nature portfolio

Peer Review File

Explainable artificial intelligence of DNA methylation-based brain tumor diagnostics

Corresponding Author: Dr Volker Hovestadt

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Developing AI algorithms to automate complex tasks in tumor classification holds great potential for clinical diagnosis. This study by Dr. Benfatto is based on a large and well annotated data set of pediatric brain tumors. The new classifiers should theoretically facilitate the interpretation of brain tumor classifiers and facilitate the identification of new epigenetic signatures that can better serve as new diagnostic markers or biological determinants. Their finding on the global patterns of differential probe usage is interesting, particularly the negligible contributions of a majority of probes and it was the limited number of probes that were highly informative to distinguish between tumor classes (although the reference of “class” here needs clarification). Another novel result is that the probes located distal to promoter regions (e.g., in enhancers. Gene bodies) are more informative for classification of most cases. Illustrating the prob-associated genes from the shinyMNP web application can be very useful to the field not only in establish probe-gene connections but also potentially with mRNA expression. The data figures were nicely prepared. There are, however, some concerns:

Major concerns:
1. While the authors claim that “we have developed a machine learning classifier that enables fast, accurate and affordable classification…”. There were no description/report of the classifiers. Were they for tumor subtypes? How many are they? Are they similar or different from previous reports?

2. The “classes“ on line 178 page 6 should be defined. Since the aim of this study is to support clinical application, the definitions and terms should ideally be clinically used or relevant. The overall strategy in the current report is heavily focused on the bio-informatic analysis with much less clinical relevance. Without proper definition of “class”, the impact the reported studies/algorithm/website is dramatically reduced.

3. The identification of the contribution of a majority of probes is negligible while few probes are highly informative to distinguish between tumor classes. Again, is this referring to different types of brain tumors (i.e., glioma, medulloblastoma, ependymoma) or subgroups of the same pathological diagnosis?

4. One potential limitation of this study is that it was based on the data generated with the 450K DNA methylation arrays. With the increased density to 850k or whole genome DNA methylation sequencing, the findings may or may not hold true. Can the current algorithm be applied to these high-resolution analysis?

Minor concerns”
1. The rationale of using DHS, LADs and heterochromatin domains in the probe usage in functional genomic regions was not well presented.

2. Many abbreviations were not defined (e.g., line 159, on page 6).

3. How were the “10,000 selected probes” (Line 176, page 6) selected?

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
This paper presents an extension of the authors' previous work, which introduced a machine learning classifier aimed at predicting the presence or absence of brain tumors in patients, along with identifying the type of tumors among 82 classes. The original paper, published in 2018, trained the classifier on 2,801 samples consisting of 428,799 features, primarily methylation status DNA sites. In this current work, the authors propose an interpretation of the classifier's outcomes by: 1) investigating the 10,000 most important features used by the classifier in decision-making, 2) elucidating the primary probes necessary for classifying each tumor class, and 3) offering a web application for dataset exploration. The paper discusses the potential impact of this study on designing genetic tests to accurately identify brain tumors in patients. In my assessment, amid the ongoing development of genetic tests, such studies represent valuable efforts in creating tools applicable within a clinical context. Overall, the paper is well-written and easy to follow.

My critique of this paper will primarily focus on the machine learning approach rather than the biological insights into tumor biology, as I lack sufficient expertise in the latter. Regarding the machine learning approach, the study employs a random forest classifier, noted in the paper as one of the most commonly used methods in the literature for high-dimensional datasets. The classifier comprises 10,000 trees trained on the dataset, with features subsequently extracted from these trees. However, the methodology for feature extraction lacks clarity and appears tailored specifically for this study. The paper states that, due to the inaccessible nature of this information from a trained RF model object, the authors devised a specialized algorithm for feature extraction. While this may be a peculiarity of the code used for the Heidelberg brain tumor classifier, it seems notably simpler to extract estimators, at least in Python using scikit-learn, from a random forest classifier. I am curious as to why this is not feasible in this case.

Following feature extraction, the pipeline involves tallying probe usage in decision trees aggregated by class combinations for each probe. The top 10,000 probes are extracted, with results indicating their contribution to 61.2% of the probe combinations, thereby explaining the classifier's decisions. My primary concern here lies in the stability of the proposed method. The 10,000 most insightful probes are derived from a single classifier trained with specific parameters (not delineated in this paper but retrievable from the original study). The paper should assess the stability of the extracted probes across multiple trainings, possibly employing different parameters. A parameter study of the main random forest hyperparameters (initially the number of estimators, but also considering maximum depth, maximum number of features in splits, etc.) would be beneficial. Given the inherent randomness in the random forest classifier at various levels (e.g., feature and sample selection for training different decision trees), it is crucial, in my opinion, to statistically evaluate whether the 10,000 primary probes remain consistent or are highly dependent on the random choices made by the machine learning algorithms.

While these 10,000 probes influence the decision trees, it remains unclear to what extent they determine the differentiation between various classes. Training the random forest solely with these 10,000 probes to evaluate their impact on classification metrics (e.g., accuracy, recall) would be informative in assessing their predictive capabilities.
Probe usage is then assessed to cluster the tumor classes using unsupervised learning techniques. The clustering, based on K-nearest neighbors to construct Shared Nearest Neighbor (SNN) graphs and cluster identification using the Louvain algorithm, results in 88 distinct clusters, appearing well-separated in the t-SNE representation provided in Figure 3. However, the accompanying text for the figure is somewhat vague in describing the obtained clusters relative to the tumor classes: "Interestingly, we found that the majority of clusters were associated with a single tumor class, indicating high class specificity of most probes." Clarification is needed regarding the precise definition of "the majority of clusters," as a numerical estimation would facilitate understanding the method's coverage.

In summary, while the paper is highly intriguing, the lack of evaluation of the stability of the proposed methods and the absence of sensitivity analysis of hyperparameters render this work challenging to gauge in terms of the reliability of the obtained 10,000 primary probes and subsequent analyses. Additionally, while the web application code is available, it is regrettable that the code for the machine learning pipeline is not accessible for review.

(Remarks on code availability)
Only the code of the webapp is share. The provided coded is well commented and organized. An example is provided with a simple classifier data to use the app and it seems (did not verify though) to connect a new classifier.

The code of the main machine learning analysis is not provided and would be helpful to fully review and evaluate the viability of the approach.

Reviewer #3

(Remarks to the Author)
The authors take advantage of a machine learning classifier that allows for classification of brain tumors based on genome-wide DNA methylation profiles. They attempt to describe an interpretable framework to explain the classifier’s decisions. They show that functional genomic regions are employed to distinguish between tumor classes and describe the nature of the methylation sites that are the basis of the classification. Overall, the motivation of the work is strong and the methods are robust. I believe a couple of issues if addressed would improve the impact of this work.

