Methylation regioselectivity prediction through machine learning

Eros Reij

Supervisors:

Aalt-Jan van Dijk, Marnix Medema, David Meijer

Abstract

Natural products, biosynthesized by plants and microbes, exhibit diverse structures and functionalities, playing key roles in various fields. These compounds often undergo complex biosynthetic pathways involving tailoring enzymes. One such tailoring enzyme group are the methyltransferases, which methylate molecular scaffolds with high regioselectivity. Though methylation is a minor modification, it significantly influences the physiochemical properties of the compound. Here, we explore the possibility of predicting methylation sites in natural products by integrating molecular and protein data. We employed a random-forest classifier alongside various molecular and sequence-based protein featurization methods. Additionally, we curated a new dataset of entries outside of traditional databases, including reactions documented in literature. This literature-based data collection offers significant potential for expanding dataset size and gaining valuable insights. Our models demonstrated strong atom-level predictions in model training but struggled at predicting sites outside of the scope of the training data. Future efforts with integration of molecular and protein data shows potential allow for accurate prediction of methylation sites.

Introduction

In nature, a rich variety of specialized metabolites is biosynthesized by plants and microbes, commonly referred to as natural products. Attributed to their varying levels of structural complexity and scaffold diversity (Atanasov et al., 2021), natural products convey a broad spectrum of functionality such as endogenous defence mechanisms and interactions with other organisms (Sparks et al., 2017). This functional diversity of natural products and their derivatives makes them a multipotential source of valuable compounds. For instance, applicable as agrochemicals (Sparks et al., 2017) and antibiotics (Cragg & Newman, 2013).

While certain synthesis pathways depend on a single core enzyme for substrate modification, others, which are more complex, operate similarly to assembly lines. These pathways, include non-ribosomal peptide synthases, as well as various enzyme complexes found within biosynthetic gene clusters (Medema et al., 2015). In these cases, a compound is formed modularly. Initially, a molecule scaffold is formed followed by, covalent modifications on the scaffold by tailoring enzymes (Walsh, 2023). A distinctive property of tailoring enzymes is their regioselectivity (Mu et al., 2020).-Regioselectivity is the property of enzymes to catalyse a reaction at a specific site on a substrate molecule. Tailoring enzymes therefore produce specific structural isomers based on their structure.

One such tailoring reaction is methylation. Methyltransferases substitute a hydrogen atom with a methyl group (figure 1) Methyltransferases exclusively replace hydrogen atoms on nucleophilic atoms with carbon, nitrogen, oxygen, and sulphur atoms being the most common targets (Abdelraheem et al., 2022). Methylation serves as a pivotal process in epigenetics by regulating gene expression (Moore et al., 2013) and plays an important role in the modification of natural products.

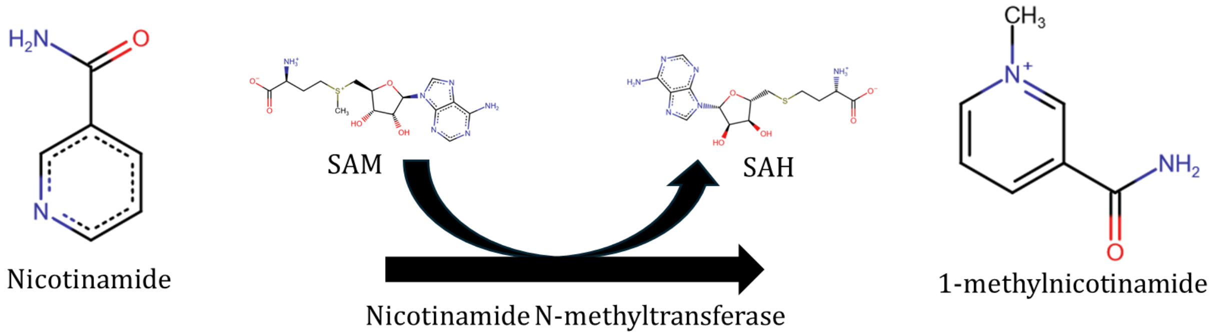


Figure 1: Schematic methylation of Nicotinamide. Schematic visualization of the methylation of Nicotinamide to 1-methylnicotinamide by Nicotinamide N-methyltransferase. The cofactor S-Adenosyl-methionine (SAM) donates a methyl group, and is converted into S-Adenosyl homocysteine (SAH).

While methylation is a minor modification of the compound, it is critical to the physiochemical properties of the resulting product. Similar pairs of compounds may exhibit large differences in effectiveness (Stumpfe et al., 2019). There are several mechanisms through which methylation exerts its effects in a biological system. For example, compound methylation impacts the molecular conformation and its physicochemical properties. These conformational changes can directly lead to increased potency in receptor binding or influence compound solubility(Pinheiro et al., 2023). Substrate methylation can increase its binding affinity with a target protein a thousand-fold (Leung et al., 2012).-

Click or tap here to enter text.Machine learning techniques have already been applied for prediction of substrate regioselectivity, such as the prediction of regioselectivity in electrophilic aromatic substitutions (Ree et al., 2021) as well as prediction of enzyme interaction sites with drug-like molecules (Öeren et al., 2022). These methodologies promising results in precisely substrate regioselectivity, attaining high accuracy whilst using only on compounds data without factoring in potential enzyme influences. This indicates the potential for employing machine learning to predict enzymatic methylation by integrating molecular and protein data.

In a previous in-house study by Yingjie Shao, a random-forest classifier model was developed for predicting natural product methylation (Shao, 2023).The modelling strategy aimed to predict the methylation state on a per-atom basis within a substrate, using both substrate and protein sequence inputs, with promising results**.** Here, we further explore the potential of machine learning by exploring various molecular and enzyme featurization techniques to predict the site of methylation in natural products. Should accurate prediction become achievable, it could enable the screening of genomes for potential valuable natural products based solely on sequence and compound data. A prospect that holds significant promise for advancing drug development efforts.

Methods

In this study, we opted for an approach that ensures data quality by only using reviewed sources, though this inherently limits the available data. Building on previously reported success in a the house study (Shao, 2023), we utilized Random Forest-based modelling as our sole modelling method. However, using Random Forest introduces the limitation of requiring fixed-length featurization approaches. This requirement constrains the range of applicable featurization methods, making the modelling process dependent on creating fixed-length feature vectors.

Data collection

In order to ensure quality of the reactions used in model training, curated reactions were retrieved from the Rhea database (Bansal et al., 2022). Reactions catalysed by methyltransferases (ENZYME class 2.1.1.-) were retrieved from the Rhea database (Release 127). Collected reactions targeting nucleotides and transport reactions were filtered out in order to retrieve only reactions on metabolites. Associated reactants of collected reactions were retrieved from the ChEBI database (Hastings et al., 2016) (Release 221) as Simplified Molecular Input Line Entry System (SMILES) using Rhea identifiers. SMILES are compact representations of chemical structures, encoding molecular information in a text string format that is both human-readable and machine-interpretable. Reactants were filtered to retain substrate and product. Using the Rhea identifiers UniProtKB reviewed (Swiss-Prot) a total of 14394 enzyme sequences were downloaded from the UniProt database (Bateman et al., 2023).

To assess the generalizability of the model, a separate holdout set was created from putative methylation reactions. A list of MIBiG (Version 3.1) (Terlouw et al., 2023) biosynthetic gene clusters containing at least one putative methyltransferase and associated compound and possible methylation sites was provided. Subsequently sites were validated by their associated literature and similar proteins where possible. Validation was done using a self-defined rule based decision flow (appendix A.1). Selected reactions were included in a finalized list of 42 entries (appendix A.2).

Data preprocessing

Collected substrate-product SMILES pairs were filtered to only include reactions where hydrogen atoms are substituted by a methyl groups. This filtering process excluded reactions involving methyltransferases that induce more than a single substitution, such as the acetylation by perrocorin-6A synthase (appendix B.1). Fingerprinting requires molecules to be of finite lengths. Therefore reactions involving polymeric substrates of variable length were reduced to their respective monomers (appendix B.2). This action was taken to ensure their inclusion in the dataset.

Substrate and product molecules were filtered from each reaction, and the relative sites of methylation on substrates were identified using RDKit (Release_2023.03.3)(Landrum et al., 2023). SMILES representations of the molecules were canonicalized to ensure consistency, indexed, and then searched to pinpoint the specific sites of methylation on the substrate. Methylation sites that could not be computably determined were manually reviewed and adjusted for inclusion or exclusion, particularly in cases where there was a difference in charge between the substrate and product. Consequently, the SMILES representations were corrected for substructure matching.

A single reaction can be catalysed by a multitude of similar proteins. To counteract the imbalance caused by more proteins associated with each reaction compared to molecules, proteins for each reaction were randomly subsampled to a maximum of five resulting in a total of 2454 used proteins and 187 unique reactions.

Atom environment fingerprinting

Substrate molecules were fragmented to local environments by creating fragments taking each heavy atom as a centre using RDKit (Release_2023.03.3)(Landrum et al., 2023). Local environments consist of subgraphs of molecules using each heavy atom and a range of connected heavy atoms similar to the fingerprinting radius. This results in in a collection of local environments equal to the amount of heavy atoms in a molecule. Molecules fragmented with a two extended connections, yields fragments equivalent to the size of the Morgan fingerprint. Additionally, an extended connectivity of four was chosen to allow for less dense feature vectors (Appendix C.1).With larger fragment sizes, there is a greater diversity in substructures, resulting in fewer non-zero elements in the fingerprint.

Resulting fragments were labelled if the centre atom is a site of methylation. Subsequently extended connectivity fingerprints (ECFP), functional-class fingerprints (FCFP) and MinHashed atom-pair fingerprints (Map4) were generated for each fragment. Molecular fingerprinting using ECFP and FCFP was conducted with RDKit. Map4 fingerprints were generated using map4 (version 1.0) (Capecchi et al., 2020). Similarly to the sizes of the local environments, different sizes were used for fingerprinting these environments. Local environments with two extended connections were fingerprinted with a radius of two. Local environments with four extended connections were fingerprinted with radii of both two and four. These varying sizes allowed for different levels of environmental sparseness. All fingerprints were folded in bit strings of 2048 elements. Larger bit strings length decreases the likelihood of bit collision, and therefore decreases the information loss at the cost of increased computation time and storage space.

Full molecule fingerprinting

Atom environments do not take the full structure of the molecules they originate from into consideration. Therefore, an additional featurisation strategy was employed. Substrate molecules were collected, and new molecules were generated by replacing hydrogen atoms with methyl groups, ensuring chemical validity. Newly generated molecules were then matched against the known products, with matches labelled as true positives and unmatched sites considered negatives. Similarly, the molecules were fingerprinted using ECFP, FCFP, and MAP4 fingerprinting methods with radii of two, three, and four, resulting in nine sets of fingerprints. Due to the larger fragment size in this method, a broader range of fingerprint sizes was employed.

Multiple sequence alignment based encoding

Since Random Forest requires fixed-length features, MSA (Multiple Sequence Alignment) was chosen to enforce equal length among the sequences used. Additionally, MSA maximizes information retention by introducing gaps without losing any data from the sequences. Collected Sequences were aligned using MAFFT (version 7.511) (Katoh & Standley, 2013) using the default BLOSUM62 scoring matrix. One hot encoding using amino acid identity was done to represent sequences to binary a vector, allowing for the preservation of sequence information while maintaining the fixed-length.

AAindex (Release 9.2) (Kawashima et al., 1999) was used to retrieve values for charge, polarity, hydrophobicity, and size in Ångström (appendix D). Collected AAindex values were used to categorize amino acids based on threshold values (appendix D), resulting in an encoding similar to one-hot encoding with the distinction that closely related amino acids, such as leucine and isoleucine, were encoded the same . Collected AAindex values were also used as direct numerical encodings for amino acids offering denser feature representations, although they are not directly interpretable.

ProteinBert embedding of protein sequences

Low protein counts associated with certain types of reactions can pose challenges for training predictive models. For example, there are only 38 sulphur methyltransferase proteins in the collected data. ProteinBert, a deep-language model for protein sequences, can be used to create equal-sized fingerprints of protein sequences of different lengths. (Brandes et al., 2022). Embeddings were generated of the collected and filtered methyltransferase sequences using ProteinBert (version 1.0.0). The non fine-tuned pretrained model, pretrained on ~106M proteins derived from UniProtKB/UniRef90, was used to generate local embeddings. Sequences were embedded to a length of 1562, two longer than the longest sequence in the training data, with all other parameters kept at their default values. Local representations were collected, and subsequently reduced through global average pooling and layers flattened to a single 1562 feature vector.

Data preparation

Random Forest requires fixed-length tabular data. Therefore, we concatenated protein and molecular representations. Though other combination methods are possible, concatenation allows for simple implementation, interpretability and consistent feature length. In total there were four featurization methods for proteins, three for local environments, and four for full molecule fingerprints. Concatenating each unique protein-molecule representation a total of 84 unique combinations were tested. Of which 36 combinations were local environment based and 48 combinations were based on full molecule fingerprints (appendix E).

Prior to modelling splits were generated of the training data for five-fold cross validation. Splits were stratified on the atom types that are methylated to ensure equal representation of each atom type across cross folds. Additionally entries sharing the same reaction were grouped, as to not have reactions being present in both the train and validation set. For each fold of, training data was min-max normalized using the scikit-learn (version 1.3.0) (Pedregosa et al., 2011)

MinMaxScaler. Scaling was applied to all combinations of features, this is only relevant in the combinations where values were used outside of the binary encoding.

Scaling the ProteinBert embeddings and AAindex property values ensures that their ranges are between 0 and 1. While this scaling may not directly impact Random Forest models, it enables the featurization to potentially be utilized in other distance-based prediction models. The normalization transformers obtained from this process were then used to normalize the test data of each fold, ensuring consistency in data processing across all folds.

