Detection of native and mirror protein structures based on Ramachandran plot analysis by interpretable machine learning models

Julia Abel¹ , Marika Kaden¹ , Katrin Sophie Bohnsack^{1†}, Mirko Weber^{1†}, Christoph Leberecht^{2‡}, Thomas Villmann^{1* }

1 University of Applied Sciences Mittweida, Saxony Institute for Computational Intelligence and Machine Learning, Mittweida, Germany

2 PharmAI, Dresden, Germany

These authors contributed equally to this work.

^{†‡} These authors also contributed equally to this work.

* thomas.villmann@hs-mittweida.de

Abstract

In this contribution the discrimination between native and mirror models of proteins according to their chirality is tackled based on the structural protein information. This information is contained in the Ramachandran plots of the protein models. We provide an approach to classify those plots by means of an interpretable machine learning classifier - the Generalized Matrix Learning Vector Quantizer. Applying this tool, we are able to distinguish with high accuracy between mirror and native structures just evaluating the Ramachandran plots. The classifier model provides additional information regarding the importance of regions, e.g. α -helices and β -strands, to discriminate the structures precisely. This importance weighting differs for several considered protein classes.

Introduction

Proteins can have two conformations: the L-enantiomeric conformation and the D-enantiomeric conformation. The first represents the natural form (further known as native) of a protein, whereas the latter represents an exact mirror-image of it [21, 52, 60, 63]. The chemical general term for two enantiomeric conformations is chirality. By definition these enantiomers are mirror-images of each other and not superimposable [62].

Although the two enantiomeric forms of one protein exhibit more or less the same behavior concerning physical and chemical properties as well as substrate specificity [39, 52, 63], mirror proteins can show a more stable behavior towards natural enzyme degradation and reduced immunogenicity [21, 41, 60].

1

2

3

4

5

6

7

8

9

10

11

The differentiation of native and mirror proteins is crucial for further analysis and research in the fields of drug discovery/development and synthetic biology. [59,61]. Especially for the latter holds that mirror proteins give the opportunity to create a fully working molecular system, which would not be dependent on natural compounds [60] and, therefore, orthogonal systems would be possible [30]. Furthermore, mirror proteins hold the possibility of creating mirror life [30]. The presence of native and mirror models is of particular importance in the structure reconstruction of proteins from amino acid sequences. Contact maps can be used as a set of constraints for *ab initio* structure prediction. They show predicted contacts based on co-evolving residues which imply their spatial proximity and are derived from a multiple sequence alignment [6]. Contact map based approaches may lead to either native or mirror models since a contact map contains no information regarding chirality and both, native and mirror proteins exhibit the same inter-atomic distances [27]. In order to screen out such mirror models, a variety of approaches were published, which either intervene in the process of model generation [16,32] or filter mirror models subsequently [11,55].

KURCZYNSKA AND KOTULSKA tried in their work [28] to distinguish between native and mirror models in an automated way (Automated Mirror-Native-Discrimination – AMND). For this purpose, they used the energy terms obtained by PyRosetta, a Python-based interface to Rosetta for molecular modeling [9]. The discrimination results of 69% in average are respectable but far away to be satisfying. Further, it was not possible for them to find one uniform rule based on that energy terms for all the proteins classes **A-G** of SCOPe [18,31], which were studied in this investigation. Albeit in [2] it becomes apparent that the calculation of the energy terms is rather complex. Alternatively, they also took into account to consider the differences within Ramachandran plots (R-plots), and thus the dihedral angles distribution of a protein's backbone by depicting them in a toroidal plot, because these plots are known to provide structural information [43]. However, they explained that these plots are not feasible for automated mirror/native discrimination.

Nonetheless, there are three things a R-plot definitely can give information about: a protein's secondary structure elements, the favored/allowed regions of those and also the handedness of helices (right-handed and left-handed) [3,48]. Following the International Union of Pure and Applied Chemistry's (IUPAC) definition of handedness being the same as chirality [37], the distinctive feature for differentiating between native and mirror models would be the handedness of helices.

Since this handedness can also be found as a characteristic of R-plots, they represent a simple but important tool for finding structural differences in proteins [3]. Assuming that native and mirror proteins display different structures the resulting differences in the R-plots should be observable.

To tackle the AMND problem, appropriate mathematical and statistical tools are required, which reflect the mathematical structure of the R-plots and can evaluate them in an automatic way with sufficient certainty. Therefore machine learning tools like (deep) artificial neural networks come into play [13], which are powerful tools to analyze complex data like R-plots. However, these networks and their inferences are frequently not explainable [45]. An alternative to deep networks provide interpretable machine learning approaches like prototype based models [5,56], which rely on the nearest prototype principle for data representation [38]. Frequently, these networks make use of specific data properties [58] and, hence, their results are better to interpret and may provide additional knowledge drawn from the training data.