1. The authors use the original 2018 (“version 11”) classifier with 91 classes as the basis for their work. The filed has moved to the next version (“version 12”) classifier, that has more than double the number of classes. Given that, their analysis seems outdated and is less relevant, since it applies to a classifier that is no longer the gold standard. The authors should apply their pipeline to the version 12 classifier so that conveys relevance to the current understanding of brain tumor methylation-based classification.
2. It would be important to know if the principles learned from brain tumor classification are unique to brain tumors, or instead whether they generally apply to methylation cancer classifiers. Some of the authors of this work were involved in the sarcoma classifier. It would be of interest to know the extent to which main principles unveiled in their work from the brain tumor classifier also apply to the classification of sarcomas—this would provide important insights that will contribute to the understanding of what will likely become and ever-expanding number of methylation-based classifiers across many cancers.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #3

(Remarks to the Author)
I had stated in my original review that the authors apply their analyses and models to an outdated ("Version 11") version of the classifier I had suggested that the authors apply their analyses to the version of the classifier that people are actually using now ("Version 12"). The authors state that use of the Version 12 classifier is "outside the focus of the present study", but I think it actually is within scope. I do believe that applying these analyses to the current version of the classifier is well within the scope of the work, since the readers will want it to be currently applicable and as impactful as possible.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
In my opinion, the points raised by reviewer 1 have been adequately addressed. However, from the brain tumour biology side of things, I am not sure the conceptual advances/novelty of the data is very high. Firstly, I agree with reviewer 3 that it seems a missed opportunity that this work has been carried out on an outdated version of the classifier. Given the emphasis on developing and sharing this approach to enable novel discoveries in the brain tumour field, surely this work should be based on the most recent classifier. Moreover, despite the authors’ claims no evidence is provided that novel insights in brain tumour biology can be gained with this approach. The three lines of validation provided (SHPRH, PWWP3A and TBX19) are confirmatory of existing knowledge and while this is an essential first step, I would have expected at least one example of a novel finding to be shown, including its functional validation to substantiate the claim that this approach and the resource shared will indeed be impactful.

(Remarks on code availability)

Reviewer #5

(Remarks to the Author)
The authors have nicely addressed the concerns of the reviewer.

(Remarks on code availability)

Reviewer #6

(Remarks to the Author)
I appreciate that this is already a revised manuscript and I have read the reply to the reviewers by the authors, but I'm a new reviewer brought to review this manuscript and naturally I got some fresh questions for the authors.

Overall this is a very interesting work which makes a clear effort to extract knowledge from the machine learning models rather than just building ML pipelines, evaluating them and report predictive performance, which is commendable. Having said this, the knowledge extraction effort performed in this paper is quite simple, and could be much better.

For instance, there are methods to infer supervised networks from the structure of interpretable machine learning models trained on omics data (e.g. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-469 https://biodatamining.biomedcentral.com/articles/10.1186/s13040-016-0106-4), these networks are quite different from those presented here as nodes would be probes rather than patients.

Moreover, there are Explainable AI methods able to extract very nuanced knowledge from machine learning models, much more refined than feature importance estimates. The SHAP methood comes to mind as the more popular of such methods. It integrates very nicely with scikit-learn models. I'm not so sure on the R implementation of RF.

Moreover, the use of t-SNE (or UMAP) as a step before clustering omics data is quite controversial, please see https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011288

From the machine learning methodology i'm surprised that the authors did not consider adding a feature selection step to the pipeline. While I agree that RF scales well to large feature spaces, it can be helped by specialised feature selection algorithms as the literature has shown many times. For instance, Recursive Feature Elimination (of course eliminating a percentage of features in each step rather than one by one) was originally designed with omics data in mind. Adding a feature selection could make the choice of features more robust across repetitions, which based on the results reported in the reply letter, is still not great.

Finally, it's not clear what evaluation methodology was used. How was the data split into training and test sets? Where the 10000 probes selected on the whole dataset or from the traininng partition of the data? (as they should). Where any preprocessing decisions taken based on the training data (as they should) or from the whole dataset? I also struggle to find a clear reporting of predictive performance of these models. If the models are not good at making predictions, any knowledge extraction is futile.

(Remarks on code availability)
Unclear from the code what kind of model evaluation methodology was used, or what preprocessing of data happened on top of the raw data.

Version 3:

Reviewer comments:

Reviewer #4

(Remarks to the Author)
The authors haven't addressed any of my concerns, the work continues to be based on an old and outdated version of the Heidelberg classifier and no functional validation of the findings have been provided.

(Remarks on code availability)

Reviewer #6

(Remarks to the Author)
I'm fine with the current version of the manuscript.

(Remarks on code availability)

Reviewer #7

(Remarks to the Author)

(Remarks on code availability)

Reviewer #8

(Remarks to the Author)

(Remarks on code availability)

Reviewer #1 (Remarks to the Author): expertise in CNS tumour methylation analysis

Developing AI algorithms to automate complex tasks in tumor classification holds great potential for clinical diagnosis. This study by Dr. Benfatto is based on a large and well annotated data set of pediatric brain tumors. The new classifiers should theoretically facilitate the interpretation of brain tumor classifiers and facilitate the identification of new epigenetic signatures that can better serve as new diagnostic markers or biological determinants. Their finding on the global patterns of differential probe usage is interesting, particularly the negligible contributions of a majority of probes and it was the limited number of probes that were highly informative to distinguish between tumor classes (although the reference of “class” here needs clarification). Another novel result is that the probes located distal to promoter regions (e.g., in enhancers. Gene bodies) are more informative for classification of most cases. Illustrating the prob-associated genes from the shinyMNP web application can be very useful to the field not only in establish probe-gene connections but also potentially with mRNA expression. The data figures were nicely prepared. There are, however, some concerns:

We thank the reviewer for their generally positive evaluation of our manuscript and their helpful comments which have allowed us to clarify some of the key concepts of our work.

Major concerns:
1. While the authors claim that “we have developed a machine learning classifier that enables fast, accurate and affordable classification…”. There were no description/report of the classifiers. Were they for tumor subtypes? How many are they? Are they similar or different from previous reports?

Thank you for raising this important point. In our work we refer to the Heidelberg CNS tumor classifier previously published by us and colleagues (Capper et al., Nature 2018) that is now widely employed to support diagnostic routines and in research. The classifier comprises 91 molecular classes (as defined by genome-wide DNA methylation profiling) that encompass all major entities and age groups. We have updated the manuscript to further clarify this point.

2. The “classes“ on line 178 page 6 should be defined. Since the aim of this study is to support clinical application, the definitions and terms should ideally be clinically used or relevant. The overall strategy in the current report is heavily focused on the bio-informatic analysis with much less clinical relevance. Without proper definition of “class”, the impact the reported studies/algorithm/website is dramatically reduced.

We would like to refer to our original study (Capper et al., Nature 2018) for a detailed definition of the 91 methylation classes and their relation to WHO-based CNS tumor entities and subtypes. This molecular reference was the result of a significant effort that integrated unsupervised analysis of DNA methylation profiles with histological, cytogenetic, mutational, transcriptional, and clinical datasets. While most of our methylation classes reflect defined tumor types or their subtypes (e.g., Group 3 medulloblastoma), we opted for the term “class” as it relates to the machine learning classifier that is used to classify query samples. We agree that a

clear definition of our 91 methylation classes is essential for our study and have clarified this point in our manuscript.

3. The identification of the contribution of a majority of probes is negligible while few probes are highly informative to distinguish between tumor classes. Again, is this referring to different types of brain tumors (i.e., glioma, medulloblastoma, ependymoma) or subgroups of the same pathological diagnosis?

This depends on the definition of the methylation class. Some methylation classes reflect tumor types (e.g., ETMR), while most classes reflect molecular subtypes of a given tumor type (e.g., H3.3-G34R-mutant GBM, WNT medulloblastoma, IDH-mutant oligodendroglioma). Please also see our response to comment 2 above.

4. One potential limitation of this study is that it was based on the data generated with the 450K DNA methylation arrays. With the increased density to 850k or whole genome DNA methylation sequencing, the findings may or may not hold true. Can the current algorithm be applied to these high-resolution analysis?