Modelling

Due to the extensive number of feature combinations and the previously reported performance in the house study by Yingjie Shao (Shao, 2023)we opted for the random forest modelling approach. Scikit-learn (version 1.3.0) was used for model generation. Random forest models were made for each split and each concatenated feature combination. For the local environments, each environment was predicted if the center atom either methylated or not. For the combinations containing the full molecule featurizations, it was predicted whether the molecule was identical to the product (appendix C).

Due to the predictions being imbalanced, as there is a multitude more non-methylated sites compared to actual methylated sites, balanced class weights were used. Default settings were used for all other parameters. The best performing model of each molecule featurization strategy was saved and subsequently used to predict the putative methylation reactions collected from MIBiG.

Model evaluation

As the models predict methylation based on atom by atom basis, assessing on accuracy alone is insufficient. Predictions performance per atom is assessed by using the macro average F1 score consideration of the inherent class imbalance in the dataset. The standard deviation of F1 scores was computed to assess the variability of model performance across different splits. To estimate the applicability of the models in a broader context, we aggregated all atom-level predictions per reaction. These predictions were labelled as either correct, if all predictions were correct, or incorrect if one or more were mislabelled. Although this approach is conservative, it offers insight into the current applicability of the generated models. To further evaluate the performance of our models, we generated ROC-AUC curves for the predictions of the final two models on the putative methylation reactions collected from MIBiG.

Feature dimensionality reduction

To contextualize the performance on the putative methylation reactions collected from MIBiG, we extracted all labelled features (enzyme, atom) combinations from the training and test sets. Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) was used to visualize the entries. UMAP is chosen for handling highly dimensional data because it excels in preserving both local and global structures while reducing dimensionality. It works by constructing a low-dimensional representation of data points by optimizing a topological representation of the data. 15 nearest data points to each point were considered when constructing the local neighbourhood graph. This is a moderate setting, balancing between capturing global and local structures. Data was reduced to two dimensions for visualization purposes. UMAP embeddings were visualized and labelled according to atom type and inclusion in train or final holdout set. For a detailed explanation of the algorithm, see the paper on UMAP by McInnes et al. (2018).

Feature importance

For interpretation of the performance on the final holdout set, feature importance was extracted from the two best performing models during the training cross-validation. The concatenated features were separated back into their respective molecule and protein components. The respective MSA importances, spread across multiple indices due to one-hot encoding, were mapped back to their original indices. The feature importances, now aligned with their respective indices in the concatenated feature vector, were visualized.

Code and packages used.

All figures were created using R, and analyses were performed in Python. Resources were collected from publicly available sources. An exhaustive list of used packages, package versions, scripts, and data acquisition details are available on the project's GitHub repository: https://github.com/Eros-R/thesis_rsmt

Results and Discussion

Strong atom-level predictions, weak molecule-level performance in training data.

In total, 84 different combinations of molecule and substrate featurization were explored. To assess the quality of the various featurization strategies, out-of-the-box performance was evaluated across a five-fold split using random forest modelling. Macro F1 scores of predictions were calculated for each site (Figure 2).

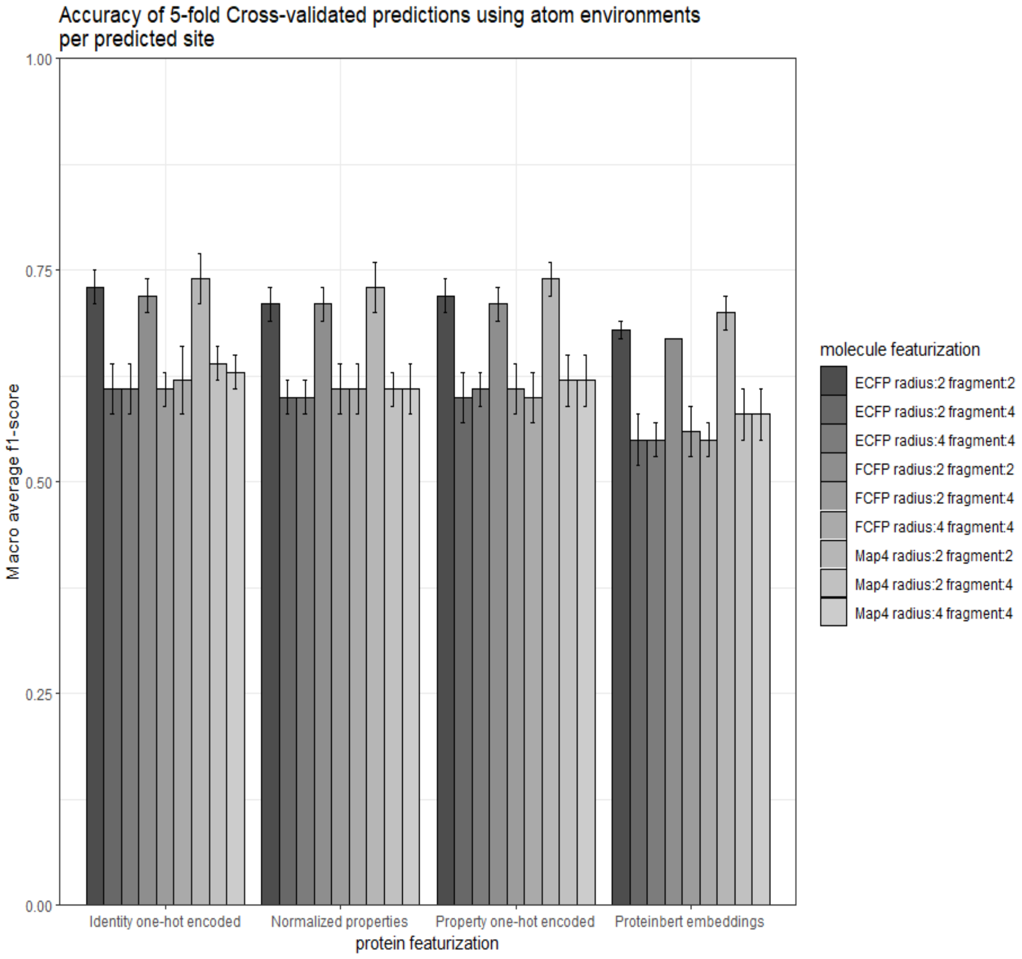


Figure 2: Macro average f1-score of site level predictions of the local environment based featurization strategy.

The best performance was observed on atom environments using small radii (Figure 2). Map4-based fingerprinting consistently yielded the best results. Full molecule featurization predictions occasionally outperform the environment based method. However, environment based generally outperform these (appendix F).

In Map4 global information is conserved by using atom-pairs combined with features from circular substructures, which likely has more variety than the ECFP with the same radius .In MAP4, the global structure is preserved up to the fragment radius, allowing more information to be stored within the same number of bits. This enhances the ability to differentiate between similar fragments (Capecchi et al., 2020).

Relatively little difference was noted between protein featurization methods, with exception of the ProteinBert embeddings, which performed the worst. However this lack of performance of the embeddings could be attributed to the lack of fine-tuning.

Overall, combinations involving small local environments outperformed those with larger fragments and fingerprint radii. The better performance of small environments is likely due to two factors: large environments result in more similar substructures due to greater overlap. Using large environments there is also the possibility that it encompasses the entire molecule, which is nearly the case with nicotinamide (appendix C). Most errors in the predictions occurred in precision and recall at true methylation sites (Appendig G). Larger fragments particularly suffered from low recall, suggesting overfitting because they capture too much redundant information. Interesting is the nearly identical performance performance between local environments with a radius of 4 and 2. This is likely caused by that increasing fragment size does not necessarily yield more unique substructures. In highly symmetrical molecules, multiple identical substructures can be derived from different centre atoms, leading to identical fingerprints. So in addition to encompassing the entire molecule, it is also difficult to use highly symmetrical molecules.

To bring the site-level predictions in context, all predictions associated with a single protein-reaction combination were aggregated. If one prediction in these aggregated results was wrong the entire reaction was flagged as wrong. Penalized accuracy scores of both fragment based and molecule based featurizations are displayed in figure 3. Clear is the discrepancy between methods, as well as the low accuracy of all methods. Though low, simulating using random choice of methylation given range results in an accuracy around ~0.05 (appendix G).

It suffices that though some feature combinations perform better than random guessing, given the results on site levels, the models lack sufficient predictive properties to be usable even in the cross-validation. Similar to the site based prediction, the small fragments are the best. Notable is the smaller differences in accuracy between fragment sizes given the permuted molecules. This is caused by the high rate of similarity between predictions, as molecular features are nearly identical, only being one methyl group apart.

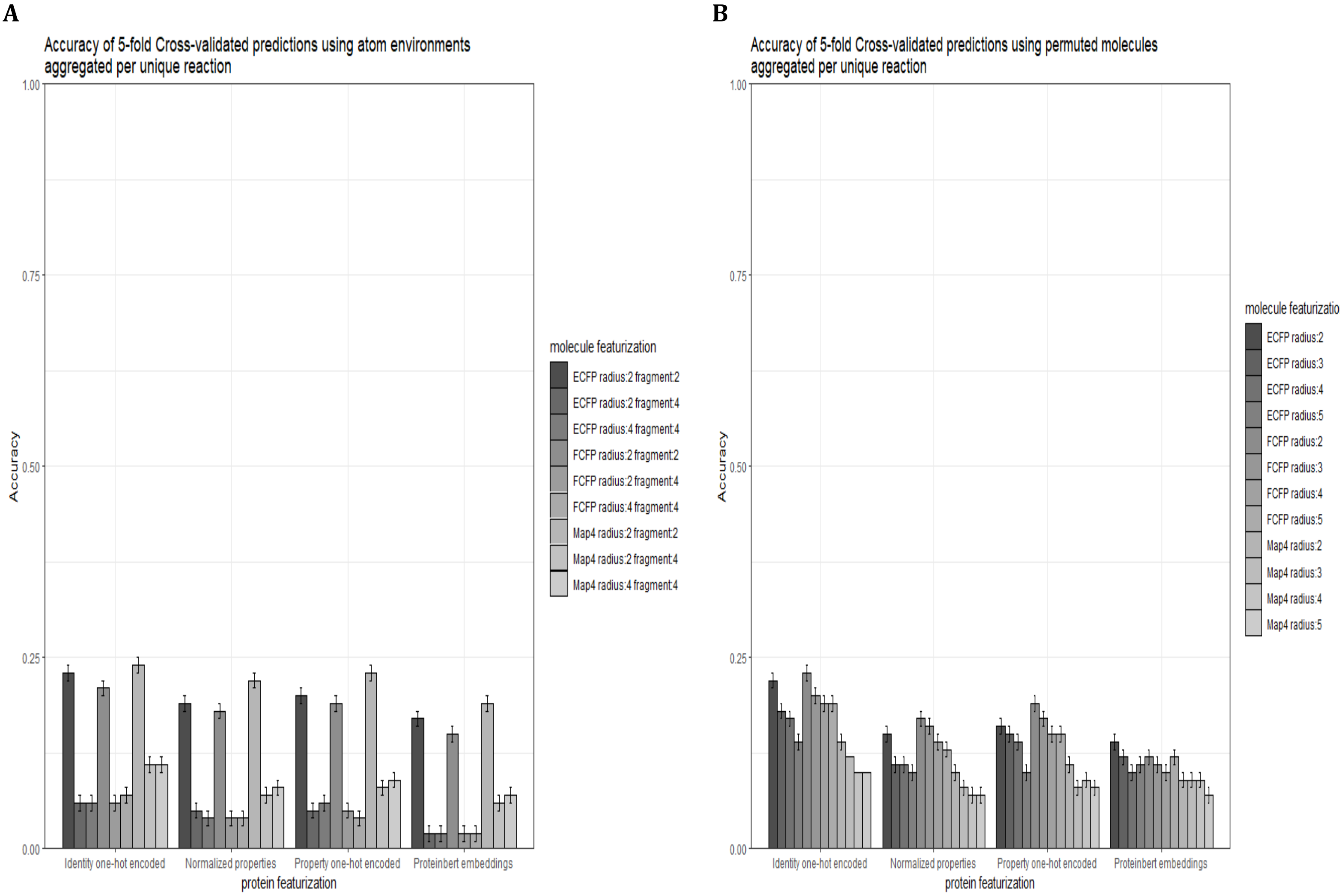


Figure 3: Accuracy scores of aggregated predictions across different featurization strategies. Site based predictions associated with a reaction were collected. If all predictions were correct, they were labelled as true, otherwise treated as a negative. A). aggregated scores for fragment based featurization. B) aggregated scores for permuted molecule based featurization.

Poor generalizability of modelling strategy.

The two top performing models were used to predict a collection of putative reactions from MIBiG. One fragment based using Map4 fingerprints with a radius of 2, and one based on the permuted molecules using EFCP fingerprints with a radius of 2. Both of these models used the one-hot encoded MSA representation. Predictions were made, and precision and recall were calculated for assessing the performance of the models (figure 4).

