Analyzing Cysteine Site Neighbors in Proteins to Reveal Dimethyl Fumarate Targets

Rosa AC[1], Benetti E[1], Gallicchio M[1], Boscaro V[1], Cangemi L[1], Dianzani C[1], Miglio G[1,2].

[1]*Dipartimento di Scienza e Tecnologia del Farmaco, Università degli Studi di Torino, Turin, 10125, Italy.*

[2]*Centro di Competenza sul Calcolo Scientifico C3S, Università degli Studi di Torino, Turin, 10125, Italy.*

Analyzing Cysteine Site Neighbors in Proteins to Reveal Dimethyl Fumarate Targets.

Proteomics. 2019 Feb;19(4):e1800301. doi: 10.1002/pmic.201800301. Epub 2019 Jan 25.

Analyzing cysteine site neighbors in proteins to reveal dimethyl fumarate targets

Arianna Carolina Rosa[1], Elisa Benetti[1], Margherita Gallicchio[1], Valentina Boscaro[1], Luigi Cangemi[1], Chiara Dianzani[1], Gianluca Miglio*[,1,2]

[1]Dipartimento di Scienza e Tecnologia del Farmaco, Università degli Studi di Torino, Turin, Italy

[2]Centro di Competenza sul Calcolo Scientifico C[3]S, Università degli Studi di Torino, Turin, Italy

***Corresponding Author**: Prof Gianluca Miglio, PhD. *Dipartimento di Scienza e Tecnologia del Farmaco and Centro di Competenza sul Calcolo Scientifico C[3]S, Università degli Studi di Torino, Turin (Italy)*. Via Pietro Giuria, 9, Turin (Italy). Phone: +39 0116707150; E-mail: gianluca.miglio@unito.it

Abbreviations:

CART, classification and regression tree; CIT, conditional inference tree; DMF, dimethyl fumarate; isoTOP-ABPP, isotopic tandem orthogonal proteolysis-activity-based protein profiling; KEAP, kelch-like ECH-associated protein; KNN, *k*-Nearest Neighbors; LDA, linear discriminant analysis; MC, modifiable cysteine; MMF, monomethyl fumarate; NMC, non-modifiable cysteine; NNET, Neuronal Network, NNET; Nrf2, nuclear factor (erythroid-derived 2)-like 2; PLS, partial least square; RF, random forest; RIN, residue interaction network; SVM, support vector machine.

Keywords: Cysteine reactivity; computational method; residue interaction network; machine learning.

Total number of words: 3296

Abstract

This work proposes a novel approach by which to consistently classify cysteine sites in proteins in terms of their reactivity toward dimethyl fumarate (DMF) and fumarate. Dimethyl fumarate-based drug products have been approved for use as oral treatments for psoriasis and relapsing-remitting multiple sclerosis in recent years. The adduction of DMF and its (re)active metabolites to certain cysteine residues in proteins is thought to underlie their effects. However, only a few receptors for these compounds have been discovered to date. Our approach takes advantage of the growing number of known DMF- and fumarate-sensitive proteins and sites to perform analyses by combining the concepts of network theory, for protein structure analyses, and machine learning procedures. Wide-ranging and previously unforeseen variety was found in the analysis of the neighborhood composition (the first neighbors) of cysteine sites found in DMF- and fumarate-sensitive proteins. Furthermore, neighborhood composition has shown itself to be a network-type attribute that is endowed with remarkable predictive power when distinct classification algorithms are employed. In conclusion, when adopted in combination with other target identification/validation approaches, methods that are based on the analysis of cysteine site neighbors in proteins should provide useful information by which to decipher the mode of action of DMF-based drugs.

Significance of the study

The adduction of dimethyl fumarate (DMF) and its (re)active metabolites to certain cysteine residues in proteins is thought to underlie the effects of DMF-based drug products. However, many more pharmacological targets (“acceptors” and “receptors”) for DMF-based drugs, than those that have been characterized thus far, should exist. This proof-of-concept study proposes a novel approach by which to consistently classify cysteine sites in terms of their reactivity towards DMF and fumarate. Our approach takes advantage of the growing number of known DMF- and fumarate-sensitive proteins and sites to perform analyses by combining concepts of network theory, for protein structure analyses, and machine learning procedures. In particular, neighborhood composition has been shown to be a network-type attribute that is endowed with remarkable predictive power. These findings could be instrumental in developing high-throughput and system-oriented methods and tools by which to find pharmacological targets for DMF-based drug products.