In this context, a key observation regarding R-plots is that we can take them as approximated density plots or histograms of dihedral angle pairs (Φ, Ψ) in the

two-dimensional plane. Thus, the data analysis and statistical evaluation have to deal with the comparison of densities and their discrete representations. Thus, we apply an adaptive vector quantization based classifier model – the so-called Generalized Matrix Learning Vector Quantizer (GMLVQ) [19, 50]. After training this classifier provides structural knowledge regarding the data, which supports the classification decision. In particular, a classification correlation matrix is delivered, which describes correlations between data [5, 56]. GMLVQ is known to be robust and easy to interpret [46, 58]. In biomedical context it was successfully applied to analyse flow cytometry data and to detect early folding residues during protein folding [4, 7].

The structure of this paper is as follows: First the data set in use is described more detailed as well as the corresponding data preprocessing. Afterwards we give a brief introduction of R-plots as well as the machine learning data analysis tool GMLVQ, which is an interpretable artificial neural network. Subsequently, we state the general workflow for the AMND to distinguish between native and mirror samples based on R-plot analysis by means of GMLVQ.

Finally we present the numerical results for classification performance and the extracted knowledge provided by the interpretable model together with its direct biological explanation.

Materials and methods

Ramachandran Plots

Ramachandran plots (R-plots) display the dihedral angles Φ and Ψ of a protein's backbone to visualize their distribution [43]. R-plots provide an easily inspectable tool to detect underlying properties of the secondary structure in that protein [3]. However it has to be kept in mind, that the measured angles describe the current state at the time of measurement of the respective dynamic atoms in the backbone [33].

Mathematically, R-plots can be taken as approximated probability distributions for pairs (Φ, Ψ) of dihedral angles. Accordingly, if we partition the plot by an $N \times N$ cell array as depicted in S1 Fig, we can describe each plot as a vector

$\mathbf{x} = (x_1, \dots, x_n)^T \in \mathbb{R}^n$ with $n = N^2$, where $x_k \geq 0$ is the (relative) number of dihedral pairs in the cell with the coordinates (j, l) where $k = (j - 1) \cdot N + l$. Thus, $\mathbf{x}$ is a relative histogram vector also denoted as probability vector, i.e. $\sum_{k=1}^n x_k = 1$.

S1 Fig shows a generic R-plot of pairs (Φ, Ψ) , for a backbone. The horizontal axis is related to Φ , whereas the vertical axis is related to Ψ . The allowed/favored areas of secondary structures are shown. Particularly, we indicate the regions of β -sheets (β -strands), as well as right- and left-handed α -helices. Additionally we have added a 6×6 cell grid to the R-plot, so that we can specifically address cells of the R-plot by their coordinates.

For example cells (1,4) and (2,4) as well as cells (4,1) and (4,2) characterize the favored regions for right-handed α -helices (α) and left-handed α -helices (L_α), respectively. The cells (1,2) and (2,1) constitute the favored regions for β -strands (β).

The Generalized Learning Vector Quantizer – An Adaptive, Interpretable Classifier in Machine Learning

The *Generalized Learning Vector Quantizer* (GLVQ, [47]) is a prototype-based machine learning method for classification derived from a heuristic approach proposed by T.

KOHONEN [22]. The mathematically well justified model optimizes the classification performance by means of stochastic gradient descent learning (SGDL, [14, 44]) and is known to be robust and interpretable. The underlying cost function to be minimized during training approximates the overall classification error for a given training data set $T = \{(\mathbf{x}_j, c(\mathbf{x}_j)) \in X \times \mathcal{C}, j = 1 \dots N\}$, where $X \subseteq \mathbb{R}^n$ is the set of training vectors and $c(\mathbf{x}_j) \in \mathcal{C} = \{1, \dots, C\}$ is a training data class label out of a set of available class labels [19]. For this purpose, the GLVQ model assumes a set of prototypes $W = \{\mathbf{w}_k \in \mathbb{R}^n, k = 1 \dots M\}$ with class labels $c(\mathbf{w}_k) \in \mathcal{C}$ such that at least one prototype is assigned to each class. Thus, a partition of the prototype set $W = \bigcup_{j=1}^C W_j$ with $W_j = \{\mathbf{w}_k \in W | c(\mathbf{w}_k) = j\}$ is obtained, i.e. $W_j \neq \emptyset$ for all $j = 1 \dots C$ and $W_i \cap W_j$ for $i \neq j$ are valid.

Further, a dissimilarity measure $d(\mathbf{x}, \mathbf{w})$ is supposed to evaluate the similarity between data and prototypes. This measure is required to be differentiable at least with respect to the second argument to ensure SGDL.

For a given GLVQ-configuration, i.e. a fixed prototype set W , a new data point $\mathbf{x}$ is classified by the assignment

$$\mathbf{x} \mapsto c(\mathbf{w}_{\omega(W)}) \quad (1)$$

with

$$\omega(W) = \operatorname{argmin}_{\mathbf{w}_j \in W} d(\mathbf{x}, \mathbf{w}_j) \quad (2)$$

known as winner-takes-all in nearest prototype classification [38]. The prototype $\mathbf{w}_\omega$ is denoted as winner of the competition.