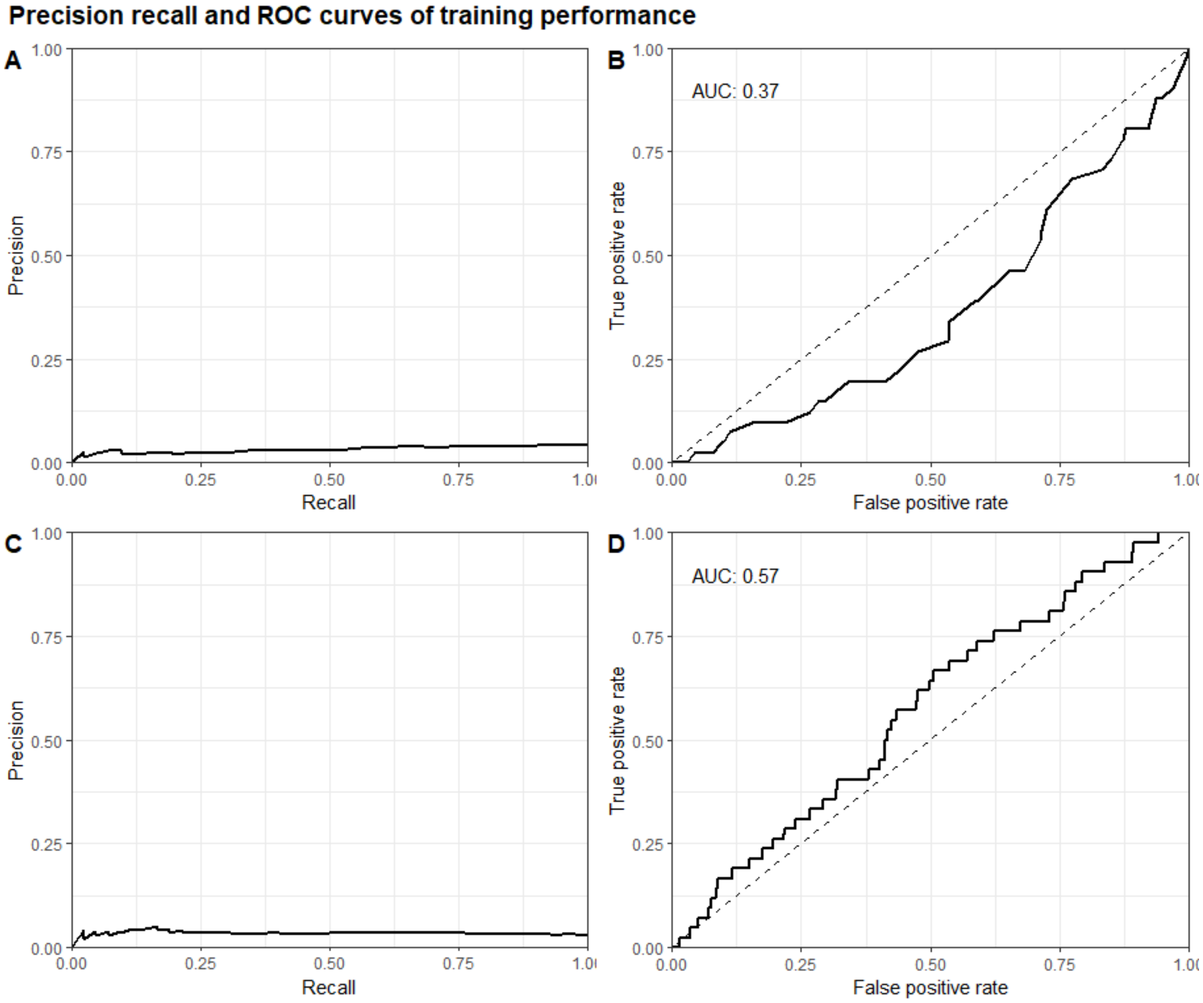


Figure 4: ROC-AUC and precision recall curves.. A) Precision recall curve of the holdout predictions using radius 2 Map4-fingerprints. B) ROC-AUC curve of the holdout predictions using radius 2 Map4 fingerprints. C) Precision recall curve of the holdout predictions using radius 2 ECFP-fingerprints on permuted molecules. B) ROC-AUC curve of the holdout predictions using radius 2 ECFP fingerprints on the permuted molecules.

It is interesting to note that the permutation based model, which performs worse in cross-validation, still outperforms the fragment-based method. However, this improvement is only marginally better than random guessing. In retrospect, a high rate of overfitting is expected given the complexity and shallowness of the modelling strategy, as the random forest models were used out of the box with no additional hyperparameter tuning.

The nature of the training data, characterized the large number of features given the MSA compared to the amount of features makes overfitting a likely possibility. Modelling was used out-of-the-box, without finetuning. and this, combined with the large number of features, likely led to the interpretation of consistent noise as strong predictors. To combat this overfitting, one approach would be to use features that aptly represent the highly dimensional data in fixed lengths when using Random Forest. Alternatively, other modelling strategies that do not rely on fixed-length features could be considered.

UMAP representation of featurized entries

In order to contextualize this poor generalizability, featurized holdout data was visualized using UMAP for both tested models (figure 5).

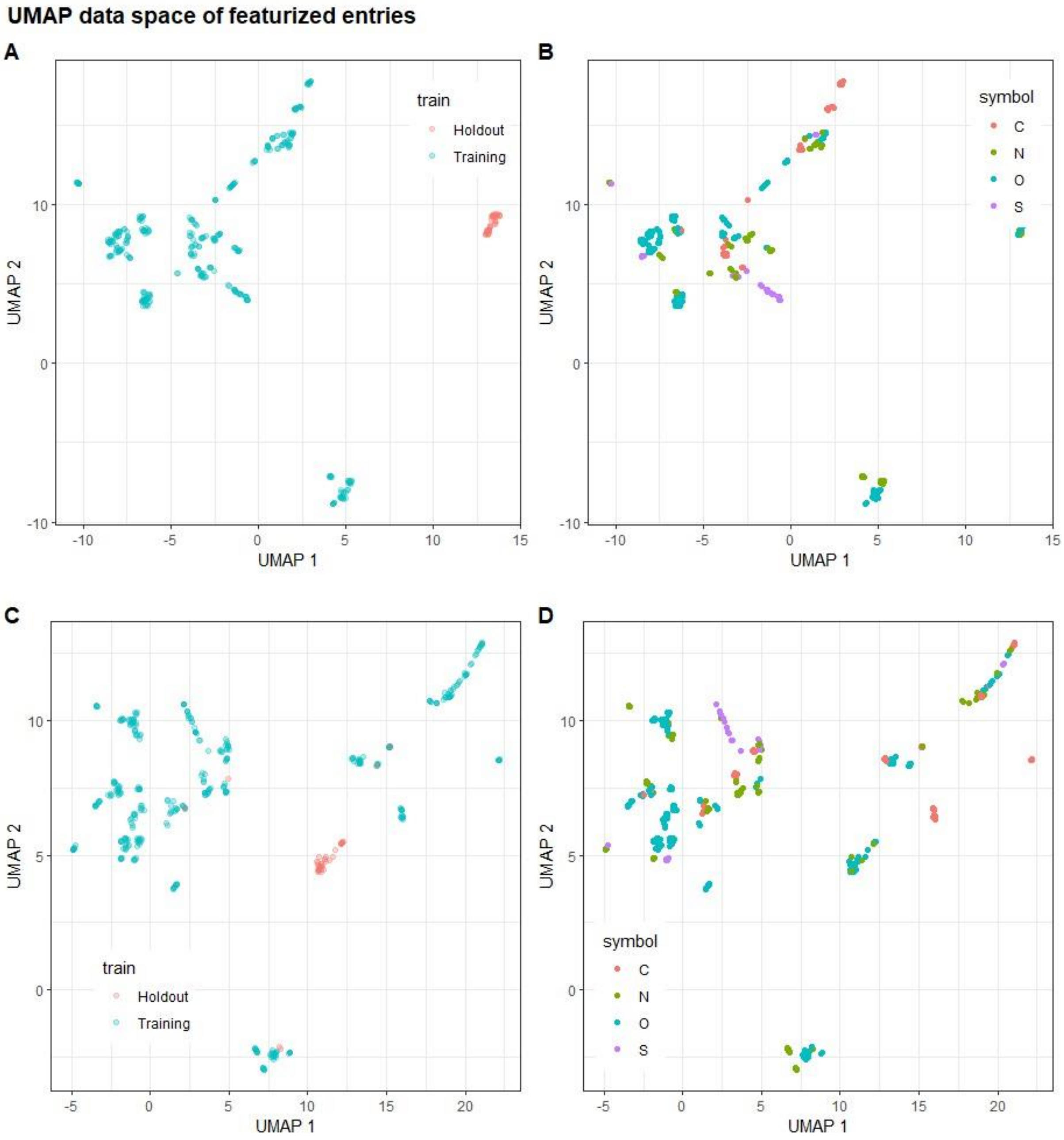


Figure 5: UMAP of featurized entries using the featurization methods of the best performing models using fragment based (A&B) and permutation based (C&D) fingerprinting methods. Both methods used one-hot encoded MSA representations.

Though this is a low space representation of large feature vectors, some clustering patterns can be observed. Despite different substrates, proteins, and methylated atoms, isolated clusters of atom of methylation do not always cluster together. The isolated cluster of the holdout data is particularly noteworthy, as this isolation likely explains the poor performance of the holdout set. This raises the question of whether the feature space is properly representable in the context of the problem.

Caution should be taken interpreting these representations. As fewer samples are used, the local and global structures of the data may be represented differently in the lower-dimensional space. This can lead to variations in how clusters are formed and visualized, affecting the overall interpretation of the data's structure and relationships. This is apparent when comparing the UMAP of entries used in the model (figure 5) with a UMAP using identical settings on all available data without down sampling to five proteins maximum (appendix H). In the actual feature space these may not be as similar as they seem in the plot. Though not explored in the scope of this project it could be that the clustering of functionally different methyltransferases are related to the evolutionary origin of the protein.

In contrast to DNA-methyltransferases, which possess conserved motifs for recognition of common features of the macromolecular substrates, methyltransferases of natural products do not possess widely conserved structural features associated with substrate recognition (Liscombe et al., 2012). Assuming this, the approach of employing global sequence alignment is likely to greatly inflate the feature vector, due to the high variability of the proteins, and possibly the inconsistent protein architecture on sequence level.

Wong & Erin-Lopez (Wong & Eirin-Lopez, 2021) describe that, in animals, methyltransferase like proteins lineages likely originated from single independent ancestors which suggests that they were subject to strong selective constraints driving its structural and/or functional specialization. Though this is in animals, the data collected entails all kingdoms, so it is possible that the functional class of methyltransferases are overgeneralized in this context.

Limitations in random forest prediction of methylation regioselectivity

In order to further analyse the models' performance in context of the poor performance, feature importances of the models used for the final tests were extracted (figure 6). In contrast to what the cross validation seemed to imply, the majority of the prediction leverages the featurized protein. Observable clusters of important features are evident (figure 6) near the tail ends of the MSA. Manual inspection of sequences revealed low occupancy at important sites (appendix I), further reinforcing the notion that the model interprets noise as significant predictors.

Feature extraction

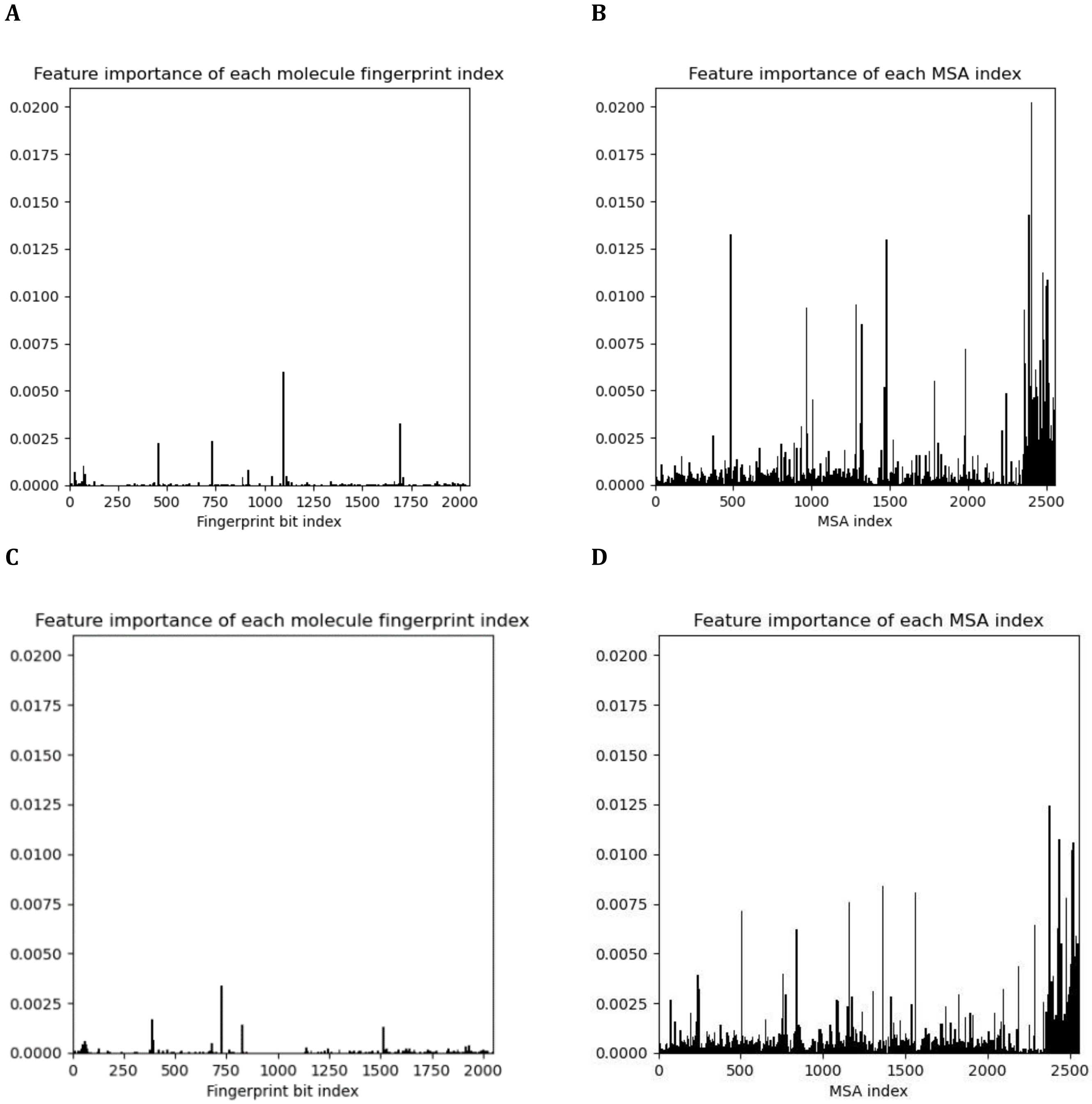


Figure 6: Feature importances of best performing models using fragment based (A&B) and permutation based (C&D) fingerprinting methods. One-hot encoded MSA features were aggregated per 20.

