Title: Machine learning to classify mutational hotspots from molecular dynamic simulations

Authors: James Davies, Georgina Menzies

Address details:

Sir Martin Evans Building, Museum Avenue, Cardiff, CF10 3AX

MenziesG@cardiff.ac.uk

Abstract

Benzo[*a*]pyrene, a notorious DNA-damaging carcinogen, belongs to the family of polycyclic aromatic hydrocarbons commonly found in tobacco smoke. Surprisingly, nucleotide excision repair (NER) machinery exhibits inefficiency in recognising specific bulky DNA adducts including Benzo[*a*]pyrene Diol-Epoxide (BPDE), a Benzo[*a*]pyrene metabolite. While sequence context is emerging as the leading factor linking the inadequate NER response to BPDE adducts, the precise structural attributes governing these disparities remain inadequately understood. We therefore combined the domains of molecular dynamics and machine learning to conduct a comprehensive assessment of helical distortion caused by BPDE-Guanine adducts in multiple gene contexts. Specifically, we implemented a dual approach involving a random forest classification-based analysis and subsequent feature selection to identify precise topological features that may distinguish adduct sites of variable repair capacity. Our models were trained using helical data extracted from duplexes representing both BPDE hotspot and non-hotspot sites within the *TP53* gene, then applied to sites within *TP53*, *cII*, and *lacZ* genes.

We show our optimised model consistently achieved exceptional performance, with accuracy, precision, and f1 scores exceeding 91%. Our feature selection approach uncovered that discernible variance in regional base pair rotation played a pivotal role in informing the decisions of our model. Notably, these disparities were highly conserved among *TP53* and *lacZ* duplexes and appeared to be influenced by the regional GC content. As such, our findings suggest that there are indeed conserved topological features distinguishing hotspots and non-hotpot sites, highlighting regional GC content as a potential biomarker for mutation.

Author Summary

Although much is known about DNA repair processes, we are still lacking some fundamental understanding relating to DNA sequence and mutation rates, specifically why some sequences mutate at a higher rate or are repaired less than others. We believe that by using a combination of Molecular Simulation and Machine Learning (ML) we can measure which structural features are present in sequences which mutate at higher rates in cancer gene and lab-based test assays frequently used to investigate toxicology.

Here we have run Molecular Dynamics on five sets of DNA sequences with and without a carcinogen found in cigarette smoke to allow us to study the mutation event that would need to be repaired. We

have measured their helical and base stacking properties. We have used ML to successfully differentiate between low and high mutating sequences using this model allowing us to begin to elucidate the structural features these groups have in common.

We believe this method could have wide reaching uses, it could be applied to any gene context and mutation event and indeed the knowledge of the structural features which are best repaired gives us insight into the biophysics of DNA repair adding knowledge to the drug design pipeline.

Introduction:

Mutations serve as the fundamental source of genetic variation, fostering adaptive evolution, whilst simultaneously contributing to the development of diseases such as cancer and age-related illnesses [1,2]. Though random, a nucleotides mutation rate, defined by its probability to undergo mutation, exhibits considerable heterogeneity throughout the genome. This variation has been demonstrated at various levels, with the rate of site-to-site mutation shown to vary by more than 100-fold across the genome, prompting sequence dependant contribution to both evolution and genetic pathologies [3]. Analysing the precise features that promote this dependency, as such, may unlock valuable insights into the mutagenic mechanisms underlying various human diseases.

It is widely accepted that site-specific mutational frequencies are contingent upon three generalised factors: (i) nucleotide stability and their vulnerability to mutagenesis; (ii) the fidelity of DNA replication processes; and (iii) the efficiency of DNA repair machinery [4]. Consequently, a natural gradient manifests, revealing the degree to which mutational occurrences are influenced by the underlying sequence composition. For instance, regions of repetitive DNA such as homonucleotide runs and microsatellite repeats serve as an active source of hypermutability through polymerase slippage events, emphasising motif recurrence over precise sequence context [5,6]. In contrast, alternative mutation spectra exhibit pronounced sequence specificity, often guided by intricate local and occasionally distant sequence contexts [7-11]. While the underlying causes of the latter remain subjects of substantial debate, emerging evidence indicates that this interdependence forms the foundation for numerous disease-associated mutation spectra. This is notably evident in the case of the *TP53* mutation spectrum within the context of lung cancer [12,13].

TP53 is a gene that is responsible for encoding the p53 tumour suppressor protein, commonly referred to as the “Guardian of the Genome” [14]. Its principal biological role involves safeguarding DNA integrity through the inhibition of cell proliferation and the promotion of apoptosis in response to cellular stress [15]. This renders *TP53* a pivotal target for inactivation in diseases, consequently establishing it as the most commonly mutated gene in human cancer [12]. *TP53* mutations are particularly prevalent in lung cancer, with approximately 50% of tumours containing a mutation [16]. Mutation distribution, however, is shown to vary substantially between lung and most other tumour types, with a higher frequency of G:C>T:A transversions in the former [12,16,17]. While the influence

of sequence context on the varying *TP53* mutation spectrum continues to be a subject of intense debate, a consensus exists regarding the underlying cause: the formation of DNA adducts induced by polycyclic aromatic hydrocarbons (PAH) found in cigarette smoke [18-20].

Benzo[a]pyrene (B[*a*]P) emerges as the predominant PAH stemming from the incomplete combustion of organic material within tobacco smoke. A study by Vu *et al.,* established that one cigarette can yield an intake of 13-38ng of B[*a*]P [21]. This compound is then metabolically processed to generate a series of carcinogenic derivatives, notably including the trans(+)anti-benzo[a]pyrene diol epoxide, hereon referred to as BPDE (Figure 1A) [22]. Empirical evidence suggests that a combination of preferential DNA damage and poor repair is responsible for the lung associated mutation hotspots in *TP53* [23]. Recurrently mutated codons 157, 158, 245, 248, 273, and 282 harbour methylated CG dinucleotides and therefore represent favourable locations for BPDE binding while concurrently exhibiting resistance to nucleotide excision repair (NER) mechanisms [24]. Notably, the (K)-7S,8R,9R,10S□+□anti-B[*a*]PDE enantiomer (10S) DNA adduct dominates, accounting for an overwhelming majority (70.8-92.9%) of the total adducts at these specific sites (Figure 1B-C) [25]. Interestingly, however, Denissenko *et al.,* report other known targets of BPDE, including 170, 186, 202, 213, 267 and 290, whose mutation is either uncommon or silent in lung cancer [23,26,27]. Given the broad specificity of BPDE and accounting for selective processes, the induction of *TP53* mutations by BPDE likely the combined result of differential adduct availability and varying repair capabilities at distinct adduct sites.

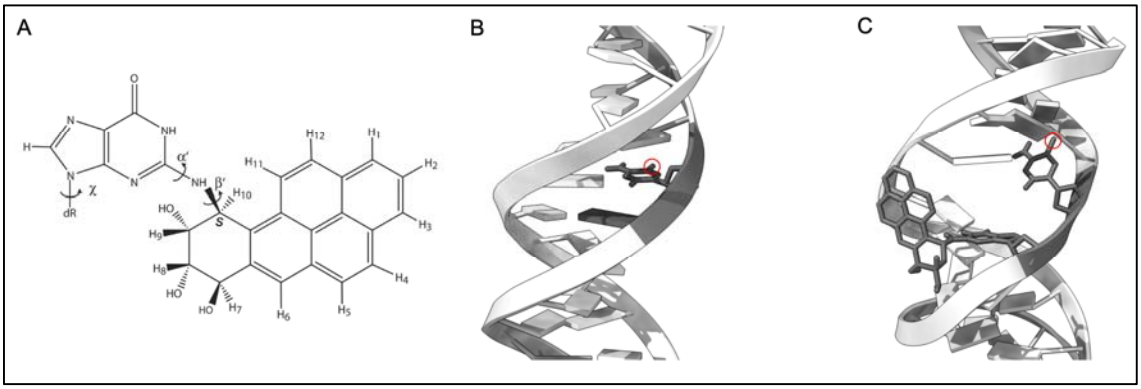

Figure 1: Chemical structure and mutagenic adduct formation of Benzo[*a*]pyrene diolepoxide. (**A**) Chemical structure of 7S,8R,dihydroxy-9R,10S-epoxy-7S,8R,9R,10S (+)-trans-anti-B[*a*]PDE; (**B**) regular unadducted DNA with methyl group (highlighted by a red circle) on the cytosine at base 6; (**C**) adduct of BPDE (sticks) bound to guanine in a CpG site in minor groove directed towards the 5′ end of the DNA sequence.