During the learning, GLVQ adapts the prototypes minimizing the overall cost function

$$E_T[E(\mathbf{x}_k, W)] = \frac{1}{\#T} \sum_{x \in T} E(\mathbf{x}_k, W)$$

by SGDL with respect to the prototypes realizing a competitive learning scheme [1], where

$$E(\mathbf{x}_k, W) = f(\mu(\mathbf{x}_k)) \quad (3)$$

is the local classification error and $\#T$ denotes the cardinality of the training data set. The local error $E(\mathbf{x}_k, W)$ depends on the choice of the monotonically increasing squashing function f and the classifier function

$$\mu(\mathbf{x}_k) = \frac{d^+(\mathbf{x}_k) - d^-(\mathbf{x}_k)}{d^+(\mathbf{x}_k) + d^-(\mathbf{x}_k)} \in [-1, 1] \quad (4)$$

where $d^\pm(\mathbf{x}_k) = d^\pm(\mathbf{x}_k, \mathbf{w}^\pm)$ and $\mathbf{w}^+ = \mathbf{w}_{\omega(W_{c(\mathbf{x}_k)})}$ is the so-called best matching correct prototype and $\mathbf{w}^- = \mathbf{w}_{\omega(W \setminus W_{c(\mathbf{x}_k)})}$ is the corresponding best matching

incorrect prototype. Frequently, the squashing function f is chosen as a parametrized sigmoid: $f_\sigma(z) = \frac{1}{1 + \exp(-\sigma z)}$ with the slope parameter σ [57].

The classifier function $\mu(\mathbf{x}_k)$ becomes negative for correctly classified training data. In fact, the cost function $E_T[E(\mathbf{x}_k, W)]$ is an approximation of the overall classification error. The approximation precision depends on the specific choice of the squashing function and their parameters. However, in case of the mentioned parametrized sigmoid function, the GLVQ is robust regarding moderate variations of the steepness parameter. For a detailed consideration we refer to [20].

In standard GLVQ, the squared Euclidean distance $d_E(\mathbf{x}, \mathbf{w}) = (\mathbf{x} - \mathbf{w})^2$ is used. 140
Another popular choice is the squared *Euclidean mapping distance* 141

$$\begin{aligned} d_{\Omega}(\mathbf{x}, \mathbf{w}) &= (\Omega(\mathbf{x} - \mathbf{w}))^2 \\ &= (\mathbf{x} - \mathbf{w})^T \Omega^T \Omega (\mathbf{x} - \mathbf{w}) \end{aligned} \quad (5)$$

proposed in [50] depending on the mapping matrix $\Omega \in \mathbb{R}^{m \times n}$. Here, m is the mapping 142
dimension usually chosen as $m \leq n$ [8]. Thus, the data are first mapped linearly by Ω 143
and afterwards the Euclidean distance is calculated in the mapping space $\mathbb{R}^m$. 144

The SGDL for training performs an attraction-repelling scheme in the data space as the 145
common feature for all LVQ-schemes [23]: For a given training data $\mathbf{x}_j$ with known class 146
label $c(\mathbf{x}_j)$, the best matching correct prototype $\mathbf{w}^+$ is moved a trifle towards $\mathbf{x}_j$ 147
whereas the best matching incorrect prototype $\mathbf{w}^-$ is slightly repelled. Iterative 148
application of that scheme with random training data selection realizes SGDL for the 149
cost function $E_T[E(\mathbf{x}_k, W)]$. 150

Interestingly, the mapping matrix Ω can also be optimized by SGDL to achieve a better 151
separation of the classes in the mapping space [51]. Note that SGDL for Ω -optimization 152
usually requires a careful regularization technique [49]. The respective algorithm is 153
known as *Generalized Matrix LVQ* (GMLVQ) [50]. 154

After training, the adjusted mapping matrix Ω provides additional information about 155
the data features with respect to the classification task. More precisely, the resulting 156
matrix $\Lambda = \Omega^T \Omega \in \mathbb{R}^{n \times n}$ allows an interpretation as *classification correlation matrix*, 157
i.e. the matrix entries Λ_{ij} reflect the strength of the correlations between the data 158
features i and j contributing to the class discrimination [5, 56]. The diagonal entries Λ_{ii} , 159
which are all non-negative, are the relevances $r_i = \Lambda_{ii}$ delivering the importance of the 160
 i th data feature for class separation: The higher the relevance value r_i the more 161
important is the corresponding feature. Relevance values $r_i \approx 0$ indicate that the 162
respective features are not important for class separation or that the information is 163
already provided by other features [15]. All relevance values r_i are collected in the 164
relevance profile vector $\mathbf{r} = (v_1, \dots, v_n)^T$. The process of relevance vector adaptation is 165
also denoted as *relevance learning*. 166