In 1959, Schweckendiek W. first described the beneficial effects exerted by fumarate-related compounds on *psoriasis vulgaris*.[1] Recent decades have seen further compelling evidence as to the therapeutic value of these agents being collected.[2-4] Moreover, drug products based on fumarate-related compounds have been developed and registered by many regulatory authorities. For example, Tecfidera® (Biogen-Idec) and Skilarence® (Almiral) are two distinct dimethyl fumarate (DMF)-only drug products that are available for the treatment of relapsing-remitting multiple sclerosis and psoriasis, respectively, in many countries. However, the mode of action of these agents still remains elusive.

Dimethyl fumarate is converted into a set of metabolites following oral administration.[5] Monomethyl fumarate (MMF) and fumarate, which are formed from DMF via hydrolysis, have been shown to mimic responses to their parent compound both *in vitro* and *in vivo*,[3-11] and thus these compounds may contribute together to the effects of the DMF-based drug products. In addition, the pharmacological activity of these related compounds has been associated, at least in part, to their reactivity toward the thiol group in the side chain of certain cysteine residues in proteins.[10,11] In fact, cell exposure to DMF has been shown to increase the abundance of cysteine residues modified by DMF and/or its metabolites.[7,8,12,13] However, only few of these modifications have been clearly related to the effects of the DMF-based drugs thus far. For example, nuclear factor (erythroid-derived 2)-like 2 (Nrf2) activation, which follows DMF adduction to certain cysteine residues in kelch-like ECH-associated protein (KEAP)-1-Nrf2 protein complexes, has been associated with the inhibition of dendritic cell maturation.[14] Moreover, the same type of covalent modification of cysteine residues in glyceraldehyde-3-phosfate dehydrogenase has been demonstrated to decrease enzymatic activity and could limit the development/activation of pro-inflammatory immune cells.[12] As for the KEAP-1-Nrf2- and glyceraldehyde-3-phosfate dehydrogenase-dependent responses, the adduction of DMF,

MMF and/or fumarate to cysteine sites in other proteins may also contribute to the effects exerted by DMF-based drug products. This hypothesis is consistent with the discovery of a number of cysteine sites in proteomes that are sensitive to these compounds.[8,12,13,15-17] However, the abundance of cysteine residues in proteomes,[18] means that there should be many more targets ("acceptors" and "receptors") for DMF-related compounds than those that have been verified thus far. Therefore, further effort must be dedicated to comprehensively mapping the DMF-based drug-sensitive proteome.

Our understanding of post-translational cysteine modifications has been enhanced by combining the methods and tools of computational and experimental proteomics. In particular, *in silico* analyses may well lead to intriguing hypotheses for more traditional studies. Therefore, the primary aim of this study is to demonstrate the performance of a novel approach for the classification (class prediction) of cysteine sites in proteins, according to their reactivity toward DMF and/or fumarate, with the long-term goal of better understanding the mode of action of DMF-based drug products. Reactivity of a cysteine site toward electrophiles depends on a number of site-specific attributes. Some of them are related to features of the microenvironment surrounding the cysteine site, which themselves are linked to the protein structure in a hierarchical manner. For example, the activating effect often demonstrated for basic residues in close proximity to cysteines is consistent with the evidence that the thiolate form of a thiol group is much more nucleophilic than its protonate counterpart, and is much more readily alkylated by electrophiles.[19-21] Thereby, sequence- and structure-based attributes have been investigated in order to discover novel predictors useful for developing classification methods.[21] For example, sequence-based method has been shown to predict fumarate adduction to cysteine sites,[22] while a structure-based method has been demonstrated to provide even more accurate predictions when applied to the same type of cysteine

modification.[23] This paper proposes a novel approach that uses the study of site neighbors to predict cysteine reactivity toward DMF and/or fumarate. It was based on the adoption of machine learning procedures to process data generated from the analysis of protein structure represented as residue interaction network (RIN; Figure 1).[23-30]

