, Richard L Anderson, Michael A Johnston, and David J Bray. What can digitisation do for formulated product innovation and development? *Polymer International*, 70(3):248–255, 2021.
38. Greg Landrum. Rdkit: Open-source cheminformatics.
39. Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. *Journal of cheminformatics*, 3(1):1–14, 2011.
40. Ibm project photoresist. https://research.ibm.com/interactive/photoresist/. accessed 14 March 2022.
41. Vassilis Vassiliadis, Michael A Johnston, and James L McDonagh. Fast, transparent, and high-fidelity memoization cache-keys for computational workflows. In *2022 IEEE International Conference on Services Computing (SCC)*, pages 174–184. IEEE, 2022.
42. John J. Irwin, Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman. ZINC: A free tool to discover chemistry for biology. *Journal of Chemical Information and Modeling*, 52(7):1757–1768, June 2012. doi: 10.1021/ci3001277.
43. Larry Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5(1):501–532, 2018. doi: 10.1146/annurev-statistics-031017-100045.
44. Marc Offroy and Ludovic Duponchel. Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry. *Analytica Chimica Acta*, 910:1 – 11, 2016. ISSN 0003-2670. doi: https://doi.org/10.1016/j.aca.2015.12.037.
45. Gurjeet Singh, Facundo Mémoli, and Gunnar E. Carlsson. Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *SPBG*, 2007.
46. Monica Nicolau, Arnold J. Levine, and Gunnar Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1102826108.
47. Laxmi Parida, Niina Haiminen, David Haws, and Jan Suchodolski. Host trait prediction of metagenomic data for topology-based visualization. In Raja Natarajan, Gautam Barua, and Manas Ranjan Patra, editors, *Distributed Computing and Internet Technology*, pages 134–149, Cham, 2015. Springer International Publishing. ISBN 978-3-319-14977-6.
48. Jessica L. Nielson, Jesse Paquette, Aiwen W. Liu, Cristian F. Guandique, C. Amy Tovar, Tomoo Inoue, Karen-Amanda Irvine, John C. Gensel, Jennifer Kloke, Tanya C. Petrossian, Pek Y. Lum, Gunnar E. Carlsson, Geoffrey T. Manley, Wise Young, Michael S. Beattie, Jacqueline C. Bresnahan, and Adam R. Ferguson. Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nature Communications*, 6:8581,

2015.
49. Abbas H Rizvi, Pablo G Camara, Elena K Kandror, Thomas J Roberts, Ira Schieren, Tom Maniatis, and Raul Rabadan. Single-cell topological rna-seq analysis reveals insights into cellular differentiation and development. *Nature Biotechnology*, 35:551 – 560, 2017. doi: 10.1038/nbt.3854.
50. Wei Guo and Ashis G. Banerjee. Identification of key features using topological data analysis for accurate prediction of manufacturing system outputs. *Journal of Manufacturing Systems*, 43:225 – 234, 2017. ISSN 0278-6125. doi: https://doi.org/10.1016/j.jmsy.2017.02.015. High Performance Computing and Data Analytics for Cyber Manufacturing.
51. Leo Carlsson, Gunnar Carlsson, and Mikael Vejdemo-Johansson. Fibres of failure: Classifying errors in predictive processes. *CoRR*, abs/1803.00384, 2018.
52. Tamal K. Dey, Facundo Mémoli, and Yusu Wang. *Multiscale Mapper: Topological Summarization via Codomain Covers*, pages 997–1013. doi: 10.1137/1.9781611974331.ch71.
53. Youjia Zhou, Methun Kamruzzaman, Patrick Schnable, Bala Krishnamoorthy, Ananth Kalyanaraman, and Bei Wang. Pheno-mapper: An interactive toolbox for the visual exploration of phenomics data. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384506. doi: 10.1145/3459930.3469511.
54. Youjia Zhou, N. Chalapathi, Archit Rathore, Yaodong Zhao, and Bei Wang. Mapper interactive: A scalable, extendable, and interactive toolbox for the visual exploration of high-dimensional data. *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*, pages 101–110, 2021.
55. Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, June 2016. doi: 10.1007/s10107-016-1030-6.
56. Aaron Defazio, Francis R. Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *CoRR*, abs/1407.0202, 2014.
57. Hsiang-Fu Yu, Fang-Lan Huang, and Chih-Jen Lin. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1-2):41–75, November 2010. doi: 10.1007/s10994-010-5221-8.
58. Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, August 1997. doi: 10.1006/jcss.1997.1504.
59. Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class AdaBoost. *Statistics and Its Interface*, 2(3):349–360, 2009. doi: 10.4310/sii.2009.v2.n3.a8.
60. Christopher K Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.
61. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct): 2825–2830, 2011.
62. Alistair J Sterling, Stamatia Zavitsanou, Joseph Ford, and Fernanda Duarte. Selectivity in organocatalysis—from qualitative to quantitative predictive models. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, page e1518, 2021.