Importantly, while NER is capable of repairing a diverse range of bulky lesions, its efficiency varies dramatically according to the chemical structure, stereochemistry, confirmation, and sequence context surrounding the lesion site [28-32]. Various studies, both experimental and theoretical, highlight how variable sequence context can influence NER efficacy by impacting DNA structure and subsequent

recognition processes [29,32-34]. Paul *et al.,* reveal that distinct resistances to NER-induced untwisting/bending renders certain contexts more likely to access conformations required for successful repair, thereby proposing a sequence-dependent propensity and trajectory for DNA "unwinding" [35]. Furthermore, we previously reported sequence-dependent variability in DNA distortion at BPDE adduct sites linked to lung cancer-associated *TP53* mutation sites [36]. Thus, the potential disparity in repair capacity within *TP53* across BPDE adduct sites may arise from sequence-driven modifications in DNA structural parameters. Despite these notable strides, however, the true extent to which sequence context influences the distortional profiles induced by BPDE adducts, and consequently, NER, remains a subject of ongoing debate, prompting the need for continued scientific scrutiny.

In this study, we sought to integrate the realms of molecular dynamics (MD) and machine learning (ML) to formulate a model capable of defining the precise features that promote the differing mutational pattern observed in *TP53* during lung tumorigenesis. We hypothesised that by the inclusion of ML methodologies we would significantly amplify the range of hotspots sites under analysis, thereby enabling the comparison of mutation spectra within single or multiple gene contexts. Our initial aim was to determine whether helical information acquired from MD simulation could indeed be used to isolate sequence dependant features that aberrate the NER of BPDE lesions in *TP53*. Upon achieving this, our subsequent goal was to ascertain the model's efficacy in utilising this information in predicting the mutation status of unseen helical data from alternative gene contexts that are recognised to undergo BPDE-induced mutations including *cII* and *lacZ*. The knowledge gained could help provide more information surrounding the aetiology of the disease, and allow for the develop innovative anti-cancer therapies aimed at reinstating healthy NER function at these sites.

Results

2.1 Data Retrieval and Definition of Spectral Investigation

Our analysis was focused on mutational spectra of three distinct gene contexts – *TP53*, *cII*, and *lacZ* – each exhibiting variable repair of BPDE-induced lesions as defined by [37], [38], and [39]. Mutational data obtained for *TP53* was used to assemble three further datasets. The datasets, named "TP53", and a secondary dataset obtained from unique simulations of corresponding *TP53* sequences, termed "Val," were constructed based on *TP53* duplexes with site-restricted methylation, serving as the training and validation sets, respectively, for our ML model. Subsequently, a third dataset named "Meth" was compiled to assess the impact of more distal methylation on *TP53* dynamics, through homomethylation. Mutational profiles derived from the reporter genes *cII* and *lacZ* served as the foundation for subsequent investigations, allowing us to infer which components (if any) appear to have a conserved role or association with sample characteristics, such as repair phenotype or adduct binding state.

2.2 Conformational Stability and Sequency

Root mean square deviation (RMSD) and total energy analysis provided confidence about the stability of the simulations. Elevated RMSD values typically suggest that flexibility and movement during a simulation are likely characterised by instability. For *TP53*, total energy calculations confirmed the stability of all simulations conducted (Figure 3A). Average RMSD values for each sequence, relative to the initial structure, provided valuable insights into the variations in flexibility observed throughout the 100ns simulation period. As depicted in Figure 3B, unbound *TP53* duplexes showed a mean RMSD of 3.83Å ± 0.81, while adducted sequences exhibited a higher mean RMSD of 5.50Å ± 1.10. Fluctuations around the mean remained stable throughout the simulation period of 10-100ns, which prompted the exclusion of the initial 10ns from all subsequent analyses. Additionally, the results obtained from the RMSD analysis demonstrated a consistent trend of increased flexibility in the adducted sequences compared to the non-adducted controls. This observation aligns with our expectations, as BPDE adducts are known to disrupt the helical structure, resulting in increased fluctuation and mobility. All differences in the distributions of RMSD values between adducted and control sequences were significant ($P < 0.000$).

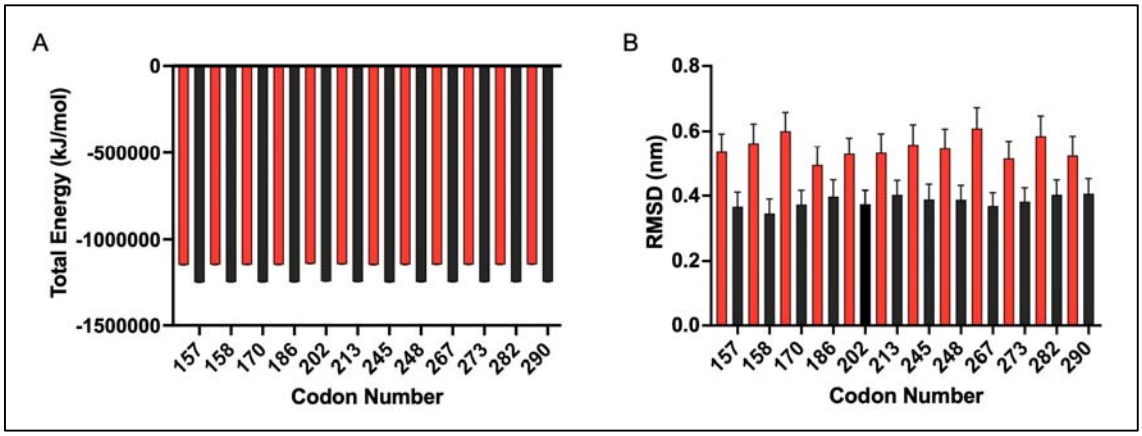

*Figure 2: Conformational stability and flexibility of adducted and control TP53 sequences. Total energy (**A**) in KJ/mol and RMSD values (**B**) in nm for control sequences (black) and adducted sequences (red); error bars show ± standard deviation. Corresponding values for Val, Meth, cII, and lacZ datasets are depicted in Supplementary Figures S1-S4.*

2.3 Classifier Comparison

To select an appropriate classifier, we first compared the prediction results of 15 unique classification algorithms on the TP53 dataset. Each classifier was instantiated with default hyperparameters, and model performance was evaluated through the quantification of accuracy, precision, recall, f1, and area under the curve (AUC) and receiver operating characteristic (ROC) scores. Among the 15 algorithms, decision tree-based models (Decision Tree, Extra Trees, and Random forest (RF)) were by far the most prominent, demonstrating advanced results ranging from 0.84 to 0.91 across each evaluation metric (Figure 4A-D). Kernel machines (Polynomial (support vector machine) SVM, radial basis function (RBF) SVM, and Gaussian Process), conversely, encountered substantial challenges

during the classification process, resulting in a notable decline in accuracy, precision, recall, f1, and, in the case of Polynomial SVM, considerable variation in ROC AUC (Figure 4E). As shown in Figure 4A-F, the RF algorithm exhibited the best discriminatory power, with all mean evaluation scores within the range of 0.88-0.99. ROC AUC values displayed minimal range (±0.035), highlighting the models consistent and robust discrimination across different scenarios (Figure 4E). We observed that, in comparison to all other well-performing algorithms, the RF model exhibits the smallest training-test gap (Figure 4F). This finding confirms the superior generalisation capacity of the RF model, solidifying its standing as the optimal classifier for our study.

