Deep Learning models to predict lung transplant rejections using cell-free DNA fragmentomics

Major research project

Masters in Bioinformatics and Biocomplexity

Student: Madhupreetha Vivekanandan (2988291)

Daily supervisor: Lucía Barbadilla Martínez

Examiner: Dr. Jeroen de Ridder

TABLE OF CONTENTS

ABSTRACT

Chronic respiratory diseases impose a significant burden to society, and lung transplant remains the last-resort treatment option. However, long-term survival rates for lung transplant recipients are hindered by a high risk of rejection, leading to Chronic Lung Allograft Dysfunction (CLAD) and eventual death. Traditional methods for diagnosing rejection are invasive and often prove too late. Thus, there is an urgent need for diagnostic methods for timely detection of rejections to improve lung transplant outcomes. The levels of donor-derived cell-free DNA (dd-cfDNA) in the recipient's blood hold great promise as a biomarker for diagnosing transplant rejections. Current measurement methods involve the selective amplification of dd-cfDNA using donor and recipient SNPs. However, this method is limited by the number of SNP differences between the donor and the recipient. We introduce a novel method for measuring dd-cfDNA levels by employing Deep Learning (DL) models to distinguish between donor- and recipient-derived cfDNA based on their tissue of origin. To this end, three DL models were developed: a feed-forward neural network trained on epigenetic features of cfDNA (extracted using Enformer), a convolutional neural network (CNN) focused on sequence motifs, and a second CNN model utilizing both epigenetic features and sequence motifs for classification. The models, trained and evaluated on a dataset of cfDNA samples collected from 47 lung transplant recipients, achieved only marginal improvement over a random classifier, with the best-performing model achieving an area under the ROC curve of just 0.524. Baseline logistic regression and dimensionality reduction analysis pointed to either highly complex or absent signals in the data. High accuracy achieved by these models on simulations with artificially embedded signals showed that given enough signals, the models could learn to distinguish dd-cfDNA from rd-cfDNA, further underscoring the complexity of the real dataset. Contrary to common consensus, there was a poor correlation between dd-cfDNA percentage and clinical signs of rejection in the training dataset, raising questions about training label accuracy. Poor data quality, inaccuracies in training labels, and incomplete hyperparameter optimization are potential contributing factors to the suboptimal performance of the models. While the models we developed did not perform well enough to be considered reliable for diagnosis, the approach of using dl models to measure dd-cfDNA levels is novel and holds great promise. With higher quality training data, selective sample incorporation into the training set, and a more extensive hyperparameter search, these DL models could serve as highly effective non-invasive tools for transplant rejection diagnosis, with potential applications in areas like prenatal diagnosis and liquid biopsy.

LAYMAN'S SUMMARY

Chronic respiratory diseases are a huge societal burden and a lung transplant is the last resort to treat them. However, many patients who undergo lung transplantation don’t survive for long because the transplanted lung is rejected by the patient’s immune system. The usual methods to detect rejection are painful for the patient, costly, and often too late. In this study, we have explored a new way to identify rejection early, using small fragments of DNA, called cell-free DNA (cfDNA), in the patient’s blood.

cfDNA is released by cells in the body when they die. Most cfDNA in the blood comes from blood cells. After a lung transplant, the transplanted lung also releases some cell-free DNA into the patient’s blood, called the donor-derived or dd-cfDNA. Patients who undergo rejection have more of these dd-cfDNA fragments in the blood compared to healthy transplant recipients. This is because the recipient’s immune system perceives the lung as a foreign object and starts killing its cells, making them release cfDNA. So, we can diagnose whether the recipient is having a rejection by measuring how much dd-cfDNA is present in their blood.

To measure dd-cfDNA, we first need a way to tell it apart from the patient’s own cfDNA. The current methods have limitations, so we developed a new approach to differentiate between dd-cfDNA and the recipient’s own cfDNA using advanced predictive algorithms called Deep Learning (DL). DL allows computers to automatically learn patterns from examples (training), and to use them to make predictions on new data. In our case, we trained these algorithms to tell the difference between dd-cfDNA and the recipient’s own cfDNA. We then used these trained models to predict if an unknown cfDNA is dd-cfDNA. The recipient’s cfDNA comes from blood cells whereas dd-cfDNA comes from the lung, so our models were meant to tell whether a cfDNA fragment came from lung or blood cells.

We trained and tested three different DL models on raw cfDNA sequencing data from 47 lung transplant recipients. The idea was to find the model that performs the best out of the three and use its predictions to count the number of dd-cfDNA fragments for each patient. From this number, we calculated the percentage of dd-cfDNA. If the percentage was above a certain threshold at any time for a patient, we could diagnose that the patient was having a rejection.

However, none of the three models we developed performed as well as expected. To understand why, we trained the models with artificial data and found that they do poorly when the data has a lot of noisy samples i.e. samples that don’t have useful patterns that the model can learn. Based on this, we think that the real data we used for training has a lot of noise and that is why the model did not perform very well.

We have a few theories on why the real data is noisy. One of them is that dd-cfDNA and the cfDNA from the recipient may not be that different. If that was the case, we could focus on specific samples that we know would be different if they originated from the lung or the recipient’s blood. Our second suspicion is that the dataset of 47 lung transplant patients we used for training our models could be inaccurately named dd-cfDNA or recipient's cfDNA. In the future, we suggest using a different dataset that is verified to have accurate labels.

Even though our models did not perform well enough for clinical usage, we believe that with the proposed improvements, they could become a powerful tool to detect transplant rejection early, making the process less invasive and more cost-effective for patients. This innovative approach of using DL to tell apart dd-cfDNA and the recipient's own cfDNA could have many applications beyond transplant rejections, opening doors for various medical advancements.

1. INTRODUCTION

Chronic respiratory diseases (CRD) are the third leading cause of mortality, responsible for 4 million deaths globally every year (Momtazmanesh, S., et al., 2023). A lung transplant remains the preferred option for treating patients with CRD when all other treatment options have failed. However, the long-term survival rates for lung transplant recipients are low due to complications arising from acute rejections and infections post-transplant. The survival rate five years after a lung transplant can be as low as 58%, with a median survival time of 5.3 years, the lowest among all types of transplants (Lund, Lars H., et al., 2014). The primary cause of death is Bronchitis Obliterans Syndrome, a form of Chronic Lung Allograft Dysfunction (CLAD) disease characterized by reduced lung capacity (Sundaresan, Sudhir, et al., 1995). The origin of CLAD can be traced to the initial post-transplant period, where acute rejection episodes orchestrated by T-cell immune responses against Human Leukocyte Antigen (HLA) lead to allograft injury. Accumulated allograft injury from such repeated episodes of rejection is a strong risk factor for CLAD, often resulting in the death of the patient (Gauthier, Jason M., Ramsey R. Hachem, and Daniel Kreisel, 2016).

The current gold standard for diagnosis of acute rejection such as transbronchial biopsies and histopathological studies are invasive, painful, expensive, and potentially dangerous (Herout, Vladimir, et al., 2019). These approaches have limited predictive value (Arcasoy, S. M., et al., 2011), and clinical symptoms indicative of rejection often appear only after the onset of CLAD. Once diagnosed, CLAD responds poorly to treatment (Mrad, Ali, and Chakraborty, R. K. 2020). With timely detection and intervention through increased doses of immunosuppressants and steroids, it is possible to delay or prevent the onset of CLAD and improve lung transplant outcomes (Mrad, Ali, and Chakraborty, R. K. 2020). Therefore, there is a pressing need for non-invasive, cost-effective biomarkers capable of predicting early-stage acute rejection in lung transplant recipients.

Studies indicate elevated levels of cell-free DNA (cfDNA) in the blood of patients experiencing acute rejection (Magnusson, Jesper M., et al., 2022) (Jang, Moon Kyoo, et al., 2021). cfDNA are small DNA fragments (120-220 bps) released into the bloodstream, as a result of DNA digestion during processes like apoptosis and necrosis. In healthy patients, most fragments are of hematopoietic origin. But in diseased patients and extraordinary circumstances (such as organ transplantation and pregnancy), fragments from the affected tissue also enter the bloodstream (Lo, YM Dennis, et al., 1998., Lo, YM Dennis, et al., 1997). Following lung transplantation, the allograft releases donor-derived cfDNA (dd-cfDNA) which accounts for a small percentage of the recipient's cfDNA. In the event of a transplant rejection, the lung allograft releases more dd-cfDNA due to tissue damage (Fig 1). A threshold of >1% dd-cfDNA in the blood has demonstrated 100% sensitivity and 73% specificity for detecting moderate to severe acute rejection (De Vlaminck, Iwijn, et al., 2015). Notably, the levels of dd-cfDNA were elevated months before the onset of clinical symptoms for acute rejection, thus making early detection possible (Jang, Moon Kyoo, et al., 2021). Taking all this evidence into account, dd-cfDNA levels in the recipient's blood are a good biomarker for early detection of acute rejection and resulting lung infections.

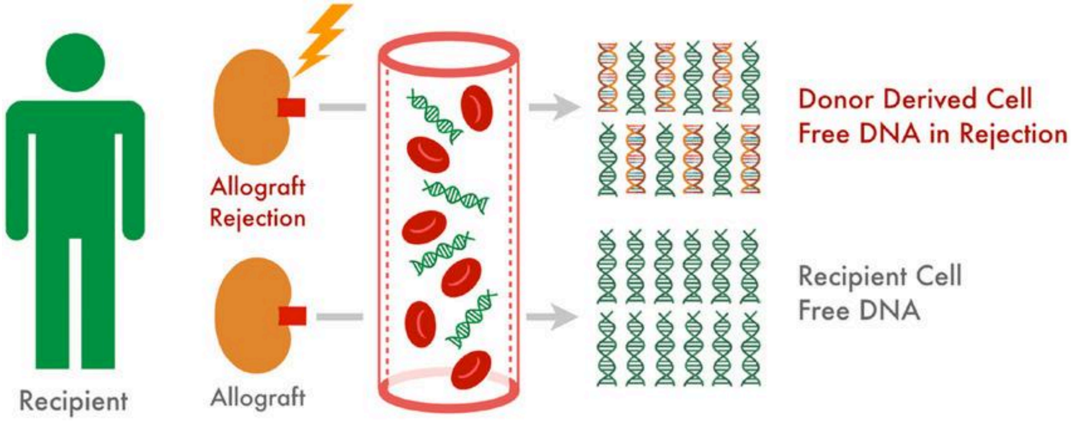


Fig 1: The concept of donor-derived cfDNA (dd-cfDNA) as a marker for allograft injury (Paul, Rohan S., et al.)

Quantification of dd-cfDNA is traditionally done using PCR-based assays, utilizing markers specific to donor and recipient cfDNA for selective amplification through methods like ddPCR, quantitative PCR, and shotgun sequencing. Early biomarkers for selective amplification leveraged sex differences and HLA mismatches between the donors and recipients, but had limited applicability (Keller, M., & Agbor-Enoh, S. 2021, Lo, YM Dennis, et al., 1998., Zou, Jun, et al., 2017). More recent techniques use donor and recipient-specific SNPs as biomarkers (De Vlaminck, Iwijn, et al., 2015, Keller, M., & Agbor-Enoh, S. 2021). This process involves performing a whole genome genotyping of the donors and recipients to create a library of SNPs unique to the donor and using those for selective amplification. However, in practice genome information of the donor is often not available and genotyping can be expensive. Still other methods developed later use computational models to selectively amplify dd-cfDNA using only the recipient's SNPs (Sharon, Eilon, et al., 2017, Keller, M., & Agbor-Enoh, S. 2021) or publicly available SNP libraries (Grskovic, Marica, et al., 2016, Keller, M., & Agbor-Enoh, S. 2021). Dd-cfDNA quantification methods based on SNP genotyping are now commercially used but suffer from the drawback that only a small portion of cfDNA fragments in the sample have SNPs that are useful for selective amplification. Further, the SNPs that do differ may fall in low cfDNA coverage areas, making them unusable unless sequencing is deep (Keller, M., & Agbor-Enoh, S. 2021). So there is a need to explore methods to distinguish between donor-derived and recipient-derived cfDNA that do not involve whole genome genotyping and SNP identification.

In recent years, Deep Learning (DL), a field of Artificial Intelligence has revolutionized the field of medicine and healthcare due to its ability to solve complex biological problems like protein fold prediction (Alphafold2, Omegafold) (Jumper, J., et al. 2021, Wu, R., et al. 2022), SNP detection (DeepVariant) (Poplin, R., et al. 2018), drug discovery (Keshavarzi Arshadi, A., et al. 2020), and cross-prediction of omics layers, such as epigenetics from the sequence (Enformer) (Avsec, Ž., et al. 2021). DL models use layers of artificial neurons, called hidden layers, to progressively extract higher-level features from the training data. This hierarchical feature extraction from multiple hidden layers makes DL models particularly powerful and capable of learning complex, subtle, and even wholly unsuspected patterns from biological datasets (Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012).

In response to the limitations of current SNP-based methods for quantifying dd-cfDNA, we developed a novel quantification method that utilizes DL models to differentiate between donor- and recipient-derived cfDNA based on their tissue of origin. Due to the predominantly hematopoietic origins of normal circulating cfDNA, the majority of the recipient-derived cfDNA are from blood cells, whereas donor-derived cfDNA originates from the transplanted lung allograft (Lui, Y. Y. N., et al. 2002). Consequently, the models were trained to learn tissue-specific patterns as a proxy for discriminating between donor- and recipient-derived fragments.

Given that all tissue types share the same DNA sequence, sequence information alone is not sufficient to differentiate between lung- and blood-derived fragments. Additional features that vary across tissue types are required for the DL models to perform the classification. Due to the uneven fragmentation process that forms cfDNA, lung and blood-derived cfDNA originate from different regions of the genome. The fragmentation process is uneven because the regions that are bound to nucleosomes and other DNA binding proteins, including transcription factors are protected from enzymatic digestion. Thus, when DNA is broken down by enzymes like DNAse during cfDNA formation, these protected regions become cfDNA (Fig 2) (Snyder, M. W., et al. 2016). In this way, cfDNA contains evidence about the in-vivo epigenetic landscape of their parent tissue, such as the locations of epigenetic regulators like nucleosomes and transcription factors (Snyder, M. W., et al. 2016). Their locations vary across tissue types since they facilitate cell-type-specific gene expression (Brahma, S., & Henikoff, S. 2020). As a result, the regions of the genome that give rise to cfDNA vary across tissue types.