Yes, our algorithm to determine probe usage could be applied to random forest classifiers that are trained on higher resolution 850k methylation arrays or whole genome DNA methylation sequencing datasets. In addition, our algorithm is not limited to DNA methylation classifiers but could be applied to other data modalities.

Minor concerns”
1. The rationale of using DHS, LADs and heterochromatin domains in the probe usage in functional genomic regions was not well presented.

Thank you for this comment. We have edited the manuscript to better explain our rationale for investigating these genomic regions: “Other functional genomic regions that often show differential DNA methylation between cell types and disease states are enhancer regions and large-scale heterochromatic domains. Hypomethylation within enhancer regions is generally associated with enhancer activation and transcription factor binding. Heterochromatic domains, which frequently are positioned at the nuclear lamina, are condensed and transcriptionally silent regions in which DNA methylation is believed to be lost in a passive process in many different types of cancer. To investigate probe usage in these genomic regions of the RF classifier, …”

2. Many abbreviations were not defined (e.g., line 159, on page 6).

We have now carefully edited the manuscript to ensure that all abbreviations are defined upon first usage. Abbreviations of methylation classes are provided in the Extended Fig. 1a and in Supplementary Table 1.

3. How were the “10,000 selected probes” (Line 176, page 6) selected?

The 10,000 probes were selected from an “outer” random forest classifier that was trained on all 428,799 high quality probes. Probes were selected by assessing their per-class importance (MeanDecreaseAcuarcy value), i.e. we determined the rank of each probe within every class, and subsequently selected the 10,000 probes with the lowest rank across classes (low rank = high MeanDecreaseAcuarcy). We have clarified this point in the manuscript.

Reviewer #2 (Remarks to the Author): expertise in explainable AI

This paper presents an extension of the authors' previous work, which introduced a machine learning classifier aimed at predicting the presence or absence of brain tumors in patients, along with identifying the type of tumors among 82 classes. The original paper, published in 2018, trained the classifier on 2,801 samples consisting of 428,799 features, primarily methylation status DNA sites. In this current work, the authors propose an interpretation of the classifier's outcomes by: 1) investigating the 10,000 most important features used by the classifier in decision-making, 2) elucidating the primary probes necessary for classifying each tumor class, and 3) offering a web application for dataset exploration. The paper discusses the potential impact of this study on designing genetic tests to accurately identify brain tumors in patients. In my assessment, amid the ongoing development of genetic tests, such studies represent valuable efforts in creating tools applicable within a clinical context. Overall, the paper is well-written and easy to follow.

We thank the reviewer for their helpful comments and suggestions that helped to significantly improve our manuscript.

My critique of this paper will primarily focus on the machine learning approach rather than the biological insights into tumor biology, as I lack sufficient expertise in the latter. Regarding the machine learning approach, the study employs a random forest classifier, noted in the paper as one of the most commonly used methods in the literature for high-dimensional datasets. The classifier comprises 10,000 trees trained on the dataset, with features subsequently extracted from these trees. However, the methodology for feature extraction lacks clarity and appears tailored specifically for this study. The paper states that, due to the inaccessible nature of this information from a trained RF model object, the authors devised a specialized algorithm for feature extraction. While this may be a peculiarity of the code used for the Heidelberg brain tumor classifier, it seems notably simpler to extract estimators, at least in Python using scikit-learn, from a random forest classifier. I am curious as to why this is not feasible in this case.

Thank you for raising this important point. The randomForest package in R provides the MeanDecreaseGini value as a measure of global feature importance (i.e., a single value across all classes). Using importance=TRUE, the algorithm also provides the MeanDecreaseAccuracy value across all classes (1 dimension: probe), and for each individual class versus all others (2 dimensions: probe x class). A main difference of our approach is that probe usages are extracted for all pairwise class combinations (3 dimensions: probe x class x class). We believe

this is very useful for identifying probes that distinguish related classes, as 2D and 1D measures tend to be dominated by class comparisons that are very distinct (e.g., IDH-mutant gliomas vs. all others). If needed, probe usages can be reduced to two or one dimension (e.g., in Fig. 2a and 2b, respectively). Calculating the usage in a pairwise manner reflects the nature of the random forest algorithm, in which binary decisions are made at each splitting node. Probe usage values are therefore a very direct measure of probe importance. In addition, probe usage values can be extracted from already trained classifiers (even from those that were trained without importance=TRUE). To our understanding, the scikit-learn implementation in Python provides feature importances similar to the MeanDecreaseGini and MeanDecreaseAccuracy values, but does not provide pairwise probe usages. While our code is specific to the randomForest package in R, we believe that the approach is applicable to other implementations of the random forest algorithm. We now include a more detailed description of the methodology for extracting probe usages in the Methods.

Following feature extraction, the pipeline involves tallying probe usage in decision trees aggregated by class combinations for each probe. The top 10,000 probes are extracted, with results indicating their contribution to 61.2% of the probe combinations, thereby explaining the classifier's decisions. My primary concern here lies in the stability of the proposed method. The 10,000 most insightful probes are derived from a single classifier trained with specific parameters (not delineated in this paper but retrievable from the original study). The paper should assess the stability of the extracted probes across multiple trainings, possibly employing different parameters. A parameter study of the main random forest hyperparameters (initially the number of estimators, but also considering maximum depth, maximum number of features in splits, etc.) would be beneficial. Given the inherent randomness in the random forest classifier at various levels (e.g., feature and sample selection for training different decision trees), it is crucial, in my opinion, to statistically evaluate whether the 10,000 primary probes remain consistent or are highly dependent on the random choices made by the machine learning algorithms.

We thank the reviewer for this important comment. We would like to clarify that the 10,000 probes of the inner classifier were selected based on the per-class MeanDecreaseAccuracy value of the outer classifier that was trained on all 428,799 probes (as in Capper et al., 2018). In our present study, we have applied our new explainable AI approach to the outer and the inner classifier of the original study. The reported probe contribution of 61.2% (Fig. 2a) reflects the newly calculated probe usage of the top 10,000 probes of the outer classifier. These probes are similar but not identical to the 10,000 probes selected for the inner classifier in the original study (overlap: 6,913 probes). Subsequent analyses (Fig. 3 and Fig. 4) were performed by assessing the probe usage of the inner classifier. We have edited the manuscript to clarify this point.

We agree with the reviewer that it is crucial to evaluate the selection of the 10,000 primary probes. In response to the reviewer’s comment, we have now re-trained the outer random forest classifier to assess the stability of probe selection for different numbers of trained trees per forest. Looking at the per-class MeanDecreaseAccuracy value of the ETMR class as an example, we found that two separate forests each consisting of 1,000 trees (generated with

different seeds) yielded a low correlation coefficient of 0.662 (Fig. R1a,b). Forests consisting of 10,000 trees each (as used in the original study) yielded a much higher correlation of 0.955. Doubling the number of trees to 20,000 resulted in a moderate increase in correlation to 0.977.

To select the 10,000 probes for the inner classifier, the original study ranked the MeanDecreaseAccuracy values (high value = low rank) within every class to select the probes with the lowest rank across all classes. We evaluated the stability of probe selection across all classes for different numbers of trained trees. Between two separate forests consisting of 10,000 trees, we observed a Jaccard index of 0.779 (77.9% of the 10,000 selected probes were identical between forests; Fig. R1c). Doubling the number of trained trees to 20,000 resulted in a Jaccard index of 0.841. We also investigated the MeanDecreaseAccuracy value across all classes. We found that values were strongly correlated between two forests consisting of 10,000 trees (r=0.954). Considering that training a single tree on this dataset takes ~90 seconds, we conclude that 10,000 trees are a reasonable choice for reproducible probe selection.