A growing number of cysteine sites that are sensitive to DMF-related compounds have been verified using experimental proteomic methods over the past few years, and computational analyses have been carried out in order to fully take advantage of these findings. A dataset of DMF-sensitive cysteine sites (DMF-DS) was prepared using the findings reported by Blewett et al.[13] A large set of DMF-sensitive sites have been discovered in primary human T cell proteins using the isotopic tandem orthogonal proteolysis-activity-based protein profiling (isoTOP-ABPP) method. An isoTOP-ABPP ratio of at least 3 was adopted as the cut-off in this work to ensure that we include a valuable number of sites and exclude those endowed with low reactivity toward DMF. All of the collected proteins and sites were further assessed for eligibility, which was defined by the following criteria: 1) proteins with an established 3D structure; 2) sequence identity of at least 80%, as reported by The Protein Modal Portal (www.proteinmodelportal.org). The same criteria were also applied to a dataset of fumarate-sensitive cysteine sites found in eukaryotic proteins (FUM-DS), which was revised and updated from a previous version.[23] A total of 32 and 43 RIN were obtained from the 3D structure of the corresponding 32 and 42 proteins (see Experimental section) included in the DMF-DS and FUM-DS, respectively (Table 1 and Table S1 and S2, Supporting Information). X-Ray diffraction was the method more often used to establish the 3D structure of these proteins. Human interleukin-1 receptor-associated kinase 4 (PDB ID: 3MOP) was found in both datasets. A total of 588 cysteine sites were found in these RIN (308 and 280 sites, in DMF-DS and FUM-DS, respectively). All these sites were divided into two subclasses according to their reactivity

toward DMF and/or fumarate: modifiable cysteine (MC; 35 and 52, for DMF-DS and FUM-DS, respectively) and non-modifiable cysteine (NMC; 273 and 228, respectively) sites. Notably, NMC sites are either cysteine residues with no evidence of reactivity toward DMF/fumarate, or endowed with an isoTOP-ABPP ratio below the cut-off. Analysis of the collected RINs allows to identify 4,327 cysteine neighbors (2,221 and 2,106 neighbors, for the DMF-DS and FUM-DS, respectively). No significant difference was determined when the numbers of MC- and NMC-neighbors were compared between the two datasets (Table 1). In contrast, significant differences were found when the number of MC- and NMC-neighbors were compared within each dataset (*P*-value, 2.1×10^{-6} and 1.7×10^{-5}, for the DMF-DS and FUM-DS, respectively; Wilcoxon rank sum test). Thereby, the existence of an association between cysteine reactivity toward DMF and/or fumarate and composition of the site neighborhood can be hypothesized. In order to assess this hypothesis, data on neighborhood composition were analyzed. First, clustering analysis algorithms were used to highlight the existence of specific patterns in these data. Unpredictably broad variety was found when DMF-DS and FUM-DS were examined (Figure S1 and S2, respectively, Supporting Information). Moreover, the lack of consistency in the results on the optimal number of clusters, which were computed using three different methods (Elbow, average silhouette and gap statistic; data not shown), mean that no obvious pattern of interactions/factors can be defined. Then, the predictive power of neighborhood composition was further assessed to evaluate whether it is an attribute by which MC and NMC sites can be classified. Preliminarily, the synthetic minority oversampling technique[31] was applied to attenuate the bias resulting from use of class-imbalanced datasets. The resulting datasets were then divided into a training (75% of the total sites) and a test subset (25%) by random selection, and analyzed using eight classification algorithms/models (Classification and Regression Tree, CART; Conditional Inference Tree,

CIT; *k*-Nearest Neighbors, KNN; Linear Discriminant Analysis, LDA; Neuronal Network, NNET; Partial Least Square, PLS; Random Forrest, RF; Support Vector Machine, SVM).[32] As shown in Figure 2, values higher than 0.8 were obtained for most algorithms/models when accuracy, sensitivity and specificity were computed. The highest and lowest performances were obtained using SVM and CIT, respectively. Moreover, a comparable performance was determined for SVM, RF and NNET. These results confirm previous findings on the reliability of these algorithms/models when employed in classification tasks,[32] including those in the field of cysteine site reactivity.[23] Moreover, they support the adoption of neighborhood composition as a potential predictor to estimate cysteine reactivity toward DMF-related drugs. In order to improve the clarity of the description, neighborhood composition of explicit examples, cysteine sites found in three proteins, were analyzed using SVM, LDA and CIT, which are endowed with high, intermediate and low performance, respectively. As shown in Figure 3, cysteine residues found in human probable DNA dC→dU-editing enzyme (3VOW; reactivity toward DMF, panel A; Figure S3, Supporting Information),[13] human tyrosyl-tRNA synthetase, cytoplasmic (1NTG; reactivity toward fumarate, panel B; Figure S4, Supporting Information),[17] and human interleukin-1 receptor-associated kinase 4 (3MOP; reactivity toward DMF, panel C; reactivity toward fumarate, panel D; Figure S5, Supporting Information)[13,17] were classified precisely when SVM was employed. In contrast, some errors were detected when LDA and CIT were tested.