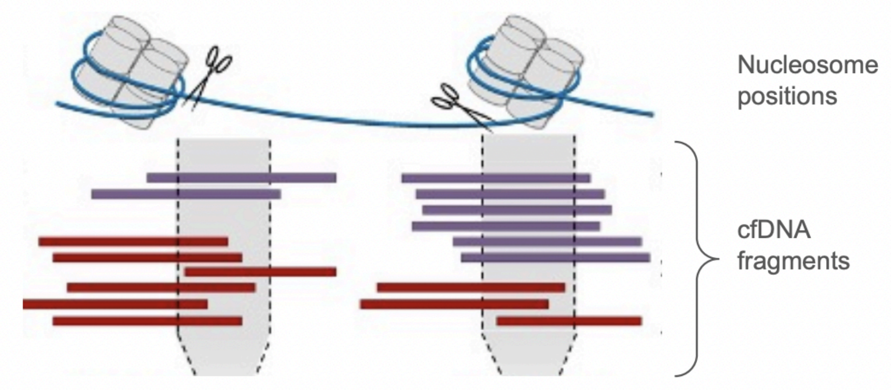


Fig 2: Protected regions (bound to nucleosomes and transcription factors) become cfDNA (Snyder, M. W., et al., 2016).

As a result of originating from different genomic regions, we hypothesized that donor- and recipient-derived fragments would exhibit distinct epigenetic characteristics like chromatin structure, DNA accessibility, epigenetic modifications, and transcription factor binding motifs. However, these properties need to be translated into numerical values, to serve as features for training our DL models. To achieve this, we utilize Enformer, an existing DL model that was trained on human and mouse genomes to predict epigenetic features from DNA sequence alone. Enformer predicts cell-type specific epigenetic features in the form of 5,313 'tracks', where each track corresponds to an epigenetic feature for a single cell line (Avsec, Ž., et al. 2021).

In addition to epigenetic feature variations, donor and recipient-derived fragments would likely exhibit distinct sequence motifs as a result of originating from different genomic regions. Motifs are recurring patterns of nucleotides that are usually associated with functional elements like promoters, enhancers, or coding regions (D'haeseleer, P. 2006). Variations in functional elements, single nucleotide polymorphisms (SNPs), and other epigenetic modifications like histone methylation across the genome would lead to variations in sequence motifs between different genomic regions. Thus, motifs extracted from the DNA sequence could serve as the second feature set for training our DL models.

In this project, we aim to predict the transplant status of 47 lung transplant recipients whose cfDNA fragment sequences were sourced from a study conducted by De Vlaminck et al. in 2015. To this end, we developed three different DL models to classify between donor- and recipient-derived fragments on a per-fragment level : (i) a feedforward neural network trained on Enformer-generated epigenetic features (ii) a convolutional neural network that utilizes patterns in extracted sequence motifs and (iii) a convolutional neural network that uses a combination of extracted motifs and epigenetic features. We utilized the fragment-level predictions from these models to calculate the % dd-cfDNA for these lung transplant recipients, and in turn, predict whether they are undergoing a transplant rejection.

2. RESULTS

2.1 Pipeline for fragment level classification

The workflow for fragment-level classification, depicted in Fig 3, begins with the raw sequencing data from cell-free DNA (cfDNA) fragments of 47 lung transplant recipients at various time points post-transplant. Each cfDNA fragment sequence is defined by its chromosome number, and genomic start and end coordinates.

We first carried out initial investigations to evaluate the separability of the dataset, like fragment length distribution analysis, dimensionality reduction, and performance evaluation of simple models like logistic regression classifiers trained on the dataset. Then, the dataset was split into training, validation, and test sets before training the DL models, to ensure the models are assessed independently of their training data. Subsequently, three different DL models were trained to predict whether each sequence in the dataset is donor- or recipient-derived. The model with the best performance was chosen as the one with the largest area under the ROC curve for the validation set.

For the final evaluation of our dd-cfDNA quantification method, we predicted the transplant status for the test patients using predictions from the best-performing DL classifier. The individual fragment level predictions were aggregated to get the dd-cfDNA percentage for each test patient. Based on the degree of correlation between dd-cfDNA percentages for test patients and clinical signs of rejection, we set out to define a % dd-cfDNA threshold for classifying the transplant status of patients as a rejection.

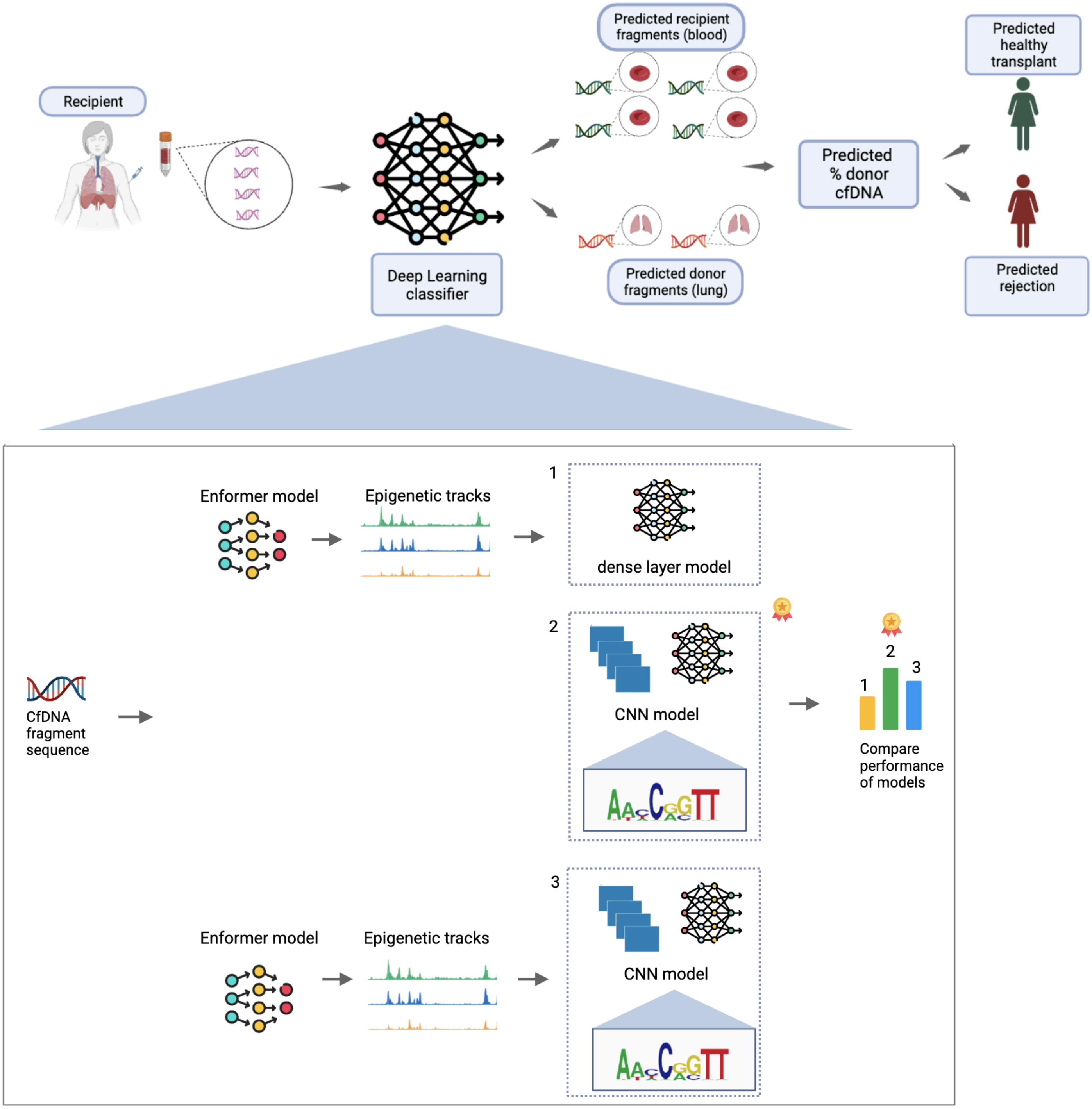


Fig 3: Workflow to predict transplant status for patients

2.2 Fragment size alone is not sufficient to classify cfDNA as donor-derived

We first examined whether the length distribution of cfDNA fragments could be used to classify cfDNA as donor-derived cfDNA (dd-cfDNA) or recipient-derived cfDNA (rd-cfDNA). Since hematopoietically derived cfDNA is longer than cfDNA from other tissues, we hypothesized that dd-cfDNA would, on average, be shorter than rd-cfDNA (Zheng, Yama WL, et al., 2012). Previous studies such as the one by Pedini, P., et al. (2023), achieved an area under the ROC curve (AUCROC) of 0.96 in predicting lung transplant status by classifying shorter cfDNA fragments (80 - 120 bps) as dd-cfDNA. However, our examination of length distribution plots of donor- and recipient-derived fragments for our dataset did not reveal any such distinctive patterns (Fig 4; Kolmogorov-Smirnov test p-value=0.62; See Supplementary Information 7.2.1).

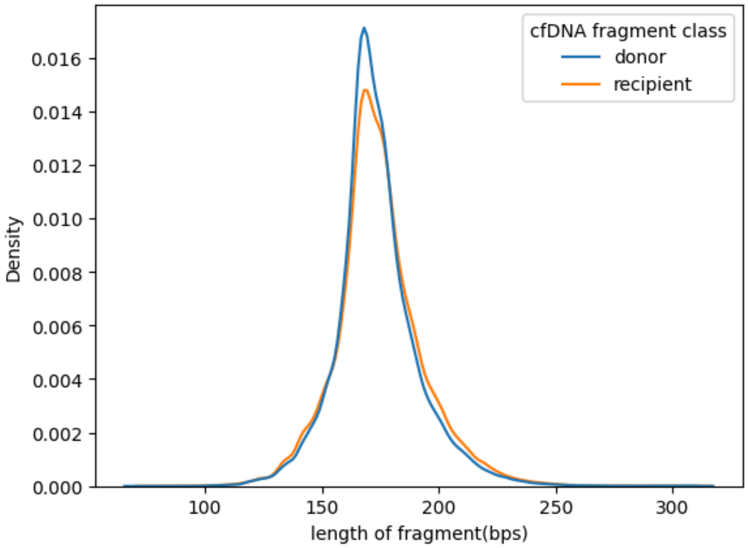


Fig 4: Length distribution of donor- and recipient-derived cfDNA

Additionally, we trained a logistic regression model to distinguish between donor- and recipient-derived fragments using only the length of the fragment. The model yielded an AUC score of 0.52 (Supplementary Figure 1). Relying solely on length as the classification feature only resulted in a marginal 0.02 improvement in the AUC score compared to a random classifier.

2.3 (non-)linear dimensionality reduction on epigenetic features does not separate dd-cfDNA from rd-cfDNA

Since using length alone as a classification feature only yielded a 0.02 improvement in AUC scores over perfectly random performance (i.e. in the absence of signal), we explored whether epigenetic features extracted from cfDNA fragments could offer better classification potential. Since we only have sequence information, we used an existing pre-trained DL model, Enformer (Avsec et al., 2021), to generate 5,313 epigenetic features for each cfDNA fragment (See Methods: 3.4.1). Enformer predicts epigenomic tracks such as DNAse hypersensitivity regions, histone modifications, etc. for many different cell lines from DNA sequence alone. Subsequently, we applied dimensionality PCA (Principal Component Analysis) and t-SNE (t-distributed Stochastic Neighbor Embedding) dimensionality reduction on Enformer-generated epigenetic tracks and visualized the results. We expected that clear epigenetic signals differentiating between dd- and rd-cfDNA, if present, would show up as distinct donor and recipient clusters in the 2D projection, and simple classifiers trained on these signals would perform well.

2.3.1 PCA:

We used PCA to reduce the 5,313 epigenetic features to 100 principal components that in total explained 97% of the variance in the data. Visualization of the two principal components that explained the maximum variance did not reveal clusters separating dd-cfDNA from rd-cfDNA (Fig 5a). Notably, the first two principal components account for only 0.5% of the total variance in the data. This indicates that a) the data has many noisy features that may be irrelevant to the difference between dd- and rd-cfDNA and/or b) the lower-dimensional representation may not faithfully capture the higher-dimensional patterns in the data.

To address the first concern, we only retained the 73 predicted epigenomic tracks associated with lung and blood cell types, anticipating these tracks to differ the most between dd-cfDNA and rd-cfDNA. PCA was performed again on these selected tracks. However, the percentage of variance covered by the first two principal components only improved by 0.3 percentage points compared to the PCA results from epigenetic features for all cell types. Visualization of PCA results after feature selection again showed no clear clusters separating dd-cfDNA and rd-cfDNA (Fig 5b).

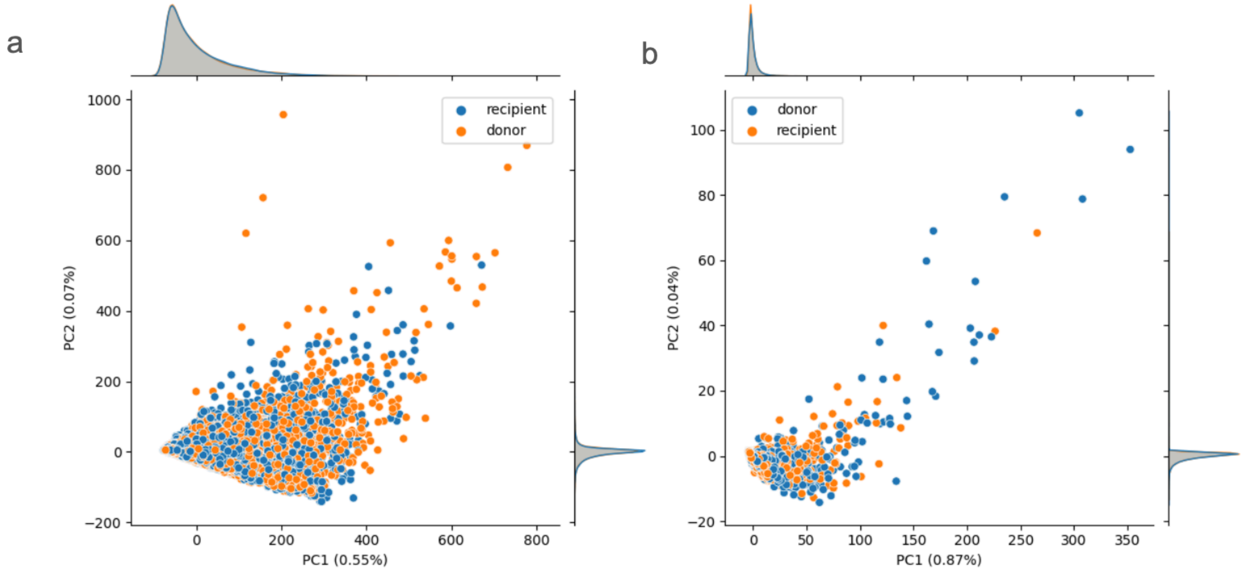


Fig 5: PCA plots for Enformer-generated epigenetic features. Top and side: kernel density-estimated distribution of the projection coordinates. Neither plot shows a separation of cfDNA fragments into donors and recipients. Plots were constructed using 25,000 donor- and recipient-derived cfDNA fragments each.
a. PCA plot of all 5,313 epigenetic features.
b. PCA plot for 73 epigenetic tracks associated with lung and blood cell types.