Further, this approach allows to detect outliers due to the bounded minimum distance 167
principle in contrast to class assignments via decision hyperplanes [12]. Accordingly, 168
outliers are rejected because the model validity cannot guaranteed for these data 169
(*outlier-reject*). Further, data $\mathbf{x}$ with uncertain classification decision can be rejected 170
evaluating the difference between the distance $d(\mathbf{x}, \mathbf{w}_{\omega(W)})$ to the overall winner $\mathbf{w}_{\omega(W)}$ 171
and the distance $d(\mathbf{x}, \omega(W^*))$ with $W^* = W \setminus W_{c(\mathbf{w}_{\omega(W)})}$. In case of a significant 172
deviation, which corresponds to an uncertain decision, the query regarding $\mathbf{x}$ is rejected 173
(*classification-reject*). 174

The Data Set 175

We used exactly the same data set as given in [28], which was generated and analyzed 176
there to detect native and mirror structures. Therefore, we give only a short summary 177
of the data generation and preprocessing, for a more detailed description it is referred 178
to [28]: 179

In order to emulate the necessary procedures in structure modeling from contact maps, 180
the authors derived these maps [17] from a set of 1,305 representative domains of 181

SCOPe superfamilies. Those maps were in turn the base for reconstructing full-atom proteins of different configurations [24, 25, 29, 54], approximately 50 native and 50 mirror models for each domain. Consequently the whole data set comprises 130,500 models, from which the dihedral angles were calculated [10].

Further all of the models in the data set can be grouped into one of the classes **A-G** of SCOPe, representing seven distinctive data sets. Class **A** (*all* α) describes proteins that are predominantly made of α -helices, whereas class **B** (*all* β) mainly consists of β -strands. Proteins with alternations of α -helices and β -strands are put into class **C**. In contrast, proteins in class **D** show segregated sections of α -helices and β -strands. So this four classes are categorized due to secondary structure features of the proteins, whereas the following three are not. The so-called multi-domain proteins are grouped into class **E**. Proteins and peptides located on surfaces or in membranes are grouped into class **F**. Small proteins, that have little or no secondary structure, are grouped into class **G** [18, 28, 31, 48]. In Tab. 1 the sizes of these groups are depicted. All of these models were condensed into the class **ALL**.

Table 1. Numbers of samples for each of the described protein classes (domains) **A – G**

protein class	protein models
A	343
B	233
C	149
D	368
E	21
F	78
G	113

Data Preprocessing and Machine Learning Settings

Each of the aforementioned data sets determines a specific learning task to distinguish mirror and native samples. To do so a relative R-plot histogram vector $\mathbf{x} \in \mathbb{R}^n$ with $n = N \times N$ with $N = 6$ as grid resolution was extracted for each sample. These data vectors served as training data for GMLVQ according to the considered tasks.

For each learning task **A-G**, and **ALL** we trained a separate GMLVQ model with three prototypes per class (mirror/native). Additionally to the prototypes we also adapted the mapping matrix Ω with the mapping dimension $m = 36$. The reported classification results are achieved as the averaged test performances obtained by 50 independent runs. Each run was done as a five-fold cross-validation procedure.

Further, for visual inspection and evaluation, we generated a *summarized R-plot*, which is collecting all pairs (Φ, Ψ) of dihedral angles for all samples of the considered task in a single R-plot but separately for native and mirror samples. Thus, these R-plots can be seen as estimated dihedral angles densities in the (Φ, Ψ) -plane for mirror and native samples, allowing a visual inspection of those (Φ, Ψ) -distributions, see S2 Fig–S4 Fig.

Results and Discussion

By taking advantage of the inherent interpretable nature of GMLVQ, particularly by considering the relevances of the data features, we can extract insightful knowledge regarding the classification decision. Especially the connection of the relevant features

with the favored and allowed regions in the R-plots for each protein class offers a biological interpretation. Tab. 2 summarizes the areas of the R-Plot of each protein class which are considered most relevant by the model.

As previous publications suggest [28, 55], native and mirror conformations of proteins rich in helices (belonging to class **A**), can be distinguished based on chirality derived information, or in more detail: right-handed α -helices are favoured in native conformations [40]. We can confirm this finding, since the features corresponding to the left- and right-handed α -helices predominantly contribute to the class discrimination of mirror and native (see S3 Fig and Tab. 2). The accuracy for this class is 86.57% (see Tab. 3 for all accuracies). Interestingly, our model achieves an accuracy of 92.56% for class **B**, which obviously collides with the statement of those native and mirror models being indistinguishable [28, 55] and, furthermore, even exceeds the accuracy of class **A**. The relevant features for class discrimination are those corresponding to β -strands in the R-plot (see S3 Fig). In detail, the important underlying secondary structures in this case might be the right-handed triple helices (collagen) and parallel β -strands [26]. However, the confirmation of the actual underlying secondary structure as well as its relation to chirality are still pending.