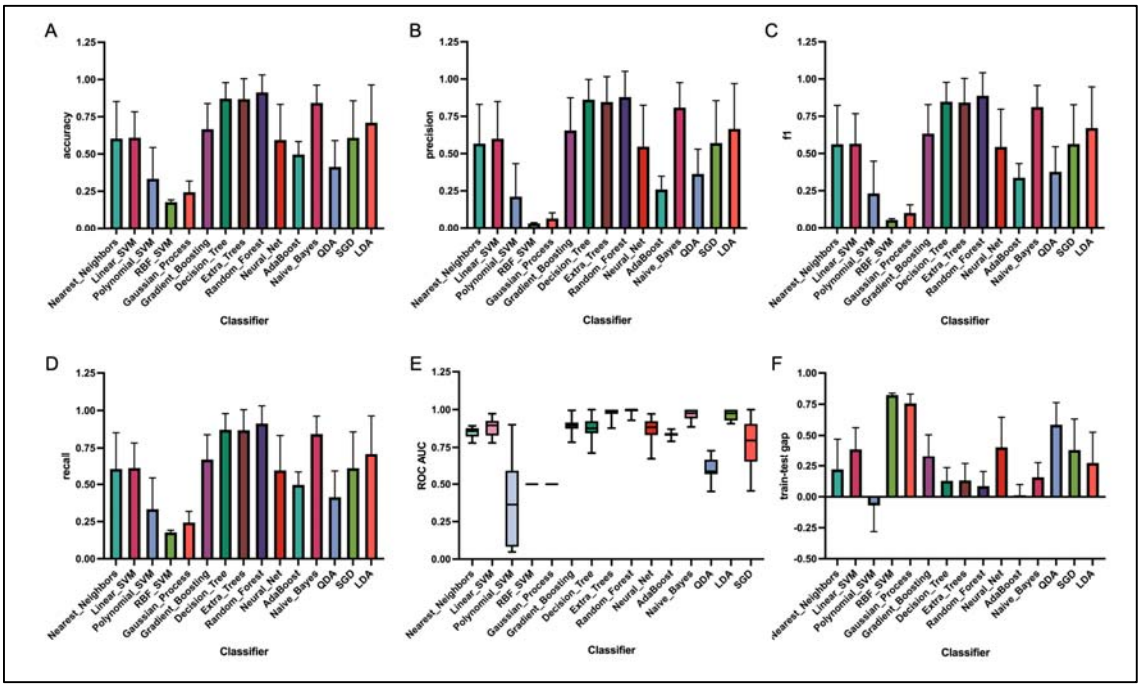

Figure 3: Evaluation results of fifteen classification algorithms using accuracy (A), precision (B), f1 (C), recall (D), ROC AUC, and training-test gap scores (F). Error bars represent ± standard deviation for panels (A) to (D) and indicate the range of ROC AUC scores in panel (E), with the median score depicted by a solid black line.

2.4 Grid Search Optimisation

To identify the optimal hyperparameter configuration of our model, we employed Grid Search Optimization (GSO). Crucial parameters, including bootstrap, min_sample_leaf, max_features, and n_estimators, significantly influence the performance of the RF classifier. Utilising GSO, we transformed these parameters into a grid with a defined spatial range and explored all points to obtain the global optimal solution.

Incorporating novel GSO parameters into the existing RF algorithm was shown to enhance model performance, yielding a 3.06% increase in model accuracy and 12.63% improvement in precision

(Figure 4A,B). Hyperparameter tuning also induced modest improvements in recall, f1, and training-test gap scores (Figure 4C-E). This not only bolstered overall accuracy but also honed the model's ability to detect positive cases while mitigating instances of false negatives and false positives. Further to this, GSO was shown to significantly improve the model's operating speed (P<0.0001) (Figure 4F). This is to be expected given that GSO typically accelerates convergence during model training.

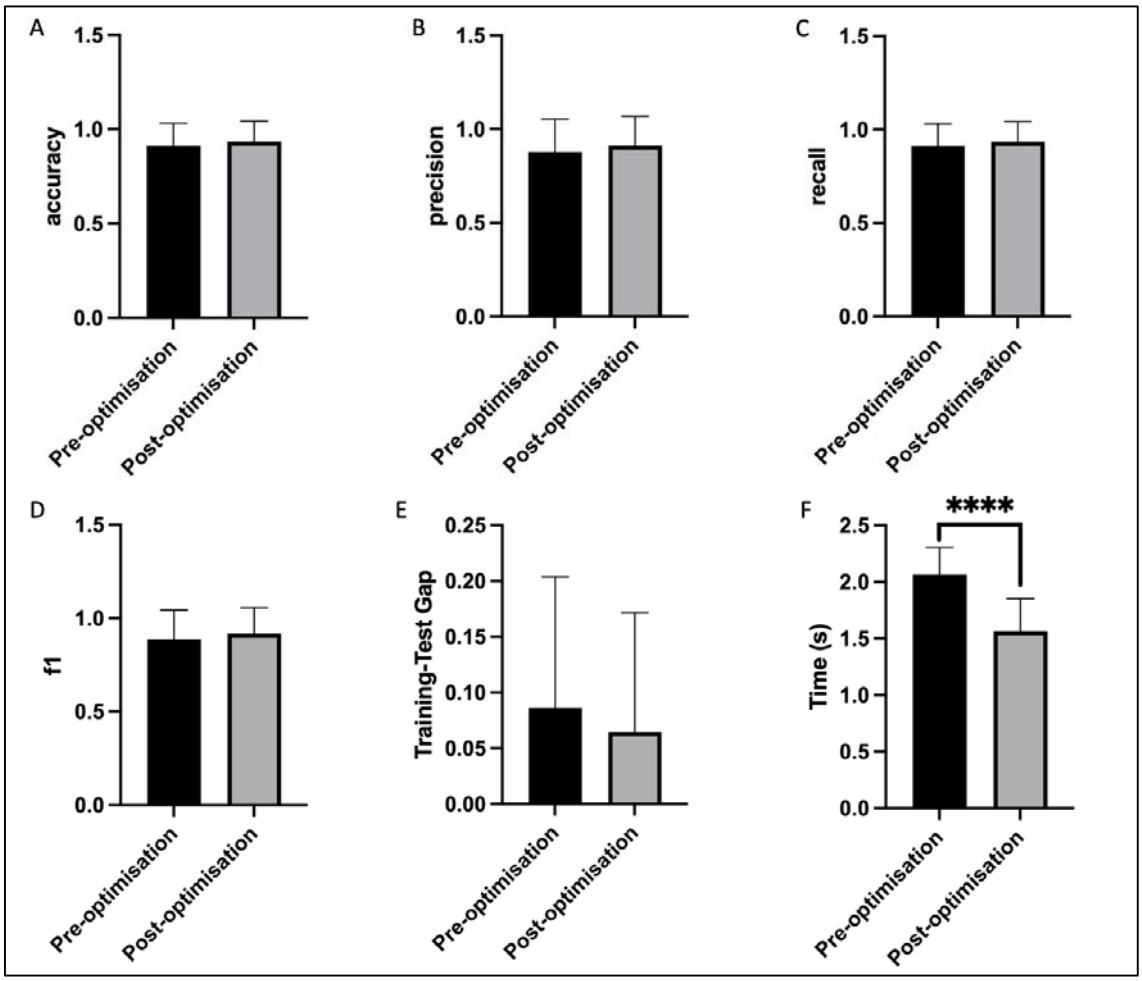

Figure 4: RF optimisation by GSO. Bar chart Illustrating the averaged accuracy (A), precision (B), recall (C), f1 (D), train-test gap (E), and execution time (F) scores before and after optimisation. Pre-optimisation scores are shown in black, while post-optimisation scores are depicted in grey. Error bars indicate ± standard deviation.

2.5 *TP53* Mutation

To elucidate the potential relationship between structural distortion and mutation frequency within the context of *TP53*, we initially deployed the optimised RF model. This approach evaluated the model's ability to leverage the knowledge acquired during the original training phase to classify two previously

unseen datasets. One dataset, sourced from the re-simulation of previously utilised *TP53* starting structures, constituted the designated validation dataset. Additionally, we simulated novel starting structures representative of the same *TP53* sequences, characterised, however, by full CpG methylation, thereby creating the methylated dataset. Once more, model performance was scrutinised using evaluation metrics accuracy, precision, f1 and recall, the values of which are depicted in Table 1. Feature scoring was then applied exclusively to adduct-bound sequences to deduce the predictive utility of each helical parameter for sample classification.

Table 1: Mean accuracy, precision, f1 and recall scores for RF model using 10-fold cross-validation.