2.3.2 T-SNE:

To rule out the possibility that the variance in the dataset is not captured by linear combinations of features, we used the non-linear dimensionality reduction method t-SNE to reduce the dimensionality to 2 dimensions with a focus on retaining local high-dimensional neighbors. We again saw no separation between donor- and recipient-derived cfDNA, irrespective of the perplexity values used (Fig 6 and Supplementary Figure 2).

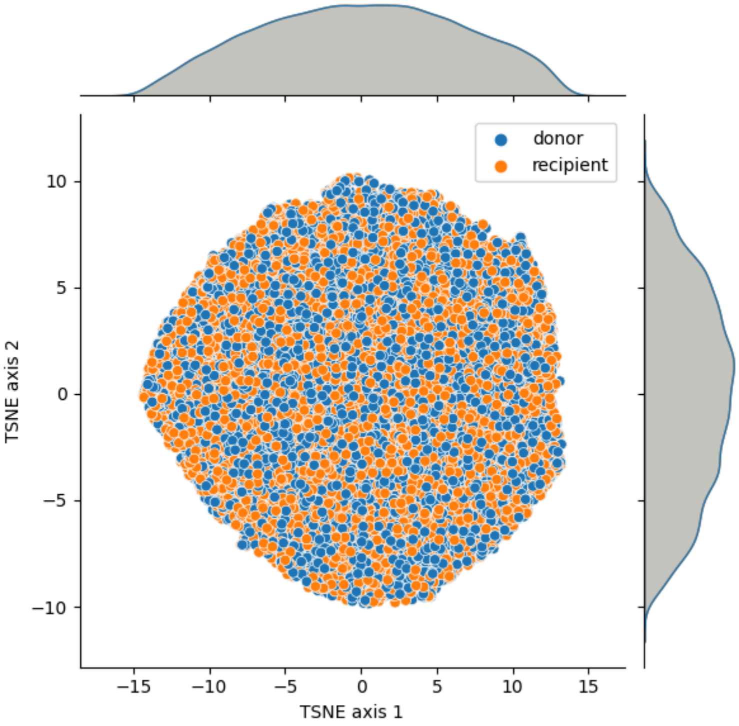
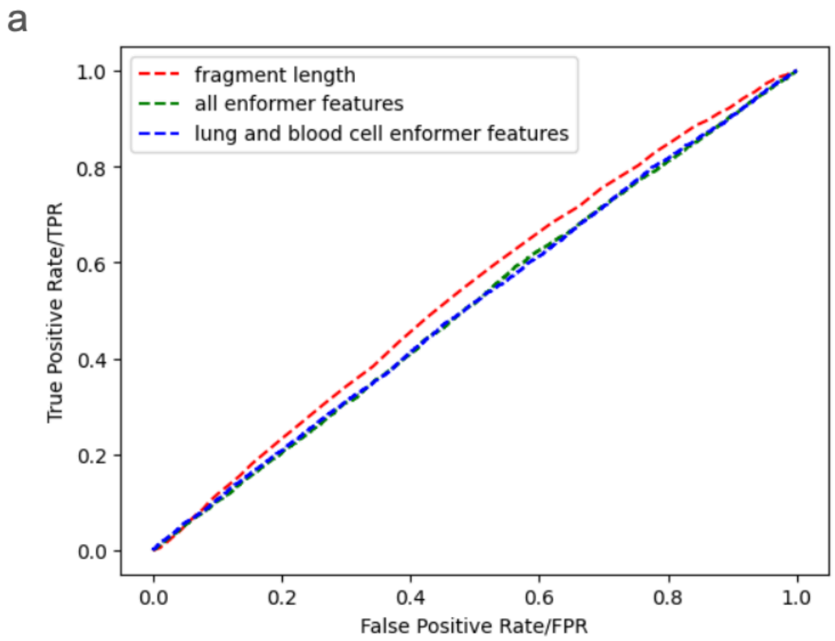


Fig 6: t-SNE visualization of Enformer epigenetic tracks

t-SNE plot of 5,313 Enformer epigenetic features (with perplexity 40) did not reveal clustering of donor- and recipient-derived cfDNA fragments.
Top and side: kernel density estimation plots of t-SNE projection coordinates. Refer to Supplementary Figure 2 for t-SNE plots with other perplexity values.

2.4 A logistic regression model cannot distinguish dd- from rd-cfDNA based on Enformer epigenetic tracks

We trained a baseline logistic regression model that classified fragments as donor- or recipient-derived using the first 100 principal components (97% variance explained). This model was intended to serve as a baseline for comparing the performance of the DL modes. The model performed poorly, with an AUC of 0.5, no better than random chance (Fig 7). We also trained a logistic regression model using only the 73 epigenetic tracks belonging to lung and blood cell types as features. This model, with an AUC of 0.51 offered no meaningful performance improvement over the previous one (Fig 7). Given the inseparability of the data using dimensionality reduction techniques, and the poor performance of our baseline model, we next investigated whether DL models could learn a suitable classification function on this complex data.



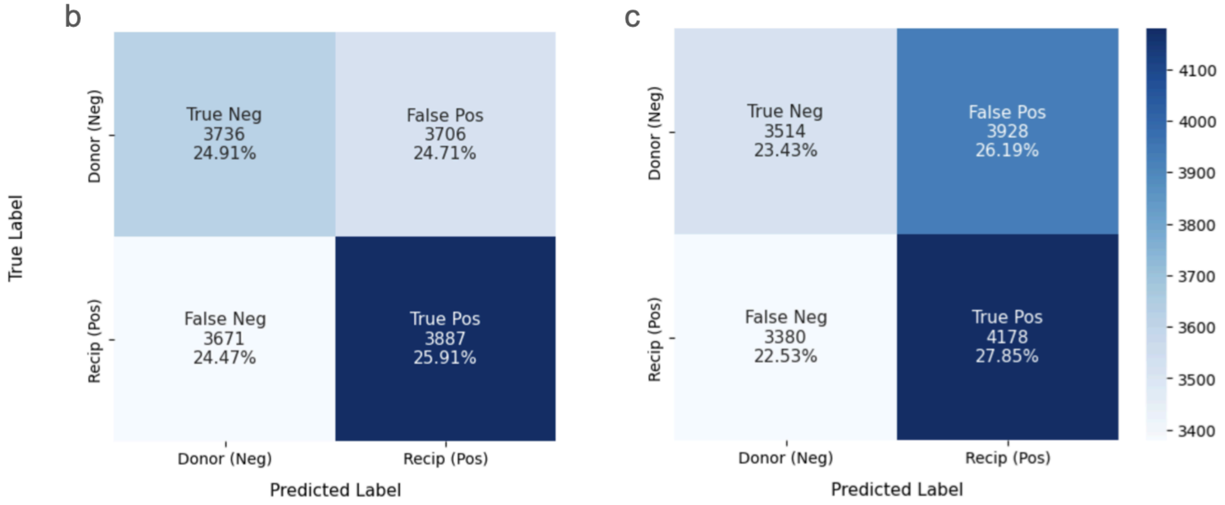


Fig 7: Performance of the logistic regression model trained on epigenetic features

- **a.** Comparison of ROC curves for logistic regression model trained on fragment length (AUC 0.54), all 5,313 epigenetic features (AUC 0.50), and 73 lung & blood cell-specific epigenetic features (AUC 0.51).
- **b.** Confusion matrix for logistic regression model trained on all 5,313 epigenetic features. Predictions are almost perfectly random (25% in each quadrant).
- **c.** Confusion matrix for logistic regression model trained on Enformer features associated with lung and blood cell types. There is a slight bias towards classifying fragments as recipient-derived.

2.5 A fully-connected feedforward neural network model trained on Enformer epigenetic tracks had low AUC scores.

Since the logistic regression model trained on epigenetic features only resulted in a small improvement in AUC score (0.01) compared to a random classifier, we suspected that the underlying patterns separating the two classes might be too complex for simple linear models. Since DL models are known to excel at learning representations for complex and noisy data due to their layered structure, we trained a DL model on 5,313 Enformer-generated epigenetic features to distinguish between dd-cfDNA and rd-cfDNA (refer to Methods 3.4.2 and 3.4.3 for details on the architecture and training process), using a fully-connected feed-forward neural network (FFNN) architecture.

However, the FFNN model did not yield improvements in the AUC score compared to the logistic regression model, with training and validation AUCs of merely 0.51 and 0.50 respectively (Fig 8 a). Interestingly, the Cross-Entropy Loss, a measure of the disparity between the model's predictions and the true labels, dropped sharply after the first epoch of training, giving the impression that some meaningful patterns were learned by the model (Fig 8 b). However, this decline in loss can be attributed to the shift in predicted probabilities from more extreme values to about 0.5 after the first epoch. Hence, the model primarily learned to be uncertain about its predictions as a strategy to minimize the loss.

We hypothesized that the model is not converging to a solution since Cross Entropy loss does not reduce after the first epoch. To address this problem, we explored various combinations of hyperparameters including learning rates, number of layers, and the number of neurons within a layer.

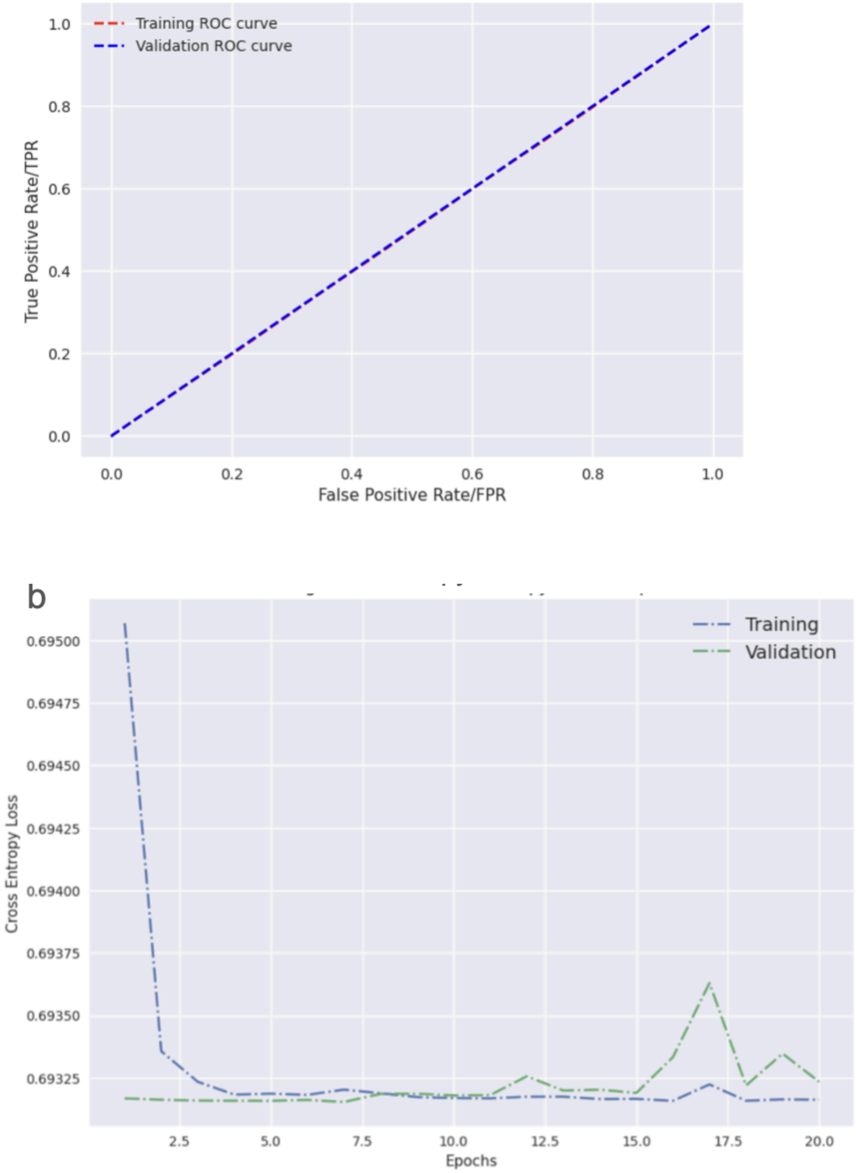


Fig 8: Performance of the fully-connected neural network trained to classify cfDNA fragments as donor- or recipient-derived using 5,313 Enformer-generated epigenetic features.

- **a.** Training(AUC: 0.51) and Validation(AUC 0.50) ROC curves
- **b.** Training and validation Cross-Entropy loss plot over epochs. Loss drops sharply after the first training epoch but does not reduce further during training.

2.5.1. Learning rate optimizations did not improve AUC scores

We trained the FFNN model on six different learning rates ranging from 0.1 to 0.0000001. AUC score from ROC curves was used as the metric for comparing the performance of all these models (Fig 9).

Fig 9a depicts the training and validation AUC scores for models where the learning rates were kept constant throughout the training process. Models trained using lower learning rates had higher AUC scores. However, even the best-performing model trained with a learning rate of 1e-7 with an AUC of 0.506 only yielded a marginal improvement over a random classifier. Next, the learning rates were progressively reduced during the training process from the initial value to nearly zero, following a cosine curve (Refer to Supplementary Figure 3) (Loshchilov, I., & Hutter, F., 2016). A comparison of the AUC scores for various initial learning rates shows that once again, there are no significant improvements in AUC scores (Fig 9b). Notably, models trained using a Cosine learning rate scheduler even slightly underperformed those trained with a constant learning rate.

Through learning rate optimizations, we were unable to find a learning rate that improved the validation set AUC score compared to the baseline dense layer model. So, we conclude that an incompatible learning rate is not the reason for the model's inability to converge to a solution.

