ChemPubSoc Europe

DOI: 10.1002/cbic.201800392

CHEMBIOCHEM Communications

The General Explanation Method with NMR Spectroscopy Enables the Identification of Metabolite Profiles Specific for Normal and Tumor Cell Lines

Klemen Pečnik,[a] Vesna Todorović,[e] Maša Bošnjak,[e] Maja Čemažar,[e] Igor Kononenko,[d] Gregor Serša,[e] and Janez Plavec*[a, b, c]

Machine learning models in metabolomics, despite their great prediction accuracy, are still not widely adopted owing to the lack of an efficient explanation for their predictions. In this study, we propose the use of the general explanation method to explain the predictions of a machine learning model to gain detailed insight into metabolic differences between biological systems. The method was tested on a dataset of ^{1}H NMR spectra acquired on normal lung and mesothelial cell lines and their tumor counterparts. Initially, the random forests and artificial neural network models were applied to the dataset, and excellent prediction accuracy was achieved. The predictions of the models were explained with the general explanation method, which enabled identification of discriminating metabolic concentration differences between individual cell lines and enabled the construction of their specific metabolic concentration profiles. This intuitive and robust method holds great promise for in-depth understanding of the mechanisms that underline phenotypes as well as for biomarker discovery in complex diseases.

Searching for metabolic biomarkers that would discriminate sample classes (i.e., cell types, diseases states, and drug effects) and provide better insight into metabolic mechanisms, response to treatment, and early diagnosis is an active area in metabolomics research.[1–4] Identifying and quantifying low-molecular-weight metabolites in biological samples is based on a variety of spectroscopic techniques, including NMR spectroscopy.[5] Finding potential biomarkers by using the metabolic NMR fingerprints of biological samples requires rigorous data analysis.[6] Prior steps consist in preprocessing of the NMR spectra and their segmentation into small regions called bins.[7] Binned spectral regions are viewed as a set of "features" with their respective "feature values" (i.e., integrated areas under resonance signals in binned regions). A data matrix consisting of features and their corresponding feature values for a large number of samples represents a sophisticated and highly convoluted dataset. Methods such as principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA) are predominantly used to analyze such datasets with the goal to discriminate between sample classes and to uncover the most important features.[8,9] Remarkably, machine learning models such as random forests (RF), artificial neural networks (ANNs), and support vector machines achieve great prediction accuracy, but their prediction processes are obscured and do not reveal the feature values that are used for predictions.[10–12] In exploring the properties of biological specimens, it is crucial to be able to explain predictions of different machine learning models and to extract feature values that enable the discrimination of sample classes. Some of the machine learning models utilize model-specific explanation methods to explain their predictions. Unfortunately, machine learning models that have the potential to achieve the best prediction accuracy do not enable any explanation of their predictions.[12] By using the general explanation method (GEM),[13] predictions made by machine learning models can be efficiently and intuitively explained.

GEM is a sensitivity analysis-based method that is used to explain prediction models and can be applied to any type of classification or regression model. Its advantage over existing explanation methods is that all subsets of the input features are perturbed, so interactions and redundancies between features are taken into account. GEM explains the prediction of a machine learning model as a list of contributions of individual feature values. The importance of a given feature value for a prediction is expressed as a "contribution value". Feature values with high contribution values indicate a large influence on the model's prediction (note that the contribution value can be either positive, supporting the prediction, or negative). GEM was previously tested on different machine learning models and was compared with existing explanation methods;

[a] K. Pečnik, Prof. Dr. J. Plavec
Slovenian NMR Centre, National Institute of Chemistry
Hajdrihova 19, SI-1000 Ljubljana (Slovenia)
E-mail: janez.plavec@ki.si

[b] Prof. Dr. J. Plavec
EN-FIST Centre of Excellence
Trg OF 13, 1000 Ljubljana (Slovenia)

[c] Prof. Dr. J. Plavec
Faculty of Chemistry and Chemical Technology, University of Ljubljana
Večna pot 113, 1000 Ljubljana (Slovenia)

[d] Prof. Dr. I. Kononenko
Faculty of Computer and Information Science, University of Ljubljana
Večna pot 113, 1001 Ljubljana (Slovenia)