The relatively small number of proteins and MC sites discovered in the experimental studies that fulfil the inclusion criteria limits the possibility to generalize the findings of this study. The likelihood of the conclusions is however supported by the evidence that well-known features of cysteine residues in proteins (e.g., relative abundance) are met by those investigated in this study. Moreover, near equal results, in term of algorithm/model

performance, were obtained by analyzing two datasets of cysteines residues characterized by a minimal overlap and including sites gathered from different sources. The inclusion of additional examples from future studies would allow stronger conclusions to be drawn. As for other computational approaches, predictions obtained by analyzing cysteine neighbors needs to be interpreted with caution. Indeed, discrepancies could be determined by comparing theoretical and experimental findings. For instance, when cysteine-152 in human glyceraldehyde-3-phosfate dehydrogenase was assessed for its reactivity toward DMF, it was classified as a probable NMC site (data not shown). This prediction is consistent with the data reported by Blewett and colleagues.[13] However, this cysteine residue has also been reported as a relevant DMF-target site.[12] As for the inconsistencies resulting from the comparisons involving experimental data, even those regarding predicted *vs* measured cysteine reactivity could be attributed, at least in part, to the differences in the methods used to assess this feature. In fact, reactivity of a cysteine site toward DMF-related electrophiles has been demonstrated to depend on the experimental conditions.[13] Therefore, some sites endowed with low reactivity (e.g., an isoTOP-ABPP ratio below the cut-off) could even be modified when an higher cell exposure to the DMF-related compounds, than those already studied, are adopted. Future studies, including those on specimens obtained from patients treated with DMF-based drug products, could provide intriguing new data on this point.

In conclusion, the modification of a large proportion of cysteine sites in proteomes has led to electrophiles that are derived from DMF-based drug products exerting pleiotropic actions. The development and adoption of high-throughput and system-oriented methods and tools should be useful in interpreting the pharmacological profiles of these agents. As indicated by this proof-of-concept study, network theory, for protein structure analyses, can be combined with machine learning techniques to estimate the propensity of

a cysteine residue to be modified by DMF and/or fumarate. This thus provides a novel approach by which to find novel “acceptors” and “receptors” for DMF-based drug products.

Experimental section

The 3D structure of the included proteins were retrieved from the RCSB-PDB repository (http://www.rcsb.org/pdb/home/home.do) and analyzed to create the corresponding 2D representations as RIN. Briefly, according to the method proposed by Doncheva et al.,[30] each PDB file was visualized using UCSF Chimera (1.11.2) software (http://www.cgl.ucsf.edu/chimera/) and converted into the corresponding RIN using RINalyzer (http://www.rinalyzer.de), a Cytoscape-plugin for protein structure network assessment. Standard amino acids and natural ligands were considered nodes, while either the presence of a covalent bond between two residues, or a distance of at least 5 Å between two Cα were adopted as criteria to establish the node connectivity.[23] Data on the first neighbors of a cysteine site were collected using a two-step procedure. In the first step, each site was represented as a numerical vector. The vector elements were the counts of the 20 natural amino acids found in the neighborhood of that site. In the second step, the resulting numerical vectors were combined in a $n \times 20$-matrix, were n is the total number of cysteine sites in a dataset. The matrices were analyzed using the *k*-means clustering algorithm to assess the associations between neighborhood composition and site reactivity. Moreover, as described by Kuhn and Johnson,[32] the predictive power of neighborhood composition was studied using eight classification models (Classification and Regression Tree, CART; Conditional Inference Tree, CIT; *k*-Nearest Neighbors, KNN; Linear Discriminant Analysis, LDA; Neuronal Network, NNET; Partial Least Square, PLS; Random Forrest, RF; Support Vector Machine, SVM). A ten-fold cross-validation

procedure was adopted, while the following metrics were computed to quantify model performance:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

where *TP* = true positive, *TN* = true negative, *FP* = false positive, and *FN* = false negative. Data were prepared, analyzed, and visualized using the R software (The R Project for Statistical Computing; https://www.r-project.org/).