Concatenation of features was performed for east of implementation and based on previously reported results (IN HOUSE). However, this process likely introduced a significant feature bias towards the proportionally larger protein representation, which likely caused overfitting. In addition with the concatenation, both substrate and protein are treated equally. Intuitively, the exact amino acids should be less important if the functional property is preserved, as is the case with multiple proteins associated with a single reaction. In contrast, molecules remain static within the context of the chemical reaction.

Moreover, the results using these concatenated feature vectors of different data types could be misleading, as suboptimal predictor variables may be artificially preferred in variable selection (Strobl et al., 2007). Additionally, the extrapolation capabilities of tree-based learners are generally poor, although this aspect was not explored in this context.

The data availability is also a strong factor in the models' performance. All reactions are retrieved from Rhea, which by design only hosts curated reactions, resulting in limited availability of methylation reactions, particularly in the context of natural products. For instance, the methylation of halides in *A. thaliana* has long been known for the enzyme's promiscuity towards different halides(Schmidberger et al., 2010). However, the Rhea reaction (RHEA: 28014) only assumes the methylation of thiocyanate.

While compared to the fragment based approach the small fragments outperform in the cross-validation, there are still some important caveats to consider. A drawback of using small environments is that this strategy allows for identical environments to be associated with different labels. Given the 1082 unique fragments with a radius of 2 that can be generated from the collected substrates, 74 atom environments share an identical SMILES with differing labels.

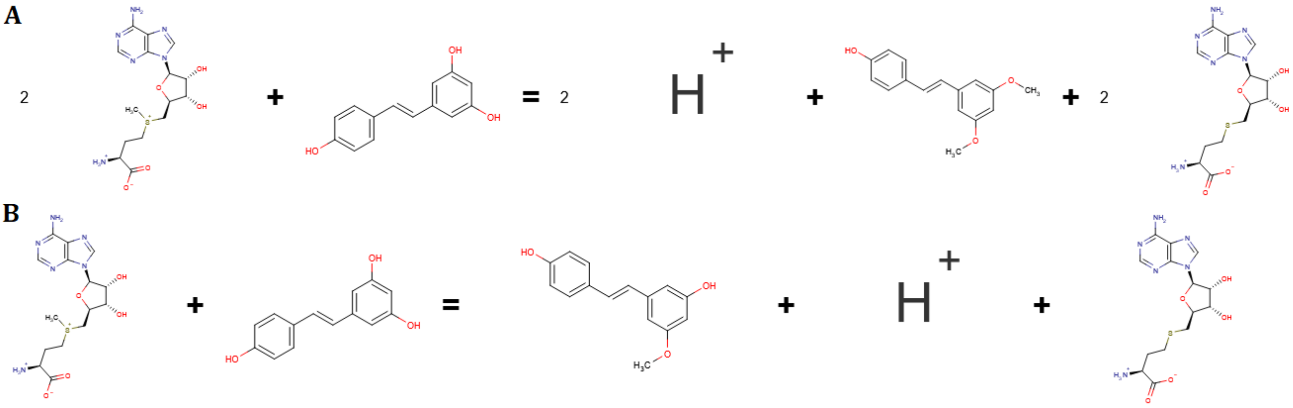


Figure 7: Two different Rhea reactions of biosynthesis of pterostilbene in grapevine. (A) and European hops (B). Though reactions are nearly identical, seuqences differ higlighting the importance of the proteins.

In addition to this, when looking at the indexed associated index, this drops to 38 cases where fragments share a label and protein, yet differ in labels. For example, in the case of pterostilbene in grapevine methylation (Figure 7) parts of the substrate molecule, highlighting the importance of the inclusion of proteins to deal with some of the complex nuances of methyltransferase.

Part of this indistinguishability may also be caused by the fingerprinting algorithm, as radial fingerprints are known to struggle with representing highly symmetrical compounds, as with natural products (Orsi & Reymond, 2024). A way to combat this would be to employ more contextualized representations of the molecules. One way would be to use specialized fingerprints for natural products, such as the ones proposed by Seo et al.(Seo et al., 2020). Though this comes with its own challenges, as this excludes the possibility of the fragment environments, likely causing overfitting. Additionally this would limit the featurization by observed states. Unobserved structures would not be considered, which could limit the generalizability of the model.

Conclusion

In this project, we explored the possibility of predicting natural product methyltransferase regioselectivity using random forest. A dataset of curated natural product reactions containing sequence and reactant data was compiled, in addition to a collection of putative reactions from MIBiG. Several feature combinations were explored, yet these efforts yielded mediocre results on cross-validation data. Furthermore, the tested approaches appear insufficient for predicting the regioselectivity of methyltransferases on unseen data. The relatively poor-performing model outperforms random guessing. However, this form of guessing also considers any heavy atom, whereas sequence information could enable a more informed approach. For instance, leveraging protein domains could help identify the specific atom type that is methylated. Filtering of data is essential, but this results in less usable data. Prioritizing the reduction of overfitting may enhance model performance if this approach is continued; however, it is likely better to consider alternative methodologies.

The claim by Liscombe et al. (2012) that natural products lack widely conserved structural features associated with substrate recognition is reinforced by the clustering of featurized entries in low dimensions. This indicates that MSA-based approaches overlook the inherent complexity of methyltransferase proteins. Each tested molecular representation has its merits and drawbacks, with small molecular fragments appearing to offer the best strategy, though permuted fragments may prove to be more generalizable. It is important to note that due to the poor performance the optimal method is unclear. Furthermore, concatenation of features disregards the inherent qualities of molecules and proteins in relation to their reactions.

To reduce noise in protein features induced by the MSA, a new approach could be considered where enzymes are not be treated as a single class but rather divided into subfamilies based on similarity. This approach would necessitate developing numerous specialized prediction models. Which in turn results in less available data overall.

Additionally to improve future modelling efforts, employing active learning can make the completeness of the feature space more clear and provide insights in the completeness. There should be more room for tuning hyperparameters to optimize model performance, especially to take caution for overfitting. Considering the complexity of each class, it may be more beneficial to differentiate between similar methyltransferases, potentially including the organism of origin or the presence of specific domains. Then more precise and less generalized models could be made for more accurate and contextualized predictions. However, this could limit the generalizability of the model due to reduced data availability for each class. Untuned ProteinBERT embeddings were used in this study, and it may be worthwhile to explore this approach more thoroughly. Including structural data and fine-tuning models could be a valuable method to explore. An end-to-end model using graph-represented molecules may offer a more comprehensive prediction at the molecular level.

The scarcity of exhaustive resources regarding natural product synthesis increases the difficulty of developing any predictive model, especially given the complexity of methyltransferases as a broad class. Here we showed that additional data can be sourced from public sources, where we were able to include 42 methyl transferase reactions reasonable certainty, compared to the 187 entries from professionally curated sources. Literature-based data collection offers significant potential for expanding dataset size and gaining valuable insights. In addition to this, it would be beneficial to establish an interlinked, database containing reactants and associated resources, including those of putative reactions.

Data and Code

All figures were created using R, and analyses were performed in Python. Resources were collected from publicly available sources. An exhaustive list of used packages, package versions, scripts, and data acquisition details are available on the project's GitHub repository: https://github.com/Eros-R/thesis_rsmt

References

Abdelraheem, E., Thair, B., Varela, R. F., Jockmann, E., Popadić, D., Hailes, H. C., Ward, J. M., Iribarren, A. M., Lewkowicz, E. S., Andexer, J. N., Hagedoorn, P., & Hanefeld, U. (2022). Methyltransferases: Functions and Applications. *ChemBioChem*, *23*(18). https://doi.org/10.1002/cbic.202200212

Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Research*, *28*(1), 304–305. https://doi.org/10.1093/nar/28.1.304

Bansal, P., Morgat, A., Axelsen, K. B., Muthukrishnan, V., Coudert, E., Aimo, L., Hyka-Nouspikel, N., Gasteiger, E., Kerhornou, A., Neto, T. B., Pozzato, M., Blatter, M.-C., Ignatchenko, A., Redaschi, N., & Bridge, A. (2022). Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Research*, *50*(D1), D693–D700. https://doi.org/10.1093/nar/gkab1016

Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., da Costa Gonzales, L. J., Hatton-Ellis, E., Hussein, A., Ignatchenko, A., … Zhang, J. (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, *51*(D1), D523–D531. https://doi.org/10.1093/nar/gkac1052

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., & Linial, M. (2022). ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, *38*(8), 2102–2110. https://doi.org/10.1093/bioinformatics/btac020

Capecchi, A., Probst, D., & Reymond, J.-L. (2020). One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*, *12*(1), 43. https://doi.org/10.1186/s13321-020-00445-4

Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., & Steinbeck, C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Research*, *44*(D1), D1214–D1219. https://doi.org/10.1093/nar/gkv1031

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kawashima, S., Ogata, H., & Kanehisa, M. (1999). AAindex: Amino Acid Index Database. *Nucleic Acids Research*, *27*(1), 368–369. https://doi.org/10.1093/nar/27.1.368

Landrum, G., Tosco, P., Kelley, B., Ric, Cosgrove, D., sriniker, gedeck, Vianello, R., NadineSchneider, Kawashima, E., N, D., Jones, G., Dalke, A., Cole, B., Swain, M., Turk, S., AlexanderSavelyev, Vaucher, A., Wójcikowski, M., … strets123. (2023). *rdkit/rdkit: 2023_03_3 (Q1 2023) Release*. Zenodo. https://doi.org/10.5281/zenodo.8254217

Leung, C. S., Leung, S. S. F., Tirado-Rives, J., & Jorgensen, W. L. (2012). Methyl Effects on Protein–Ligand Binding. *Journal of Medicinal Chemistry*, *55*(9), 4489–4500. https://doi.org/10.1021/jm3003697

Liscombe, D. K., Louie, G. V., & Noel, J. P. (2012). Architectures, mechanisms and molecular evolution of natural product methyltransferases. *Natural Product Reports*, *29*(10), 1238. https://doi.org/10.1039/c2np20029e

McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*.

Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., de Bruijn, I., Chooi, Y. H., Claesen, J., Coates, R. C., Cruz-Morales, P., Duddela, S., Düsterhus, S., Edwards, D. J., Fewer, D. P., Garg, N., Geiger, C., Gomez-Escribano, J. P., Greule, A., … Glöckner, F. O. (2015). Minimum Information about a Biosynthetic Gene cluster. *Nature Chemical Biology*, *11*(9), 625–631. https://doi.org/10.1038/nchembio.1890

Moore, L. D., Le, T., & Fan, G. (2013). DNA Methylation and Its Basic Function. *Neuropsychopharmacology*, *38*(1), 23–38. https://doi.org/10.1038/npp.2012.112

Öeren, M., Walton, P. J., Suri, J., Ponting, D. J., Hunt, P. A., & Segall, M. D. (2022). Predicting Regioselectivity of AO, CYP, FMO, and UGT Metabolism Using Quantum Mechanical Simulations and Machine Learning. *Journal of Medicinal Chemistry*, *65*(20), 14066–14081. https://doi.org/10.1021/acs.jmedchem.2c01303

Orsi, M., & Reymond, J.-L. (2024). One chiral fingerprint to find them all. *Journal of Cheminformatics*, *16*(1), 53. https://doi.org/10.1186/s13321-024-00849-6

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Pinheiro, P. de S. M., Franco, L. S., & Fraga, C. A. M. (2023). The Magic Methyl and Its Tricks in Drug Discovery and Development. *Pharmaceuticals*, *16*(8), 1157. https://doi.org/10.3390/ph16081157

Pommié, C., Levadoux, S., Sabatier, R., Lefranc, G., & Lefranc, M. (2004). IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *Journal of Molecular Recognition*, *17*(1), 17–32. https://doi.org/10.1002/jmr.647

Ree, N., Göller, A. H., & Jensen, J. H. (2021). RegioSQM20: improved prediction of the regioselectivity of electrophilic aromatic substitutions. *Journal of Cheminformatics*, *13*(1), 10. https://doi.org/10.1186/s13321-021-00490-7

Schmidberger, J. W., James, A. B., Edwards, R., Naismith, J. H., & O'Hagan, D. (2010). Halomethane Biosynthesis: Structure of a SAM-Dependent Halide Methyltransferase from *Arabidopsis thaliana*. *Angewandte Chemie International Edition*, *49*(21), 3646–3648. https://doi.org/10.1002/anie.201000119

Seo, M., Shin, H. K., Myung, Y., Hwang, S., & No, K. T. (2020). Development of Natural Compound Molecular Fingerprint (NC-MFP) with the Dictionary of Natural Products (DNP) for natural product-based drug development. *Journal of Cheminformatics*, *12*(1), 6. https://doi.org/10.1186/s13321-020-0410-3

Shao, Y. (2023). *Predicting Methyltransferase Regioselectivity based on Substrate and Enzyme Information by Using Random Forest*. Wageningen University & Research.