[e] Dr. V. Todorović, Dr. M. Bošnjak, Prof. Dr. M. Čemažar, Prof. Dr. G. Serša
Institute of Oncology
Zaloška cesta 2, 1000 Ljubljana (Slovenia)

Supporting Information and the ORCID identification numbers for the authors of this article can be found under https://doi.org/10.1002/cbic.201800392.

it was shown that its intuitive explanation of models' predictions improved the user's understanding of the models.[12–14]

Various immortalized cell lines are widely used as models of more complex biological systems. Gaining insight into their metabolic differences is essential for drug development and for the prediction of clinical response to treatment.[15,16] In the current study, five cell lines that differ not only in their status (tumor vs. normal) but also in their morphology and tissue of origin (epithelial and fibroblast cells from lung and mesothelium) were specifically selected (Table 1). The initial step involved "training" the RF and ANN models on a dataset of 63 ^{1}H NMR spectra acquired on normal and tumor cell lines. In the following steps, we set out to investigate the ability of the GEM to uncover the most important feature values that the RF and ANN models use for prediction with the goals of identifying discriminating concentration differences of metabolites amongst the five cell lines and constructing their specific metabolic concentration profiles.

Table 1. Overview of cell lines.

	A549	WI-38	MeT-5A	MSTO-211H	NCI-H2052
cell type	epithelial	fibroblast	epithelial	fibroblast	epithelial
disease	carcinoma	normal	normal	mesothelioma	mesothelioma
no. of samples	23	23	8	4	5
sample nos.	1–23	24–46	47–54	55–58	59–63

Intelligent bucketing was performed in the spectral region between $\delta = 0.1$ and 9.5 ppm on all ^{1}H NMR spectra, which reduced their dimensionality to 235 binned regions, termed features. Feature values (integrated areas of individual features) were normalized to a "constant sum" equal to 100 and were organized as a data matrix. To acquire features with distinct feature values that discriminate cell lines, the dataset was initially analyzed with the RF and ANN models, which both achieved prediction accuracy of 95% as tested with the "leave one out" cross validation method (Tables S1 and S2 in the Supporting Information). High prediction accuracy demonstrates the ability of the RF and ANN models to learn successfully of feature values that discriminate the five cell lines. The GEM was applied to calculate the contribution values of the individual feature values for each sample. Contribution values were then averaged across samples from the same cell line type to highlight the most important features (Tables S3 and S4).

For the RF model, features were considered important if their average contribution values were higher than 0.02 for all cell lines. In the case of the ANN model, the average contribution values of the features that were considered important were higher than 0.02 for the A549 cell line, 0.03 for the WI-38 cell line, 0.05 for the MeT-5A and MSTO-211H cell lines, and 0.04 for the NCI-H2052 cell line. Using both models, 25 important features were identified and are shown in the ^{1}H NMR spectrum of one of the WI-38 samples as an example (Figure 1).

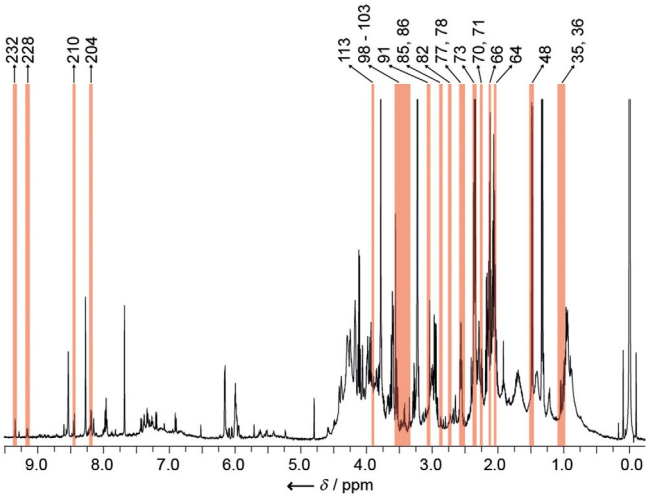

Figure 1. Representative ^{1}H NMR spectrum of the WI-38 cell line sample (index 32). Orange bars represent 25 important features with the highest average contributions indicated with numbers at the top of the spectrum.