As protein classes **C** and **D** structurally show a combination of the aforementioned two classes *all-alpha* and *all-beta*, the relevant features for class discrimination also do. Class **C** shows the best of all investigated accuracies. This result concurs with the findings in [28]. Among the protein classes which were not categorized due to their secondary structure, the multi-domain class **E** by far shows the best accuracy with 91.8%, whereas classes **F** and **G** do not exceed 80%. Even though class **E** has got such a high accuracy it has to be treated with care, since there has not been enough data for this protein class. The relatively poor accuracy for class **F** is most likely due to the fact that membrane proteins propose some difficulties in structure elucidation [34, 42, 64] and this results in low resolutions [36]. A poor resolution in turn may lead to inaccurate atom coordinates. That means, the calculations of the dihedral angles cannot be correct either and, therefore, complicate the classification. As for class **G** the obtained low accuracy is probably due to the fact that small proteins do not have that many amino acids, less than 100 [35, 53], and therefore show less α -helices or β -strands than other proteins.

The R-plots for classes **E-G** together with the corresponding relevance profiles are depicted in S3 Fig .

In order to assess the models suitability for a more general problem we considered all protein classes for training as well as for testing and achieved an overall accuracy of 88.09%. Pursuing this approach, which is more general and more considerable for application, we investigated the behavior of our model by training with all protein classes but testing with only one protein class at a time. The achieved accuracies are in good agreement with those of the single classes as Tab. 3 shows.

Conclusion

In the present contribution we offer a valid approach for distinguishing mirror and native conformations of proteins based on structure information. The approach is based on the evaluation of the respective R-plots by means of an interpretable machine learning model. This model, the Generalized Learning Matrix Vector Quantizer, is known to be robust and highly interpretable according to the underlying reference principle. Moreover, according to the integrated relevance learning metric adaptation, the approach provides beside the classification ability additional knowledge regarding

Table 2. Most relevant cells for differentiation of mirror and native models based on the corresponding relevance profiles.

protein class	cells for α -helices	cells for β -strands
ALL	(1,4), (2,4), (4,1)	(1,2), (2,1)
class A	(1,3), (1,4), (2,3), (2,4), (4,1), (4,2)	—
class B	—	(1,2), (2,1)
class C	(1,4), (2,4), (4,1), (4,2)	(2,1)
class D	(1,4), (2,4), (4,1), (4,2)	(1,2), (2,1)
class E	(1,4), (2,4), (4,1), (4,2)	(2,1)
class F	(1,4), (2,4), (4,1), (4,2)	—
class G	(2,4)	(1,2), (2,1)

Table 3. Obtained averaged (test) accuracies (A), obtained accuracies using the general model for specific classes together with their respective standard deviations for the protein classes (results refer to 50 runs of cross-validated GMLVQ). For comparison, additionally the best classification results A_K from [28] are given (for all classes as average of the others)

protein class/results	A	σ_A	$A_{overall}$	$\sigma_{A_{overall}}$	A_K
ALL	88.09%	0.03	88.09%	0.03	69%
class A	86.57%	1.48	85.34%	0.03	71%
class B	92.56%	1.23	90.78%	0.03	76%
class C	98.50%	0.70	98.46%	0.01	78%
class D	93.04%	0.39	92.76%	0.01	70%
class E	91.80%	8.92	92.42%	0.03	70%
class F	65.84%	3.03	67.38%	0.03	60%
class G	75.45%	2.48	75.50%	0.01	61%

the classification decision. In the context of R-plot analysis this information consists in a weighting of importance of R-plot regions regarding best mirror-native-separability.

KURCZYNSKA AND KOTULSKA state that structural features are no discriminatory property of native and mirror models. Although information regarding chirality can be used to differentiate models rich in α -helices, according to Vendruscolo et al. this does not hold for all- β structures [55]. However, we were able to show that a discrimination of native and mirror models using structural features is indeed possible.

The GMLVQ classifier achieves high separation accuracies for all protein classes except class **F** and **G**. At least for the latter one, acceptable results are obtained. In fact, the resulted accuracies for protein classes **F** – **G** show that a distinction of mirror and native structures by means of R-plots is possible with high precision and sensitivity. The interpretable model offers additional insights: In particular, the relevance profiles, weighting the regions like α -helices and β -strands of R-plots for mirror-native-discrimination, differ for the considered protein classes. The obtained relevance profiles are in good agreement with respective biological knowledge about protein structure chirality. at least for the considered data set.

Thus, the presented approach offers a successful alternative to the statistical approach based on energy levels as proposed in [28] and emphasize the importance of R-plots for structural analysis of proteins as already mentioned in [3]. Along this line, also data processing is easier in the present approach compared to the complex calculations of the energy levels [2].

Further investigation should include to improve the classification performance by finer

cell resolutions of R-plots. However, this requires more training data for sufficient learning stability. Further, reject option strategies should be included to detect outliers.

Acknowledgments

The authors greatly acknowledge helpful discussions within the CI-research group of the SICIM at the UAS Mittweida.

Supporting Information

J.A., K.S.B., and M.K. were supported by a grant for a so-called *Young Researcher Group* provided by the European Social Fund (ESF).

S1 Fig. Generic Ramachandran plot (R-plot) for dihedral angles. The allowed/favored regions of secondary structures are shown. Particularly, we indicate the areas of β -sheets/-strands (β), as well as right-handed α -helices (α) and left-handed α -helices (L_α). Additionally we have added a 6×6 cell array to the R-plot, so that we can address regions (cells) of the R-plot accordingly.