Sparks, T. C., Hahn, D. R., & Garizi, N. V. (2017). Natural products, their derivatives, mimics and synthetic equivalents: role in agrochemical discovery. *Pest Management Science*, *73*(4), 700–715. https://doi.org/10.1002/ps.4458

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*(1), 25. https://doi.org/10.1186/1471-2105-8-25

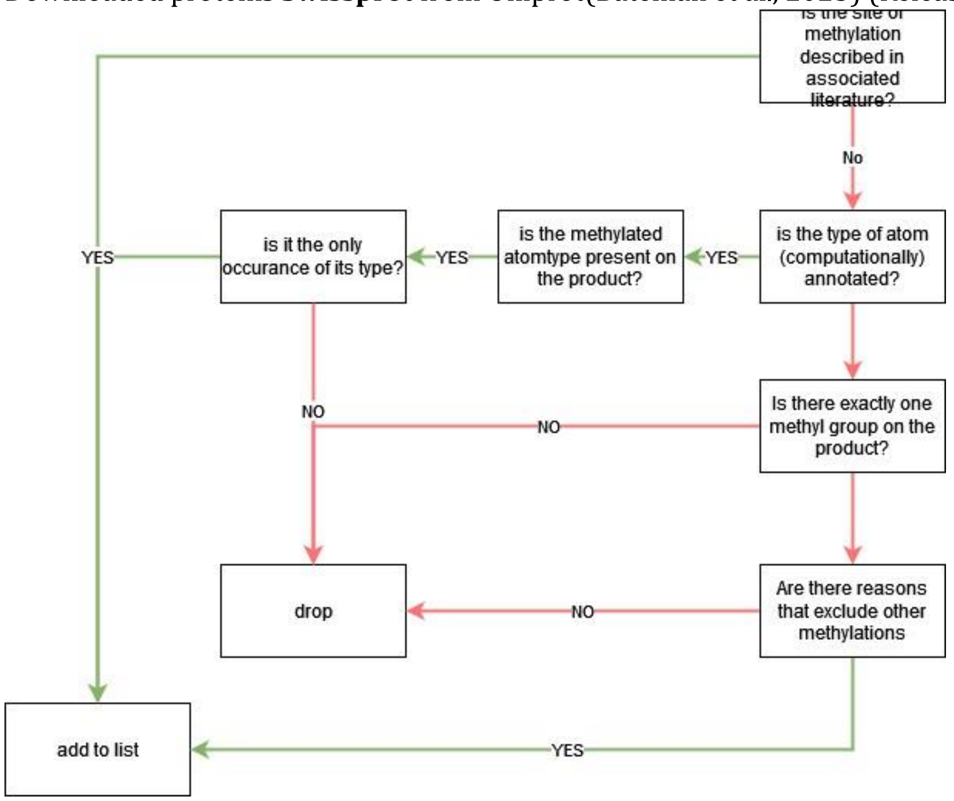
Stumpfe, D., Hu, H., & Bajorath, J. (2019). Evolving Concept of Activity Cliffs. *ACS Omega*, *4*(11), 14360–14368. https://doi.org/10.1021/acsomega.9b02221

Terlouw, B. R., Blin, K., Navarro-Muñoz, J. C., Avalon, N. E., Chevrette, M. G., Egbert, S., Lee, S., Meijer, D., Recchia, M. J. J., Reitz, Z. L., van Santen, J. A., Selem-Mojica, N., Tørring, T., Zaroubi, L., Alanjary, M., Aleti, G., Aguilar, C., Al-Salihi, S. A. A., Augustijn, H. E., … Medema, M. H. (2023). MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Research*, *51*(D1), D603–D610. https://doi.org/10.1093/nar/gkac1049

Wong, J. M., & Eirin-Lopez, J. M. (2021). Evolution of Methyltransferase-Like (METTL) Proteins in Metazoa: A Complex Gene Family Involved in Epitranscriptomic Regulation and Other Epigenetic Processes. *Molecular Biology and Evolution*, *38*(12), 5309–5327. https://doi.org/10.1093/molbev/msab267

Appendix A: Data collection

Training data flow:

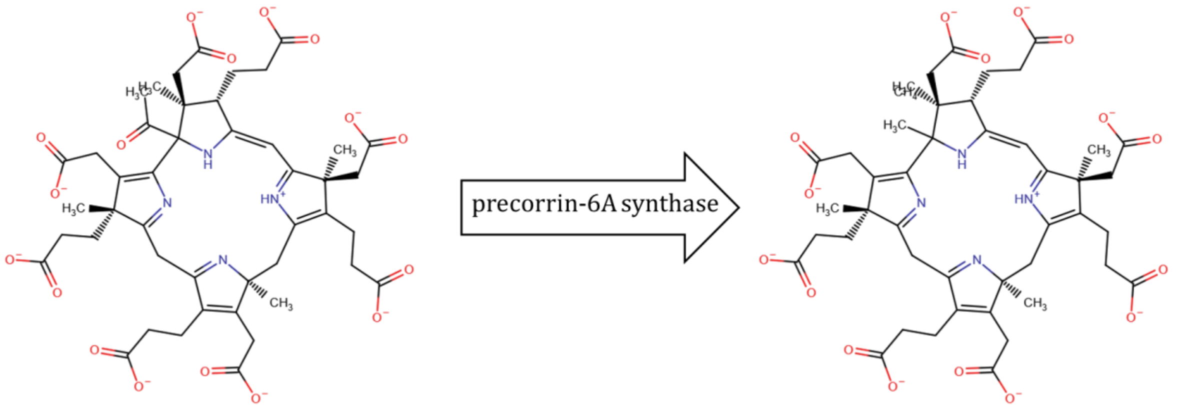
- Methyl transferase reactions from Rhea (Bansal et al., 2022) (22-5-2023)
- Methyltransferase reactions determined as containing **a protein** from ENZYME class: 2.1.1.- (Bairoch, 2000)
- Filtering reactants containing nucleotides/transport reactions.
- Downloaded SMILES from ChEBI db. Using lib ChEBIpy. (https://pypi.org/project/libChEBIpy/)
- Downloaded proteins **Swissprot** from Uniprot(Bateman et al., 2023) (Release 2023_05)
- 
is the site of methylation described in associated literature?
No
is the type of atom (computationally) annotated?
YES
is the methylated atomtype present on the product?
YES
is it the only occurance of its type?
YES
NO
NO
Is there exactly one methyl group on the product?
Are there reasons that exclude other methylations
NO
drop
YES
add to list

- Sampling max 5 proteins for each reaction.
- identification of sites of methylation.
- identification of 'class' methylation (C/N/O/S).
- Filtering and issues. (reactions more than methylation and charge.)
- Monomerization of polymers.

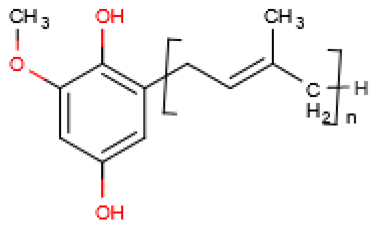
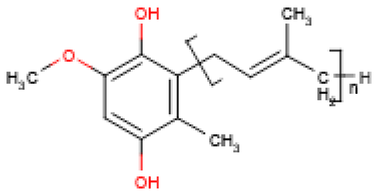
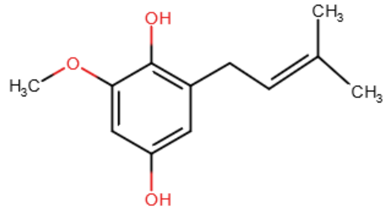
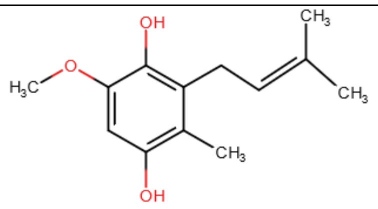
reaction_id	protein_id
BGC0000164	Q1MX76_9ACTN
BGC0000137	A8M6Z4_SALAI
BGC0000217	Q79E43_STRSQ
BGC0000455	E9LY85_AMYOR
BGC0001194	A0A0F7R6I4_9ACTN
BGC0002098	R4T3H0_9PSEU
BGC0000230	Q8KSX2_9ACTN
BGC0000257	B6VRR4_STRGD

BGC0002381	A0A6C1BZA8_9ACTN
BGC0001896	A0A3S9PUC0_STRLT
BGC0000065	E5DHP2_STRGB
BGC0000224	Q2MGC2_STRGR
BGC0000817	V9Y0Q9_9BACT
BGC0001578	B8LZ57_TALSN
BGC0001773	A0A1B1ZXZ6_9PSEU
BGC0000470	K4K984_9ACTN
BGC0000077	HPM5_HYPSB
BGC0000076	HPM5_HYPSB
BGC0000837	Q5E6K9_ALIF1
BGC0000836	A0A0H2V620_ECOL6
BGC0001266	GRA3_CLAGR
BGC0002210	PYVH_ASPV1
BGC0001677	A0A1L4BL56_PENRO
BGC0002616	A0A1L4BL56_PENRO
BGC0000104	MPAG_PENBR
BGC0001563	A0A218PFY6_9PLEO
BGC0001510	A0A1V0QSW3_9ACTN
BGC0001738	PHM5_PYRSX
BGC0001750	A0A291FH17_9BACT
BGC0002185	GRGD_PENSQ
BGC0001302	A0A0E3URN3_9ACTN
BGC0000897	D7PC21_STRLR
BGC0001046	D1GLT4_9ACTN
BGC0002261	LUC1_FUSSX
BGC0001880	BGC0001880_MT
BGC0000955	G0LWU9_MYXXA
BGC0001610	A0A1V0D7E3_9ACTN
BGC0002124	A0A7D3VQY7_9ACTN
BGC0001231	V5UVU0_9BACT
BGC0001204	A0A0B6VRG1_9ACTN
BGC0001547	A0A1L7NQ67_EMEVA
BGC0002131	BGC0002131_MT

Appendix B: non-methyl EC:2.1.1.-

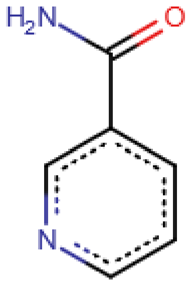
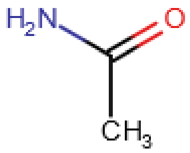
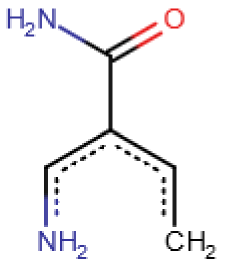


precorrin-5 precorrin-6A

	Substrate	product
Polymeric		
Monomer		

delocalization leads to a uniform electron density and consistent chemical properties regardless of how the double bonds are conceptualized or drawn within the ring structure.

Appendix C: Molecule labeling

Full substrate molecule	2 extended connections	4 extended connections
		

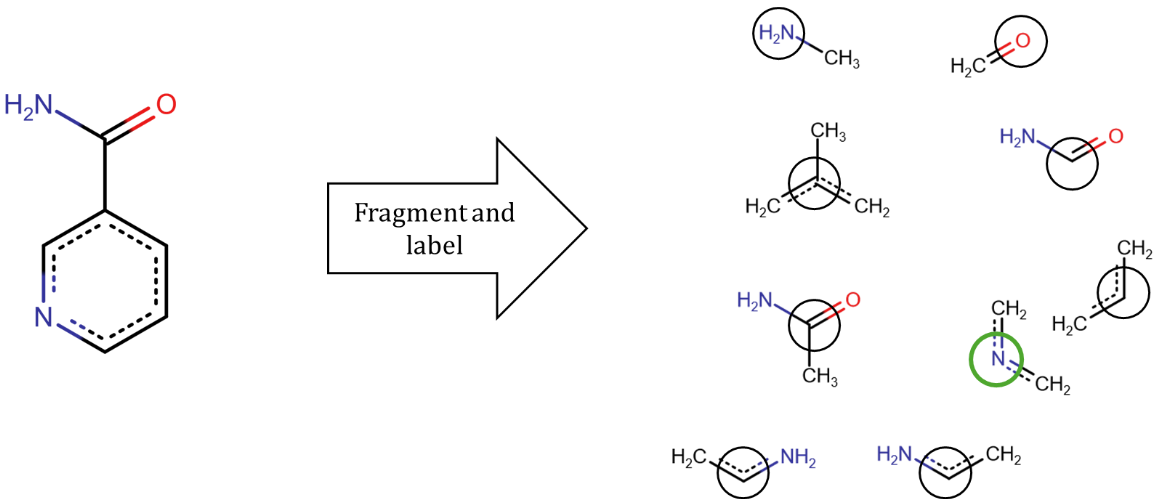


Figure 8: fragment based labelling.