Author's Contribution to the Manuscript

Study design, concept and supervision were performed by GM. ACR, EB, MG, VB, LC, and CD analyzed data, critically reviewed findings and contributed to manuscript editing.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Acknowledgements

Dr Jessica Audisio (Dipartimento di Scienza e Tecnologia del Farmaco, Università degli Studi di Torino) is gratefully acknowledged for her support in the implementation of the method. This work was supported by funds from the Università degli Studi di Torino, Ricerca Locale Ex 60% 2015 and 2016-2017 to GM.

References

[1] W. Schweckendiek, Med. Monatsschr. 1959, 13, 103.

[2] R. Lijnen, E. Otters, D. Balak, B. Thio, J. Dermatolog. Treat. 2016, 27, 31.

[3] R. Gold, D. L. Arnold, A. Bar-Or, M. Hutchinson, L. Kappos, E. Havrdova, D. G. MacManus, T. A. Yousry, C. Pozzilli, K. Selmaj, M. T. Sweetser, R. Zhang, M. Yang, J. Potts, M. Novas, D. H. Miller, N. C. Kurukulasuriya, R. J. Fox, T. J. Phillips, Mult. Scler. 2017, 23, 253.

[4] U. Mrowietz, J. C. Szepietowski, R. Loewe, P. van de Kerkhof, R. Lamarca, W. G. Ocker, V. M. Tebbs, I. Pau-Charles, Br. J. Dermatol. 2017, 176, 615.

[5] H. Ashrafian, G. Czibik, M. Bellahcene, D. Aksentijević, A. C. Smith, S. J. Mitchell, M. S. Dodd, J. Kirwan, J. J. Byrne, C. Ludwig, H. Isackson, A. Yavari, N. B. Støttrup, H. Contractor, T. J. Cahill, N. Sahgal, D. R. Ball, R. I. Birkler, I. Hargreaves, D. A. Tennant, J. Land, C. A. Lygate, M. Johannsen, R. K. Kharbanda, S. Neubauer, C. Redwood, R. de Cabo, I. Ahmet, M. Talan, U. L. Günther, A. J. Robinson, M. R. Viant, P. J. Pollard, D. J. Tyler, H. Watkins, Cell Metab. 2012, 15, 361.

[6] E. D. Deeks, Drugs 2016, 76, 243.

[7] A. M. Manuel, N. Frizzell, Amino Acids 2013, 45, 1243.

[8] G. G. Piroli, A. M. Manuel, M. D. Walla, M. J. Jepson, J. W. Brock, M. P. Rajesh, R. M. Tanis, W. E. Cotham, N. Frizzell, Biochem J. 2014, 462, 231.

[9] G. Miglio, E. Veglia, R. Fantozzi, Int. Immunopharmacol. 2015, 28, 215.

[10] U. Mrowietz, P. J. Morrison, I. Suhrkamp, M. Kumanova, B. Clement, Trends Pharmacol. Sci. 2017, 39, 1.

[11] J. Brück, R. Dringen, A. Amasuno, I. Pau-Charlesm, K. Ghoreschi, Exp. Dermatol. 2018, 27, 611.

[12] A. Hammer, A. Waschbisch, I. Knippertz, E. Zinser, J. Berg, S. Jörg, K. Kuhbandner, C. David, J. Pi, A. Bayas, D. H. Lee, A. Haghikia, R. Gold, A. Steinkasserer, R. A. Linker, Front. Immunol. 2017, 8, 1922.

[13] M. D. Kornberg, P. Bhargava, P. M. Kim, V. Putluri, A. M. Snowman, N. Putluri, P. A. Calabresi, S. H. Snyder, Science 2018, 360, 449.

[14] N. Ternette, M. Yang, M. Laroyia, M. Kitagawa, L. O'Flaherty, K. Wolhulter, K. Igarashi, K. Saito, K. Kato, R. Fischer, A. Berquand, B. M. Kessler, T. Lappin, N. Frizzell, T. Soga, J. Adam, P. J. Pollard, Cell Rep. 2013, 3, 689.