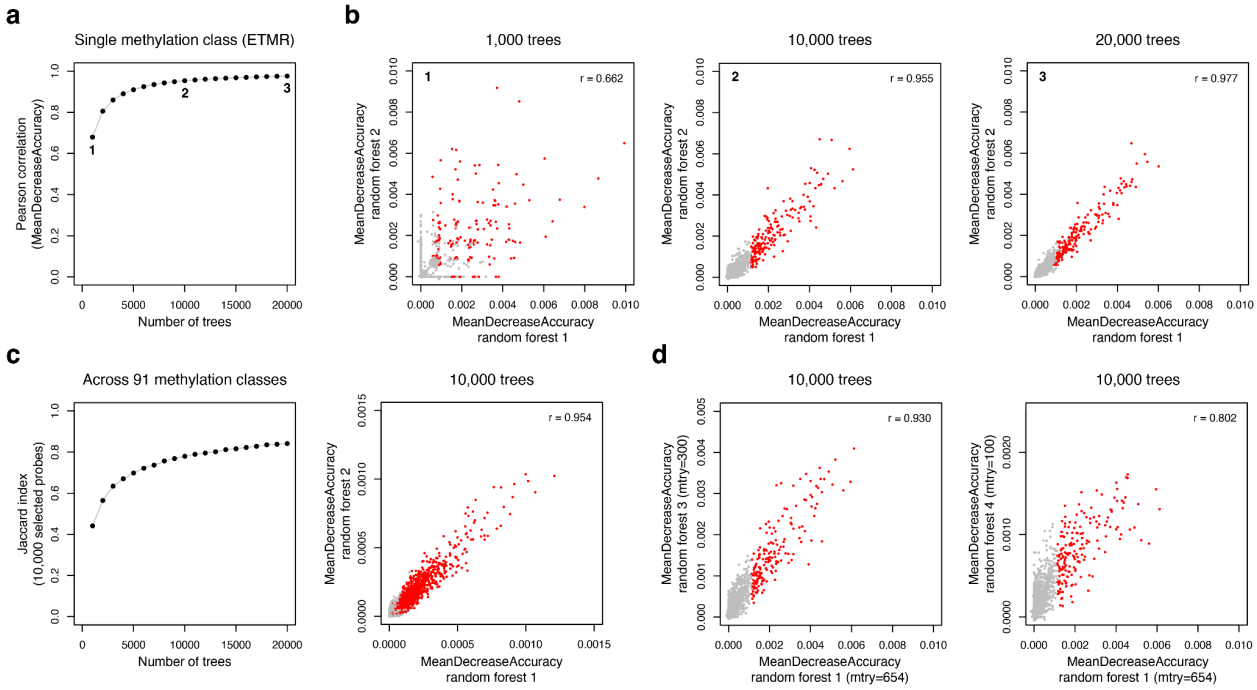

Figure R1 | Robustness of probe selection. Red color in panels **b** and **d** indicate probes selected for the ETMR methylation class in random forest 1 (lowest rank in ETMR and among the 10,000 selected probes). Red color in panel **c** indicates the 10,000 selected probes.

In response to the reviewer's suggestion, we also assessed the effect of the maximum number of considered features at each split ("mtry" parameter) on the stability of probe selection. We considered mtry values of 654 (square-root of 428,799, the default value), 300, and 100. Comparing forests of 10,000 trees, we found that the MeanDecreaseAccuracy values for the ETMR class were correlated, but to a lesser degree than in the previous analysis (r=0.930 and r=0.802 for mtry values of 654 vs. 300 and 654 vs. 100, respectively; Fig. R1d). As fewer values are considered at each split, MeanDecreaseAccuracy values are less accurate, which could explain these results.

While these 10,000 probes influence the decision trees, it remains unclear to what extent they determine the differentiation between various classes. Training the random forest solely with these 10,000 probes to evaluate their impact on classification metrics (e.g., accuracy, recall) would be informative in assessing their predictive capabilities.

We believe the reviewer is referring to the inner classifier, which has been trained using the 10,000 probes selected from the outer classifier as described in the original study. Classification metrics and predictive capabilities of the inner classifier were evaluated previously (Capper et al., 2018 and Maros et al., 2020). For example, the estimated error rate was 4.28% between all classes and 1.14% between clinically relevant groupings. The present study includes an analysis of the probe usages of the inner classifier and how they distinguish between various classes (presented in Fig. 3).
We have edited the manuscript to clarify this point: “For further analyses we focused on the inner RF model that is based on 10,000 probes selected from the outer RF model (Capper) and is used for generating predictions with the Heidelberg brain tumor classifier. Using this model, we extracted and aggregated pairwise probe usage values similarly as for the outer classier.”

Probe usage is then assessed to cluster the tumor classes using unsupervised learning techniques. The clustering, based on K-nearest neighbors to construct Shared Nearest Neighbor (SNN) graphs and cluster identification using the Louvain algorithm, results in 88 distinct clusters, appearing well-separated in the t-SNE representation provided in Figure 3. However, the accompanying text for the figure is somewhat vague in describing the obtained clusters relative to the tumor classes: "Interestingly, we found that the majority of clusters were associated with a single tumor class, indicating high class specificity of most probes." Clarification is needed regarding the precise definition of "the majority of clusters," as a numerical estimation would facilitate understanding the method's coverage.

We now describe the relation between clusters of probes with similar usage to tumor classes more precisely and have edited the manuscript to the following: “The number of probes per cluster was highly variable and ranged from 2 to 211 (median of 117; Extended Data Fig. 4b). Interestingly, we found that the majority of clusters were associated with a single tumor class, indicating high class specificity of most selected probes (Fig. 3b). Specifically, 71 of 88 clusters (80.7%) were associated with a specific tumor class. For instance, to classify ETMR samples, the model predominantly employed probes belonging to cluster 27 (hypermethylated, n=136 probes) and cluster 78 (hypomethylated, n=44 probes; Fig. 3c,d). We also identified 17 clusters (19.3%) that were associated with 2 or 3 tumor classes. Among these were clusters associated with GBM, IDH-mutant glioma, and PITAD methylation class families of closely related tumor classes (Fig. 3b). For example, probes belonging to cluster 1 (hypermethylated, n=211) were associated with classes O IDH, A IDH, and A IDH HG (Fig. 3e,f).”

In summary, while the paper is highly intriguing, the lack of evaluation of the stability of the proposed methods and the absence of sensitivity analysis of hyperparameters render this work challenging to gauge in terms of the reliability of the obtained 10,000 primary probes and

subsequent analyses. Additionally, while the web application code is available, it is regrettable that the code for the machine learning pipeline is not accessible for review.

We thank the reviewer for their encouraging words and raising the important point of the stability of our proposed method of extracting probe usages. We have now addressed this point by extracting probe usages on a separately trained outer random forest classifier (428,799 features, 10,000 trees, mtry=654). The cumulative probe usage displayed an almost identical trend compared to the original outer classifier. The top 10,000 and 25,000 probes contributed to 64% and 81% of the total probe usage, respectively. We then directly compared per-probe usages of the two models. Overall probe usages showed a high correlation (r=0.975) and, taking the ETMR tumor class as an example, similar results were observed at the per-class level (r=0.972). These results demonstrate a high stability of extracted probe usages as a metric employed in our explainable AI approach. These new analyses are now included in the Results section and in Extended Data Fig. 2c-f. For the code availability please see our response below.