Context	Accuracy	Precision	f1	recall
TP53	*0.935*	*0.913*	*0.917*	*0.935*
Val	*0.986*	*0.987*	*0.986*	*0.986*
Meth	*0.928*	*0.931*	*0.926*	*0.828*

Table 1 reveals a prominent improvement in our model's performance upon its deployment with validation data. The model achieves a near perfect performance, showcasing an average recall, accuracy, and f1 score of 0.986, coupled with an average precision score of 0.987. Notably, precision and f1 metrics exhibit an approximate 8% increase, indicating improved accuracy in positive predictions and a refined equilibrium between precision and recall. Despite the introduction of added complexity in the form of CpG methylation, our RF model maintains exceptional classification performance, achieving high scores in precision (0.931), f1 (0.926), and accuracy (0.928). Prediction recall scores, however, experience a marginal decline, registering a modest value of 0.828, indicating potential challenges in accurately identifying positive instances in the event of differentially methylated training and test datasets.

Feature scoring unveiled notable alignments in parameter types defined as pivotal during the classification process. Predominantly, parameters illustrating the rotation of base pairs/base pair axes took precedence among the top 10% of scored features, contributing to more than 61% of highlighted components (Table 2). Notably, a distinctive bias was unveiled towards the axis-displacement parameter tip, constituting 25% and 19% of the identified features in the *TP53* and validation datasets, respectively. Inclination and tip angles intricately mirror the orientation of base-pair planes concerning the helical axis. This observation naturally prompts the anticipation that parameters governing base step planarity such as roll, tilt, and buckle angles would also exhibit elevated scoring. Our dataset validates this expectation by showcasing a substantial prevalence of these parameters within the upper echelons of the top 10% of featured attributes.

Table 2: Top 10 helical parameters within TP53-related datasets identified by feature scoring. Abbreviations explained in supplementary table four.

Rank	*TP53*	*Val*	*Meth*

1	14pl	23re	14pl
2	23re	14pl	21ad
3	10ad	23tp	13pl
4	10tp	13pl	14og
5	11st	11st	11tp
6	10xp	16ad	23tp
7	14sh	10tp	6pl
8	13sg	23sr	12se
9	17ad	11pl	13ad
10	23tp	10xp	13se

Encouragingly, these trends extended to the methylation dataset, with tip, buckle, axis-bend, propel, and opening continuing to dominate as primary helical attributes. Notably this set of five parameters accounted for an impressive 72% of identified features, with tip emerging as the most influential contributor, commanding 19% of the overall proportion. Bases 10, 14, and 23 maintained their position as the highest-scoring base positions across all three *TP53* datasets. However, due to their distal location from the lesion site, it remains challenging to draw definitive conclusions regarding the precise significance of these positions in relation to the observed distortion.

2.6 Alternative Gene Contexts

After deploying our RF model and feature scoring pipeline, we employed the approach to deduce the relative discriminatory potential of our training data concerning hotspot and non-hotspot samples in alternative gene contexts, and identify potentially conserved helical parameters likely to contribute to the repair phenotype. Analogous to the previous mutational analysis, we harnessed the insights gained from the initial training event to guide the investigation of two supplementary unseen gene contexts with established BPDE-binding sites—namely, *lacZ* and *cII*. Mean accuracy, precision, f1 and recall scores obtained from 10-fold cross-validation can be seen in Table 3.

Table 3: Mean accuracy, precision, f1 and recall scores for RF model using 10-fold cross-validation.

Context	Accuracy	Precision	f1	recall
cII	*0.628*	*0.737*	*0.639*	*0.628*
lacZ	*0.788*	*0.884*	*0.800*	*0.788*

As anticipated, deploying the RF model on structural data acquired from duplexes with unseen base compositions yielded a noticeable decline in classifier performance. Specifically, in the context of *cII*, accuracy, f1, and recall demonstrated substantial reductions of approximately 43-45%. Remarkably, despite these changes, precision scores remained steadfast at 0.737, highlighting the model's classification potential amid shifting sequence contexts. Similarly, the RF model's predictive capacity encountered limitations when applied to the *lacZ* dataset. The model accurately predicted a significant

proportion of samples (78.8%), achieving a notable precision rate of 88.4%. Nonetheless, it's worth noting that recall (78.8%) slightly lagged precision, hinting at potential cases where positive instances might not have been precisely identified. These observations align with our expectations, underscoring the model's sensitivity to sequence intricacies and elucidate the potential boundaries surrounding the applicability of ML algorithms trained on individual gene contexts.

Table 4: Top 10 helical parameters within non-TP53 datasets identified by feature scoring. Abbreviations detailed in supplementary table four.

Rank	*cII*	*lacZ*
1	17rl	10in
2	15in	9be
3	13tp	17tp
4	16ad	20sh
5	8se	16tp
6	15ad	22tt
7	15tp	10be
8	11yp	17be
9	21tw	17sg
10	8in	5tp

Examining the top-ranked features in *cII* and *lacZ* datasets reaffirmed a notable emphasis on rotation-specific parameters. Parallel to that observed in *TP53*, the tip parameter remained dominant, constituting 22% and 28% of identified features in *cII* and *lacZ* datasets, respectively (Table 4). Similarly, an abundance of parameters associated with base step rotation, including buckle, inclin, propel, and opening, persisted, accounting for 58% (*cII*) and 69% (*lacZ*) of the top 36 features in each dataset. Despite minor alterations in the distribution of rotational parameters, these findings closely resembled those within *TP53*-specific contexts, accentuating the emphasis on rotational parameters as primary contributors to the discrimination process. Additionally, it's worth noting that the positioning of these parameters exhibited notable deviation from *TP53* results, aligning with our expectations due to the change in base contexts. Collectively, these insights underscore the fundamental role of rotational parameters in shaping the model's discriminative prowess across diverse gene contexts.

Discussion:

The intricacies of DNA structure and its dynamic flexibility constitute essential features of sustained eukaryotic life [40-43]. Paradoxically, the very attributes that underscore DNA's functionality can also support detrimental processes, such as those driving tumorigenesis [36,44,45]. While prior research has exposed the influence of BPDE on damaged base conformation and DNA bending, a thorough investigation into the structural factors driving mutations across multiple gene contexts has not been pursued [8,36,46]. To this end, we conducted a focused investigation to uncover how sequence

context influences the structural modifications induced by BPDE adducts at mutation sites within *TP53*, *cII*, and *lacZ* genes. This allowed us to identify conserved patterns in helical distortion among regions of elevated mutational frequency, irrespective of gene context, offering novel insights that can aid in the development of anti-cancer therapies.

Accessing and interpreting the information embedded within MD data, is non-trivial due to its high volume and dimensionality. As such, we embraced an approach based on ML principals, facilitating a comprehensive appraisal of helical structure across a substantial number of duplex samples. With various ML methodologies offering unique advantages for analysing such datasets, we embarked on an evaluation of 15 distinct ML classifiers, aiming to pinpoint the most appropriate model for our specific framework. As anticipated, decision tree-based models, including Decision Tree, Extra Trees, and RF, consistently outperformed other models across different class variables. Although the performance differences among these three models was marginal, it is noteworthy that the RF algorithm consistently achieved the highest scores for all evaluation metrics, and so became our method of choice for further analysis.

We leveraged a well-established feature in the field of cancer biology – the *TP53* mutation spectrum – to evaluate our model's proficiency in classifying previously unseen helical data [37]. When applied to the validation dataset, the model achieved the highest prediction scores, instilling confidence in the robustness and accuracy of our RF. Despite maintaining exceptional performance on the validation data, however, the classifier exhibited partial regression when applied to the methylation dataset. This variance is likely attributed to the differential methylation statuses between the training and test data, underscoring minor limitations in model generalisation. Indeed, the influence of CpG methylation on DNA flexibility and its potential repercussions on NER have been a subject of persistent inquiry. Several studies have indicated that regions with heightened cytosine methylation enhance NER function at BPDE lesions [47-49]. Given our models adaptability to methylation data, we suggest that cytosine methylation has limited influence on the observed dynamics, and therefore, on NER efficiency. This aligns seamlessly with the notion that all *TP53* CpG sites within lung tissue exhibit methylation, implying that methylation, in isolation, cannot comprehensively explain the variable DNA repair rates witnessed at BPDE binding sites [50].