[15] E. D. Merkley, T. O. Metz, R. D. Smith, J. W. Baynes, N. Frizzell, Mass Spectrom. Rev. 2014, 33, 98.

[16] M. Yang, N. Ternette, H. Su, R. Dabiri, B. M. Kessler, J. Adam, B. T. The, P. J. Pollard, Metabolites 2014, 4, 640.

[17] M. M. Blewett, J. Xie, B. W. Zaro, K. M. Backus, A. Altman, J. R. Teijaro, B. F. Cravatt, Sci. Signal. 2016, 9, rs10.

[18] L. I. Leichert, T. P. Dick, Incidence and physiological relevance of protein thiol switches. Biol. Chem. 2015, 396, 389.

[19] A. Higdon, A. R. Diers, J. Y. Oh, A. Landar, V. M. Darley-Usmar, Biochem. J. 2012, 442, 453.

[20] R. M. Lopachin, T. Gavin, A. Decaprio, D. S. Barber, Chem. Res. Toxicol. 2012, 25, 239.

[21] S. M. Marino, V. N. Gladyshev, J. Biol. Chem. 2012, 287, 4419.

[22] G. Miglio, A. D. Sabatino, E. Veglia, M. T. Giraudo, M. Beccuti, F. Cordero, F. Biochim. Biophys. Acta 2016, 1864, 211.

[23] G. Miglio, Amino Acids 2018, 5, 163.

[24] R. Blagus, L. Lusa, BMC Bioinformatics 2013, 14, 106.

[25] M. Kuhn, K. Johnson, Applied predictive modelling. Springer-Verlag, New York 2013.

[26] P. Csermely, R. Nussinov, A. Szilágyi, Curr. Top. Med. Chem. 2013, 13, 2.

[27] M. Giollo, A. J. Martin, I. Walsh, C. Ferrari, S. C. Tosatto, BMC Genomics 2014, 15, S7.

[28] L. Di Paola, A. Giuliani, Curr. Opin. Struct. Biol. 2015, 31, 43.

[29] M. Bhattacharyya, S. Ghosh, S. Vishveshwara, Curr. Protein. Pept. Sci. 2016, 17, 4.

[30] F. Fanelli, A. Felline, F. Raimondi, M. Seeber, Biochem. Soc. Trans. 2016, 44, 613.

[31] J. Salamanca, Viloria, M. F. Allega, M. Lambrughi, E. Papaleo, Sci. Rep. 2017, 7, 2838.

[32] N. T. Doncheva, Y. Assenov, F. S. Domingues, M. Albrecht, Nat. Protoc. 2012, 7, 670.

Table 1. Overview of proteins and cysteine sites analyzed.

	Data set	
	DMF	FUM
Proteins (*n*)	32	42
3D structures established by:		
X-ray diffraction (*n*)	28	37
Solution NMR (*n*)	3	5
Electron microscopy (*n*)	1	-
Electron crystallography (*n*)	-	1
RIN (*n*)	32	43
Cysteine sites (*n*)	308	280
MC:NMC (*n*)	35:273	52:228
First neighbors (*n*)	2,221	2,106
MC, *median* [*range*]	5 [2 – 12]	6 [3 – 12]
NMC, *median* [*range*]	7 [3 – 14]	8 [2 – 13]

RIN, residue interaction network; MC, modifiable cysteine; NMC, non-modifiable cysteine.

Figure Legends

Figure 1. Identification of cysteine site first neighbors. The 3D structure (top left) of a DMF- and/or fumarate-sensitive protein was retrieved from the RCSB-PDB repository and analyzed to create the corresponding 2D representations as residue interaction networks (top left). The first neighbors of a cysteine site (bottom) were visualized, identified and counted for future analyses. Adenosine deaminase (PDB ID: 3IAR) has been shown as an explicit example. Cysteine residues are depicted as red spheres (3D structure) or circles (networks).

Figure 2. Predictive power of neighborhood amino acid composition. Eight algorithms/models (Classification and Regression Tree, CART; Conditional Inference Tree, CIT; *k*-Nearest Neighbors, KNN; Linear Discriminant Analysis, LDA; Neuronal Network, NNET; Partial Least Square, PLS; Random Forrest, RF; Support Vector Machine, SVM) were tested to analyze data on neighborhood compositions of cysteine site included in the DMF-DS (A) and FUM-DS (B). Accuracy, sensitivity and specificity were computed to quantify algorithm/model performance.