The RF and ANN models each found 15 important features, of which features 35, 70, 78, 82, and 101 were important to both (Figure 2). For the RF model, the A549 cell line exhibited seven important features, whereas the WI-38 cell line exhibited four, the MeT-5A cell line exhibited three, the NCI-H2052 cell

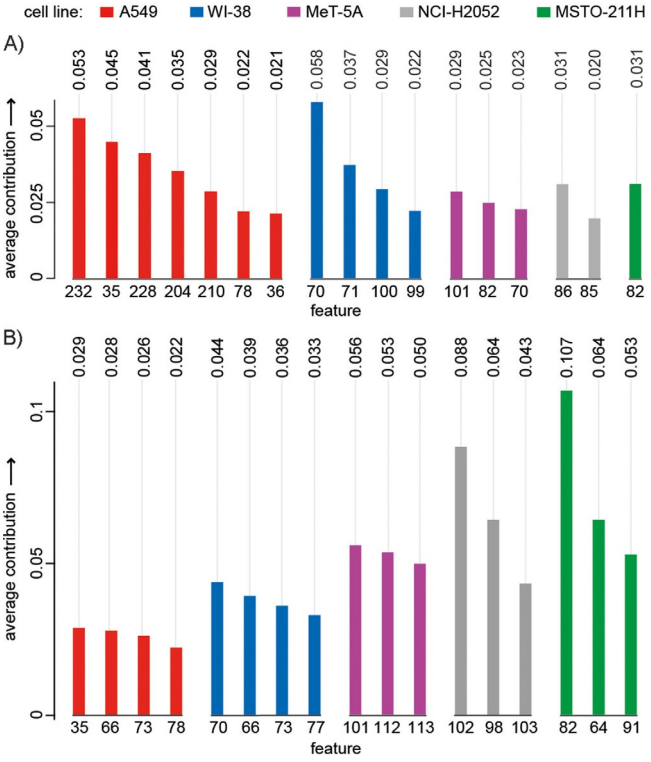

Figure 2. Average contribution values, calculated by the GEM, of important features for the A) RF and B) ANN models for all cell line types (indicated on top). Average contribution values are presented numerically at the top of the bars.

line exhibited two, and the MSTO-211H cell line exhibited one important feature (Figure 2A). Of the 15 features, 13 were specific for individual cell lines. Feature 70 was common to the normal WI-38 and MeT-5A cell lines, and interestingly, its average contribution value was higher for the WI-38 cell line than for the MeT-5A cell line (Figure 2A). Feature 82 was common to the normal MeT-5A cell line and to the tumor MSTO-211H cell line and exhibited comparable average contribution values (Figure 2A). For the ANN model, the A549 and WI-38 cell lines exhibited four important features, whereas MeT-5A, NCI-H2052, and MSTO-211H each exhibited three important features (Figure 2B). Of the 15 features, 13 were specific for individual cell lines, whereas features 66 and 73 were common to the tumor A549 cell line and the normal WI-38 cell line (Figure 2B).

Features with the highest average contribution values for the RF model exhibited distinct feature values that discriminated respective cell lines (Figure 3). The feature values of feature 35 were the highest for the A549 cell line. On the other hand, for the A549 cell line, the feature values of features 228 and 232 were the lowest (Figure 3). Noteworthy, the feature values of the above features clearly discriminated the A549 cell line from the other cell lines. The feature values of features 36, 78, 204 and 210 were less distinct, as some of them overlapped with the feature values of the other cell lines.

In the case of the WI-38 cell line, the feature values of feature 70 were distinctly higher than those of the other cell lines, which thus enabled clear discrimination. However, some feature values of features 71, 99, and 100 for the WI-38 cell line overlapped with the feature values of the other cell lines, which made discrimination less straightforward (Figure 3).

We observed that the feature values of features 70, 82, and 101 for the Met-5A cell line were passage number dependent. For example, samples 53 and 54, collected from the same passage, exhibited distinct feature values in feature 82 and overlapped in feature 101. On the other hand, samples 49, 50, 51, and 52, collected from the same passage, exhibited distinct feature values in feature 101 and overlapped in feature 82 (Figure 3). However, a combination of features 70, 82, and 101, which the RF model utilizes to make predictions, unambiguously discriminated the Met-5A cell line from the other cell lines.