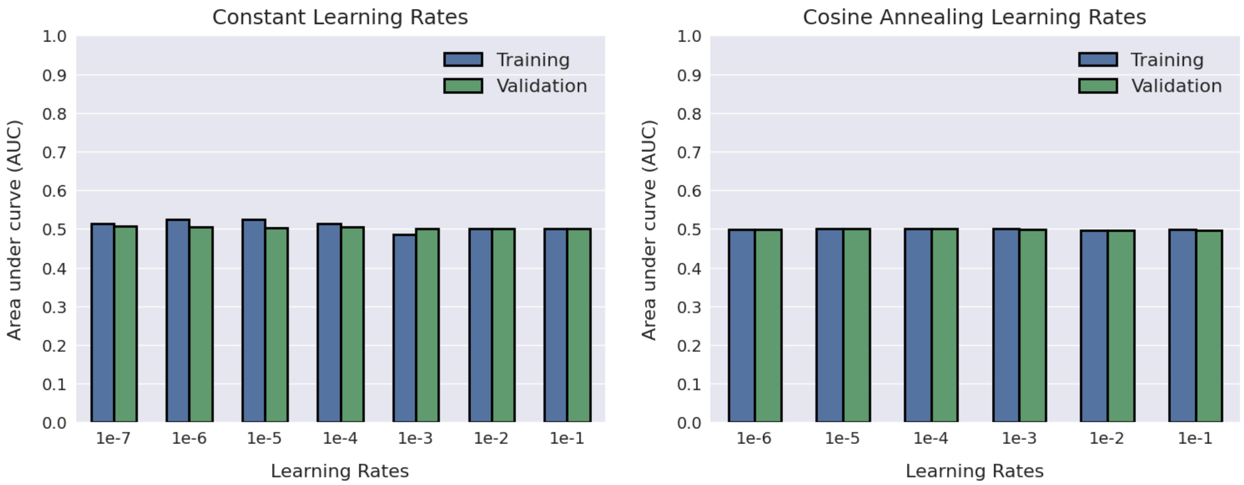


Fig 9: Comparison of ROC AUC scores for models trained using varying learning rates

- **a.** Training and validation AUC scores for constant learning rates.
- **b.** Training and validation AUC scores for learning rates varied using a cosine function scheduler. .

A constant learning rate of 1e-5 resulted in the highest training set AUC of 0.52. The highest validation AUC score of 0.506 was observed for the lowest attempted learning rate of 1e-7

2.5.2 Model architecture and training process optimizations did not improve AUC scores

The FFNN model consisting of two hidden layers was trained on 5,313 Enformer-generated epigenetic tracks as features. This involves training over 11 million weights, whereas only 500,000 samples were used for the training process. The high complexity of the model relative to the number of training samples could explain the challenges faced by the model in finding generalisable patterns rather than shortcut solutions (Geirhos, R., et al., 2020).

We tried two approaches to simplify the model. The first approach involved a form of feature selection, wherein only Enformer tracks corresponding to blood and lung cell types were retained for training. This amounted to a total of 73 input tracks and vastly decreased the number of parameters to fit. However, the validation AUC score of this model remained at 0.5, indicating that lung and blood cell tracks indeed don't have more separability than the other tracks (Fig 10). This outcome aligns with the findings from PCA analysis and logistic regression plots (Fig 5 and 7a), where restricting the input to only lung and blood cell tracks did not improve separation or classification performance, respectively. In the second approach, we changed the number of hidden layers in the architecture, removing or adding one. Neither approach yielded improved AUC scores (Fig 10).

As a final optimization step, we experimented with alternate Loss functions and optimizers. Using Binary Cross Entropy Loss (BCELoss) instead of Cross Entropy Loss did not change the AUC scores. Similarly, training using the Stochastic Gradient Descent (SGD) optimizer instead of the Adam optimizer did not yield improvements in AUC scores. A comparison of the AUC scores for all the hyperparameter optimizations is illustrated in Fig 10.

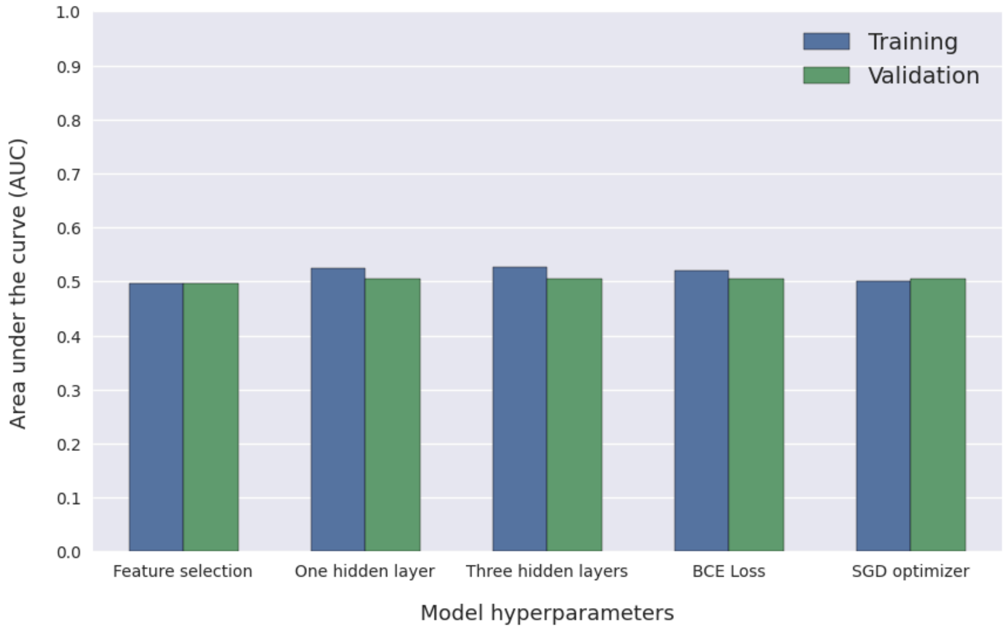


Fig 10 Comparison of AUC scores for various hyperparameter optimizations. None offer improvements on validation set performance. Feature selection: reducing 5,313 Enformer-generated epigenetic tracks to the 73 lung and blood cell-type-specific ones. BCE: binary cross-entropy. SGD: stochastic gradient descent.

2.6 Artificially embedded signals greatly increase the performance of the FFNN model

To investigate whether the low AUC scores are a result of a lack of meaningful patterns in the data or a flawed training methodology, we trained the FFNN model on simulated data with augmented signals. The simulated data consists of synthetically generated samples with different magnitudes of signal added in, that specifically distinguishes between donor- and recipient-derived fragments. In addition to investigating possible flaws in training methodology, we also aimed to explore data-related problems like the presence of noise through the simulations.

2.6.1 Creating simulated data and adding artificial signals:

A two-step process was followed to create synthetic data with embedded signals differentiating the donor and recipient-derived samples. We first created 625,000 synthetic samples by uniformly sampling between the training-data-wide minimum and maximum values for each of the 5,313 epigenomic features. These samples were then randomly assigned as donor- or recipient-derived ensuring an equal distribution of donors and recipients to create a class-balanced dataset. We next split the data into 500,000 training and 125,000 validation samples. A classifier trained on these samples should perform similarly to a random classifier, since these samples, so far, do not have any signal to discriminate between donor and recipient fragments.

In the second step, we progressively added more signals that distinguish between donor- and recipient-derived classes to the samples. Some percentage '*s'* of all samples were modified: donor-derived samples by replacing some percentage '*n'* of their features (tracks) with a value slightly exceeding the maximum observed value for that track, recipient-derived samples by replacing some of their tracks by a randomly generated value slightly lower than the minimum observed value for that track (See Supplementary Figure 4). We varied the fraction '*s'* of samples and the fraction '*n'* of features to which the signal was added (5-90%, equally divided between donor and recipient samples).

2.6.2 Comparison of performance for varying levels of artificial signals

A total of 36 synthetic datasets of training and validation data were created using the process described above. Each set had a unique combination of percentage of samples and features that were augmented, ranging from 5% to 90%. The standard two-hidden layer FFNN model was trained on all these datasets. The validation set AUC scores for a representative sample of these models are summarized in the heatmap in Fig 11. The most notable observation is that the percentage of samples with augmented signal had a greater impact on the AUC scores than the percentage of features with augmented signal. For instance, datasets featuring only 10% of samples with augmented signals consistently had lower AUC scores, regardless of the percentage of augmented features. Conversely, when datasets comprised >70 % samples with augmented signal, even if they included only 10% of the features, the model had significantly improved performance. This shows that the FFNN architecture can pick up on small but consistent discriminating signals, if present in enough of the data. It is noteworthy that AUC scores were much higher than all models trained on real data so far, even in low augmentation regimes.

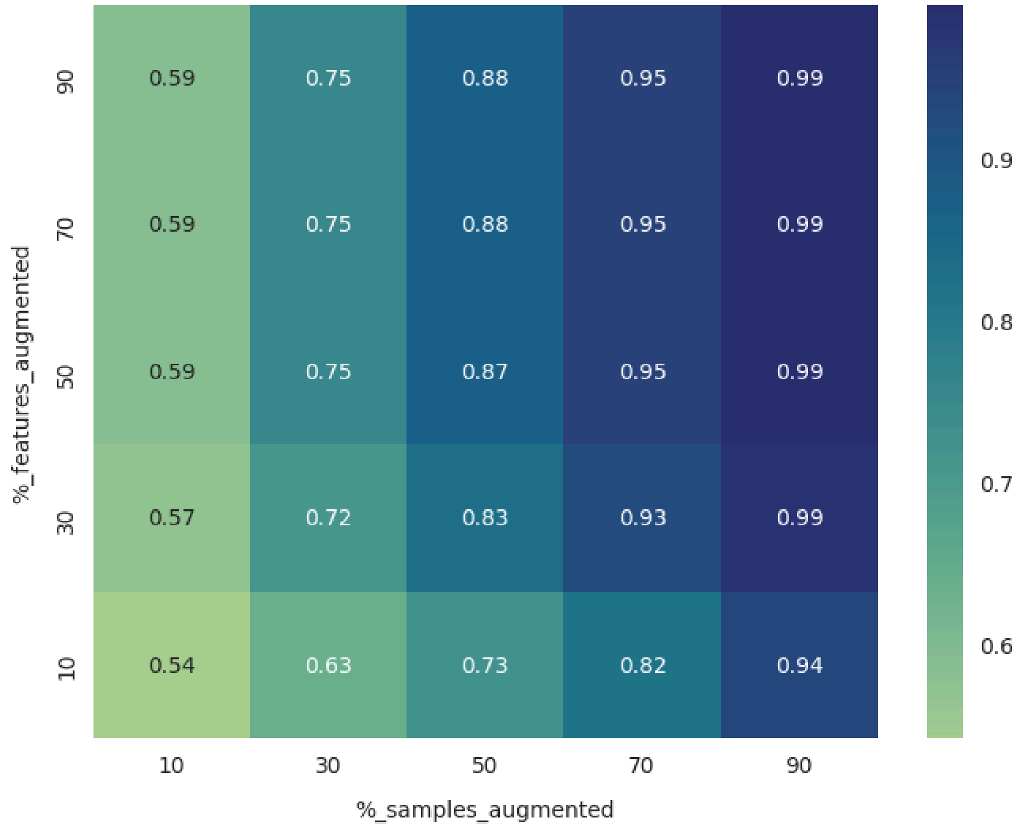


Fig 11: Heatmap of AUCs for varying percentages of augmented features and samples
The x-axis represents the percentage of samples in the training and validation set which were augmented with signals differentiating the donors from recipients. The y-axis similarly represents the percentage of features that were augmented with differentiating signals. The values in the heatmap represent the AUC scores for the models that were trained on a dataset with the given combination of %_samples_augments and %_features_augmented.

2.6.3 Interpreting the results from simulations

The randomly generated data points that are not augmented with a signal can be interpreted as noise in the dataset since they do not contain any meaningful signal for classification. Increasing the percentage of augmented samples and features is equivalent to removing noise from the data. The model achieved an AUC of 0.9 when at least 70% of the training samples contained meaningful signals for classification. This proves that the model performs well with low noise levels in the data. The low AUC scores observed on the real epigenetic features could thus be attributed to the absence of signal in the data, or signals that are too diluted or noisy for the model to learn.

The observation that increasing the percentage of augmented samples caused a greater improvement in AUC scores than increasing the percentage of augmented features means the model is more sensitive to noisy samples than noisy features.

Further, we observe that for augmented sample percentages exceeding 50, AUC scores remain favorable (> 0.75) even with only 5% of features containing useful information for classification (data not shown). This suggests that the model performs well in scenarios where enough samples carry meaningful signals, even if the majority of the features are noisy. This, combined with the model's sensitivity to noisy samples implies that the real lung transplant dataset likely suffers from the presence of noisy samples, i.e. samples that have no signals to separate donor from recipient class. However, it is important to acknowledge that the signal in real data could arise from a complex combination of features. Consequently, the required percentage of non-noisy features for achieving high AUC scores is likely understated in the simulation analysis compared to the real dataset.

2.7 CNN model could not identify sequence motifs distinct to donor and recipient fragments

We next explored whether a Convolutional Neural Network (CNN) trained on cfDNA fragment sequences could successfully classify fragments as donor- or recipient-derived. We reasoned that predicting epigenomic tracks on small cfDNA fragments might be too unreliable, whereas distinct sequence motifs might separate lung- and blood-derived cfDNA if enough of them derive from genomic locations that become cfDNA more often in one or the other tissue. The CNN model consisted of two convolutional filters to identify repetitive motifs or patterns associated with transcription factor binding sites, nucleosomes, or other functional and regulatory elements. The convolutional filters are followed by simple fully connected layers that distinguish between donor- and recipient-derived fragments based on these extracted motifs as features. However, despite extensive optimization efforts, the CNN model exhibited only marginal improvement in AUC scores over the previously trained models, with the best-performing model achieving an AUC of 0.524 on the validation set.

2.7.1 Learning rate optimizations

As part of the hyperparameter optimization process, we trained the CNN model with learning rates ranging from 0.1 to 0.0000001. Despite some improvement in training AUCs for certain learning rates, validation AUCs showed no improvement over the previous models (Fig 12).