Figure 3. Predicted reactivity of explicit examples. Probabilities (P) of a cysteine site found in human probable DNA dC→dU-editing enzyme (PDB ID: 3VOW; reactivity toward DMF, panel A), human tyrosyl-tRNA synthetase, cytoplasmic (1NTG; reactivity toward fumarate, panel B) and human interleukin-1 receptor-associated kinase 4 (3MOP; reactivity toward DMF, panel C; reactivity toward fumarate, panel D) to be a modifiable site was computed by analyzing data on their neighborhood (see Figure S3-S5) using three algorithms/models (Support Vector Machine, SVM; Linear Discriminant Analysis, LDA;

LDA or Conditional Inference Tree, CIT, CIT) trained on the corresponding sets of cysteine sites and compared to the measured reactivity (Class).

Figure 1.

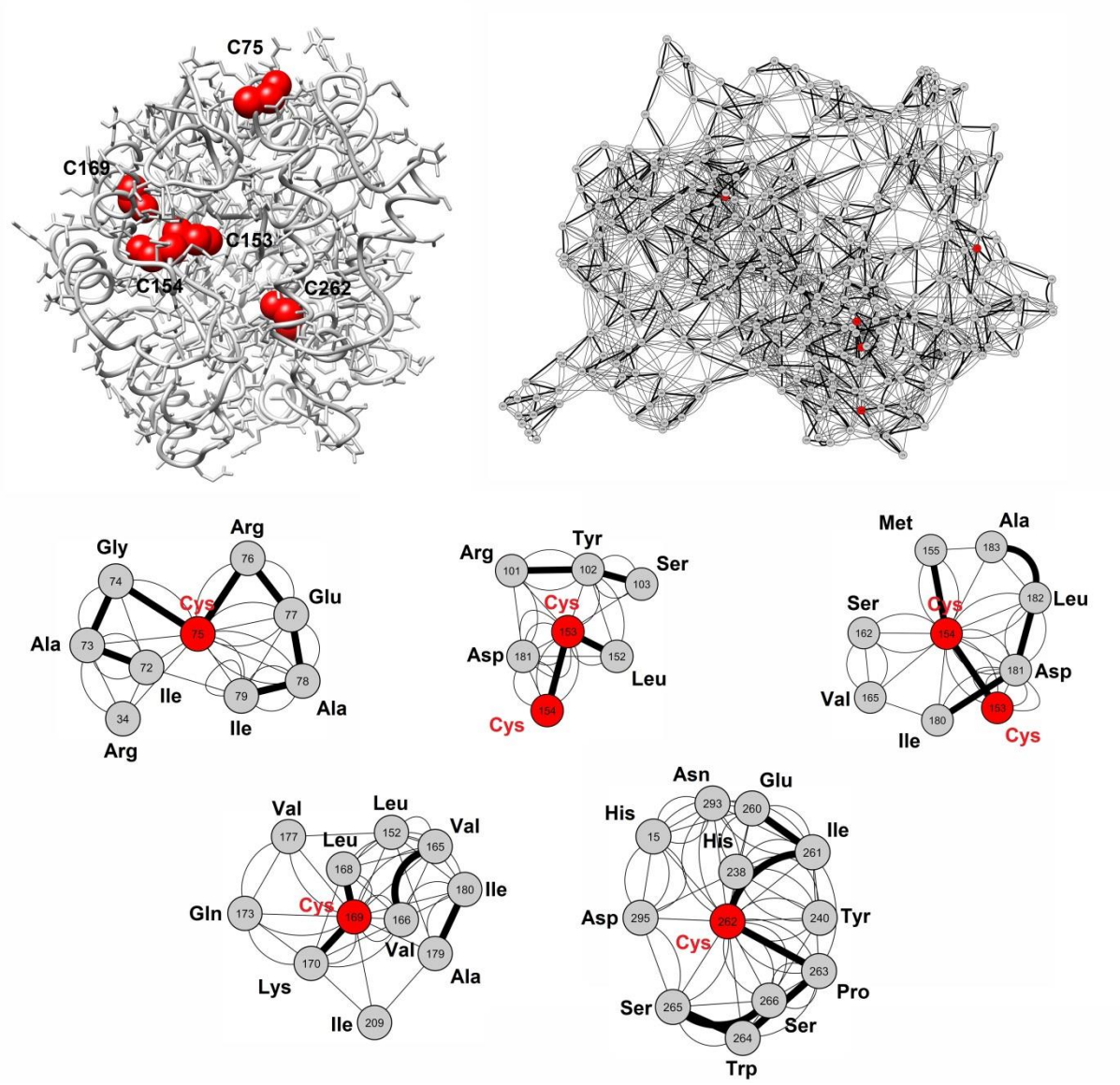

Figure 2.