The tumor NCI-H2052 cell line could be discriminated from the other cell lines by the fact that the feature values of features 85 and 86 were clearly higher for the NCI-H2052 cell line than for the other cell lines (Figure 3). Three out of four MSTO-211H cell line samples collected from the same passage exhibited feature values of feature 82 that were clearly higher than those of the other cell lines. Sample 55 collected from a different passage exhibited a lower feature value than the other three MSTO-211H cell line samples (Figure 3) and was the only sample wrongly predicted by the RF model in the studied dataset.

Features with the highest average contribution values for the ANN model also exhibited distinct feature values that discriminated the respective cell lines (Figure 4). The feature values of feature 35 were distinctly higher for the A549 cell line than for the other cell lines. The feature values of features 66, 73, and 78 for the A549 cell line were less distinct, as some of them overlapped with the feature values of the other cell lines (Figure 4).

The feature values of feature 70 were distinctly higher in the case of the WI-38 cell line than for the other cell lines, whereas

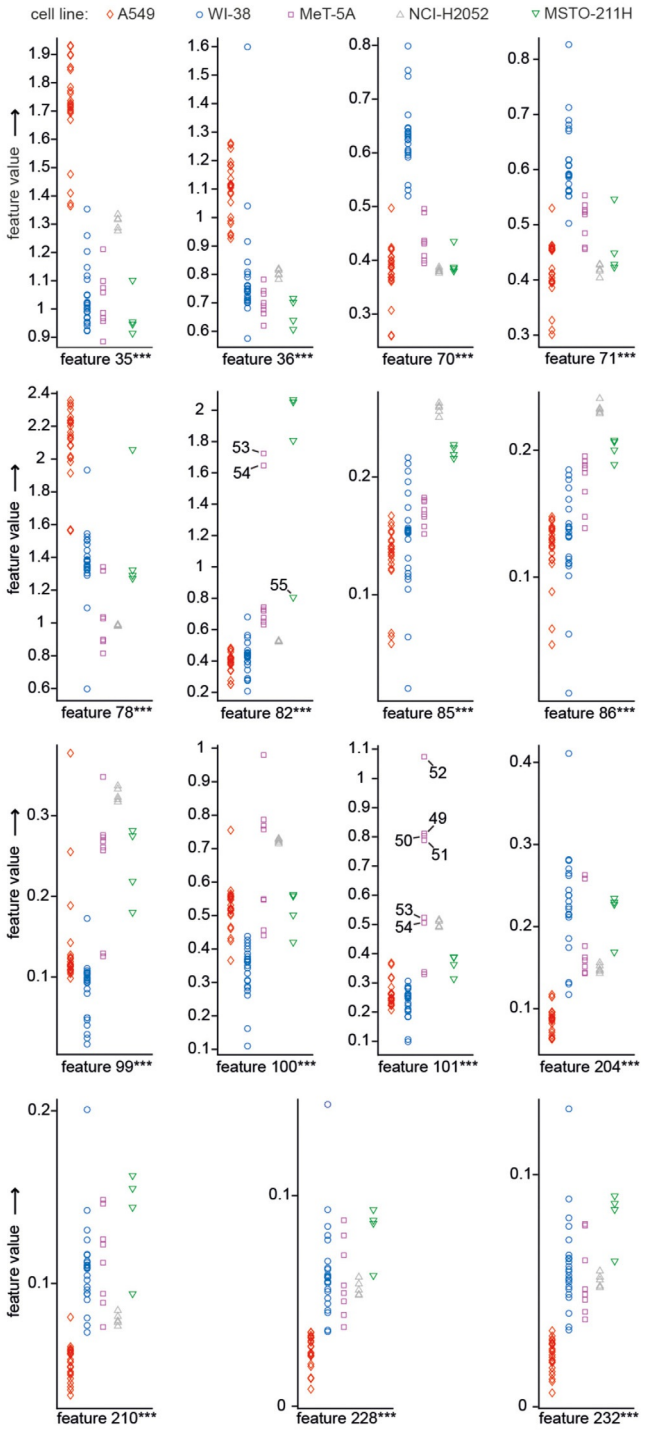

Figure 3. Feature values of the 15 important features for the RF model for all samples. Cell line types are indicated at the top. Plots for features 82 and 101 contain sample indexes indicated with numbers. *** $p < 0.001$, one-way ANOVA.