The model's effectiveness noticeably declined when confronted with helical data from different gene contexts, aligning with the notion that underscores model sensitivity to variations in contextual information. While it is commonly agreed that imbalanced datasets adversely impacts classifier performance, it is possible to mitigate population bias and improve overall model training by combining all available data [51]. However, this strategy can complicate the assessment of information specific to the training dataset. Thus, we made the deliberate choice to maintain the segregation of data. It's essential to note that our input data had some limitations, particularly in the absence of information regarding non-hotspot sites. To address this, for *lacZ* we defined non-hotspot sites based on reported codon mutational frequencies upon BPDE exposure, categorising them as

sites with lower mutational frequencies compared to those of defined hotspots [39]. Similar limitations were observed in the *cII* dataset, where information on non-hotspot sites was limited [38]. Given these constraints in our non-hotspot data, we refrain from speculating whether this decline in performance solely stems from sequence context or if it's compounded by the inherent dataset imbalance present in *cII* and *lacZ* datasets.

A recurrent and pivotal parameter during our analysis was tip, denoting the angular displacement of the base pair around its y-axis [52]. Notably, this particular parameter exhibited remarkable prominence, emerging as the preeminent feature within each of the five helical datasets. Alternative rotational parameters, including buckle, opening, and propel also garnered high scores through our combined feature scoring approach. This outcome aligns with our expectations, considering that the collective variation in individual base/base step parameters is likely to exert a substantial impact on overall axis displacement. Our findings resonate with prior molecular modelling studies, which highlight the effects of DNA bending at established sites of DNA damage. Using multivariate statistical analysis, Menzies *et al*. report that buckle and opening parameters significantly contribute to the structural differentiation of adducted sequences in *TP53* [36]. Paul *et al.,* using crystallographic and fluorescence lifetime (FLT)-based conformational studies, show that elevated values for slide, roll, and twist parameters can aid DNA opening events during NER, highlighting a distinct interplay between rotational and transitional parameters [35]. As such, we refrain from speculating that rotational parameters alone govern NER functionality, instead highlighting the importance of parameters, including tip, as a potent discriminator of hotspot and non-hotspot sites.

A global analysis of feature importance scores also offered insights into the specific positions of base pairs with significant predictive utility. Results highlighted that bases of intrinsic importance to the classification process were distributed across the entirety of the 23-mer region under analysis. Moreover, each specific gene context exhibited a distinct importance profile, emphasising the context-dependent nature of these critical bases. This observation aligns with previous studies indicating that sequence context, even more distal to adjacent bases, can exert influence over local distortion and, consequently, NER capacity [32,33]. The specific mechanism and extent to which distal sequence affect NER, however, remains a topic of extensive debate.

In the NER pathway, the crucial processes of lesion recognition and initiation are instigated by the heterodimeric XPC–RAD23B–CETN2 complex, hereafter referred to as XPC. Using thermal energy alone, XPC scans the genomic landscape for irregularities by unwinding short segments of DNA in a non-specific manner, a process aptly referred to as the interrogation mode [53]. Local distortion or destabilisation induced by a DNA lesion increases XPC residency at the damage site, allowing for the insertion of functional components and subsequent separation of the helical backbone [54-56]. Lesion recognition is an indispensable, rate-limiting step of NER, and stalled XPC is strictly required for the recruitment of transcription factor, TFIIH [57-59]. Therefore, it is reasonable to consider that various base contexts may impart distinct preferences for strand separation, thus potentially influencing the

overall effectiveness of NER and, subsequently, mutational frequency. Paul *et al.,* in evaluating the role of sequence context on DNA 'opening' by Rad4/XPC, highlight that regions of elevated van der Waals stacking energy may confer resistance to DNA opening [35]. Moreover, it is widely reported that variation in stacking stabilisation among both adduct types and sequence contexts can compensate for destabilising distortions caused by these lesions and in turn, resist NER [60]. Interestingly, it has been found that that duplex stability depends linearly on G/C content, with the free energy of G/C base steps consistently lower than those of A/T or "mixed" base pair steps (with one G-C and one A-T pair) [61]. Given that mutations are AT-biased in eukaryotes, it's conceivable that regions with high GC content might confer resistance to NER processes due to their enhanced thermodynamic stability, primarily governed by base stacking van der Waals interactions [62-64].

Analysis of *TP53*, *cII,* and *lacZ* mutational signatures appear to echo these insights. By comparing a 10bp region spanning the lesion site (base pairs 2-12), we observe significant statistical disparities in the G/C content between hotspot and non-hotspot sites in *TP53* and *lacZ* ($P < 0.005$), (supplementary tables 1,3). In contrast, no such distinction is evident within the *cII* gene (supplementary table 2) potentially contributing to the observed limitations in classifier performance during *cII* deployment. While it is vital to explore the stacking profiles of adduct-bound duplexes in the specific sequence contexts discussed, our results strongly indicate that, in the case of BPDE adducts, there exists a substantial preservation of features distinguishing hotspot and non-hotspot sites, primarily governed by variations in G/C content.

In conclusion, our investigation firmly establishes the efficacy of employing a ML-based approach to not only classify, but also predict biological features from a selectively curated set of helical data. This underscores the practical utility of such methodologies in the analysis of results derived from MD simulations. Within the constraints of the models developed, our findings reveal the significant role of regional helical topology in distinguishing mutation hotspot from non-hotspot sites. Particularly, we showcase that the RF classification of BPDE-bound duplexes is primarily steered by variations in base pair tip, likely stemming from differing thermodynamic stability linked to regional GC content. It's worth noting that while our analysis has provided insights into the distinctive features influencing mutational patterns in well-studied genes, a broader dataset encompassing a larger variety of sequences is essential to comprehensively understand the precise origins of these mutations. We highlight that this approach is not limited to the genes discussed above and could be applied to any gene subset of interest, including those associated with alternative human diseases. Furthermore, our methodology holds potential for screening mutation hotspot sites to assess the mutational impact of both novel and existing drug compounds. This extends the applicability of our approach beyond fundamental research, making it a valuable tool in drug development and toxicology studies.

Methods:

DNA Sequences

TP53, *cII*, and *lacZ* were selected as suitable models to explore the influence of sequence context on codon mutability. For each gene context, MD simulations were performed on 25-mer duplex sequences encompassing either mutation hotspot or non-hotspot sites in the presence of the 10S (+)-trans-N2-BPDE-dG adduct (Supplementary Table S1-4). Each sequence consisted of a methylated cytosine at position six, and guanine at position 7. To replicate the homomethylation observed in *TP53*-associated lung cancer, unique structures were generated wherein methylation groups were added to all CpG dinucleotides on both forward and reverse strands. For all sequences, position 7 defines a known location of BPDE binding, and subsequently, the mutation site for any hotspot sequence. Simulations were performed on both adducted and non-adducted duplexes. Sequences were defined as having a mutation hotspot at position 7 if the frequency of G:C>T:A substitutions observed at this site was significantly elevated compared to the expected frequency. *TP53* codon 282 is characterised by a G:C>A:T substitution, however, given its extremely high incidence in lung tumours, was included in our analysis. Mutational data for *TP53*, *cII*, and *lacZ* gene contexts were obtained from [37], [38], and [39], respectively.

Starting Structures

All sequences were assembled using the Builder toolbox in PyMOL [65]. This process involved the manual input of each sequence in the form of the single-letter code. Once inputted, Builder then generates the corresponding nucleotide structure, assembling coordinates for sugar, phosphate, and base groups, as well as those for any modifications present. Builder was also utilised in the assembly of the methylated cytosine bases. A nuclear magnetic resonance (NMR) structure for the 7S,8R,dihydroxy-9R,10S-epoxy-7S,8R,9R,10S (+)-trans-anti-B[*a*]PDE DNA adduct (PDB ID: 1AXO) was utilised to assemble carcinogen-bound duplexes [66]. This PDB file contains coordinates for six structures obtained from a relaxation matrix refinement, and the first structure was chosen. The sequence present within the obtained structure was remodelled to resemble hotspot and non-hotspot sites within each gene context using the Builder toolbox as described above.

Molecular Dynamics Simulations (MD)

All simulations were carried out in triplicate using the GROMACS package and Amber99 force field with backbone torsion modifications, bsc1 [67,68]. Forcefield parameters for BPDE-bound to guanine are detailed in [46]. Partial charges for the methylated cytosine base were calculated using Atomic Charge Calculator and subsequently added to the forcefield [69]. All structures were placed in a cubic box, solvated using the explicit water model TIP3P, and neutralised with the appropriate number of Na^+ ions prior to simulation. Temperature coupling thermostat was applied using v-rescale, and Particle-mesh Ewald (PME) was used for long-range electrostatics. Simulations were carried out using the NPT ensemble, with periodic boundary conditions, at a temperature of 310.15°K, and a pressure of 1atm. All simulations were performed, in triplicate, using three-stage process: steepest descent energy minimisation with a tolerance of 1000 kJ mol^{-1} nm^{-1}, followed by a two-stage equilibration process, each one 50,000 steps in length with a time step integration of 0.002ps, making a total of 100ps; and an MD stage run for a total of 100ns.