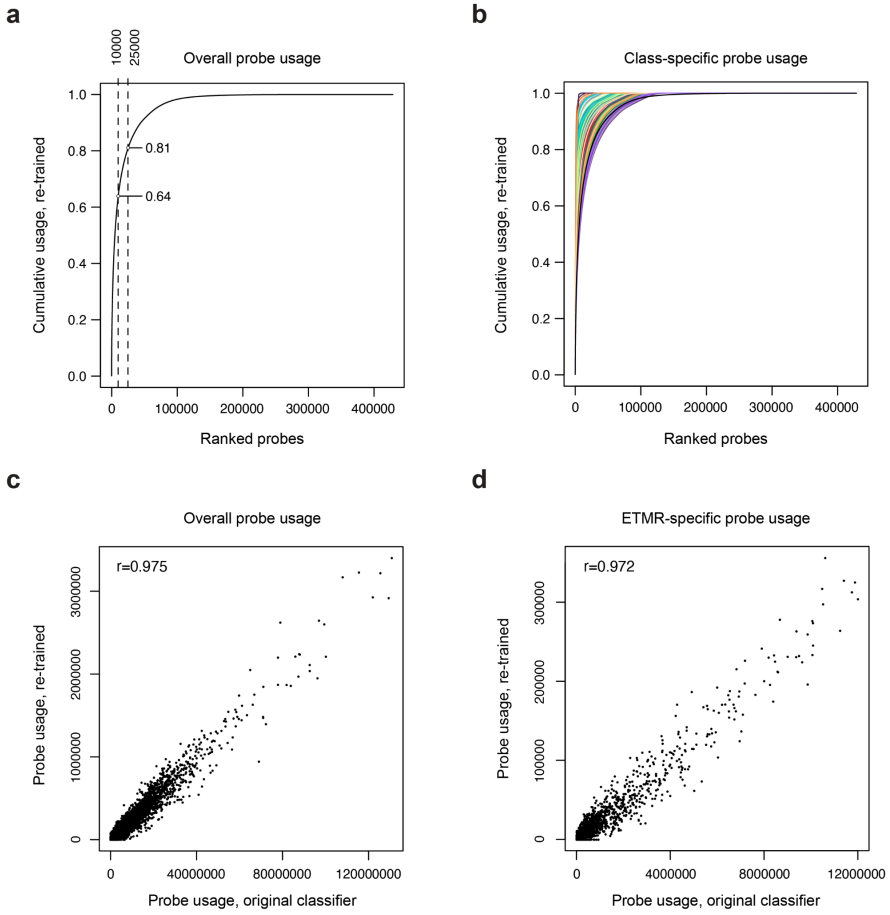

Figure R2 | Stability of probe usage values. a Plot shows the overall probe usage of a re-trained independent RF model with probes ranked from the most to least used. Vertical lines indicate the fraction of the total usage of the first 10,000 and 25,000 probes. **b** Plot shows the cumulative probe usage of an independent RF model for each tumor class. The probes are ranked by their usage. **c** Scatterplot shows the overall probe usages of the re-trained RF model against the original RF model. **d** Same as **c** for the ETMR tumor class.

Reviewer #2 (Remarks on code availability):

Only the code of the webapp is share. The provided coded is well commented and organized. An example is provided with a simple classifier data to use the app and it seems (did not verify though) to connect a new classifier.

The code of the main machine learning analysis is not provided and would be helpful to fully review and evaluate the viability of the approach.

We thank the reviewer for taking the time to investigate the code of our explainable AI approach. We would like to highlight that the code for the main machine learning analysis has been shared previously (https://github.com/mwsill/mnp_training) following the publication of the original study (Capper et al., 2018).

Reviewer #3 (Remarks to the Author): clinical expertise in CNS tumour diagnosis

The authors take advantage of a machine learning classifier that allows for classification of brain tumors based on genome-wide DNA methylation profiles. They attempt to describe an interpretable framework to explain the classifier's decisions. They show that functional genomic regions are employed to distinguish between tumor classes and describe the nature of the methylation sites that are the basis of the classification. Overall, the motivation of the work is strong and the methods are robust. I believe a couple of issues if addressed would improve the impact of this work.

We thank the reviewer for their positive remarks on our work and for their helpful suggestions which we have addressed below and in the revised version of our manuscript.

1. The authors use the original 2018 ("version 11") classifier with 91 classes as the basis for their work. The filed has moved to the next version ("version 12") classifier, that has more than double the number of classes. Given that, their analysis seems outdated and is less relevant, since it applies to a classifier that is no longer the gold standard. The authors should apply their pipeline to the version 12 classifier so that conveys relevance to the current understanding of brain tumor methylation-based classification.

We agree that applying our explainable AI approach to the version 12 CNS tumor classifier would be of interest. However, while the most recent version of the classifier is accessible via the Heidelberg web portal, the updated molecular reference cohort and the trained random forest model have not been made public at this time and are outside the focus of the present study. We believe that the version 11 classifier remains relevant in diagnostic routines and in research and that a better understanding of its epigenetic underpinnings is of interest to the community. A recent example is the deep learning approach by Vermeulen and colleagues (Nature, 2023) that couples the version 11 reference to low-pass nanopore sequencing for intraoperative tumor classification. We would also like to highlight that our study provides a

general proof-of-concept of an explainable AI approach that can be applied to any existing or future random forest-based tumor classifier (e.g., the sarcoma classifier, discussed below).

2. It would be important to know if the principles learned from brain tumor classification are unique to brain tumors, or instead whether they generally apply to methylation cancer classifiers. Some of the authors of this work were involved in the sarcoma classifier. It would be of interest to know the extent to which main principles unveiled in their work from the brain tumor classifier also apply to the classification of sarcomas—this would provide important insights that will contribute to the understanding of what will likely become and ever-expanding number of methylation-based classifiers across many cancers.

We would like to thank the reviewer for this important suggestion. To investigate if the principles learned from the brain tumor classifier are generalizable across other DNA methylation-based tumor classifiers, we applied our explainable AI approach to a random forest classifier of the recently published sarcoma cohort (Koelsche et al., Nature Communications 2021). This dataset contains 1,077 samples belonging to 65 methylation classes. Using our approach, we calculated the pairwise probe usage for the 10,000 probes of the inner random forest classifier. Results are now included in the Discussion, in Extended Data Fig. 6, and in Fig. R3 below. In accordance with results obtained from the CNS tumor classifier, we observed that few probes are employed more frequently than others. Looking at the class-specific probe usage, some classes showed a much more pronounced usage inequality, i.e. few probes that are selected most of the time. We then performed dimensionality reduction of the 10,000 probes based on their pairwise usage between all 65 classes. The analysis revealed 73 clusters of probes with similar usage patterns, revealing a high redundancy of probes. These results demonstrate that the main principles unveiled from the CNS tumor classifier also apply to other DNA methylation-based tumor classifiers.