S2 Fig. Summarized R-plots and respective cell relevance for mirror/native models for learning of all data (ALL). The summarized R-plots for native (left) and mirror (right) samples. The plots are estimators for the dihedral angles densities in the (Φ, Ψ) -plane. The relevance profile vector $\mathbf{r} = (r_1, \dots, r_{36})^T$ is depicted in the middle, arranged accordingly to the cells. The higher (darker) the r_i -value the more important is the respective cell for AMND. The achieved GMLVQ classification accuracy for this data set is 88.09%.

S3 Fig. Summarized R-plots and respective cell relevance for mirror/native models for learning of the classes A–D. The summarized R-plots for native (left) and mirror (right) samples. The plots are estimators for the dihedral angles densities in the (Φ, Ψ) -plane for samples. The obtained relevance profile vectors $\mathbf{r} = (r_1, \dots, r_{36})^T$ from relevance learning in GMLVQ are depicted in the middle, arranged accordingly to the cells of the R-plot. The higher (darker) the r_i -value the more important is the respective cell for AMND. The achieved GMLVQ classification accuracies for these data sets are given in Tab. 3.

S4 Fig. Summarized R-plots and respective cell relevance for mirror/native models for learning of the class E–F. The summarized R-plots for native (left) and mirror (right) samples. The plots are estimators for the dihedral angles densities in the (Φ, Ψ) -plane for samples. The obtained relevance profile vectors $\mathbf{r} = (r_1, \dots, r_{36})^T$ from relevance learning in GMLVQ are depicted in the middle, arranged accordingly to the cells of the R-plot. The higher (darker) the r_i -value the more important is the respective cell for AMND. The achieved GMLVQ classification accuracies for these data sets are given in Tab. 3.

S5 Tab. As an add-on we provide in Tab. 4 the sensitivities as well as the specificities corresponding to the reported accuracies from Tab. 3.

Table 4. Overall model (test) sensitivities (Se) and specificities (Sp) together with their respective standard deviations for the protein classes (results refer to 50 runs of cross-validated GMLVQ)

protein class / results	Se	σ_{Se}	Sp	σ_{Sp}
ALL	87.38%	0.09	88.77%	0.09
class A	83.31%	0.12	87.48%	0.06
class B	91.70%	0.21	89.93%	0.13
class C	98.59%	0.01	98.32%	0.03
class D	92.01%	0.06	93.43%	0.03
class E	88.51%	0.08	96.33%	0.04
class F	65.74%	0.14	69.00%	0.08
class G	75.66%	0.15	75.35%	0.13

Data Availability

The repository can be found at <https://www.ebi.ac.uk/biostudies/studies/S-BSST81> or <http://comprec-lin.iar.pwr.edu.pl/mirrorModels/>.

References

1. S. C. Ahalt, A. K. Krishnamurty, P. Chen, and D. E. Melton. Competitive learning algorithms for vector quantization. *Neural Networks*, 3(3):277–290, 1990. (document)
2. R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O’Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack, R. Das, D. Baker, B. Kuhlman, T. Kortemme, and J. J. Gray. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation*, 13(6):3031–3048, 2017. (document)
3. R. Ayoub and Y. Lee. Rupee: A fast and accurate purely geometric protein structure search. *PLoS ONE*, 14(3):1–17, 2019. (document)
4. M. Biehl, K. Bunte, and P. Schneider. Analysis of flow cytometry data by matrix relevance learning vector quantization. *PLoS ONE*, 8(3):e59401, 2013. (document)
5. M. Biehl, B. Hammer, and T. Villmann. Prototype-based models in machine learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, (2):92–111, 2016. (document)
6. S. Bittrich, M. Kaden, C. Leberecht, F. Kaiser, T. Villmann, and D. Labudde. Application of an interpretable classification model on Early Folding Residues during protein folding. *BioData Mining*, 12(1):16, dec 2019. (document)
7. S. Bittrich, M. Kaden, C. Leberecht, F. Kaiser, T. Villmann, and D. Labudde. Application of an interpretable classification model on early folding residues during protein folding. *BioData Mining*, 12(1), 2019. (document)
8. K. Bunte, P. Schneider, B. Hammer, F.-M. Schleif, T. Villmann, and M. Biehl. Limited rank matrix learning, discriminative dimension reduction and visualization. *Neural Networks*, 26(1):159–173, 2012. (document)