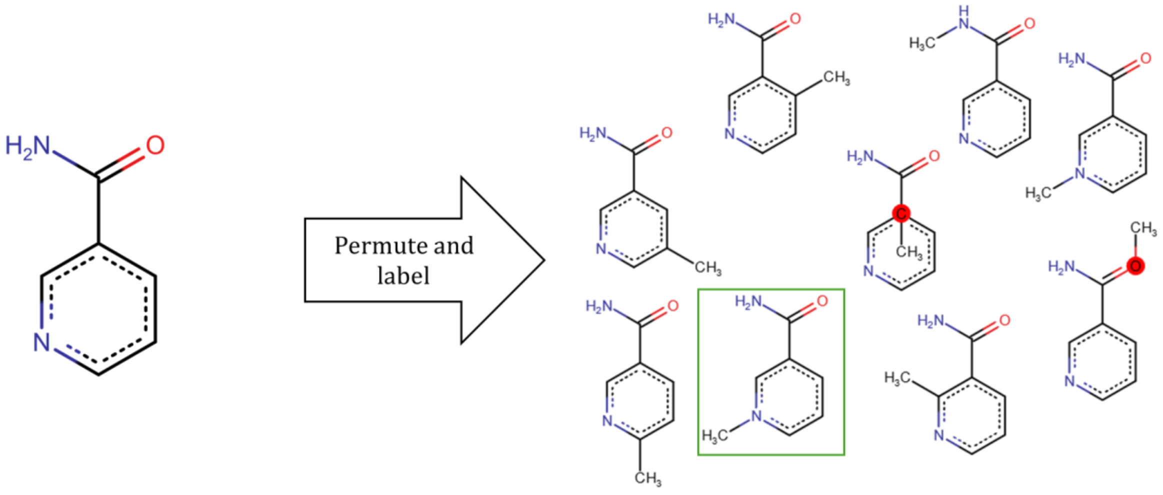


Figure 9: molecule permutation

Appendix D: Collected AAindex values

GRAR740102 Polarity (Grantham, 1974)

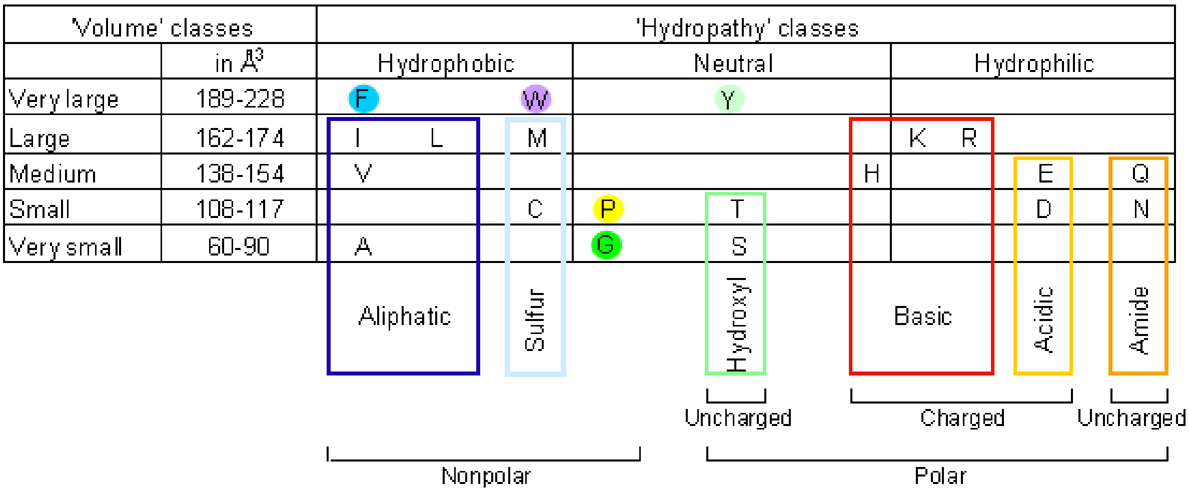
FAUJ880111 Positive charge (Fauchere et al., 1988)

FAUJ880112 Negative charge (Fauchere et al., 1988)

FASG890101 Hydrophobicity index (Fasman, 1989)

SIZEimgt Size in Angstrom (Pommié, C. et al., 2004)

Refer to encode_aadict.py for exact cutoffs of one-hot encoding. Encoding cutoffs are based on the IMGT 'Physicochemical' classes of the 20 common amino acids(Pommié et al., 2004)



'Volume' classes		'Hydropathy' classes		
	in $Å^3$	Hydrophobic	Neutral	Hydrophilic
Very large	189-228	F W	Y	
Large	162-174	I L M		K R
Medium	138-154	V	H	E Q
Small	108-117	C	P T	D N
Very small	60-90	A	G S	

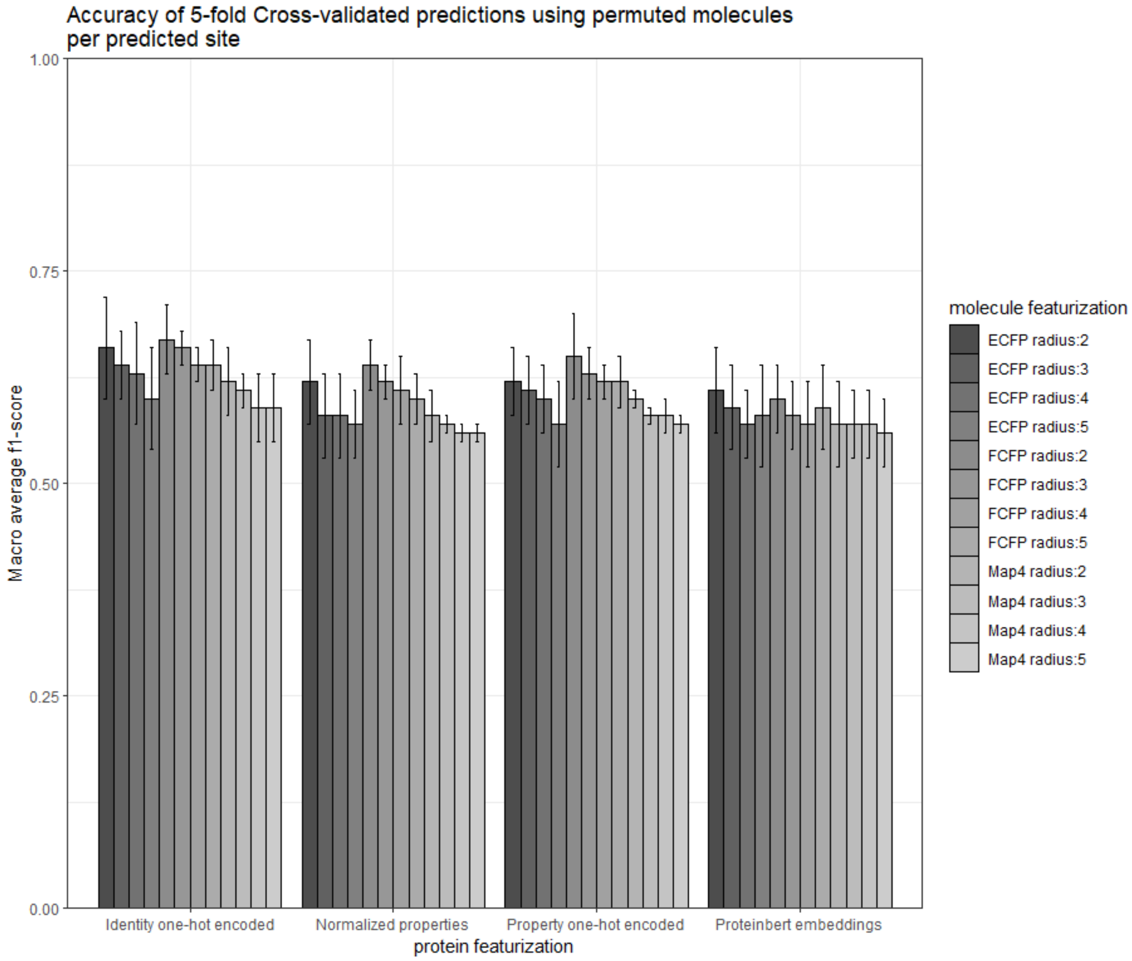
Aliphatic
Sulfur
Hydroxyl
Basic
Acidic
Amide
Uncharged
Charged
Uncharged
Nonpolar
Polar

Appendix E: Visualization model combinations

	Map4			ECFP			FCFP		
Identity									
property									
proteinBERT									

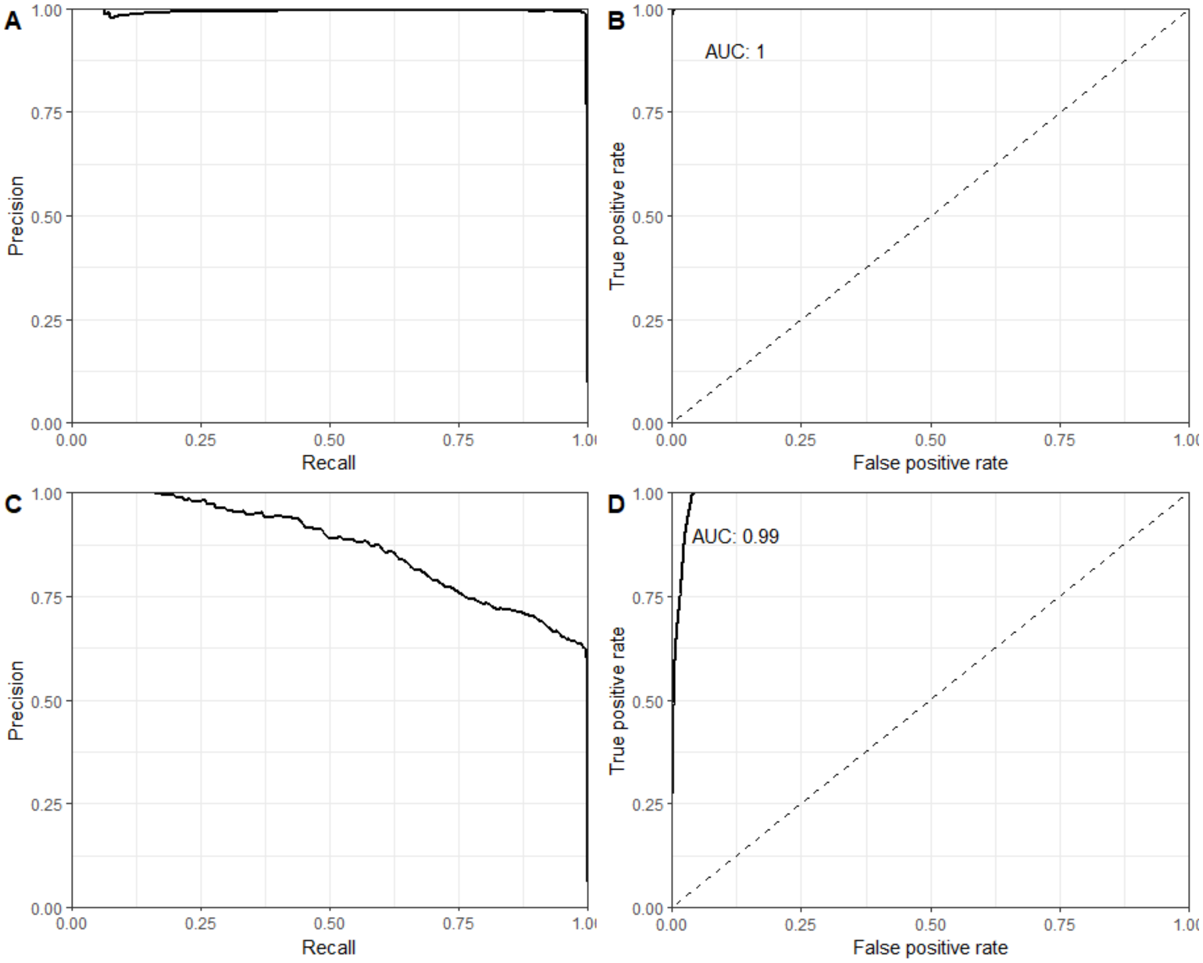
	Map4				ECFP				FCFP			
Identity												
property												
Norm. properties												
proteinBERT												

Appendix F:Cross validation performance figures



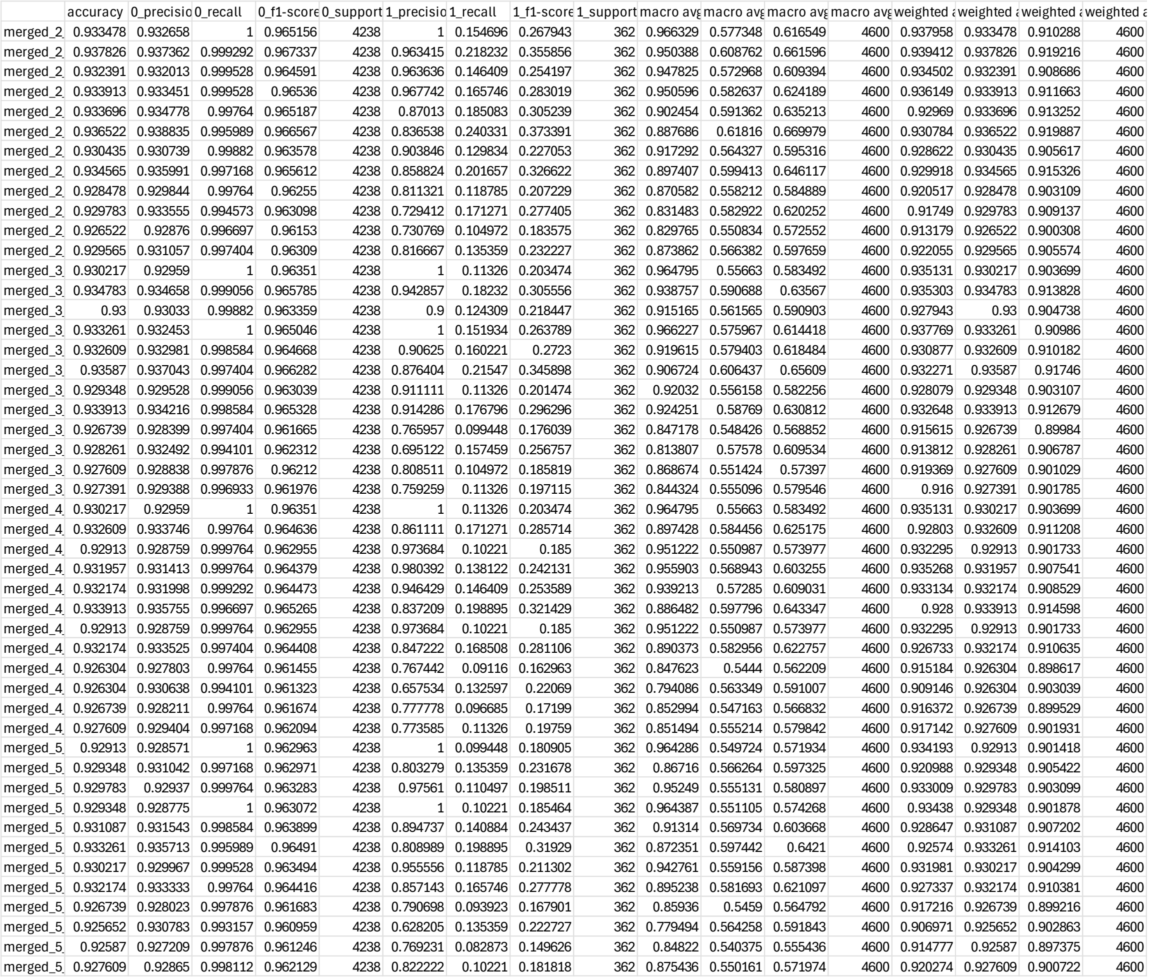
Macro average f1-scores of the predictions across 5-cv from the full molecule approach.