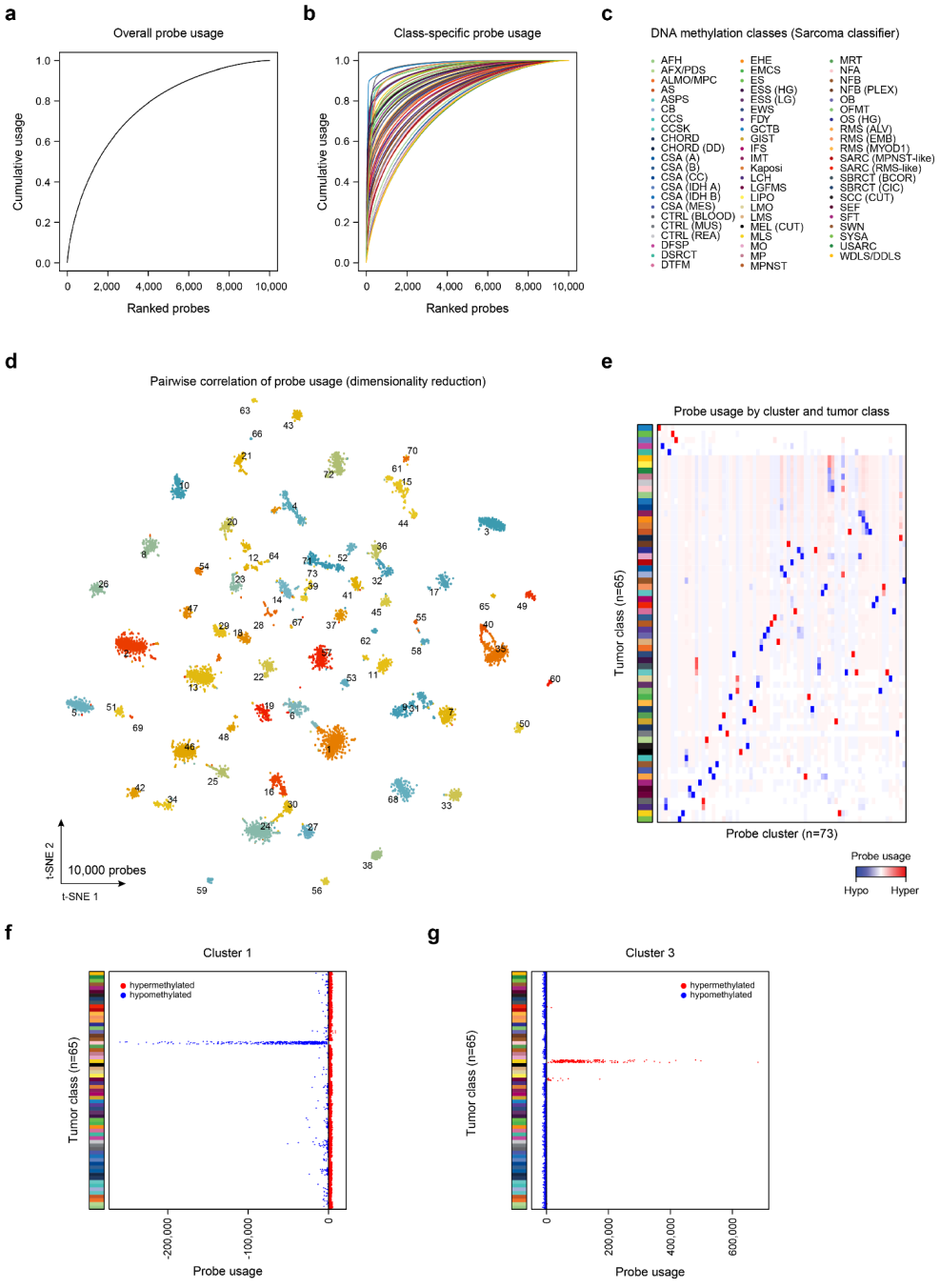

Figure R3 | XAI of a sarcoma classifier. a Plot shows the overall probe usage of a sarcoma RF model with probes ranked from the most to least used. **b** Plot shows the cumulative probe usage for each tumor class. The probes are ranked by their usage. **c** Tumor class legend of the sarcoma cohort. **d** t-SNE dimensionality reduction and unsupervised clustering of 10,000 selected probes according to their usage (Pearson's distance). A total of 73 clusters are identified. Individual clusters of probes are numbered. **e** Heatmap shows the total probe usage by class and by cluster. Blue color indicates clusters in which probes are predominantly hypomethylated in a given tumor class, red color indicates clusters which are predominantly hypermethylated. Clusters (columns) and tumor classes (rows) are ordered by hierarchical clustering. **f** Scatterplot showing the usage of probes belonging to cluster 1 for each tumor class. **e** Scatterplot showing the usage of probes belonging to cluster 3.

Reviewer #3 (Remarks to the Author):

I had stated in my original review that the authors apply their analyses and models to an outdated ("Version 11") version of the classifier I had suggested that the authors apply their analyses to the version of the classifier that people are actually using now ("Version 12"). The authors state that use of the Version 12 classifier is "outside the focus of the present study", but I think it actually is within scope. I do believe that applying these analyses to the current version of the classifier is well within the scope of the work, since the readers will want it to be currently applicable and as impactful as possible.

We agree with the reviewer that analyzing the Version 12 classifier would be of interest for our explainable AI study. However, we (the corresponding and the first author) have not been involved in the generation of the Version 12 classifier, and as of yet there has not been a publication describing its construction nor has its reference dataset been publicly released. Therefore we believe that the Version 12 classifier is outside the focus of our explainable AI study, which instead builds on the publicly available Version 11 reference, ensuring reproducibility of our study. We also object to the notion that the Version 11 reference is outdated, as it enables comparisons between all major brain tumor types and subtypes, and is the foundation for other recent work, such as the “Sturgeon” nanopore-based brain tumor classifier (Vermeulen, 2023). We hope the reviewer agrees that our study in its current form bears conceptual novelty and will be of sufficient scientific impact to warrant its publication.

Reviewer #4 (Remarks to the Author):

In my opinion, the points raised by reviewer 1 have been adequately addressed. However, from the brain tumour biology side of things, I am not sure the conceptual advances/novelty of the data is very high. Firstly, I agree with reviewer 3 that it seems a missed opportunity that this work has been carried out on an outdated version of the classifier. Given the emphasis on developing and sharing this approach to enable novel discoveries in the brain tumour field, surely this work should be based on the most recent classifier. Moreover, despite the authors' claims no evidence is provided that novel insights in brain tumour biology can be gained with this approach. The three lines of validation provided (SHPRH, PWWP3A and TBX19) are confirmatory of existing knowledge and while this is an essential first step, I would have expected at least one example of a novel finding to be shown, including its functional validation to substantiate the claim that this approach and the resource shared will indeed be impactful.

We thank the reviewer for their time reviewing our revised manuscript. For their first comment regarding the Version 12 classifier, we would like to point the reviewer to our response to reviewer #3 above. For their second comment, the main objectives of our study were (i) to provide the computational methodology to perform intuitive explainable AI analyses of random forest classifiers and (ii) to provide researchers with a resource to nominate genes of interest for future study and/or to investigate the tumor type specificity and epigenetic regulation of genes a researcher is already interested in. The four example genes highlighted in our manuscript (*SHPRH*, *PWWP3A*/*MUM1*, *TBX19*, *RET*) were first identified by using the interactive web portal, and we then queried external gene expression datasets and performed literature search. While we agree that functional follow-up is required to substantiate their role, we would argue that our results already provide novel insights into their biology (e.g., their tumor type specificity, epigenetic regulation) that will guide future study. This is particularly important for tumor types that are not well studied and in which model systems are limited (e.g., *SHPRH* in ETMR, *PWWP3A* in HGNET with MN1 activation, *RET* in HGNET with BCOR activation). With these considerations, we hope the reviewer agrees that our resource will be of sufficient impact for the research community to warrant its publication.

Reviewer #5 (Remarks to the Author):

The authors have nicely addressed the concerns of the reviewer.