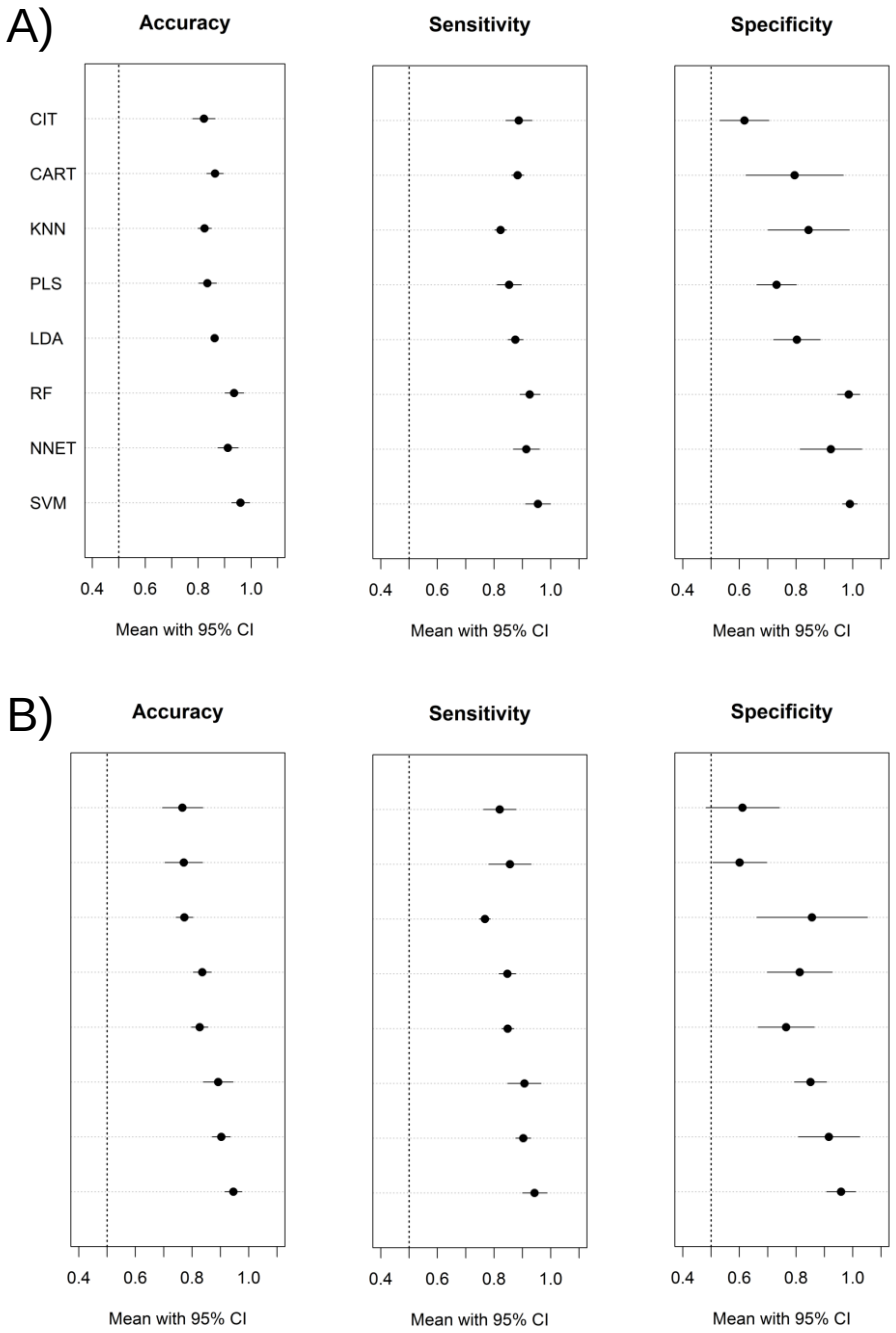

Figure 3.