DNA Structural Parameters

To evaluate structural distortion resulting from adduct formation, we assessed seventeen distinct DNA structural parameters for each Molecular Dynamics (MD) trajectory. These parameters encompassed both intra-base pair characteristics—shear, propeller, buckle, stretch, stagger, and opening—as well as inter-base pair attributes—twist, roll, tilt, rise, shift, and slide. Supplementing these, two translational base pair axis parameters (X and Y displacement), two rotational parameters (inclination and tip), and an axis-bend parameter were calculated to define base pair geometry relative to the helical axis. Parameters were measured using open source Curves+ and Canal software [52].

Classifier Comparison

To select an optimal classifier, we compared the predictive capabilities of 15 distinct classification algorithms. We compared the performance of Nearest Neighbours, Linear support vector machine (SVM), Polynomial SVM, Radial Basis Function (RBF) SVM, Gaussian Process, Gradient Boosting, Decision Tree, Extra Trees, Random Forest, Neural Network, AdaBoost, naïve Bayes, Quadratic Discriminant Analysis (QDA), Stochastic Gradient Descent (SGD), and Linear Discriminant Analysis (LDA). Each classifier was instantiated with default parameters and trained on 80% of the *TP53* dataset. Trained models were then used to predict the target variable for the testing set, with metrics accuracy, precision, recall, f1, train-test gap, and AUC ROC used to evaluate model performance. Helical information for base pairs 3-23 only was used as ML input data.

Random Forest

Each decision tree in the RF was trained to predict the repair phenotype and carcinogen status of the duplexes described above, based on their structural dynamics (Figure 5). For model evaluation, 80% of the *TP53* dataset was randomly selected as training data, with the remaining data being separated for model testing. Utilising identical training events, the classifier was then deployed to various mutational models including *cII*, *LacZ*, and novel *TP53* datasets, where 95% of the unseen data was used for testing, ensuring a comprehensive representation of RF applicability. Optimal hyperparameters were selected using 10-fold GSO. Each RF model were implemented using the RandomForestClassifier module from the Python machine learning library scikit-learn [70]. RF models were ran using the same 10 random seeds to ensure reproducibility across each validated model.

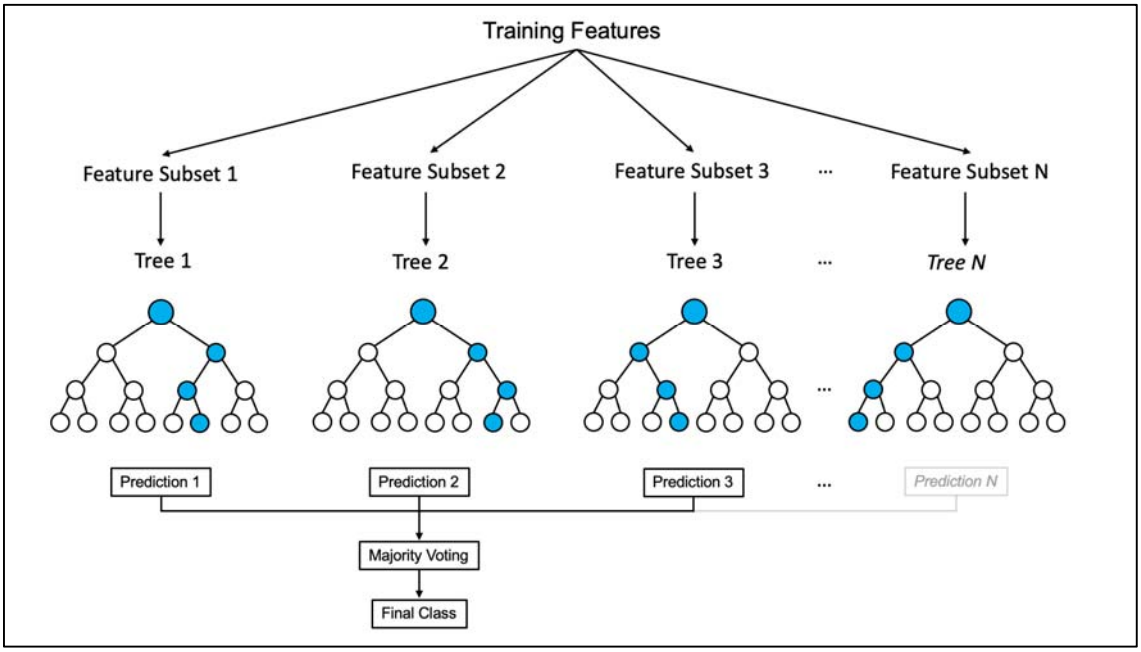

Figure 5: Framework of a Random Forest algorithm.

Feature selection was conducted to identify a subset of relevant features essential for constructing robust learning models and informing the classification algorithm. Multiple feature selection techniques were employed to ensure comprehensive coverage and insights into feature importance. Leveraging RF feature importance, recursive feature elimination, and lasso regularization, we identified the most informative features for our analysis. From each technique, we extracted the top 10% (36) features. These features were assigned scores ranging from 36 (highest importance) to 0 (lowest importance), excluding features with a score of 0. The resultant scores from each technique were amalgamated across datasets, resulting in a comprehensive list of the most significant features for our investigation.

References

1. Alexandrov, L. B. et al. 2013. Signatures of mutational processes in human cancer. *Nature* 500(7463), pp. 415-421. doi: 10.1038/nature12477
2. Bae, T. et al. 2018. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science* 359(6375), pp. 550-555. doi: 10.1126/science.aan8690
3. Ellegren, H., Smith, N. G. and Webster, M. T. 2003. Mutation rate variation in the mammalian genome. *Curr Opin Genet Dev* 13(6), pp. 562-568. doi: 10.1016/j.gde.2003.10.008
4. Baer, C. F., Miyamoto, M. M. and Denver, D. R. 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat Rev Genet* 8(8), pp. 619-631. doi: 10.1038/nrg2158
5. Harfe, B. D. and Jinks-Robertson, S. 2000. Sequence composition and context effects on the generation and repair of frameshift intermediates in mononucleotide runs in Saccharomyces cerevisiae. *Genetics* 156(2), pp. 571-578. doi: 10.1093/genetics/156.2.571
6. Jackson, A. L., Chen, R. and Loeb, L. A. 1998. Induction of microsatellite instability by oxidative DNA damage. *Proc Natl Acad Sci U S A* 95(21), pp. 12468-12473. doi: 10.1073/pnas.95.21.12468