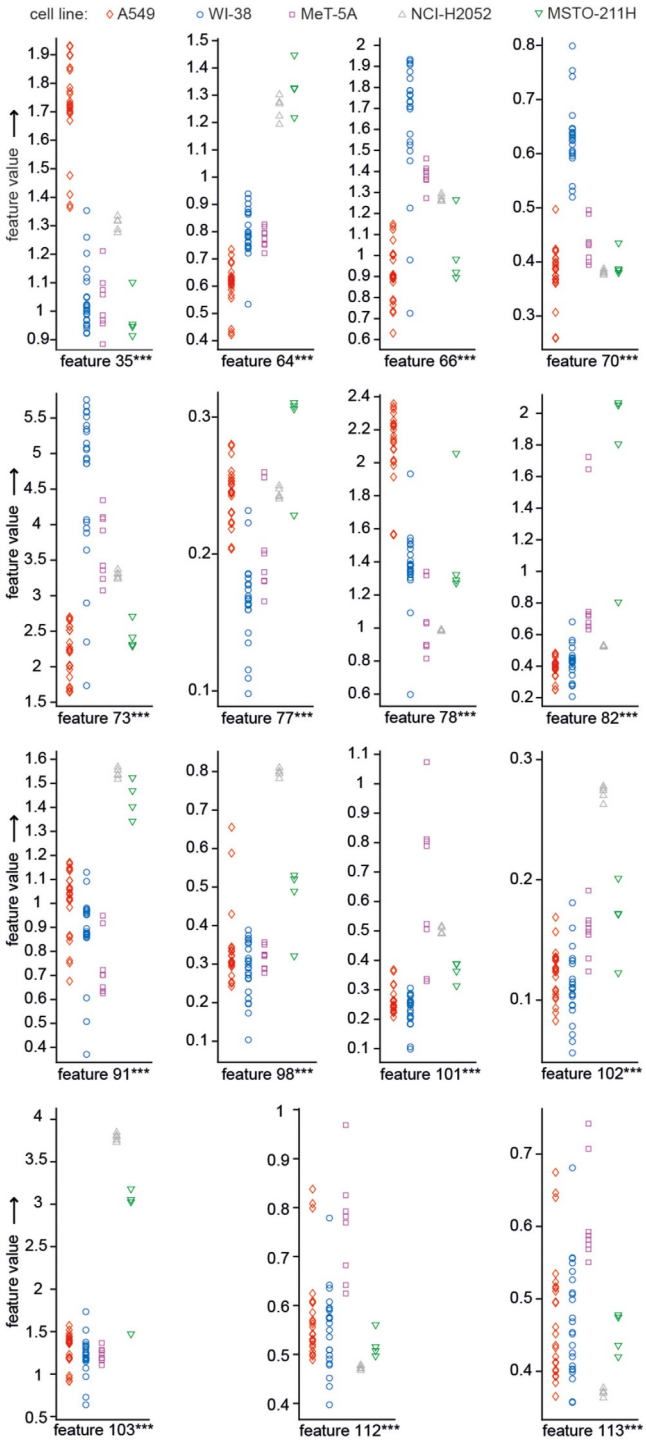

Figure 4. Feature values of 15 important features for the ANN model for all samples. Cell line types are indicated at the top. ***$p < 0.001$, one-way ANOVA.

some of the feature values of features 66, 73, and 77 overlapped with the feature values of the other cell lines (Figure 4).

The feature values of features 101, 112, and 113 for the Met-5A cell line overlapped with the feature values of the other cell lines. However, a combination of features 101, 112, and 113 was able to discriminate the Met-5A cell line from the other cell lines (Figure 4).

Clearly higher feature values of features 98, 102, and 103 unambiguously discriminated the tumor NCI-H2052 cell line from the other cell lines (Figure 4).

All four MSTO-211H cell line samples exhibited clearly higher feature values of features 64 and 91 than the A549, WI-38 and Met-5A cell lines. The feature values of feature 82 for three out of the four MSTO-211H cell line samples collected from the same passage were clearly higher than the feature values of the other cell lines (Figure 4).