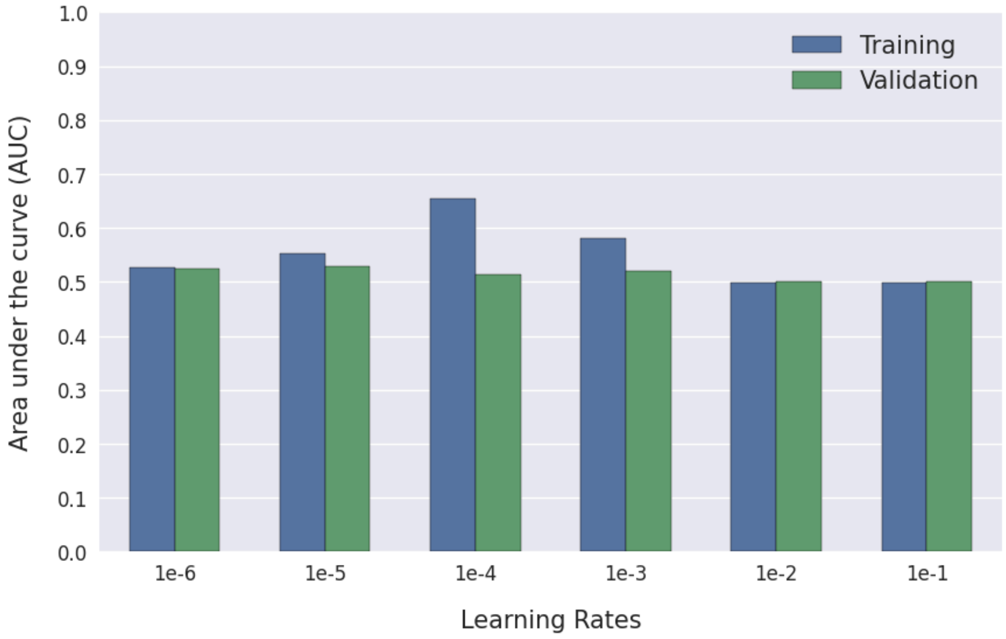


Fig 12: Comparison of AUC scores for various learning rates - CNN model

2.7.2 Regularization methods to address overfitting

Comparison of training and validation AUC scores show that, on average AUC scores are higher for the training set compared to the validation set (Fig 9a, Fig 10, Fig 13). This implies overfitting, since the model performs better on the training set, but fails to generalize well on unseen data. To address this issue, we implemented regularization techniques such as dropout and weight decay.

Dropout involves randomly setting a fraction of neurons to zero at each update during training (Srivastava, N., et al., 2014). This introduces noise into the training process and prevents the model from perfectly fitting to the training data. We implemented dropout by inserting drop-out layers, where the output of a fraction of neurons from the previous layer are set to zero, after every convolutional filter and dense layer. We trained the CNN model for dropout probabilities ranging from 0.1 to 0.9, where dropout probability is the proportion of neurons that were set to zero at each dropout layer. However, the addition of dropout layers only led to a deterioration of performance on the training set, without corresponding improvements on the validation set (Fig 13a). With the highest AUC of 0.5, the CNN model did not perform better than a random classifier for any of the dropout probabilities.

Weight decays prevent overfitting by discouraging the model from assigning excessive importance to any specific feature (Loshchilov, I., & Hutter, F., 2017). This is achieved by adding a regularization term that is a function of the magnitude of weights to the loss function, hence penalizing large weights. We trained the CNN model with weight decay ranging from 0.0001 to 0.1, where weight decay refers to the weight assigned to the regularization term. While the weight decay mitigated some overfitting (seen by the reduced training AUC scores), there were no corresponding improvements in the validation performance. As with dropout layers, for all weight decays, the CNN model only had an AUC of 0.5 (Fig 13b).

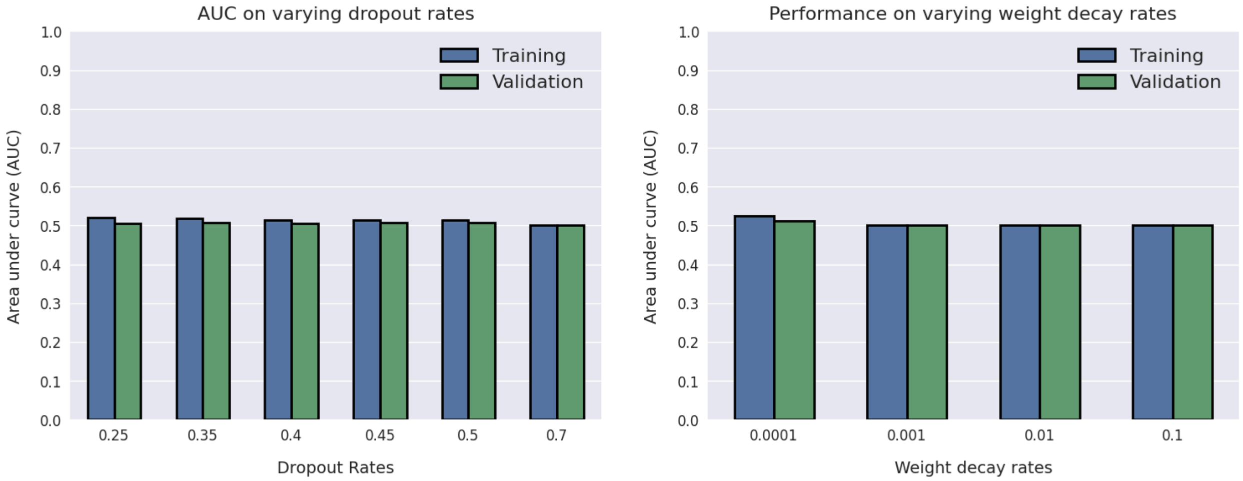


Fig 13: Comparison of AUC scores for various types of regularization

- **a.** Comparison of training and validation AUC scores for varying drop-out rates. We don't see any significant improvement in the validation AUC scores for any drop -out rate
- **b.** Comparison of training and validation AUC scores for varying rates of weight decay. Again, although the training performance worsened, we don't see any improvements in the validation AUC scores for any weight decay rate.

Given that attempts to address overfitting did not increase the validation AUC scores, we concluded that there is likely a lack of generalizable signals in the sequence motifs that separate the donor- and recipient-derived fragments.

2.7.3 Performance on simulated data :

To investigate whether the low validation AUC scores result from a lack of meaningful signals or flawed training methodology, we trained the CNN model with simulated data. To create the simulated data, we first generated random DNA sequences matching the CNN input size and randomly assigned them as donor- or recipient-derived. The last 4 bases in all donor-derived fragments were replaced with T's, while for the recipient-derived fragments, they were replaced with A's. The model achieved 83% accuracy on the simulated dataset, suggesting that it performs well in the presence of clear signals that separate the two classes. Based on these simulations, we conclude that the low AUC scores for the CNN model are likely attributed to the lack of a clear signal separating dd-cfDNA and rd-cfDNA in the real dataset.

2.8 CNN model trained on a combination of epigenetic features and sequence motifs yielded low AUC scores

To address the lack of clear signals in epigenetic features and sequence motifs when used separately, we investigated whether a combination of both possesses the required signals for classification. To this end, we trained a CNN model that extracts sequence motifs from cfDNA fragments and utilizes these motifs in addition to Enformer-generated epigenetic fractures for classification. However, this model did not yield significant improvements in AUC scores compared to the previous models, with training and validation AUC ROC scores of 0.574 and 0.542 respectively.

2.9 Patient-level aggregate classification does not correspond to clinical signs of rejection

Of all the three DL models that were trained for distinguishing dd-cfDNA from recipient-derived cfDNA, the CNN model that utilizes both sequence motifs and Enformer-generated epigenetic features (combined CNN model) yielded the best classification results, with an AUC score of 0.542 on the validation set (Fig 14). Hence, this CNN model was utilized for the final evaluation of our approach on the test set.

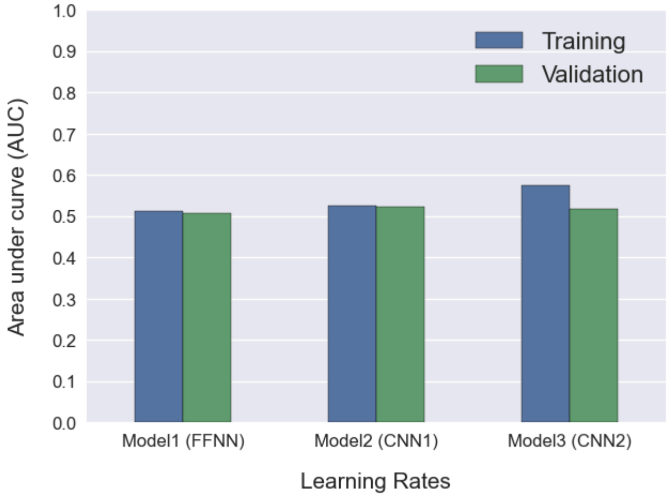
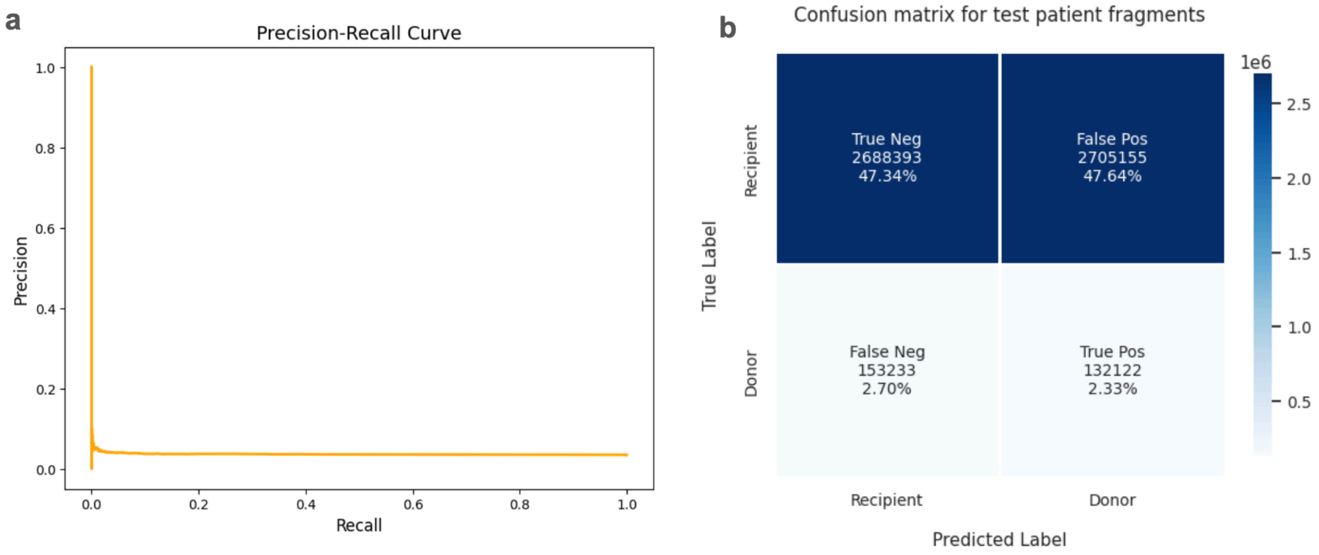


Fig 14: Comparison of AUC scores for three DL models for classification of cfDNA into donor and recipient-derived.
Model 1: FFNN trained on Enformer-generated epigenetic features (training AUC: 0.507 validation AUC: 0.506)
Model 2: CNN trained on cfDNA sequence to extract sequence motifs (training AUC: 0.53 validation AUC: 0.524)
model 3: CNN trained to utilize both sequence motifs and Enformer-generated epigenetic tracks for classification (training AUC: 0.574 Validation AUC: 0.542)

The test set consists of samples collected across various post-transplant durations(0 - 2500 days) from 30 patients. The recipient-derived fragments were not undersampled in the test set, leading to a huge class imbalance. Taking this class imbalance into account, we used the

F1 score as the metric for evaluating model performance, since the F1 score is less sensitive than the AUC score to class imbalance (Fig 16a). From the confusion matrix (Fig 16b), we see the CNN model performed similarly to a random classifier, with almost 50% of the samples predicted as dd-cfDNA, while only 0.05 % of the fragments are dd-cfDNA according to the true labels. The F1 score of the model was just 0.0854.

Subsequently, we calculated the % dd-cfDNA for all test patients, by aggregating individual fragment-level predictions of the best-performing model. The goal is to investigate whether these predicted donor percentages correlated with transplant rejection in the test patient group. At the fragment level, although the AUC indicates a performance only slightly better than random chance, the number of true positives and true negatives still exceeded false positives and false negatives. So we anticipated that predicted and true donor percentages will follow a similar pattern where patients with higher true predicted donor percentages will also have higher predicted donor percentages. However, there was almost no observed correlation (Fig 16 c) with a Pearson correlation coefficient of -0.08 (p-value = 0.562).



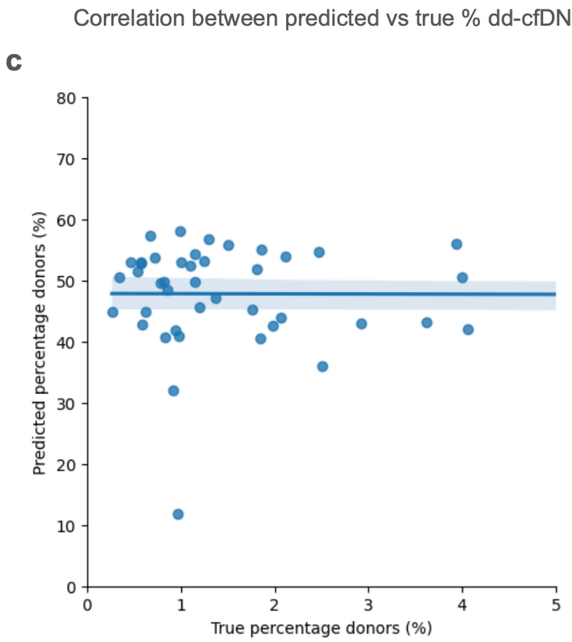


Fig 16: Performance of combined CNN model on test patients

- **a.** The precision-recall curve of the combined CNN model for test patients (F1 score: 0.0854)
- **b.** Confusion matrix of the combined CNN model for test patients.
- **c.** There was no correlation between true % dd-cfDNA and predicted % dd-cfDNA, with a Pearson correlation coefficient of -0.08

Lastly, we assessed the viability of predicted % dd-cfDNA in post-transplant days as an indicator of transplant rejection. Our approach involved identifying specific time points associated with clinical signs of rejection for each patient and analyzing patterns in predicted % dd-cfDNA during these intervals (Patient metadata, including clinical signs of rejection, were sourced from the original study, the details of which are provided in the section on Code and Data availability). Given the association between elevated donor cfDNA levels and transplant rejection, we anticipated a concurrent increase in predicted % dd-cfDNA at these identified time points. However, our dataset revealed no discernible correlation between higher % dd-cfDNA levels and clinical signs of rejection (Fig 13 d). Interestingly, this lack of correlation also extends to the % dd-cfDNA calculated from true labels, prompting suspicions about the accuracy of true labels used for training the models (Fig 14). This also raises questions about the reliability of observed clinical signs of rejection as a definitive indicator of transplant rejection. Thus, both inaccuracies in the training labels and the limited predictive capabilities of our DL models are contributing factors to the observed lack of incongruence of predicted % dd-cfDNA with clinical signs of rejection.