7. Baylin, S. B., Herman, J. G., Graff, J. R., Vertino, P. M. and Issa, J. P. 1998. Alterations in DNA methylation: a fundamental aspect of neoplasia. *Adv Cancer Res* 72, pp. 141-196.
8. Donny-Clark, K. and Broyde, S. 2009. Influence of local sequence context on damaged base conformation in human DNA polymerase iota: molecular dynamics studies of nucleotide incorporation opposite a benzo[a]pyrene-derived adenine lesion. *Nucleic Acids Res* 37(21), pp. 7095-7109. doi: 10.1093/nar/gkp745
9. Page, J. E., Zajc, B., Oh-hara, T., Lakshman, M. K., Sayer, J. M., Jerina, D. M. and Dipple, A. 1998. Sequence context profoundly influences the mutagenic potency of trans-opened benzo[a]pyrene 7,8-diol 9,10-epoxide-purine nucleoside adducts in site-specific mutation studies. *Biochemistry* 37(25), pp. 9127-9137. doi: 10.1021/bi980273v
10. Rogozin, I. B. et al. 2018. Mutational signatures and mutable motifs in cancer genomes. *Brief Bioinform* 19(6), pp. 1085-1101. doi: 10.1093/bib/bbx049
11. Hodgkinson, A. and Eyre-Walker, A. 2011. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* 12(11), pp. 756-766. doi: 10.1038/nrg3098
12. Hollstein, M., Sidransky, D., Vogelstein, B. and Harris, C. C. 1991. p53 mutations in human cancers. *Science* 253(5015), pp. 49-53. doi: 10.1126/science.1905840
13. Toyooka, S., Tsuda, T. and Gazdar, A. F. 2003. The TP53 gene, tobacco exposure, and lung cancer. *Hum Mutat* 21(3), pp. 229-239. doi: 10.1002/humu.10177
14. Lane, D. P. 1992. Cancer. p53, guardian of the genome. *Nature* 358(6381), pp. 15-16. doi: 10.1038/358015a0
15. Vogelstein, B., Lane, D. and Levine, A. J. 2000. Surfing the p53 network. *Nature* 408(6810), pp. 307-310. doi: 10.1038/35042675
16. Hainaut, P. and Hollstein, M. 2000. p53 and human cancer: the first ten thousand mutations. *Adv Cancer Res* 77, pp. 81-137. doi: 10.1016/s0065-230x(08)60785-x
17. Béroud, C. and Soussi, T. 2003. The UMD-p53 database: new mutations and analysis tools. *Hum Mutat* 21(3), pp. 176-181. doi: 10.1002/humu.10187
18. Yu, D., Berlin, J. A., Penning, T. M. and Field, J. 2002. Reactive oxygen species generated by PAH o-quinones cause change-in-function mutations in p53. *Chem Res Toxicol* 15(6), pp. 832-842. doi: 10.1021/tx010177m
19. Xu, D., Penning, T. M., Blair, I. A. and Harvey, R. G. 2009. Synthesis of phenol and quinone metabolites of benzo[a]pyrene, a carcinogenic component of tobacco smoke implicated in lung cancer. *J Org Chem* 74(2), pp. 597-604. doi: 10.1021/jo801864m
20. Rodin, S. N. and Rodin, A. S. 2005. Origins and selection of p53 mutations in lung carcinogenesis. *Semin Cancer Biol* 15(2), pp. 103-112. doi: 10.1016/j.semcancer.2004.08.005
21. Vu, A. T., Taylor, K. M., Holman, M. R., Ding, Y. S., Hearn, B. and Watson, C. H. 2015. Polycyclic Aromatic Hydrocarbons in the Mainstream Smoke of Popular U.S. Cigarettes. *Chem Res Toxicol* 28(8), pp. 1616-1626. doi: 10.1021/acs.chemrestox.5b00190
22. Baird, W. M., Hooven, L. A. and Mahadevan, B. 2005. Carcinogenic polycyclic aromatic hydrocarbon-DNA adducts and mechanism of action. *Environ Mol Mutagen* 45(2-3), pp. 106-114. doi: 10.1002/em.20095
23. Denissenko, M. F., Pao, A., Tang, M. and Pfeifer, G. P. 1996. Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in P53. *Science* 274(5286), pp. 430-432. doi: 10.1126/science.274.5286.430
24. Denissenko, M. F., Pao, A., Pfeifer, G. P. and Tang, M. 1998. Slow repair of bulky DNA adducts along the nontranscribed strand of the human p53 gene may explain the strand bias of transversion mutations in cancers. *Oncogene* 16(10), pp. 1241-1247. doi: 10.1038/sj.onc.1201647
25. Matter, B., Wang, G., Jones, R. and Tretyakova, N. 2004. Formation of diastereomeric benzo[a]pyrene diol epoxide-guanine adducts in p53 gene-derived DNA sequences. *Chem Res Toxicol* 17(6), pp. 731-741. doi: 10.1021/tx049974l
26. Chen, J. X., Zheng, Y., West, M. and Tang, M. S. 1998. Carcinogens preferentially bind at methylated CpG in the p53 mutational hot spots. *Cancer Res* 58(10), pp. 2070-2075.
27. Yoon, J. H., Lee, C. S. and Pfeifer, G. P. 2003. Simulated sunlight and benzo[a]pyrene diol epoxide induced mutagenesis in the human p53 gene evaluated by the yeast functional assay: lack of correspondence to tumor mutation spectra. *Carcinogenesis* 24(1), pp. 113-119. doi: 10.1093/carcin/24.1.113
28. Buterin, T. et al. 2000. Unrepaired fjord region polycyclic aromatic hydrocarbon-DNA adducts in ras codon 61 mutational hot spots. *Cancer Res* 60(7), pp. 1849-1856.

29. Cai, Y., Patel, D. J., Geacintov, N. E. and Broyde, S. 2009. Differential nucleotide excision repair susceptibility of bulky DNA adducts in different sequence contexts: hierarchies of recognition signals. *J Mol Biol* 385(1), pp. 30-44. doi: 10.1016/j.jmb.2008.09.087
30. Kropachev, K. et al. 2009. The sequence dependence of human nucleotide excision repair efficiencies of benzo[a]pyrene-derived DNA lesions: insights into the structural factors that favor dual incisions. *J Mol Biol* 386(5), pp. 1193-1203. doi: 10.1016/j.jmb.2008.12.082
31. Cai, Y., Patel, D. J., Geacintov, N. E. and Broyde, S. 2007. Dynamics of a benzo[a]pyrene-derived guanine DNA lesion in TGT and CGC sequence contexts: enhanced mobility in TGT explains conformational heterogeneity, flexible bending, and greater susceptibility to nucleotide excision repair. *J Mol Biol* 374(2), pp. 292-305. doi: 10.1016/j.jmb.2007.09.034
32. Cai, Y. et al. 2010a. Distant neighbor base sequence context effects in human nucleotide excision repair of a benzo[a]pyrene-derived DNA lesion. *J Mol Biol* 399(3), pp. 397-409. doi: 10.1016/j.jmb.2010.04.004
33. Cai, Y., Patel, D. J., Broyde, S. and Geacintov, N. E. 2010b. Base sequence context effects on nucleotide excision repair. *J Nucleic Acids* 2010, doi: 10.4061/2010/174252
34. Cai, Y., Geacintov, N. E. and Broyde, S. 2012. Nucleotide excision repair efficiencies of bulky carcinogen-DNA adducts are governed by a balance between stabilizing and destabilizing interactions. *Biochemistry* 51(7), pp. 1486-1499. doi: 10.1021/bi201794x
35. Paul, D. et al. 2021. Impact of DNA sequences on DNA 'opening' by the Rad4/XPC nucleotide excision repair complex. *DNA Repair (Amst)* 107, p. 103194. doi: 10.1016/j.dnarep.2021.103194
36. Menzies, G. E., Reed, S. H., Brancale, A. and Lewis, P. D. 2015. Base damage, local sequence context and TP53 mutation hotspots: a molecular dynamics study of benzo[a]pyrene induced DNA distortion and mutability. *Nucleic Acids Res* 43(19), pp. 9133-9146. doi: 10.1093/nar/gkv910
37. Forbes, S. A. et al. 2015. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 43(Database issue), pp. D805-811. doi: 10.1093/nar/gku1075
38. Yoon, J. H., Smith, L. E., Feng, Z., Tang, M., Lee, C. S. and Pfeifer, G. P. 2001. Methylated CpG dinucleotides are the preferential targets for G-to-T transversion mutations induced by benzo[a]pyrene diol epoxide in mammalian cells: similarities with the p53 mutation spectrum in smoking-associated lung cancers. *Cancer Res* 61(19), pp. 7110-7117.
39. Beal, M. A., Gagné, R., Williams, A., Marchetti, F. and Yauk, C. L. 2015. Characterizing Benzo[a]pyrene-induced lacZ mutation spectrum in transgenic mice using next-generation sequencing. *BMC Genomics* 16, p. 812. doi: 10.1186/s12864-015-2004-4
40. Yella, V. R., Vanaja, A., Kulandaivelu, U. and Kumar, A. 2020. Delving into Eukaryotic Origins of Replication Using DNA Structural Features. *ACS Omega* 5(23), pp. 13601-13611. doi: 10.1021/acsomega.0c00441
41. Abeel, T., Saeys, Y., Bonnet, E., Rouzé, P. and Van de Peer, Y. 2008. Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res* 18(2), pp. 310-323. doi: 10.1101/gr.6991408
42. Yella, V. R., Bhimsaria, D., Ghoshdastidar, D., Rodríguez-Martínez, J. A., Ansari, A. Z. and Bansal, M. 2018. Flexibility and structure of flanking DNA impact transcription factor affinity for its core motif. *Nucleic Acids Res* 46(22), pp. 11883-11897. doi: 10.1093/nar/gky1057
43. Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S. and Honig, B. 2009. The role of DNA shape in protein-DNA recognition. *Nature* 461(7268), pp. 1248-1253. doi: 10.1038/nature08473
44. DiGiovanna, J. J. and Kraemer, K. H. 2012. Shining a light on xeroderma pigmentosum. *J Invest Dermatol* 132(3 Pt 2), pp. 785-796. doi: 10.1038/jid.2011.426
45. Feng, Z. et al. 2002. Preferential DNA damage and poor repair determine ras gene mutational hotspot in human cancer. *J Natl Cancer Inst* 94(20), pp. 1527-1536. doi: 10.1093/jnci/94.20.1527
46. Menzies, G. E., Prior, I. A., Brancale, A., Reed, S. H. and Lewis, P. D. 2021. Carcinogen-induced DNA structural distortion differences in the RAS gene isoforms; the importance of local sequence. *BMC Chem* 15(1), p. 51. doi: 10.1186/s13065-021-00777-8
47. Derreumaux, S., Chaoui, M., Tevanian, G. and Fermandjian, S. 2001. Impact of CpG methylation on structure, dynamics and solvation of cAMP DNA responsive element. *Nucleic Acids Res* 29(11), pp. 2314-2326. doi: 10.1093/nar/29.11.2314