9. S. Chaudhury, S. Lyskov, and J. J. Gray. PyRosetta: A script-based interface for
implementing molecular modeling algorithms using Rosetta. *Bioinformatics*,
26(5):689–691, 2010. (document)
10. P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke,
I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. De Hoon.
Biopython: Freely available Python tools for computational molecular biology
and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009. (document)
11. J. M. Duarte, R. Sathyapriya, H. Stehr, I. Filippis, and M. Lappe. Optimal
contact definition for reconstruction of Contact Maps. *BMC Bioinformatics*,
11(283):436–440, 2010. (document)
12. L. Fischer, B. Hammer, and H. Wersing. Efficient rejection strategies for
prototype-based classification. *Neurocomputing*, 169:334–342, 2015. (document)
13. I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.
(document)
14. S. Graf and H. Lushgy. *Foundations of Quantization for Probability Distributions*,
volume 1730 of *Lect. Notes in Mathematics*. Springer, Berlin, 2000. (document)
15. B. Hammer and T. Villmann. Generalized relevance learning vector quantization.
Neural Networks, 15(8-9):1059–1068, 2002. (document)
16. T. F. Havel and M. E. Snow. A new method for building protein conformations
from sequence alignments with homologues of known structure. *Journal of*
Molecular Biology, 217(1):1–7, 1991. (document)
17. H. K. Ho, M. J. Kuiper, and R. Kotagiri. PConPy - A Python module for
generating 2D protein maps. *Bioinformatics*, 24(24):2934–2935, 2008. (document)
18. T. J. Hubbard, B. Ailey, S. E. Brenner, A. G. Murzin, and C. Chothia. SCOP: A
structural classification of proteins database. *Nucleic Acids Research*,
27(1):254–256, 1999. (document)
19. M. Kaden, M. Lange, D. Nebel, M. Riedel, T. Geweniger, and T. Villmann.
Aspects in classification learning - Review of recent developments in Learning
Vector Quantization. *Foundations of Computing and Decision Sciences*,
39(2):79–105, 2014. (document)
20. M. Kaden, M. Riedel, W. Hermann, and T. Villmann. Border-sensitive learning
in generalized learning vector quantization: an alternative to support vector
machines. *Soft Computing*, 19(9):2423–2434, 2015. (document)
21. S. B. Kent. Novel protein science enabled by total chemical synthesis. *Protein*
Science, 28(2):313–328, 2019. (document)
22. T. Kohonen. Learning Vector Quantization. *Neural Networks*, 1(Supplement
1):303, 1988. (document)
23. T. Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information*
Sciences. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997).
(document)
24. B. M. Konopka, M. Ciombor, M. Kurczynska, and M. Kotulska. Automated
procedure for contact-map-based protein structure reconstruction. *Journal of*
Membrane Biology, 247(5):409–420, 2014. (document)

25. G. G. Krivov, M. V. Shapovalov, and R. L. Dunbrack. Improved prediction of
protein side-chain conformations with SCWRL4. *Proteins: Structure, Function
and Bioinformatics*, 77(4):778–795, 2009. (document)
26. P. Kumar and A. Arya. Ramachandran plot - A simplified approach, 2019.
(document)
27. M. Kurczynska, E. Kania, B. M. Konopka, and M. Kotulska. Applying PyRosetta
molecular energies to separate properly oriented protein models from mirror
models, obtained from contact maps. *Journal of Molecular Modeling*,
22(111):1–10, 2016. (document)
28. M. Kurczynska and M. Kotulska. Automated method to differentiate between
native and mirror protein models obtained from contact maps. *PLoS ONE*,
13(5):1–19, 2018. (document), 3
29. Y. Li and Y. Zhang. REMO: A new protocol to refine full atomic protein models
from C-alpha traces by optimizing hydrogen-bonding networks. *Proteins:
Structure, Function and Bioinformatics*, 76(3):665–676, 2009. (document)
30. J. Ling, C. Fan, H. Qin, M. Wang, J. Chen, P. Wittung-Stafshede, and T. Zhu.
Mirror-Image 5S Ribonucleoprotein Complexes. *Angewandte Chemie
International Edition*, 59, 2019. (document)
31. L. Lo Conte. SCOP: a Structural Classification of Proteins database. *Nucleic
Acids Research*, 28(1):257–259, 2000. (document)
32. O. Lund, J. Hansen, S. Brunak, and J. Bohr. Relationship between protein
structure and geometrical constraints. *Protein Science*, 5(11):2217–2225, 1996.
(document)
33. R. G. C. Maack, C. Gillmann, and H. Hagen. Uncertainty-aware ramachandran
plots. *IEEE Pacific Visualization Symposium*, 2019-April(May):227–231, 2019.
(document)
34. J. Martin and A. Sawyer. Elucidating the structure of membrane proteins.
BioTechniques, 66(4):167–170, 2019. (document)
35. S. Miravet?Verde, T. Ferrar, G. Espadas?García, R. Mazzolini, A. Gharrah,
E. Sabido, L. Serrano, and M. Lluch?Senar. Unraveling the hidden universe of
small proteins in bacterial genomes. *Molecular Systems Biology*, 15(2):1–17, 2019.
(document)
36. I. Moraes, G. Evans, J. Sanchez-Weatherby, S. Newstead, and P. D. Stewart.
Membrane protein structure determination - The next generation. *Biochimica et
Biophysica Acta - Biomembranes*, 1838(1 PARTA):78–87, 2014. (document)
37. G. P. Moss. Basic terminology of stereochemistry (IUPAC Recommendations
1996). *International Union of Pure and Applied Chemistry*, 68(12):2193–2222,
1996. (document)
38. M. N. Murty and V. S. Devi. *Pattern Recognition*. Springer, London, 2011.
(document)
39. J. K. Noel, A. Schug, A. Verma, W. Wenzel, A. E. Garcia, and J. N. Onuchic.
Mirror images as naturally competing conformations in protein folding. *Journal
of Physical Chemistry B*, 116(23):6880–6888, 2012. (document)