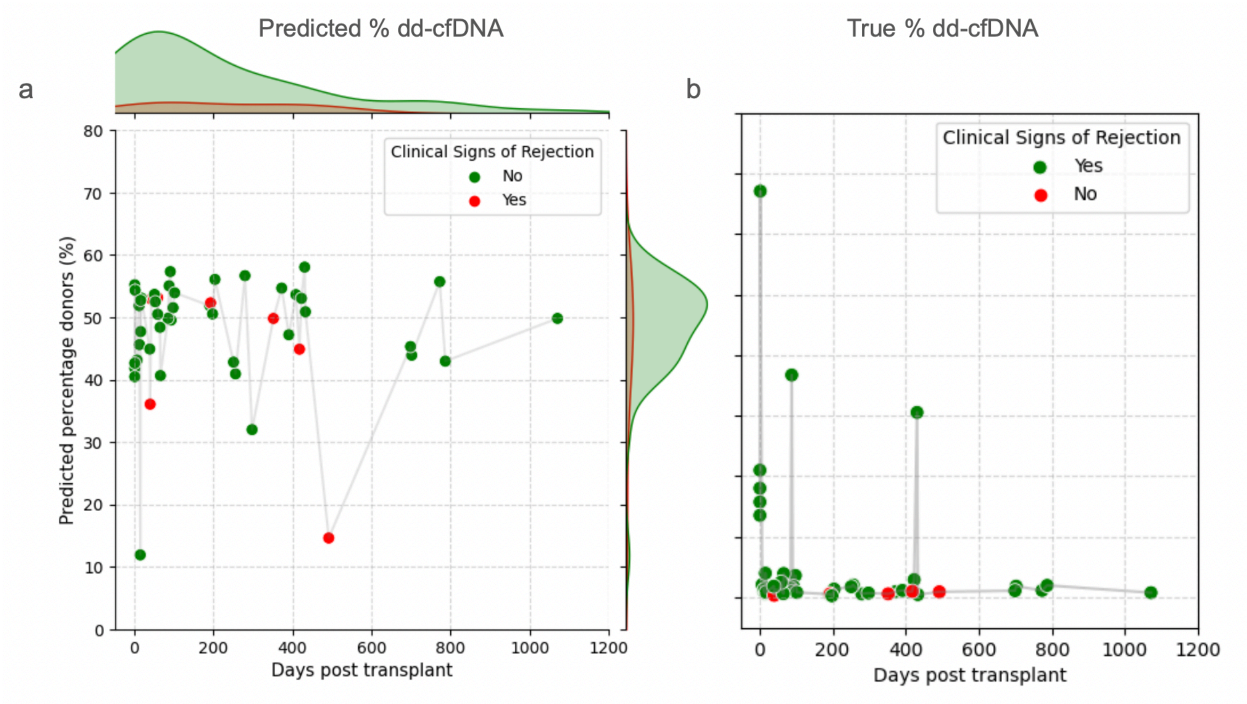


Fig 13: Correlation of predicted % dd-cfDNA with clinical signs of transplant rejection.

- **a.** Predicted % dd-cfDNA for all test patients across various days post-transplant. The predicted % dd-cfDNA does not correlate well with observed clinical signs of rejection.
- **b.** True % dd-cfDNA for all test patients across various days post-transplant. It also does not correlate well with observed clinical signs of rejection

3. METHODS

3.1 Data and label acquisition:

The data used in this project was obtained from a study conducted by De Vlaminck *et al.* involving 47 patients listed for a lung transplant at the Stanford University Hospital (Fig 14a)

The true donor and recipient labels for the cfDNA fragment dataset were obtained from the study by De Vlaminck et al. 2015, which also served as the source of our dataset. The labels were the results of a diagnostic assay developed by the authors of the study to distinguish between donor and recipient-derived cfDNA fragments. The diagnostic assay involves the following steps. First, genotyping was performed on whole blood samples collected from donors and recipients, to create donor and recipient-specific SNP libraries (Fig 14b). SNPs for the libraries were selected from single-base alleles that differed between donors and recipients and were homozygous within each individual. After transplantation, cfDNA fragments were sequenced from whole blood samples collected from the recipient. The fragments were then labeled as donor or recipient-derived based on the presence or absence of donor and recipient SNPs from the SNP library (Fig 14c) (Refer to Supplementary Information 7.2.2 for more details on sequencing and genotyping).

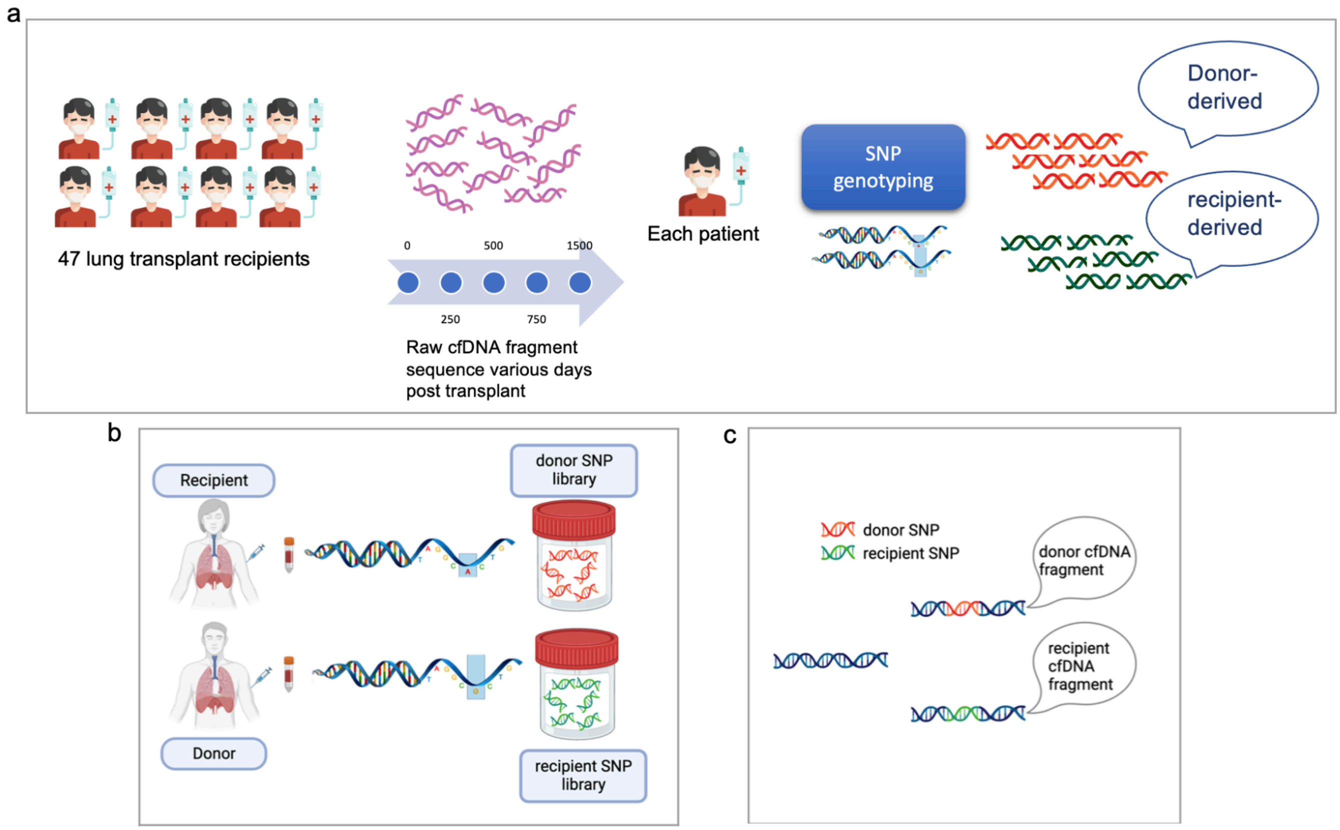


Fig 14: Assigning donor and recipient true labels to cfDNA

a. Data for the project was obtained from 47 lung transplant recipients. It consists of raw cfDNA sequences over various days post-transplant. The authors of the study labeled the fragments for each patient as donor- or recipient-derived using SNP genotyping.

b. Creation of donor and recipient SNP libraries before transplant

c. Discriminating between donor- and recipient-derived cfDNA using SNP biomarkers.

3.2 Downsampling

The dataset used for our project has a huge class imbalance since only 5% of the total samples are dd-cfDNA. Such a class imbalance can bias the models toward predicting the majority class. So, before the training process, we downsampled the recipient-derived cfDNA (rd-cfDNA) class by randomly selecting only a few rd-cfDNA fragments per patient, such that the number of donor- and recipient-derived cfDNA fragments are equal for that patient.

3.3 Train, Validation, and Test split

The dataset was split into training, validation, and test sets to ensure that the models are evaluated independently of the data they were trained on. The test set comprises samples from 10 out of 47 patients and was reserved for evaluating the final model’s performance. The test samples were excluded from training and hyperparameter optimization to prevent any inadvertent data leakage. For the remaining 41 patients, the samples were split into training and validation sets such that fragments from chromosomes 1 to 11 constituted the training set, while the validation set encompassed fragments from the remaining chromosomes. This strategic partitioning at the chromosome level aimed to distribute approximately 80% and 20% of the non-test samples to the training and validation sets respectively. Chromosome-based partitioning was a deliberate measure to prevent any risk of data leakage, given the potential overlap of cfDNA fragments between different patients.

3.4. Feedforward neural network (FFNN) trained on epigenetic features

3.4.1 Generating Enformer output

Enformer requires a one-hot encoded DNA sequence of size 196,607 bps for predicting epigenetic features. To adapt cfDNA fragments (of average length 167 bps) to Enformer input size, we symmetrically extended the genomic start and end coordinates using the human reference genome sequence (GRCh38.p14). This extension results in a sequence measuring 196,607 bps, with the original cfDNA fragment positioned in the middle. These sequences were then one-hot encoded to convert the bases into numerical values, which were provided as input to a trained Enformer model. For each input sequence, Enformer predicts 5,313 epigenetic tracks that are divided into 128 bp bins each. The predictions from the two central bins, covering 256 bps, were averaged for all the tracks to generate 5,313 features for training the feed-forward neural network. We chose to use the two central bins since together, they cover 256 bps, and hence are sufficient to encompass an average cfDNA fragment of length 167 bps.

3.4.2 Feedforward neural network architecture

The feedforward neural network featured an input layer with 5,313 neurons and an output layer with two neurons, representing the likelihood of the input being recipient-derived or donor-derived. Positioned between the input and output layers were two or three hidden layers, with Rectified Linear Unit (RELU) activation layers inserted between them to introduce non-linearity. The hyperparameters governing the number of hidden layers and neurons per layer were tuned during the training process to optimize the models' performance.

3.4.3 Feedforward neural network training

The training was conducted with a batch size of 128 for 20 epochs using the Cross-Entropy Loss function and Adam optimizer (Kingma, D. P., & Ba, J. 2014). During the training and validation process, hyperparameters like learning rates, number of hidden layers, and neurons per layer were manually tuned, by comparing model performances using the area under the ROC curve (AUCROC).

3.5 Convolutional Neural Network (CNN) model

3.5.1 Generating one-hot encoded sequences for CNN

The CNN model requires one-hot encoded cfDNA fragment sequences of a fixed length, aligning with the kernel size of 330 in the first convolutional layer. To achieve this, cfDNA genomic coordinates were extended symmetrically on both sides until they reached the specific kernel size. Subsequently, the corresponding sequences were extracted from the human reference genome (GRCh38.p14) and one-hot encoded before being input to the CNN filters.

3.5.2 CNN model architecture

The CNN model consists of two convolutional layers, followed by simple fully-connected layers. In each convolutional layer, multiple convolutional filters extract specific motifs or features from the input cfDNA sequence, creating a combined feature map. After each convolutional layer, max-pooling layers were added to downsample the number of motifs in the feature map, ensuring that the final output from the CNN maintains a size of n*5. Here, 'n' denotes the number of convolutional filters in the final layer and 5 represents the number of motifs in the final feature map after downsampling. This is followed by a fully-connected input layer with n*5 neurons and an output layer with two neurons, representing the likelihood of the input being recipient-derived or donor-derived.

3.5.3 CNN model training

Training was conducted with a batch size of 128 over 20 epochs using the Cross Entropy Loss function and Adam optimizer (Kingma, D. P., & Ba, J. 2014). We manually tuned hyperparameters like learning rate, the number of convolutional layers, the number of filters per layer, filter size, and the complexity of the fully-connected layers. Model performance was compared using the AUCROC on the validation set to assess the impact of various hyperparameter combinations.

We replicated the same architecture and training process for the third CNN model, with the only changes being the inclusion of Enformer-generated epigenetic features along with the sequence motifs as input to the fully-connected layers, resulting in a feature set of size (5,313 + n*5), where 'n' denotes the number of convolutional filters in the final layer.

4. CONCLUSIONS AND DISCUSSIONS

In conclusion, this project addresses the critical need for non-invasive diagnostic tools for early detection of rejection among lung transplant recipients, given their poor long-term survival rates. In recent times, cell-free DNA has emerged as a versatile non-invasive diagnostic tool with wide applications like prenatal diagnosis of fetal genetic abnormalities (Mortazavipour, Mohamad Mahdi, and Shirin Shahbazi, 2022) and cancer detection through liquid biopsy

(Cisneros-Villanueva, M., et al., 2022.). Various studies have shown that donor-derived cfDNA (dd-cfDNA) levels in the recipient's blood are a promising biomarker for transplant rejection diagnosis (De Vlaminck, Iwijn, et al., 2015, Ju, Chunrong, et al., 2023). Despite the potential of dd-cfDNA as a biomarker, distinguishing dd-cfDNA from recipient-derived cfDNA poses a significant challenge. Deep Learning is a subfield of Machine Learning that has found wide application in the field of biology due to its ability to automatically learn patterns and representations from large datasets (Webb, Sarah, 2018.). In this project, we set out to explore a novel method of measuring the percentage of dd-cfDNA by leveraging Deep Learning models to learn patterns that distinguish dd-cfDNA from rd-cfDNA.