48. Muheim, R., Buterin, T., Colgate, K. C., Kolbanovsij, A., Geacintov, N. E. and Naegeli, H. 2003. Modulation of human nucleotide excision repair by 5-methylcytosines. *Biochemistry* 42(11), pp. 3247-3254. doi: 10.1021/bi0268504
49. Colgate, K., Huang, X., Kolbanovskiy, A., Skorvaga, M., Van Houten, B., Amin, S. and Geacintov, N. 2003. Effects of methylation and temperature on nucleotide excision repair of stereoisomeric benzo[a]pyrene diol epoxide-N2-dG adducts in a 5-MeCG*p53 codon 273 sequence context catalyzed by UvrABC proteins from B-caldotenax *Chemical Research in Toxicology* 16(12), pp. 1676-1677
50. Tornaletti, S. and Pfeifer, G. P. 1995. Complete and tissue-independent methylation of CpG sites in the p53 gene: implications for mutations in human cancers. *Oncogene* 10(8), pp. 1493-1499.
51. C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse and A. Napolitano, "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance," in *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 40, no. 1, pp. 185-197, Jan. 2010, doi: 10.1109/TSMCA.2009.2029559.
52. Olson, W. K. et al. 2001. A standard reference frame for the description of nucleic acid base-pair geometry. *J Mol Biol* 313(1), pp. 229-237. doi: 10.1006/jmbi.2001.4987
53. Chen, X. et al. 2015. Kinetic gating mechanism of DNA damage recognition by Rad4/XPC. *Nat Commun* 6, p. 5849. doi: 10.1038/ncomms6849
54. Hoogstraten, D. et al. 2008. Versatile DNA damage detection by the global genome nucleotide excision repair protein XPC. *J Cell Sci* 121(Pt 17), pp. 2850-2859. doi: 10.1242/jcs.031708
55. Dip, R., Camenisch, U. and Naegeli, H. 2004. Mechanisms of DNA damage recognition and strand discrimination in human nucleotide excision repair. *DNA Repair (Amst)* 3(11), pp. 1409-1423. doi: 10.1016/j.dnarep.2004.05.005
56. Gunz, D., Hess, M. T. and Naegeli, H. 1996. Recognition of DNA adducts by human nucleotide excision repair. Evidence for a thermodynamic probing mechanism. *J Biol Chem* 271(41), pp. 25089-25098. doi: 10.1074/jbc.271.41.25089
57. Li, C. L., Golebiowski, F. M., Onishi, Y., Samara, N. L., Sugasawa, K. and Yang, W. 2015. Tripartite DNA Lesion Recognition and Verification by XPC, TFIIH, and XPA in Nucleotide Excision Repair. *Mol Cell* 59(6), pp. 1025-1034. doi: 10.1016/j.molcel.2015.08.012
58. Sugasawa, K., Akagi, J., Nishi, R., Iwai, S. and Hanaoka, F. 2009. Two-step recognition of DNA damage for mammalian nucleotide excision repair: Directional binding of the XPC complex and DNA strand scanning. *Mol Cell* 36(4), pp. 642-653. doi: 10.1016/j.molcel.2009.09.035
59. Volker, M. et al. 2001. Sequential assembly of the nucleotide excision repair factors in vivo. *Mol Cell* 8(1), pp. 213-224. doi: 10.1016/s1097-2765(01)00281-7
60. Mu, H. et al. 2012. Nucleotide excision repair of 2-acetylaminofluorene- and 2-aminofluorene-(C8)-guanine adducts: molecular dynamics simulations elucidate how lesion structure and base sequence context impact repair efficiencies. *Nucleic Acids Res* 40(19), pp. 9675-9690. doi: 10.1093/nar/gks788
61. Zacharias, M. 2020. Base-Pairing and Base-Stacking Contributions to Double-Stranded DNA Formation. *J Phys Chem B* 124(46), pp. 10345-10352. doi: 10.1021/acs.jpcb.0c07670
62. Lynch, M. 2010. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A* 107(3), pp. 961-968. doi: 10.1073/pnas.0912629107
63. Petrov, D. A. and Hartl, D. L. 1999. Patterns of nucleotide substitution in Drosophila and mammalian genomes. *Proc Natl Acad Sci U S A* 96(4), pp. 1475-1479. doi: 10.1073/pnas.96.4.1475
64. Lynch, M. et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A* 105(27), pp. 9272-9277. doi: 10.1073/pnas.0803466105
65. Schrödinger, L. & DeLano, W., 2020. *PyMOL*, Available at: http://www.pymol.org/pymol.
66. Feng, B., Gorin, A., Hingerty, B. E., Geacintov, N. E., Broyde, S. and Patel, D. J. 1997a. Structural alignment of the (+)-trans-anti-benzo[a]pyrene-dG adduct positioned opposite dC at a DNA template-primer junction. *Biochemistry* 36(45), pp. 13769-13779. doi: 10.1021/bi970069s
67. Páll, S., Zhmurov, A., Bauer, P., Abraham, M., Lundborg, M., Gray, A., Hess, B. and Lindahl, E. 2020. Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS. *The Journal of Chemical Physics*, **153**, 134110.
68. Case DA, Aktulga HM, Belfon K, Ben-Shalom IY, Brozell SR, Cerutti DS, Cheatham TE, III, Cruzeiro VWD, Darden TA, Duke RE, Giambasu G, Gilson MK, Gohlke H, Goetz AW, Harris R, Izadi S, Izmailov SA, Jin C, Kasavajhala K, Kaymak MC, King E, Kovalenko A, Kurtzman T,

Lee TS, LeGrand S, Li P, Lin C, Liu J, Luchko T, Luo R, Machado M, Man V, Manathunga M, Merz KM, Miao Y, Mikhailovskii O, Monard G, Nguyen H, O'Hearn KA, Onufriev A, Pan F, Pantano S, Qi R, Rahnamoun A, Roe DR, Roitberg A, Sagui C, Schott-Verdugo S, Shen J, Simmerling CL, Skrynnikov NR, Smith J, Swails J, Walker RC, Wang J, Wei H, Wolf RM, Wu X, Xue Y, York DM, Zhao S, Kollman PA. *AMBER 11.* San Francisco: University of California; 2010.

69. Raček, T., Schindler, O., Toušek, D., Horský, V., Berka, K., Koča, J., & Svobodová, R. 2020. Atomic Charge Calculator II: web-based tool for the calculation of partial atomic charges.
70. Pedregosa,F.a.V.,G,Gramfort,A. Michel,V. Thirion,B.a.G.,O. Blondel,M. Prettenhofer,P. Weiss, R.a.D.,V. Vanderplas,J. Passos,A.Cournapeau,D. Brucher,M. Perrot,M, Duchesnay,E. 2011. Scikit-learn:Machine Learning in Python. *Journal of Machine Learning Research* 12, pp.2825-2830.