Precision recall and ROC curves of training performance

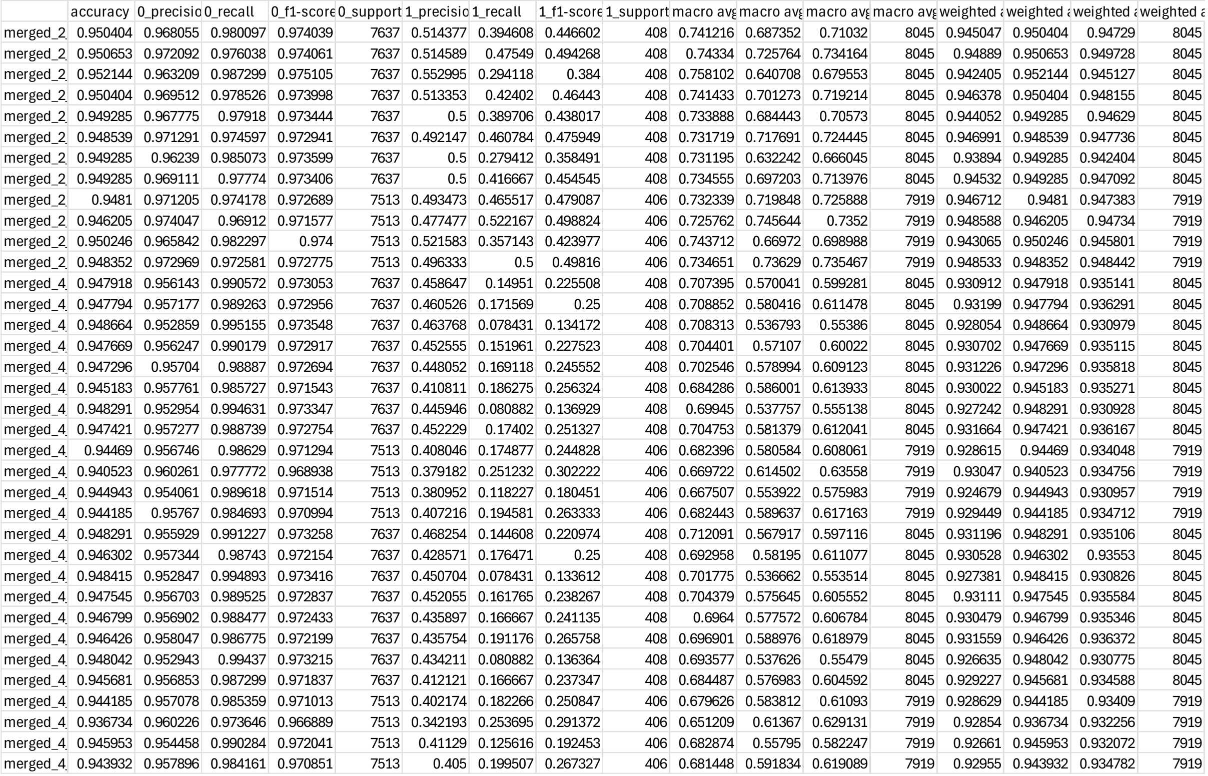


Appendix G: Cross validation performance metrics

Check git for full tables!.



	accuracy	0_precisio	0_recall	0_f1-score	0_support	1_precisio	1_recall	1_f1-score	1_support	macro avg	macro avg	macro avg	macro avg	weighted a	weighted a	weighted a	weighted a
merged_2	0.933478	0.932658	1	0.965156	4238	1	0.154696	0.267943	362	0.966329	0.577348	0.616549	4600	0.937958	0.933478	0.910288	4600
merged_2	0.937826	0.937362	0.999292	0.967337	4238	0.963415	0.218232	0.355856	362	0.950388	0.608762	0.661596	4600	0.939412	0.937826	0.919216	4600
merged_2	0.932391	0.932013	0.999528	0.964591	4238	0.963636	0.146409	0.254197	362	0.947825	0.572968	0.609394	4600	0.934502	0.932391	0.908686	4600
merged_2	0.933913	0.933451	0.999528	0.96536	4238	0.967742	0.165746	0.283019	362	0.950596	0.582637	0.624189	4600	0.936149	0.933913	0.911663	4600
merged_2	0.933696	0.934778	0.99764	0.965187	4238	0.87013	0.185083	0.305239	362	0.902454	0.591362	0.635213	4600	0.92969	0.933696	0.913252	4600
merged_2	0.936522	0.938835	0.995989	0.966567	4238	0.836538	0.240331	0.373391	362	0.887686	0.61816	0.669979	4600	0.930784	0.936522	0.919887	4600
merged_2	0.930435	0.930739	0.99882	0.963578	4238	0.903846	0.129834	0.227053	362	0.917292	0.564327	0.595316	4600	0.928622	0.930435	0.905617	4600
merged_2	0.934565	0.935991	0.997168	0.965612	4238	0.858824	0.201657	0.326622	362	0.897407	0.599413	0.646117	4600	0.929918	0.934565	0.915326	4600
merged_2	0.928478	0.929844	0.99764	0.96255	4238	0.811321	0.118785	0.207229	362	0.870582	0.558212	0.584889	4600	0.920517	0.928478	0.903109	4600
merged_2	0.929783	0.933555	0.994573	0.963098	4238	0.729412	0.171271	0.277405	362	0.831483	0.582922	0.620252	4600	0.91749	0.929783	0.909137	4600
merged_2	0.926522	0.92876	0.996697	0.96153	4238	0.730769	0.104972	0.183575	362	0.829765	0.550834	0.572552	4600	0.913179	0.926522	0.900308	4600
merged_2	0.929565	0.931057	0.997404	0.96309	4238	0.816667	0.135359	0.232227	362	0.873862	0.566382	0.597659	4600	0.922055	0.929565	0.905574	4600
merged_3	0.930217	0.92959	1	0.96351	4238	1	0.11326	0.203474	362	0.964795	0.55663	0.583492	4600	0.935131	0.930217	0.903699	4600
merged_3	0.934783	0.934658	0.999056	0.965785	4238	0.942857	0.18232	0.305556	362	0.938757	0.590688	0.63567	4600	0.935303	0.934783	0.913828	4600
merged_3	0.93	0.93033	0.99882	0.963359	4238	0.9	0.124309	0.218447	362	0.915165	0.561565	0.590903	4600	0.927943	0.93	0.904738	4600
merged_3	0.933261	0.932453	1	0.965046	4238	1	0.151934	0.263789	362	0.966227	0.575967	0.614418	4600	0.937769	0.933261	0.90986	4600
merged_3	0.932609	0.932981	0.998584	0.964668	4238	0.90625	0.160221	0.2723	362	0.919615	0.579403	0.618484	4600	0.930877	0.932609	0.910182	4600
merged_3	0.93587	0.937043	0.997404	0.966282	4238	0.876404	0.21547	0.345898	362	0.906724	0.606437	0.65609	4600	0.932271	0.93587	0.91746	4600
merged_3	0.929348	0.929528	0.999056	0.963039	4238	0.911111	0.11326	0.201474	362	0.92032	0.556158	0.582256	4600	0.928079	0.929348	0.903107	4600
merged_3	0.933913	0.934216	0.998584	0.965328	4238	0.914286	0.176796	0.296296	362	0.924251	0.58769	0.630812	4600	0.932648	0.933913	0.912679	4600
merged_3	0.926739	0.928399	0.997404	0.961665	4238	0.765957	0.099448	0.176039	362	0.847178	0.548426	0.568852	4600	0.915615	0.926739	0.89984	4600
merged_3	0.928261	0.932492	0.994101	0.962312	4238	0.695122	0.157459	0.256757	362	0.813807	0.57578	0.609534	4600	0.913812	0.928261	0.906787	4600
merged_3	0.927609	0.928838	0.997876	0.96212	4238	0.808511	0.104972	0.185819	362	0.868674	0.551424	0.57397	4600	0.919369	0.927609	0.901029	4600
merged_3	0.927391	0.929388	0.996933	0.961976	4238	0.759259	0.11326	0.197115	362	0.844324	0.555096	0.579546	4600	0.916	0.927391	0.901785	4600
merged_4	0.930217	0.92959	1	0.96351	4238	1	0.11326	0.203474	362	0.964795	0.55663	0.583492	4600	0.935131	0.930217	0.903699	4600
merged_4	0.932609	0.933746	0.99764	0.964636	4238	0.861111	0.171271	0.285714	362	0.897428	0.584456	0.625175	4600	0.92803	0.932609	0.911208	4600
merged_4	0.92913	0.928759	0.999764	0.962955	4238	0.973684	0.10221	0.185	362	0.951222	0.550987	0.573977	4600	0.932295	0.92913	0.901733	4600
merged_4	0.931957	0.931413	0.999764	0.964379	4238	0.980392	0.138122	0.242131	362	0.955903	0.568943	0.603255	4600	0.935268	0.931957	0.907541	4600
merged_4	0.932174	0.931998	0.999292	0.964473	4238	0.946429	0.146409	0.253589	362	0.939213	0.57285	0.609031	4600	0.933134	0.932174	0.908529	4600
merged_4	0.933913	0.935755	0.996697	0.965265	4238	0.837209	0.198895	0.321429	362	0.886482	0.597796	0.643347	4600	0.928	0.933913	0.914598	4600
merged_4	0.92913	0.928759	0.999764	0.962955	4238	0.973684	0.10221	0.185	362	0.951222	0.550987	0.573977	4600	0.932295	0.92913	0.901733	4600
merged_4	0.932174	0.933525	0.997404	0.964408	4238	0.847222	0.168508	0.281106	362	0.890373	0.582956	0.622757	4600	0.926733	0.932174	0.910635	4600
merged_4	0.926304	0.927803	0.99764	0.961455	4238	0.767442	0.09116	0.162963	362	0.847623	0.5444	0.562209	4600	0.915184	0.926304	0.898617	4600
merged_4	0.926304	0.930638	0.994101	0.961323	4238	0.657534	0.132597	0.22069	362	0.794086	0.563349	0.591007	4600	0.909146	0.926304	0.903039	4600
merged_4	0.926739	0.928211	0.99764	0.961674	4238	0.777778	0.096685	0.17199	362	0.852994	0.547163	0.566832	4600	0.916372	0.926739	0.899529	4600
merged_4	0.927609	0.929404	0.997168	0.962094	4238	0.773585	0.11326	0.19759	362	0.851494	0.555214	0.579842	4600	0.917142	0.927609	0.901931	4600
merged_5	0.92913	0.928571	1	0.962963	4238	1	0.099448	0.180905	362	0.964286	0.549724	0.571934	4600	0.934193	0.92913	0.901418	4600
merged_5	0.929348	0.931042	0.997168	0.962971	4238	0.803279	0.135359	0.231678	362	0.86716	0.566264	0.597325	4600	0.920988	0.929348	0.905422	4600
merged_5	0.929783	0.92937	0.999764	0.963283	4238	0.97561	0.110497	0.198511	362	0.95249	0.555131	0.580897	4600	0.933009	0.929783	0.903099	4600
merged_5	0.929348	0.928775	1	0.963072	4238	1	0.10221	0.185464	362	0.964387	0.551105	0.574268	4600	0.93438	0.929348	0.901878	4600
merged_5	0.931087	0.931543	0.998584	0.963899	4238	0.894737	0.140884	0.243437	362	0.91314	0.569734	0.603668	4600	0.928647	0.931087	0.907202	4600
merged_5	0.933261	0.935713	0.995989	0.96491	4238	0.808989	0.198895	0.31929	362	0.872351	0.597442	0.6421	4600	0.92574	0.933261	0.914103	4600
merged_5	0.930217	0.929967	0.999528	0.963494	4238	0.955556	0.118785	0.211302	362	0.942761	0.559156	0.587398	4600	0.931981	0.930217	0.904299	4600
merged_5	0.932174	0.933333	0.99764	0.964416	4238	0.857143	0.165746	0.277778	362	0.895238	0.581693	0.621097	4600	0.927337	0.932174	0.910381	4600
merged_5	0.926739	0.928023	0.997876	0.961683	4238	0.790698	0.093923	0.167901	362	0.85936	0.5459	0.564792	4600	0.917216	0.926739	0.899216	4600
merged_5	0.925652	0.930783	0.993157	0.960959	4238	0.628205	0.135359	0.222727	362	0.779494	0.564258	0.591843	4600	0.906971	0.925652	0.902863	4600
merged_5	0.92587	0.927209	0.997876	0.961246	4238	0.769231	0.082873	0.149626	362	0.84822	0.540375	0.555436	4600	0.914777	0.92587	0.897375	4600
merged_5	0.927609	0.92865	0.998112	0.962129	4238	0.822222	0.10221	0.181818	362	0.875436	0.550161	0.571974	4600	0.920274	0.927609	0.900722	4600