Our first classification model was trained on epigenetic data extracted from cfDNA fragments using Enformer. However, despite hyperparameter optimizations, the model only showed an improvement in AUC of 0.02 compared to a random classifier (Fig 9 and 10). The subsequent CNN model extracted sequence motifs from cfDNA fragments to perform the classification. This model too achieved similar performance results with an AUC of 0.52, despite hyperparameter optimizations and the application of regularization techniques to mitigate overfitting (Fig 12 and Fig 13). The third CNN model, which combines epigenetic features and sequence motif extraction, similarly offered no performance improvements over the other models (Fig 14). The poor performance of the models can be attributed to the lack of patterns that could be learned by a simple deep neural network. This was further confirmed when all the models achieved >90% accuracy on simulated datasets which were artificially augmented with signals (Fig 11).

There are several possible reasons for the poor predictive performance of the DL models trained on Enformer-generated epigenetic tracks. Deep learning models are designed to distinguish between classes by patterns representative of each class. This necessitates the presence of unifying patterns - whether simple or complex - in most samples belonging to one class, and different unifying patterns for most samples of the other class (LeCun, Y., Bengio, Y., & Hinton, G. 2015). We've reasoned that fragments from dd-cfDNA and recipient-derived cfDNA belong to different regions of the genome, and hence their epigenetic signatures will be different. This, however, does not imply that there is a common unifying pattern in epigenetic features that are shared by all the dd-cfDNA fragments originating from different genomic regions. Consequently, the low AUC scores may be attributed to the lack of consistent unifying patterns correlated with donor- or recipient-derived fragments.

Another possible reason is the premise of our hypothesis, which relies on the assumption that nucleosome and transcription factor positions vary between lung and blood cell types. However, the extent of this variation remains uncertain, given the diverse influences on nucleosome position - stemming from both cell-type specific factors and those universally conserved across cell types. The DNA sequence provides a baseline for nucleosome positioning while factors like epigenetic modifications, chromatin remodeling, transcription factors, and cellular differentiation collectively influence cell-type specific nucleosome positions (Li, Shuxiang, Yunhui Peng, and Anna R. Panchenko, 2022., Jiang, Cizhong, and B. Franklin Pugh, 2009., Chen, Taiping, and Sharon YR Dent, 2014). Although cell-type specific variations exist, the prevailing consensus is that the DNA sequence predominantly dictates nucleosome positions, creating similar nucleosome occupancy across cell types (Struhl, Kevin, and Eran Segal, 2013., Radman-Livaja, Marta, and Oliver J. Rando, 2010). For our hypothesis, this implies that a significant proportion of dd-cfDNA shares the same genomic region as rd-cfDNA, potentially lacking meaningful signals that are representative of either class. These noisy samples could be impacting the model's ability to generalize, leading to low AUC scores on the validation set. This notion is supported by simulations with synthetic data, revealing that a minimum level of informative samples is required for the deep learning models to perform well. While in theory, aggregating results from various weak classifiers trained on noisy data could lead to improved classification (Ji, C., & Ma, S. 1996.), our dataset might lack the minimum signal level required for weak classifiers to learn meaningful decision functions and outperform random classifiers. Consequently, aggregating classification results from all fragments for a patient did not yield better patient-level classification results (Fig 16c).

Future efforts to improve classification results by reducing the number of noisy samples could benefit from exclusively utilizing regions exhibiting significant tissue-specific variations in nucleosome positioning for training. DNAse Hypersensitivity (DHS) sites show high variation in nucleosome positions due to their role as regulatory elements that influence tissue-specific gene expression. Tissue-specific nucleosome spacing in DHS sites has been successfully used to infer tissue type from cfDNA fragment sequences (Snyder, M. W., et al., 2016). Applying this idea to our project, the future training set could be restricted to fragments originating from DHS sites associated with lung and blood cell types, sourced from the genome-wide index of human DHS sites (Meuleman, Wouter, et al., 2020.).

The challenges in achieving high AUC scores could also, in part, be attributed to the incomplete exploration of the hyperparameter space. Deep Learning models are known for their sensitivity to hyperparameters, and the model's performance is greatly influenced by the right combination of values for parameters such as the learning rate, number of layers, neurons per layer, initial weights, batch size and momentum (Taylor, R., et al. 2021). Although we manually tested various hyperparameter combinations, the absence of an exhaustive cross-validation approach might have limited the exploration of the complete hyperparameter landscape. Future efforts could benefit from a more comprehensive hyperparameter optimization process through methods like random search, or optimization of network architecture through automated neural architecture search (Bergstra, J., & Bengio, Y. 2012).

Additionally, we suspect that low AUC scores could be attributed to inaccuracies in the true labels used for training. Despite numerous studies conclusively establishing that dd-cfDNA levels peak during transplant rejection (Agbor-Enoh, S., et al. 2019, Jang, Moon Kyoo, et al., 2021, Ju, Chunrong, et al., 2023), it is perplexing that no correlation was observed between dd-cfDNA calculated from true labels and clinical signs of rejection (Fig 13b). This discrepancy leads us to believe that there are some inaccuracies in the true labels. While the poor correlation could in part be explained by clinical signs of rejection being an unreliable indicator of transplant rejection, the absence of distinct length distribution patterns between donor- and recipient-derived cfDNA, despite multiple studies showing that dd-cfDNA is shorter than recipient-derived cfDNA (Zheng, Yama WL, *et al.*, 2012, Pedini, P., *et al.*, 2023), suggests that label inaccuracies definitely exist. Other potential dataset problems contributing to the low AUC scores are the lack of sufficient dd-cfDNA training samples and the presence of dd-cfDNA samples in the same genomic regions as the recipient-derived fragments, necessitating deeper sequencing to identify non-overlapping areas. In the future, utilizing a different dataset for training and validation of the models could yield better classification performance.

Although the deep learning models we developed to classify between dd-cfDNA and recipient-derived cfDNA did not perform well enough to be considered reliable diagnostic tools, we envision significant potential with refinement. This could involve training with a different dataset, dataset modifications for noise reduction, and a comprehensive hyperparameter optimization process. With these improvements, our models could emerge as effective, non-invasive, and cost-effective diagnostic tools for early detection of transplant rejection. The novel approach of using Deep Learning models to measure dd-cfDNA levels holds great promise and can be generalized to any scenario involving more than one type of cfDNA, such as liquid biopsy for cancer detection (Gai, W., & Sun, K. 2019, Ray, S. K., & Mukherjee, S. 2022) and prenatal screening for chromosomal disorders (Norton, M. E., et al. 2015, Curnow, K. J., et al. 2014).

5. DATA AND CODE AVAILABILITY

The raw cell-free DNA sequence data utilized for the project was acquired from the Sequence Read Archive https://www.ncbi.nlm.nih.gov/ (Accession number: PRJNA263522). The patient metadata, including details about clinical signs of rejection and the time points when sequencing data was collected, is also included in the sequence archive (https://www.ncbi.nlm.nih.gov/Traces/study/?query_key=1&WebEnv=MCID_65969777ad40442121a49c14&o=acc_s%3Aa)

The code for initial data analysis, training and validating DL models, and generating all the plots for the report are available on GitHub at the following URL - https://github.com/vmadhupreetha/fragmentomics/tree/master.

6. REFERENCES:

1. Momtazmanesh, Sara, et al. "Global burden of chronic respiratory diseases and risk factors, 1990–2019: an update from the Global Burden of Disease Study 2019." *EClinicalMedicine* 59 (2023).
2. Lund, Lars H., et al. "The registry of the International Society for Heart and Lung Transplantation: thirty-first official adult heart transplant report—2014; focus theme: retransplantation." *The Journal of Heart and Lung Transplantation* 33.10 (2014): 996-1008.
3. Sundaresan, Sudhir, et al. "Prevalence and outcome of bronchiolitis obliterans syndrome after lung transplantation." *The Annals of thoracic surgery* 60.5 (1995): 1341-1347.

4. Gauthier, Jason M., Ramsey R. Hachem, and Daniel Kreisel. "Update on chronic lung allograft dysfunction." *Current transplantation reports* 3 (2016): 185-191.
5. Herout, Vladimir, et al. "Transbronchial biopsy from the upper pulmonary lobes is associated with increased risk of pneumothorax–a retrospective study." *BMC Pulmonary Medicine* 19 (2019): 1-6.
6. Arcasoy, S. M., et al. "Pathologic interpretation of transbronchial biopsy for acute rejection of lung allograft is highly variable." *American Journal of Transplantation* 11.2 (2011): 320-328.
7. Magnusson, Jesper M., et al. "Cell‐free DNA as a biomarker after lung transplantation: A proof‐of‐concept study." *Immunity, Inflammation and Disease* 10.5 (2022): e620.
8. Lo, YM Dennis, et al. "Presence of donor-specific DNA in plasma of kidney and liver-transplant recipients." *The Lancet*351.9112 (1998): 1329-1330.
9. Lo, YM Dennis, et al. "Presence of fetal DNA in maternal plasma and serum." *The lancet* 350.9076 (1997): 485-487.
10. De Vlaminck, Iwijn, et al. "Noninvasive monitoring of infection and rejection after lung transplantation." *Proceedings of the National Academy of Sciences* 112.43 (2015): 13336-13341.
11. Jang, Moon Kyoo, et al. "Donor-derived cell-free DNA accurately detects acute rejection in lung transplant patients, a multicenter cohort study." *The Journal of Heart and Lung Transplantation* 40.8 (2021): 822-830.
12. Zou, Jun, et al. "Rapid detection of donor cell free DNA in lung transplant recipients with rejections using donor-recipient HLA mismatch." *Human immunology* 78.4 (2017): 342-349.
13. Sharon, Eilon, et al. "Quantification of transplant-derived circulating cell-free DNA in absence of a donor genotype." *PLoS computational biology* 13.8 (2017): e1005629.
14. Grskovic, Marica, et al. "Validation of a clinical-grade assay to measure donor-derived cell-free DNA in solid organ transplant recipients." *The Journal of Molecular Diagnostics* 18.6 (2016): 890-902.
15. Zheng, Yama WL, et al. "Nonhematopoietically derived DNA is shorter than hematopoietically derived DNA in plasma: a transplantation model." *Clinical chemistry* 58.3 (2012): 549-558.
16. Mortazavipour, Mohamad Mahdi, and Shirin Shahbazi. "The current applications of cell-free fetal DNA in prenatal diagnosis of single-gene diseases: A review." *International Journal of Reproductive BioMedicine* 20.8 (2022): 613.
17. Cisneros-Villanueva, M., et al. "Cell-free DNA analysis in current cancer clinical trials: a review." *British Journal of Cancer* 126.3 (2022): 391-400.
18. Ju, Chunrong, et al. "Application of plasma donor-derived cell free DNA for lung allograft rejection diagnosis in lung transplant recipients." *BMC Pulmonary Medicine* 23.1 (2023): 37.
19. Webb, Sarah. "Deep learning for biology." Nature 554.7693 (2018): 555-557.
20. Dna methylation and histone modifications modulate nucleosome dynamics and positions
Li, Shuxiang, Yunhui Peng, and Anna R. Panchenko. "DNA methylation: Precise modulation of chromatin structure and dynamics." *Current Opinion in Structural Biology* 75 (2022): 102430.

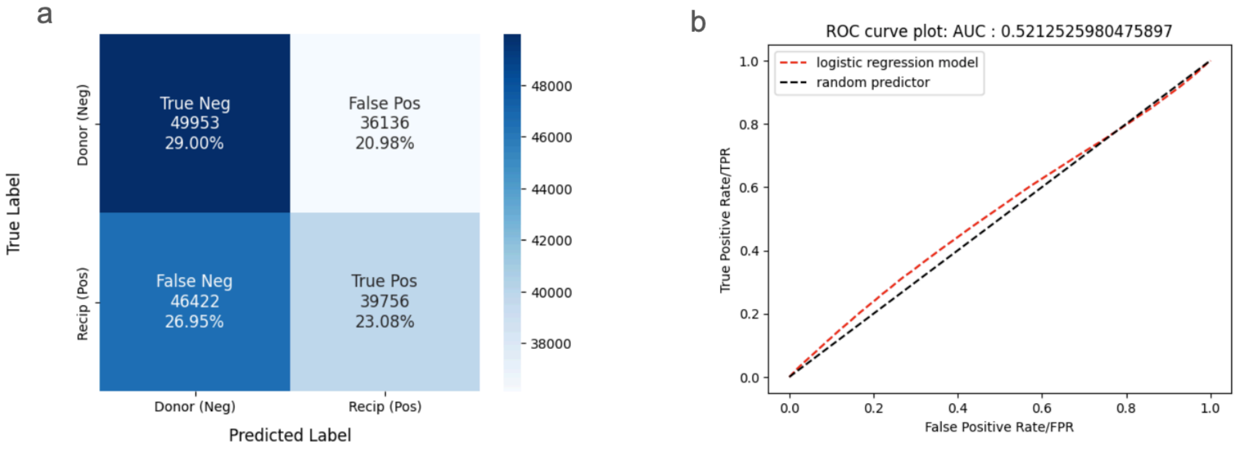
21. Nucleosome positions are dynamically determined by transcription factors due to chromatin remodelling done by TFs .Jiang, Cizhong, and B. Franklin Pugh. "Nucleosome positioning and gene regulation: advances through genomics." *Nature Reviews Genetics* 10.3 (2009): 161-172.
22. Chromatin remodelling during cellular differentiation Chen, Taiping, and Sharon YR Dent. "Chromatin modifiers and remodellers: regulators of cellular differentiation." *Nature Reviews Genetics* 15.2 (2014): 93-106.
23. Struhl, Kevin, and Eran Segal. "Determinants of nucleosome positioning." *Nature structural & molecular biology* 20.3 (2013): 267-273.
24. Radman-Livaja, Marta, and Oliver J. Rando. "Nucleosome positioning: how is it established, and why does it matter?." *Developmental biology* 339.2 (2010): 258-266.