The results of the GEM show that there are differences and similarities in interpretations of the dataset by the RF and ANN models. In the case of the A549 cell line, feature 35 exhibits a high average contribution value for both models, whereas features 228 and 232 exhibit high average contribution values only for the RF model. All three features unambiguously discriminate the A549 cell line from the other cell lines. Interestingly, features 228 and 232 are not amongst the important features for the ANN model. For the WI-38 cell line predictions, feature 70 exhibits a high average contribution value for both models. However, features 66, 73, and 78 exhibit high average contribution values for the ANN model, and features 71, 99, and 100 exhibit high average contribution values for the RF model. Interestingly, only feature 70 unambiguously discriminates the WI-38 cell line from the other cell lines and is important to both models. In the case of the Met-5A cell line, feature 101 exhibits a high average contribution value for both models, whereas features 82 and 70 exhibit high average contribution values for the RF model, and features 112 and 113 exhibit high average contribution values for the ANN model. The feature values of features 85, 86, 98, 102, and 103 all unambiguously discriminate the NCI-H2052 cell line from the other cell lines; however, only features 85 and 86 are important for the RF model, whereas only features 98, 102, and 103 are important for the ANN model. In the case of the MSTO-211H cell line, feature 82 exhibits a high average contribution value for both models. Moreover, features 64 and 91 exhibit high average contribution values only for the ANN model.

The features with the highest average contribution values for the RF and ANN models exhibit meaningful feature value differences that discriminate the five cell lines. Their respective signals in the NMR spectra were assigned and used to identify valine (features 35, 36, and 71), isoleucine (feature 35), glutamic acid (features 66 and 73), aminoadipic acid (features 70 and 71), glutathione (features 77 and 78), asparagine (features 85 and 86), creatine (feature 91), proline (feature 100), uridine diphosphate glucose (UDPG, features 101,103, 112, and 113), myoinositol (feature 103), choline (feature 103), aspartic acid (features 112 and 113), and NAD (features 204, 210, 228, and 232) with the use of 2D ^{13}C HSQC, ^{13}C HMBC, and TOCSY spectra (Figures S1–S13).

Once the metabolites corresponding to important features were identified, the metabolite concentration profiles specific for the individual cell lines, as learned by the RF and ANN models, could be constructed (Figure 5). Such specific metabolic concentration profiles of the individual cell lines enabled their further detailed analysis. For example, the tumor cell line A549 was discriminated from the other cell lines by the in-

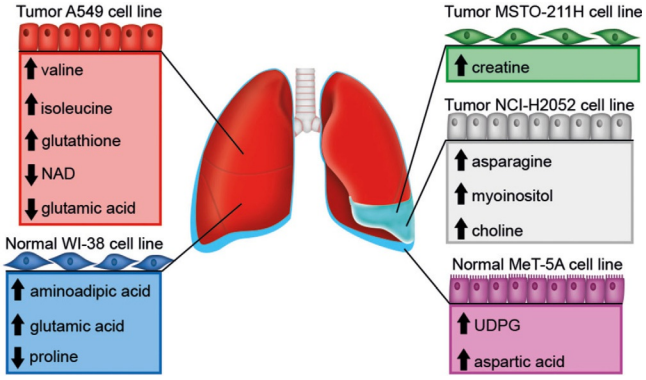

Figure 5. Changes in concentrations of the metabolites that discriminate cell lines according to the RF and ANN models. Upward-pointing arrows indicate an increased concentration of a metabolite, whereas downward-pointing arrows indicate a decreased concentration. Blue line around lungs represents mesothelium, and the cyan part on the right lung represents pleural infusion.

creased concentration of valine, isoleucine, and glutathione and decreased concentration of NAD and glutamic acid (Figure 5). Valine and isoleucine are together with leucine collectively known as the branched-chain amino acids (BCAAs), whereas NAD is one of the five coenzymes involved in the formation of branched-chain α-keto acids (BCKAs). The metabolic concentration profile of the A549 cell line is in accordance with the current understanding of the metabolism of the non-small lung carcinoma (NSCLC) cells, for which overexpression of branched-chain aminotransferase 1 (BCAT1) results in increased intracellular concentrations of BCAAs through the amination of BCKAs.[17–19] In support, both glutamic acid and glutathione have active roles in the proliferation processes of the A549 cell line, and an increased concentration of glutathione has also been observed in tumor A549 cells.[20,21] Additionally, the specific metabolic concentration profiles of the cell lines that were ascertained with the use of the GEM additionally revealed metabolites whose roles in cell metabolism are not yet understood fully (Figure 5). An example is the identification of an increased concentration of aminoadipic acid in the normal WI38 cell line, which has not been described in the literature so far.