We thank the reviewer for their assessment of the revised manuscript and our response to the previous reviewers' concerns.

Reviewer #6 (Remarks to the Author):

I appreciate that this is already a revised manuscript and I have read the reply to the reviewers by the authors, but I'm a new reviewer brought to review this manuscript and naturally I got some fresh questions for the authors.

Overall this is a very interesting work which makes a clear effort to extract knowledge from the machine learning models rather than just building ML pipelines, evaluating them and report predictive performance, which is commendable. Having said this, the knowledge extraction effort performed in this paper is quite simple, and could be much better.

We thank the reviewer for their effort reviewing our revised manuscript and their helpful suggestions which helped us to improve our work.

For instance, there are methods to infer supervised networks from the structure of interpretable machine learning models trained on omics data (e.g. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-469 https://biodatamining.biomedcentral.com/articles/10.1186/s13040-016-0106-4), these networks are quite different from those presented here as nodes would be probes rather than patients.

Thank you for these suggestions. For our revised version, we applied the FuNeL approach developed by Lazzarini and colleagues to generate a network of DNA methylation probes based on their co-prediction from a rule-based model. We compared results with our previously defined clusters of probes that are based on probe usage in our random forest model (see Figure 3a). We initially applied the FuNeL approach using recommended parameters (100 permutations, 10,000 BioHEL runs), which was not feasible because of the size of our dataset (2,801 samples belonging to 91 classes, 10,000 features/methylation probes). We drastically reduced the number of permutations and runs to 20, obtaining a sparse co-prediction network (Figure R1a).

To interpret the co-prediction network, we analyzed the probe connectivity within and between our previously defined clusters. The results indicate that for probes belonging to some of the clusters, such as cluster 1 (IDH-mutant specific) or cluster 27 (ETMR-specific), the connectivity is much higher within the cluster than to probes belonging to other clusters (Figure R1a), supporting our findings that redundant probes within a cluster are often interchangeably used to classify certain tumor classes.
For other clusters we did not observe the same trend (e.g., cluster 78, ETMR-specific). We asked if increasing the number of permutations and runs would lead to more refined results. However, we observed only minor differences in network connectivity between 20 and 50 permutations and runs (Figure R1b). Further increasing the number of permutations would be very time consuming.

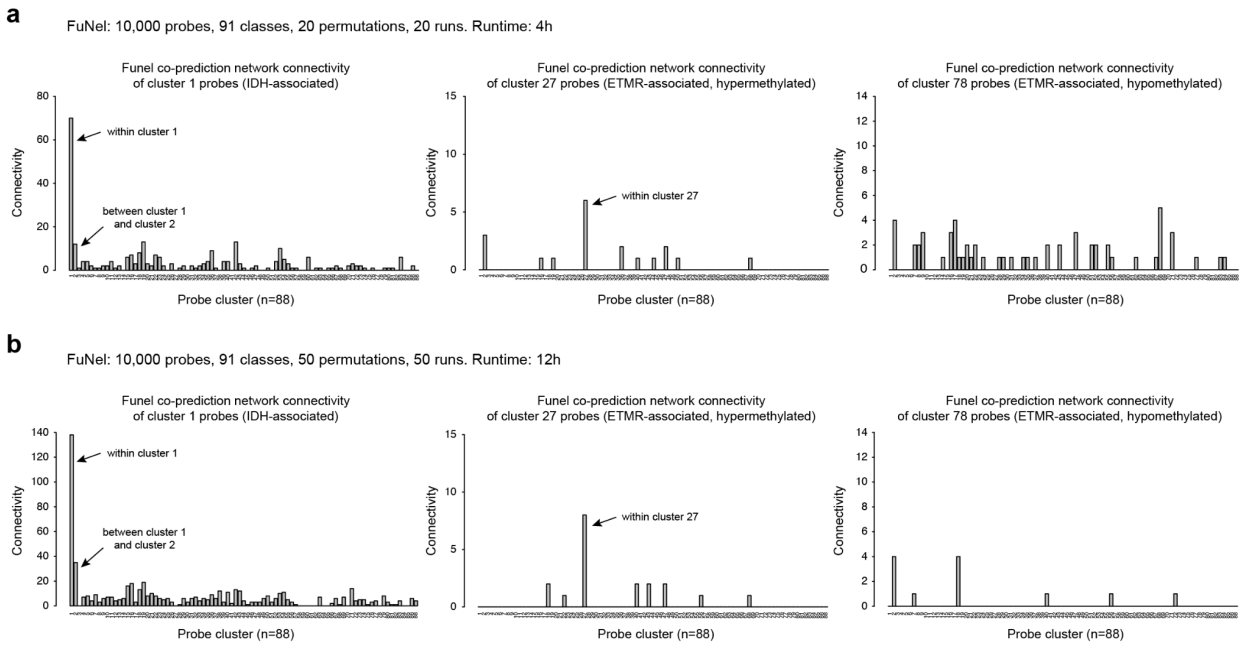

Figure R1: Results of the FuNel pipeline applied to the brain tumor classifier dataset. **a)** Left bar plot shows co-prediction network connectivity of cluster 1 probes (n=211) to other probes, grouped by their cluster association. 20 permutations and 20 runs were used to calculate the co-prediction network. Highest connectivity is observed for probes within cluster 1, indicating that they are often used interchangeably to classify IDH-mutant tumor classes, supporting our previous results. Center and right bar plots show connectivity for cluster 27 (n=136) and cluster 78 probes (n=44). **b)** Bar plots show connectivity for 50 permutations and 50 runs.

We conclude that although there are similarities between the FuNel co-prediction network and our clusters of probes based on their usage, the overlap between the results of the two approaches is not exhaustive. The reason could be that the rule-based model of FuNel is not well suited to distinguish 91 different tumor classes in a 10,000 dimensional feature space. The original publication (Lazzarini et al, BioData Mining 2016) applies the FuNel approach to distinguish between 2 class labels.

While we agree that our approach of calculating probe usage values is quite simple, it is a straightforward approach to quickly extract biological knowledge from already trained random forest models even when dealing with datasets containing a large number of different classes and/or features. We also point out that our approach computes pairwise importance between individual classes, which to our knowledge is not feasible with other tested methods.

Unfortunately, the code for the suggested SNPInterForest approach from the Yoshida et al., 2011 publication is no longer available.

Moreover, there are Explainable AI methods able to extract very nuanced knowledge from machine learning models, much more refined than feature importance estimates. The SHAP method comes to mind as the more popular of such methods. It integrates very nicely with scikit-learn models. I'm not so sure on the R implementation of RF.

We agree with the reviewer that the popular SHAP method is well-suited to generate feature importance estimates. For the revised version, we have trained a random forest model using Python scikit-learn using the same 10,000 features (out-of-bag error of 0.04) and compared SHAP values to the probe usage values defined using our method to assess their similarity.