25. Snyder, Matthew W., et al. "Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin." *Cell*164.1 (2016): 57-68.
26. D'haeseleer, Patrik. "What are DNA sequence motifs?." *Nature biotechnology* 24.4 (2006): 423-425.
27. Loshchilov, Ilya, and Frank Hutter. "Sgdr: Stochastic gradient descent with warm restarts." *arXiv preprint arXiv:1608.03983*(2016).
28. Geirhos, Robert, et al. "Shortcut learning in deep neural networks." *Nature Machine Intelligence* 2.11 (2020): 665-673.
29. Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." *The journal of machine learning research* 15.1 (2014): 1929-1958.
30. Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." *arXiv preprint arXiv:1711.05101* (2017).
31. LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature* 521.7553 (2015): 436-444.
32. Ji, Chuanyi, and Sheng Ma. "Combinations of weak classifiers." *Advances in Neural Information Processing Systems* 9 (1996).
33. Taylor, Rhian, et al. "Sensitivity analysis for deep learning: ranking hyper-parameter influence." *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2021.
34. Gai, Wanxia, and Kun Sun. "Epigenetic biomarkers in cell-free DNA and applications in liquid biopsy." *Genes* 10.1 (2019): 32.
35. Norton, Mary E., et al. "Cell-free DNA analysis for noninvasive examination of trisomy." *New England Journal of Medicine*372.17 (2015): 1589-1597.
36. Curnow, Kirsten J., et al. "Clinical experience and follow-up with large scale single-nucleotide polymorphism–based noninvasive prenatal aneuploidy testing." *American journal of obstetrics and gynecology* 211.5 (2014): 527-e1.
37. Ray, Suman K., and Sukhes Mukherjee. "Cell free DNA as an evolving liquid biopsy biomarker for initial diagnosis and therapeutic nursing in Cancer-An evolving aspect in Medical Biotechnology." *Current Pharmaceutical Biotechnology* 23.1 (2022): 112-122.
38. Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." *Nature* 596.7873 (2021): 583-589.
39. Wu, Ruidong, et al. "High-resolution de novo structure prediction from primary sequence." *BioRxiv* (2022): 2022-07.
40. Poplin, Ryan, et al. "A universal SNP and small-indel variant caller using deep neural networks." Nature biotechnology 36.10 (2018): 983-987.
41. Keshavarzi Arshadi, Arash, et al. "Artificial intelligence for COVID-19 drug discovery and vaccine development." *Frontiers in Artificial Intelligence* (2020): 65.
42. Baxi, Emily G., et al. "Answer ALS, a large-scale resource for sporadic and familial ALS combining clinical and multi-omics data from induced pluripotent cell lines." *Nature neuroscience*25.2 (2022): 226-237.
43. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems* 25 (2012).
44. Mnih, Volodymyr, et al. "Playing atari with deep reinforcement learning." *arXiv preprint arXiv:1312.5602* (2013).
45. Lui, Yanni YN, et al. "Predominant hematopoietic origin of cell-free DNA in plasma and serum after sex-mismatched bone marrow transplantation." *Clinical chemistry* 48.3 (2002): 421-427.
46. D'haeseleer, Patrik. "What are DNA sequence motifs?." *Nature biotechnology* 24.4 (2006): 423-425.
47. Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

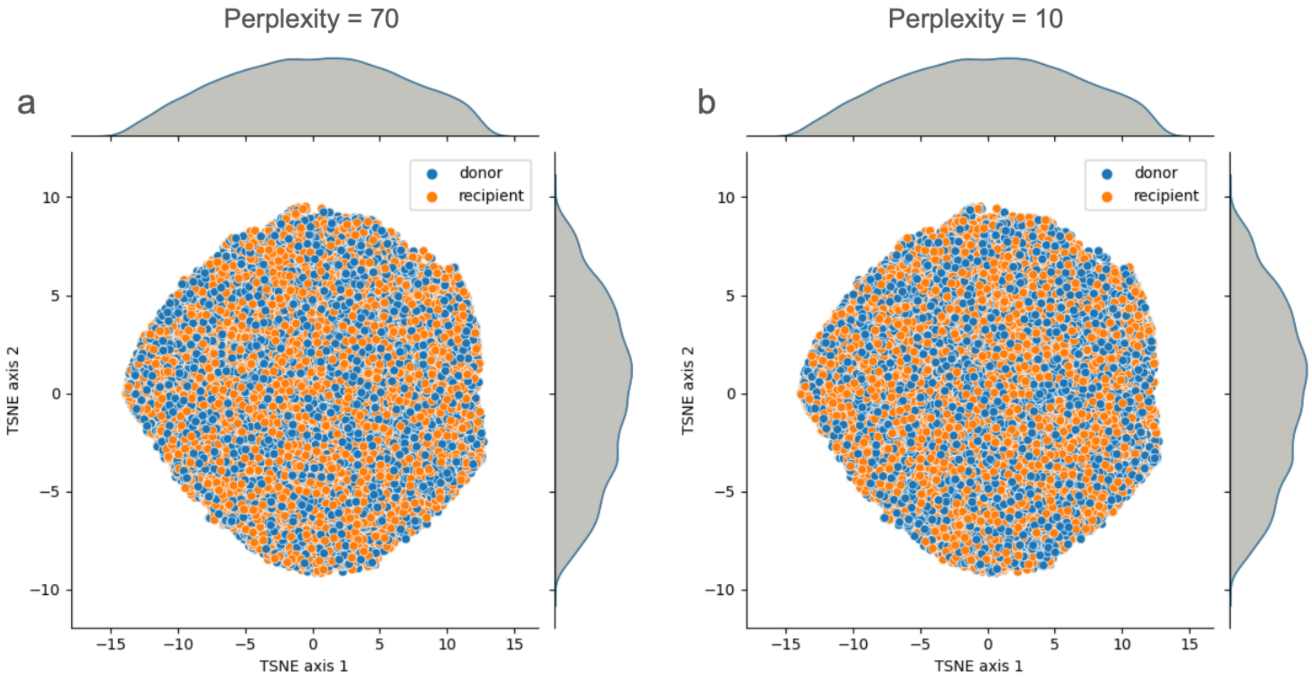
48. Bergstra, James, and Yoshua Bengio. "Random search for hyper-parameter optimization." Journal of machine learning research 13.2 (2012).
49. Agbor-Enoh, Sean, et al. "Donor-derived cell-free DNA predicts allograft failure and mortality after lung transplantation." *EBioMedicine* 40 (2019): 541-553.
50. Brahma, Sandipan, and Steven Henikoff. "Epigenome regulation by dynamic nucleosome unwrapping." *Trends in biochemical sciences* 45.1 (2020): 13-26.
51. Mrad, Ali, and Rebanta K. Chakraborty. "Lung Transplant Rejection." (2020).
52. Keller, Michael, and Sean Agbor-Enoh. "Donor-derived cell-free DNA for acute rejection monitoring in heart and lung transplantation." *Current Transplantation Reports* 8.4 (2021): 351-358.

7. SUPPLEMENTARY

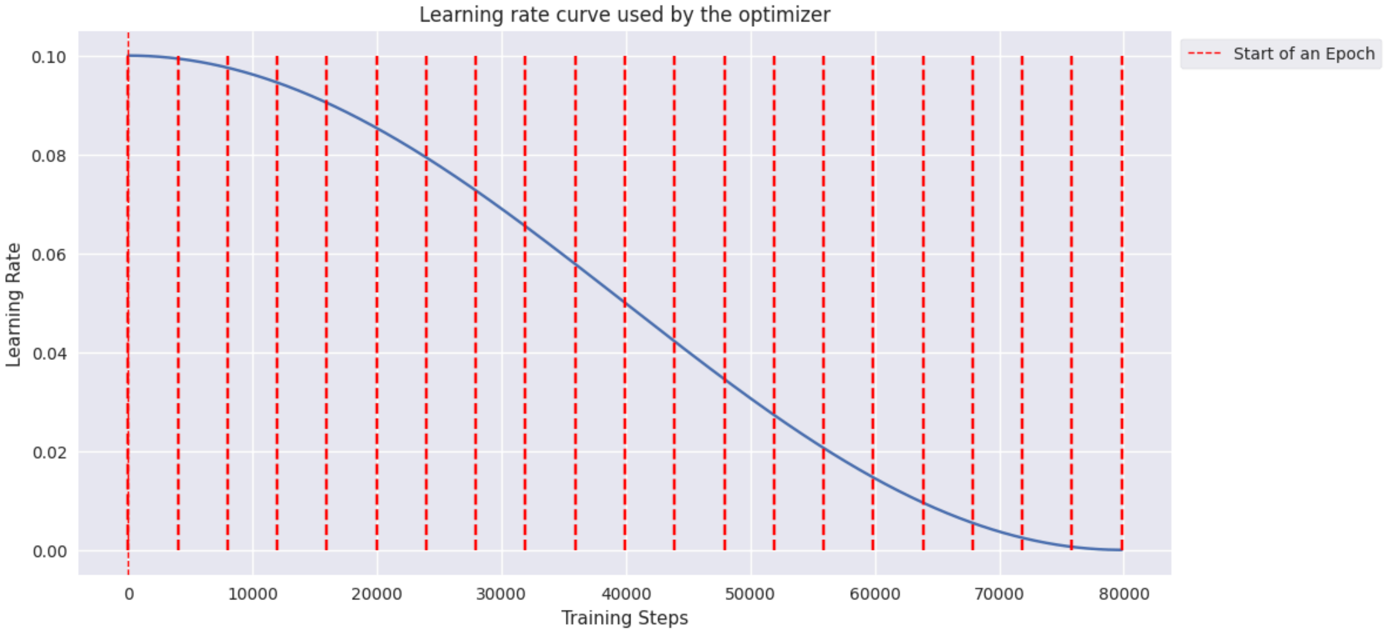
7.1 Supplementary Figures



Supplementary Figure 1: Logistic regression model trained on fragment length as sole feature, with 300,000 donor- and recipient-derived fragments each.
a. Confusion matrix of logistic regression classifier
b. ROC curve of logistic regression classifier.

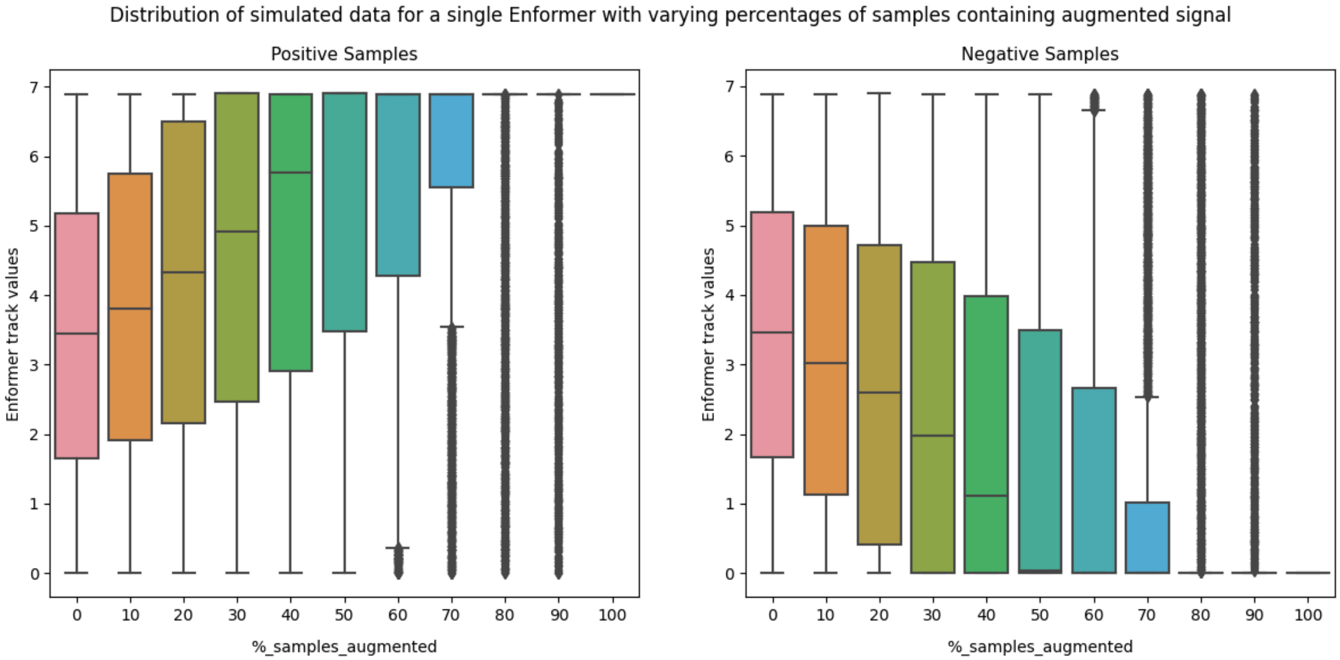


Supplementary Figure 2: t-SNE plot of donor and recipient-derived cfDNA fragments
a. T-SNE plot for perplexity 70 **b.** T-SNE plot for perplexity 10
We could observe no clear donor or recipient-specific patterns from the T-SNE plot. So, we explored different values of perplexity, which defines balance between preserving local and global structures in the data. But for all values for perplexities, donor and recipient-derived fragments clustered in the same region.



Supplementary Figure 3: The learning rate curve of the optimizer gradually reduces from the initial learning rate of 0.1 to 0 following a cosine curve. Here, each training step is a batch, the weights and learning rates are updated after every batch. The vertical lines in red mark the beginning of a new training epoch.

Reducing the learning rate during the training process, following a cosine curve helps the optimizer generalize better (avoid overfitting) and converge to a solution faster (Loshchilov, I., & Hutter, F., 2016).



Supplementary Figure 4: Distribution of augmented donor-derived (positive) samples and recipient-derived (negative) samples. The plots show the distribution of all Enformer track values for varying percentage of samples ‘s’ that were augmented with synthetic signals.

Positives (donor-derived samples): s% of positive samples were modified by replacing some Enformer tracks with a randomly generated value slightly exceeding the maximum observed value for that track. As ‘s’ increases, the mean of enformer track distribution increases, as more positive samples are replaced with higher values.

Negatives (recipient-derived samples): s% of negative samples were modified by replacing some of Enformer tracks by a randomly generated value slightly lower than the minimum observed value for that track. As ‘s’ increases, the mean of Enformer track distribution decreases, as more negative samples are replaced with lower values.

7.2 Supplementary Information

7.2.1 Kolmogorov-Smirnov (KS) Test : The KS test was done on a class balanced subset of training samples consisting of 300,000 each of donor- and recipient-derived fragments. The resulting KS statistic was found to be 0.042, with a corresponding p-value of 6.92. Since the p-value is higher than 0.05, we concluded that the difference in length distribution between donor- and recipient-derived fragments is not statistically significant.

7.2.2 Data acquisition : Plasma samples were collected from patients at specific time points post transplant : weeks 1, 2, 4, and 6, as well as months 2, 2.5, 3, 4, 5, 6, 8, 10, 12, 16, 20, and 24. The cfDNA fragments in the plasma were sequenced using Illumina HiSeq 2000, HiSeq 2500, or NextSeq 500; 1 × 50 bp or 2 × 100 bp, with sequencing libraries prepared using the the NEBNext DNA Library Prep Master Mix Set for Illumina.Genotyping was performed on Illumina whole-genome arrays (HumanOmni2.5-8 or HumanOmni1)