	accuracy	0_precisio	0_recall	0_f1-score	0_support	1_precisio	1_recall	1_f1-score	1_support	macro avg	macro avg	macro avg	macro avg	weighted a	weighted a	weighted a	weighted a
merged_2	0.950404	0.968055	0.980097	0.974039	7637	0.514377	0.394608	0.446602	408	0.741216	0.687352	0.71032	8045	0.945047	0.950404	0.94729	8045
merged_2	0.950653	0.972092	0.976038	0.974061	7637	0.514589	0.47549	0.494268	408	0.74334	0.725764	0.734164	8045	0.94889	0.950653	0.949728	8045
merged_2	0.952144	0.963209	0.987299	0.975105	7637	0.552995	0.294118	0.384	408	0.758102	0.640708	0.679553	8045	0.942405	0.952144	0.945127	8045
merged_2	0.950404	0.969512	0.978526	0.973998	7637	0.513353	0.42402	0.46443	408	0.741433	0.701273	0.719214	8045	0.946378	0.950404	0.948155	8045
merged_2	0.949285	0.967775	0.97918	0.973444	7637	0.5	0.389706	0.438017	408	0.733888	0.684443	0.70573	8045	0.944052	0.949285	0.94629	8045
merged_2	0.948539	0.971291	0.974597	0.972941	7637	0.492147	0.460784	0.475949	408	0.731719	0.717691	0.724445	8045	0.946991	0.948539	0.947736	8045
merged_2	0.949285	0.96239	0.985073	0.973599	7637	0.5	0.279412	0.358491	408	0.731195	0.632242	0.666045	8045	0.93894	0.949285	0.942404	8045
merged_2	0.949285	0.969111	0.97774	0.973406	7637	0.5	0.416667	0.454545	408	0.734555	0.697203	0.713976	8045	0.94532	0.949285	0.947092	8045
merged_2	0.9481	0.971205	0.974178	0.972689	7513	0.493473	0.465517	0.479087	406	0.732339	0.719848	0.725888	7919	0.946712	0.9481	0.947383	7919
merged_2	0.946205	0.974047	0.96912	0.971577	7513	0.477477	0.522167	0.498824	406	0.725762	0.745644	0.7352	7919	0.948588	0.946205	0.94734	7919
merged_2	0.950246	0.965842	0.982297	0.974	7513	0.521583	0.357143	0.423977	406	0.743712	0.66972	0.698988	7919	0.943065	0.950246	0.945801	7919
merged_2	0.948352	0.972969	0.972581	0.972775	7513	0.496333	0.5	0.49816	406	0.734651	0.73629	0.735467	7919	0.948533	0.948352	0.948442	7919
merged_4	0.947918	0.956143	0.990572	0.973053	7637	0.458647	0.14951	0.225508	408	0.707395	0.570041	0.599281	8045	0.930912	0.947918	0.935141	8045
merged_4	0.947794	0.957177	0.989263	0.972956	7637	0.460526	0.171569	0.25	408	0.708852	0.580416	0.611478	8045	0.93199	0.947794	0.936291	8045
merged_4	0.948664	0.952859	0.995155	0.973548	7637	0.463768	0.078431	0.134172	408	0.708313	0.536793	0.55386	8045	0.928054	0.948664	0.930979	8045
merged_4	0.947669	0.956247	0.990179	0.972917	7637	0.452555	0.151961	0.227523	408	0.704401	0.57107	0.60022	8045	0.930702	0.947669	0.935115	8045
merged_4	0.947296	0.95704	0.98887	0.972694	7637	0.448052	0.169118	0.245552	408	0.702546	0.578994	0.609123	8045	0.931226	0.947296	0.935818	8045
merged_4	0.945183	0.957761	0.985727	0.971543	7637	0.410811	0.186275	0.256324	408	0.684286	0.586001	0.613933	8045	0.930022	0.945183	0.935271	8045
merged_4	0.948291	0.952954	0.994631	0.973347	7637	0.445946	0.080882	0.136929	408	0.69945	0.537757	0.555138	8045	0.927242	0.948291	0.930928	8045
merged_4	0.947421	0.957277	0.988739	0.972754	7637	0.452229	0.17402	0.251327	408	0.704753	0.581379	0.612041	8045	0.931664	0.947421	0.936167	8045
merged_4	0.94469	0.956746	0.98629	0.971294	7513	0.408046	0.174877	0.244828	406	0.682396	0.580584	0.608061	7919	0.928615	0.94469	0.934048	7919
merged_4	0.940523	0.960261	0.977772	0.968938	7513	0.379182	0.251232	0.302222	406	0.669722	0.614502	0.63558	7919	0.93047	0.940523	0.934756	7919
merged_4	0.944943	0.954061	0.989618	0.971514	7513	0.380952	0.118227	0.180451	406	0.667507	0.553922	0.575983	7919	0.924679	0.944943	0.930957	7919
merged_4	0.944185	0.95767	0.984693	0.970994	7513	0.407216	0.194581	0.263333	406	0.682443	0.589637	0.617163	7919	0.929449	0.944185	0.934712	7919
merged_4	0.948291	0.955929	0.991227	0.973258	7637	0.468254	0.144608	0.220974	408	0.712091	0.567917	0.597116	8045	0.931196	0.948291	0.935106	8045
merged_4	0.946302	0.957344	0.98743	0.972154	7637	0.428571	0.176471	0.25	408	0.692958	0.58195	0.611077	8045	0.930528	0.946302	0.93553	8045
merged_4	0.948415	0.952847	0.994893	0.973416	7637	0.450704	0.078431	0.133612	408	0.701775	0.536662	0.553514	8045	0.927381	0.948415	0.930826	8045
merged_4	0.947545	0.956703	0.989525	0.972837	7637	0.452055	0.161765	0.238267	408	0.704379	0.575645	0.605552	8045	0.93111	0.947545	0.935584	8045
merged_4	0.946799	0.956902	0.988477	0.972433	7637	0.435897	0.166667	0.241135	408	0.6964	0.577572	0.606784	8045	0.930479	0.946799	0.935346	8045
merged_4	0.946426	0.958047	0.986775	0.972199	7637	0.435754	0.191176	0.265758	408	0.696901	0.588976	0.618979	8045	0.931559	0.946426	0.936372	8045
merged_4	0.948042	0.952943	0.99437	0.973215	7637	0.434211	0.080882	0.136364	408	0.693577	0.537626	0.55479	8045	0.926635	0.948042	0.930775	8045
merged_4	0.945681	0.956853	0.987299	0.971837	7637	0.412121	0.166667	0.237347	408	0.684487	0.576983	0.604592	8045	0.929227	0.945681	0.934588	8045
merged_4	0.944185	0.957078	0.985359	0.971013	7513	0.402174	0.182266	0.250847	406	0.679626	0.583812	0.61093	7919	0.928629	0.944185	0.93409	7919
merged_4	0.936734	0.960226	0.973646	0.966889	7513	0.342193	0.253695	0.291372	406	0.651209	0.61367	0.629131	7919	0.92854	0.936734	0.932256	7919
merged_4	0.945953	0.954458	0.990284	0.972041	7513	0.41129	0.125616	0.192453	406	0.682874	0.55795	0.582247	7919	0.92661	0.945953	0.932072	7919
merged_4	0.943932	0.957896	0.984161	0.970851	7513	0.405	0.199507	0.267327	406	0.681448	0.591834	0.619089	7919	0.92955	0.943932	0.934782	7919

Using simulation where randomly picking one site as methylated:

Script used: *sim.py*

Pseudocode

```
    count = 0
    for x in atomcounts:
        if np.random.choice(x,1) == x-1:
            count += 1
```

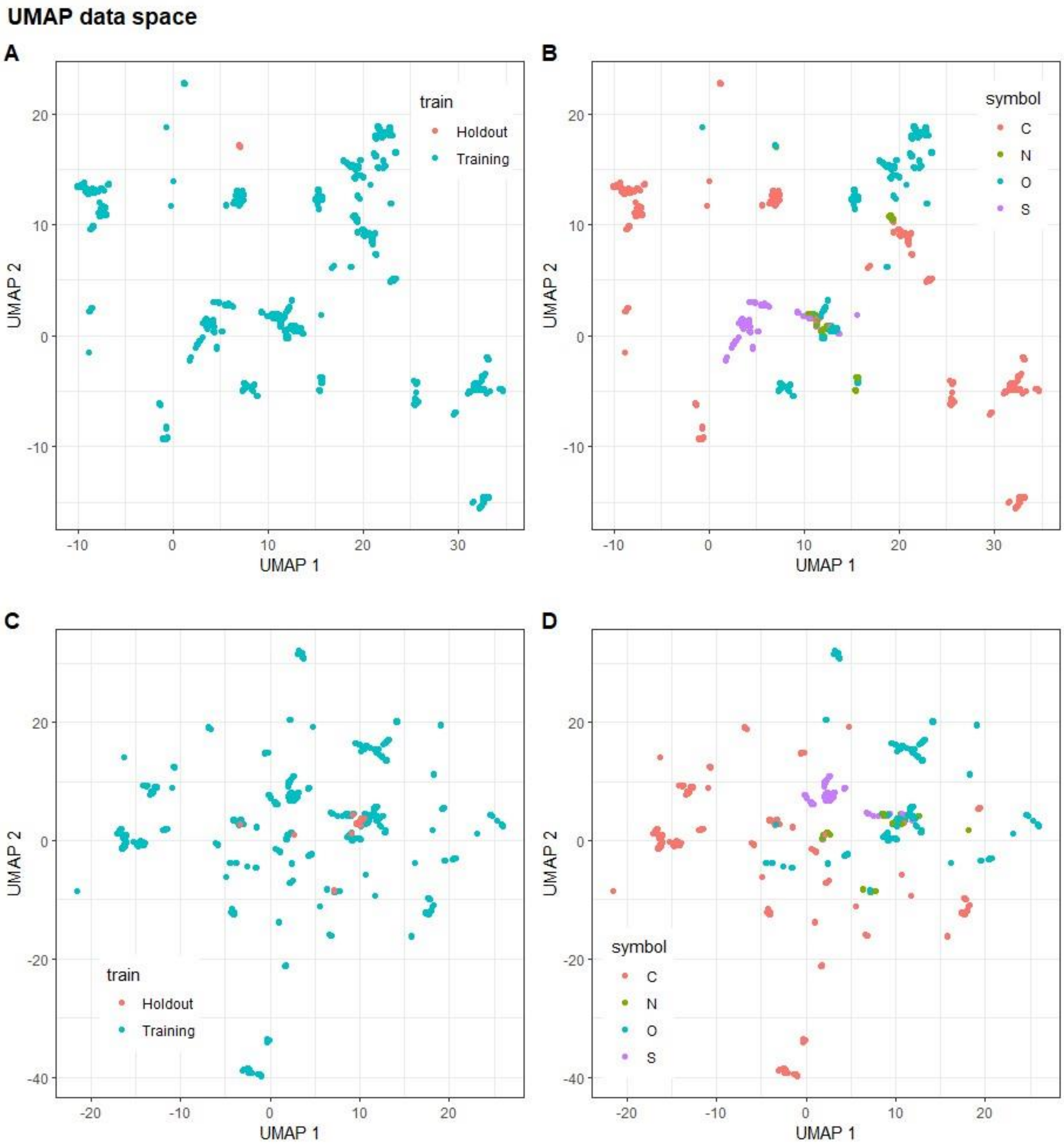
resulting:

preds = 187

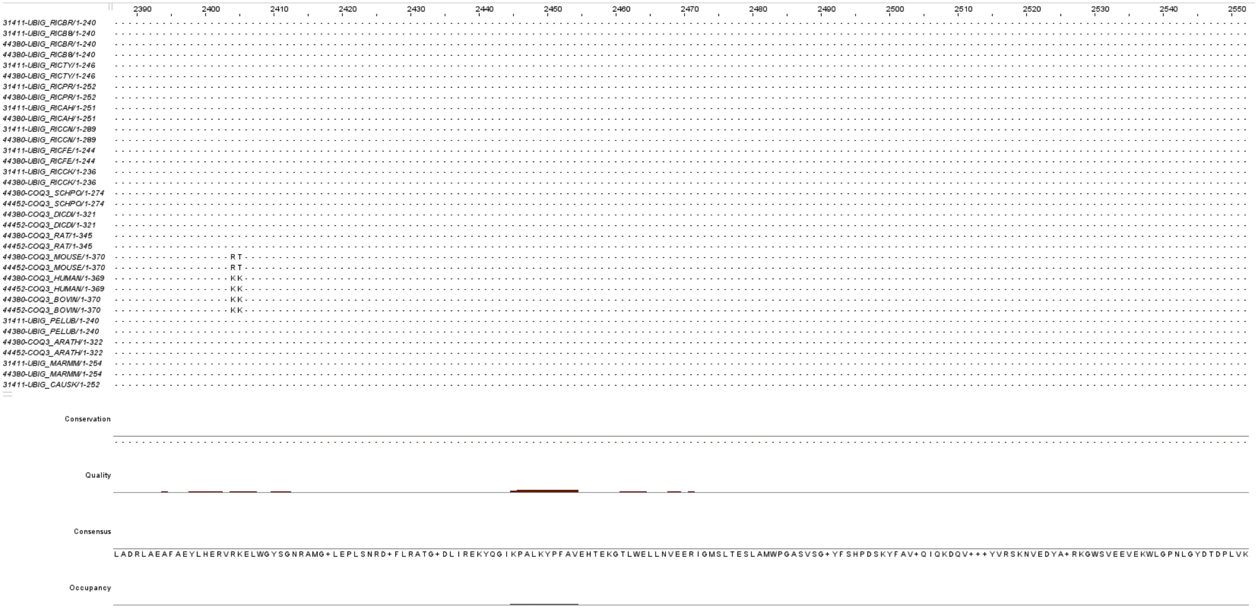
counts = 9

accuracy = 0.0481283422459893

Appendix H: UMAP dimensionality reduction without downsampling



Appendix I: MSA tail occupancy



Note the low occupancy in the range of ~2000 – 2500. Despite the strong feature importance associated with these positions they mostly consist of gaps

Appendix J: Feature importance

	Permutation		**Fragment**	
Rank	**Importance**	**Index**	**Importance**	**Index**
0	0.0202	2405	0.0124	2374
1	0.0143	2386	0.0108	2435
2	0.0132	485	0.0106	2520
3	0.0130	1479	0.0102	2510
4	0.0117	2403	0.0084	1363
5	0.0112	2476	0.0081	1561
6	0.0109	2507	0.0078	2475
7	0.0105	2498	0.0076	1159
8	0.0095	1285	0.0071	506
9	0.0094	969	0.0066	2516

Top ten most important features in the two used models. A majority of these are within the protein part of the feature vector. (> 2048).