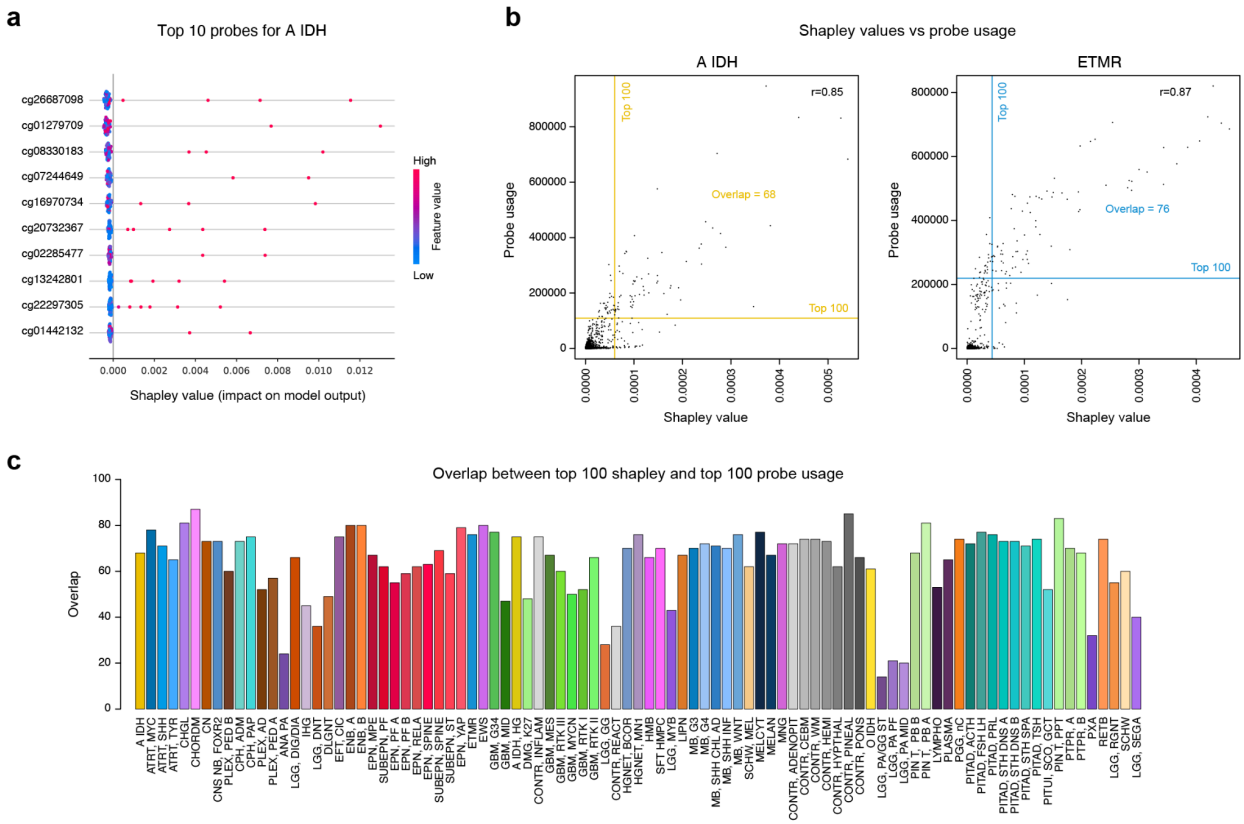

Figure R2: Results of the SHAP approach applied to the brain tumor classifier dataset. **a)** Summary plot of the top 10 probes for the A_IDH class (y-axis). Shapley values are indicated on the x-axis. Results indicate that probes hypermethylated in IDH-mutant classes are frequently selected during training of the random forest model to separate these classes. **b)** Left scatter plot compares A_IDH-specific Shapley values to our previously generated probe usage values. Results indicate a high agreement between the two approaches (Pearson's r=0.85). Overlap between the top 100 probes from each approach are indicated. Right scatter plot similarly compares ETMR-specific values. **c)** Bar plot shows overlap between the top 100 probes using the Shapley approach and our previous probe usage approach for all 91 classes. High concordance can be observed.

We initially aimed at calculating Shapley values for all samples (n=2,801), but this did not lead to any results in a reasonable timeframe due to the size of the dataset. We therefore calculated Shapley values for a single sample in each of the 91 classes (similarly using one sample per class as the background). We found that a relatively small number of probes were associated with highest Shapley importance values for each class, while the majority for probes showed minor contribution (Figure R2a), similar to our previous observations (see Figure 2 in the manuscript).

We then directly compared Shapley values for all probes to our probe usage values. We found that both metrics were highly correlated for most tumor classes. For example, the A_IDH tumor

class showed Pearson's correlation value of 0.85 (Figure R2b left). The ETMR tumor class showed a correlation value of 0.87 (Figure R2b right). We then asked what was the overlap between the top 100 probes using Shapley values and usage for all classes. We found that in 85% of the classes more than 50% of the top 100 probes were shared, with a minimum of 14 and a maximum of 87 across classes (Figure R2c).

We have now included the results of this comparison into the manuscript and in the novel Extended Data Figure 3. We also want to point out that calculating Shapley values is computationally demanding and only a limited set of samples can be used for its calculation in a reasonable amount of time. Conversely, our proposed probe utilizes all samples and its computation is very quick. Additionally, our approach can also be used to calculate pairwise probe importances between individual classes.

Moreover, the use of t-SNE (or UMAP) as a step before clustering omics data is quite controversial, please see https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1011288

It is unclear if the reviewer is referring to the t-SNE analysis presented in Figure 3a (our present study) or Figure 1a (which was initially described in Capper, Nature 2018). However, in neither analysis is the clustering (i.e., the definition of clusters/classes) based on the t-SNE, but merely overlaid as a form of visualization.

From the machine learning methodology i'm surprised that the authors did not consider adding a feature selection step to the pipeline. While I agree that RF scales well to large feature spaces, it can be helped by specialised feature selection algorithms as the literature has shown many times. For instance, Recursive Feature Elimination (of course eliminating a percentage of features in each step rather than one by one) was originally designed with omics data in mind. Adding a feature selection could make the choice of features more robust across repetitions, which based on the results reported in the reply letter, is still not great.

Finally, it's not clear what evaluation methodology was used. How was the data split into training and test sets? Where the 10000 probes selected on the whole dataset or from the traininng partition of the data? (as they should). Where any preprocessing decisions taken based on the training data (as they should) or from the whole dataset? I also struggle to find a clear reporting of predictive performance of these models. If the models are not good at making predictions, any knowledge extraction is futile.

Our study investigates an existing RF model for brain tumor classification that was previously published (Capper, 2018). To generate this RF model, a feature selection step has indeed been performed. The 10,000 most predictive probes were selected from an "outer" random forest classifier, and then used for training an "inner" classifier that showed improved performance. Our present study investigates the probe usage in both the outer (Figure 2) and the inner classifier (Figure 3 and 4).

Classification metrics and predictive capabilities of the original RF model were evaluated in cross-validation and in an independent prospective dataset, achieving high predictive accuracy (Capper, 2018). For example, the estimated error rate was 4.28% between all classes and 1.14% between clinically relevant groupings. Since then, other studies have investigated its application in diagnostic settings (e.g., Sturm, Nature Medicine 2023). Further comparisons to other machine learning algorithms were presented in Maros, Nature Protocols 2020. Data reprocessing steps included basic filtering of putative poor quality probes and correction for the type of input material (FFPE, fresh-frozen), and decisions were made on the training dataset (please also see below).

Reviewer #6 (Remarks on code availability):

Unclear from the code what kind of model evaluation methodology was used, or what preprocessing of data happened on top of the raw data.

We would like to point the reviewer to the original study (Capper, 2018) for details on the preprocessing of the raw data and the model evaluation. In brief, methylation values were generated by scaling intensities to negative and positive control probes. Probes not mapping uniquely to the genome, probes that are associated with common SNPs, or probes that are mapping to the X or Y chromosome were removed (428,799 probes were retained). The original model was evaluated in cross-validation and in a separate prospective cohort of 1,155 diagnostic CNS tumors.