The benefits of analyzing the dataset with different machine learning models are especially evident for the A549 and NCI-H2052 cell lines. In the case of the A549 cell line, the RF and ANN models uncovered increased concentrations of valine, isoleucine, and glutathione. However, an increased concentration of NAD was uncovered only by the RF model, and a decreased concentration of glutamic acid was uncovered only by the ANN model. In the case of the NCI-H2052 cell line, an increased concentration of asparagine was uncovered by the RF model, and increased concentrations of myoinositol and choline were uncovered only by the ANN model. The above results demonstrate that the RF and ANN models individually provide unique and crucial information about concentration differences that discriminate cell lines and that only the combined information from both models enables construction of specific metabolic concentration profiles for individual cell lines in a thorough manner (Figure 5). As the GEM can be easily applied to any machine learning model and explain its predictions, analyzing biological datasets with different machine learning models may be advantageous.

To compare the GEM approach, our dataset of 63 ^{1}H NMR spectra acquired on normal and tumor cell lines was also analyzed with the PCA and PLS-DA methods, which are frequently used in metabolomics studies. PCA was unable to uncover clusters that could separate cell lines. Instead, it showed that samples predominantly clustered according to their "passage number"; this indicated that most of the variability in the metabolite concentrations arose from the degree of subculturing (Tables S5–S9 and Figure S14 A). Next, the dataset was analyzed with the PLS-DA method, which showed good separation of the NCI-H2052 cell line, three MSTO-211H cell lines, and most of the A549 cell lines but failed to separate the WI-38 and Met-5A cell lines (Figure S15 A). Features 44, 48, 73, 95, 103, and 109 were the most "influential" features towards the separation of the cell lines in the PLS-DA scores plot (Figure S15 B). Analysis of their feature values (Figure S16) revealed that only feature 103 had distinctly higher feature values for the NCI-H2052 cell line, which enabled its discrimination from the other cell lines. The feature values of the remaining "influential" features overlapped amongst cell lines, which made their discrimination less straightforward. Features 35, 70, 85, 86, 98,102, 228, and 232 uncovered by the GEM for the RF and ANN models that unambiguously discriminated cell lines were overlooked by both the PCA and PLS-DA methods.

Furthermore, the intrinsic stability of the GEM for the RF model[22] was evaluated and compared with the RF model's specific explanation methods, mean decrease accuracy (MDA) and mean decrease Gini (MDG) (Table S10). The average Spearman coefficient for MDA was very weak for the MeT-5A and MSTO-211H cell lines and weak for the A549, WI-38, and NCI-H2052 cell lines, whereas MDG exhibited weak average Spearman coefficients for all cell lines. The average Spearman coefficient for the GEM was strong for the A549, WI-38, and NCI-H2052 cell lines and moderate for the MeT-5A and MSTO-211H cell lines. These results demonstrate that the intrinsic stability of the GEM is much higher than that of MDA and MDG. Consequently, the calculations of the feature contributions in repeated runs are consistent with the GEM. The important features that are found by GEM are thus more accurate and reliable than these found by MDA or MDG.

In conclusion, we have demonstrated that by using the GEM to explain the predictions of the RF and ANN models on a dataset of ^{1}H NMR spectra, concentration differences between the metabolites were identified that discriminated individual normal and tumor cell-line types. We believe that uncovering such intricate metabolic differences, not only between cell lines but also between other biological systems, is key for accurate diagnosis, efficient drug development, and overall understanding of the mechanisms that underlie phenotypes. Moreover, promising advances in spectroscopic techniques and machine learning models affirm that the GEM can become an indispensable tool for metabolic phenotyping, biomarker discovery, and drug target detection of complex diseases.