63. L Jay Field, Donald D MacDonald, Susan B Norton, Christopher G Ingersoll, Corinne G Severn, Dawn Smorong, and Rebekka Lindskoog. Predicting amphipod toxicity from sediment chemistry using logistic regression models. *Environmental Toxicology and Chemistry: An International Journal*, 21(9):1993–2005, 2002.
64. Jingxia Cui, Wenze Li, Chao Fang, Shunting Su, Jiaoyang Luan, Ting Gao, Lihong Hu, Yinghua Lu, and Guanhua Chen. Adaboost ensemble correction models for tddft calculated absorption energies. *Ieee Access*, 7:38397–38406, 2019.
65. Volker L Deringer, Albert P Bartók, Noam Bernstein, David M Wilkins, Michele Ceriotti, and Gábor Csányi. Gaussian process regression for materials and molecules. *Chemical Reviews*, 121(16):10073–10141, 2021.
66. James L McDonagh, Ardita Shkurti, David J Bray, Richard L Anderson, and Edward O Pyzer-Knapp. Utilizing machine learning for efficient parameterization of coarse grained molecular force fields. *Journal of chemical information and modeling*, 59(10):4278–4288, 2019.
67. Matthew J Burn and Paul LA Popelier. Creating gaussian process regression models for molecular simulations using adaptive sampling. *The Journal of Chemical Physics*, 153(5): 054111, 2020.
68. Olga Obrezanova and Matthew D Segall. Gaussian processes for classification: Qsar modeling of admet and target activity. *Journal of chemical information and modeling*, 50(6): 1053–1061, 2010.
69. Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006. doi: 10.1016/j.patrec.2005.10.010.
70. B.W. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451, October 1975. doi: 10.1016/0005-2795(75)90109-9.
71. Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.
72. James L McDonagh, Neetika Nath, Luna De Ferrari, Tanja Van Mourik, and John BO Mitchell. Uniting cheminformatics and chemical theory to predict the intrinsic aqueous solubility of crystalline druglike molecules. *Journal of chemical information and modeling*, 54 (3):844–856, 2014.
73. Hirotomo Moriwaki, Yu-Shi Tian, Norihito Kawashita, and Tatsuya Takagi. Mordred: a molecular descriptor calculator. *Journal of Cheminformatics*, 10(1), February 2018. doi: 10.1186/s13321-018-0258-y.
74. H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5 (2):107–113, May 1965. doi: 10.1021/c160017a018.
75. Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, September 2002. doi: 10.1021/ci010132r.
76. Alice Capecchi, Daniel Probst, and Jean-Louis Reymond. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*, 12(1), June 2020. doi: 10.1186/s13321-020-00445-4.
77. Dask Development Team. *Dask: Library for dynamic task scheduling*, 2016.
78. Ridha Ben Said, Joel Motaka Kolle, Khaled Essalah, Bahoueddine Tangour, and Abdelhamid Sayari. A unified approach to co2–amine reaction mechanisms. *ACS omega*, 5(40): 26125–26133, 2020.
79. Saeed Danaei Kenarsari, Dali Yang, Guodong Jiang, Suojiang Zhang, Jianji Wang, Armistead G Russell, Qiang Wei, and Maohong Fan. Review of recent advances in carbon dioxide separation and capture. *Rsc Advances*, 3(45):22739–22773, 2013.
80. Matthew Rocklin. Dask: Parallel computation with blocked algorithms and task scheduling. In Kathryn Huff and James Bergstra, editors, *Proceedings of the 14th Python in Science Conference*, pages 130 – 136, 2015.
81. Gilles Richner. Promoting co2 absorption in aqueous amines with benzylamine. *Energy Procedia*, 37:423–430, 2013.
82. William Conway, Yaser Beyad, Gilles Richner, Graeme Puxty, and Paul Feron. Rapid co2 absorption into aqueous benzylamine (bza) solutions and its formulations with monoethanolamine (mea), and 2-amino-2-methyl-1-propanol (amp) as components for post combustion capture processes. *Chemical Engineering Journal*, 264:954–961, 2015.
83. Jonathan GM Conn, James W Carter, Justin JA Conn, Vigneshwari Subramanian, Andrew Baxter, Ola Engkvist, Antonio Llinas, Ekaterina L Ratkova, Stephen D Pickett, James L McDonagh, et al. Blinded predictions and post hoc analysis of the second solubility challenge data: Exploring training data and feature set selection for machine and deep learning models. *Journal of Chemical Information and Modeling*, 2023.
84. Thomas MJ Fruchterman and Edward M Reingold. Graph drawing by force-directed placement. *Softw. Pract. Exp.*, 21(11):1129–1164, 1991.
85. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. 16:321–357, June 2002. doi: 10.1613/jair.953.
86. Hirotomo Moriwaki, Yu-Shi Tian, Norihito Kawashita, and Tatsuya Takagi. Descriptor list, Accessed: 28/10/2021.
87. Sebastian Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*, 3 (24), April 2018. doi: 10.21105/joss.00638.
88. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
89. Mdl information systems, inc., 14600 catalina street, san leandro, ca 94577.
90. Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280, 2002.
91. Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.