Acknowledgements

The authors acknowledge financial support from the Slovenian Research Agency (research core funding no. P1-0242 and P3-0003). The authors would like to thank Dr. Marjeta Knez for her helpful advice and discussions on various topics related to this study.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: cancer · general explanation method · machine learning · metabolomics · NMR spectroscopy

[1] M. S. Palmnas, H. J. Vogel, *Metabolites* **2013**, *3*, 373 – 396.
[2] M. Guma, S. Tiziani, G. S. Firestein, *Nat. Rev. Rheumatol.* **2016**, *12*, 269 – 281.
[3] W. J. Griffiths, T. Koal, Y. Wang, M. Kohl, D. P. Enot, H.-P. Deigner, *Angew. Chem. Int. Ed.* **2010**, *49*, 5426 – 5445; *Angew. Chem.* **2010**, *122*, 5554 – 5575.
[4] C. H. Johnson, J. Ivanisevic, G. Siuzdak, *Nat. Rev. Mol. Cell Biol.* **2016**, *17*, 451.
[5] A. C. Dona, M. Kyriakides, F. Scott, E. A. Shephard, D. Varshavi, K. Veselkov, J. R. Everett, *Comput. Struct. Biotechnol. J.* **2016**, *14*, 135 – 153.
[6] S. Ren, A. A. Hinzman, E. L. Kang, R. D. Szczesniak, L. J. Lu, *Metabolomics* **2015**, *11*, 1492 – 1513.
[7] A. Smolinska, L. Blanchet, L. M. Buydens, S. S. Wijmenga, *Anal. Chim. Acta* **2012**, *750*, 82 – 97.
[8] A. Alonso, S. Marsal, A. Julià, *Front. Bioeng. Biotechnol.* **2015**, *3*, 23.
[9] B. Worley, R. Powers, *Curr. Metabolomics* **2013**, *1*, 92 – 107.
[10] P. S. Gromski, H. Muhamadali, D. I. Ellis, Y. Xu, E. Correa, M. L. Turner, R. Goodacre, *Anal. Chim. Acta* **2015**, *879*, 10 – 23.
[11] P. J. Trainor, A. P. DeFilippis, S. N. Rai, *Metabolites* **2017**, *7*, E30.
[12] E. Štrumbelj, I. Kononenko, *Knowl. Inf. Syst.* **2014**, *41*, 647 – 665.
[13] E. Štrumbelj, I. Kononenko, *J. Mach. Learn. Res.* **2010**, *11*, 1 – 18.
[14] E. Štrumbelj, Z. Bosnić, I. Kononenko, B. Zakotnik, C. Grašič Kuhar, *Knowl. Inf. Syst.* **2010**, *24*, 305 – 324.
[15] S. V. Sharma, D. A. Haber, J. Settleman, *Nat. Rev. Cancer* **2010**, *10*, 241.
[16] S. Hayton, G. L. Maker, I. Mullaney, R. D. Trengove, *Cell. Mol. Life Sci.* **2017**, *74*, 4421 – 4441.
[17] J. R. Mayers, M. E. Torrence, L. V. Danai, T. Papagiannakopoulos, S. M. Davidson, M. R. Bauer, A. N. Lau, B. W. Ji, P. D. Dixit, A. M. Hosios, et al., *Science* **2016**, *353*, 1161 – 1165.
[18] J. R. Mayers, M. G. van der Heiden, *Cancer Res.* **2017**, *77*, 3131 – 3134.
[19] E. A. Ananieva, A. C. Wilkinson, *Curr. Opin. Clin. Nutr. Metab. Care* **2018**, *21*, 64.
[20] D. R. Sappington, E. R. Siegel, G. Hiatt, A. Desai, R. B. Penney, A. Jamshidi-Parsian, R. J. Griffin, G. Boysen, *Biochim. Biophys. Acta Gen. Subj.* **2016**, *1860*, 836 – 843.
[21] Y. J. Kang, Y. Feng, E. L. Hatcher, *J. Cell. Physiol.* **1994**, *161*, 589 – 596.
[22] W. Huazhen, Y. Fan, L. Zhiyuan, *BMC Bioinformatics* **2016**, *17*, 60.

Manuscript received: July 13, 2018
Accepted manuscript online: August 1, 2018
Version of record online: September 14, 2018