40. M. Novotny and G. J. Kleywegt. A survey of left-handed helices in protein structures. *Journal of Molecular Biology*, 347(2):231–241, 2005. (document)
41. M. Peplow. A Conversation with Ting Zhu. *ACS Central Science*, 4(7):783–784, jul 2018. (document)
42. G. Postic, Y. Ghouzam, V. Guiraud, and J. C. Gelly. Membrane positioning for high- and low-resolution protein structures through a binary classification approach. *Protein Engineering, Design and Selection*, 29(3):87–91, 2015. (document)
43. G. N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations, 1963. (document)
44. H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Stat.*, 22:400–407, 1951. (document)
45. W. Samek, G. Monath, A. Vedaldi, L. Hansen, and K.-R. Müller, editors. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, number 11700 in LNAI. Springer, 2019. (document)
46. S. Saralajew, L. Holdijk, M. Rees, and T. Villmann. Robustness of generalized learning vector quantization models against adversarial attacks. In A. Vellido, K. Gibert, C. Angulo, and J. Guerrero, editors, *Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization – Proceedings of the 13th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization, WSOM+2019, Barcelona*, volume 976 of *Advances in Intelligent Systems and Computing*, pages 189–199. Springer Berlin-Heidelberg, 2019. (document)
47. A. Sato and K. Yamada. Generalized learning vector quantization. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference*, pages 423–9. MIT Press, Cambridge, MA, USA, 1996. (document)
48. R. D. Schaeffer and V. Daggett. Protein folds and protein folding. *Protein Engineering, Design and Selection*, 24(1-2):11–19, 2011. (document)
49. P. Schneider, K. Bunte, H. Stiekema, B. Hammer, T. Villmann, and M. Biehl. Regularization in matrix relevance learning. *IEEE Transactions on Neural Networks*, 21(5):831–840, 2010. (document)
50. P. Schneider, B. Hammer, and M. Biehl. Adaptive relevance matrices in learning vector quantization. *Neural Computation*, 21:3532–3561, 2009. (document)
51. P. Schneider, B. Hammer, and M. Biehl. Distance learning in discriminative vector quantization. *Neural Computation*, 21:2942–2969, 2009. (document)
52. T. N. Schumacher, L. M. Mayr, D. L. Minor, M. A. Milhollen, M. W. Burgess, and P. S. Kim. Identification of D-peptide ligands through mirror-image phage display. *Science*, 271(5257):1854–1857, mar 1996. (document)
53. M. Su, Y. Ling, J. Yu, J. Wu, and J. Xiao. Small proteins: Untapped area of potential biological importance, 2013. (document)
54. M. Vassura, L. Margara, P. Di Iena, F. Medri, P. Fariselli, and R. Casadio. FT-COMAR: Fault tolerant three-dimensional structure reconstruction from protein contact maps. *Bioinformatics*, 24(10):1313–1315, 2008. (document)

55. M. Vendruscolo, E. Kussell, and E. Domany. Recovery of protein structure from contact maps. *Folding and Design*, 2(5):295–306, 1997. (document)
56. T. Villmann, A. Bohnsack, and M. Kaden. Can learning vector quantization be an alternative to SVM and deep learning? *Journal of Artificial Intelligence and Soft Computing Research*, 7(1):65–81, 2017. (document)
57. T. Villmann, J. Ravichandran, A. Villmann, D. Nebel, and M. Kaden. Investigation of activation functions for Generalized Learning Vector Quantization. In A. Vellido, K. Gibert, C. Angulo, and J. Guerrero, editors, *Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization – Proceedings of the 13th International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization, WSOM+2019, Barcelona*, volume 976 of *Advances in Intelligent Systems and Computing*, pages 179–188. Springer Berlin-Heidelberg, 2019. (document)
58. T. Villmann, S. Saralajew, A. Villmann, and M. Kaden. Learning vector quantization methods for interpretable classification learning and multilayer networks. In C. Sabourin, J. Merelo, A. Barranco, K. Madani, and K. Warwick, editors, *Proceedings of the 10th International Joint Conference on Computational Intelligence (IJCCI), Sevilla*, pages 15–21, Lissabon, Portugal, 2018. SCITEPRESS - Science and Technology Publications, Lda. ISBN: 978-989-758-327-8. (document)
59. Z. Wang, W. Xu, L. Liu, and T. F. Zhu. A synthetic molecular system capable of mirror-image genetic replication and transcription. *Nature Chemistry*, 8(7):698–704, 2016. (document)
60. J. Weidmann, M. Schnölzer, P. E. Dawson, and J. D. Hoheisel. Copying Life: Synthesis of an Enzymatically Active Mirror-Image DNA-Ligase Made of D-Amino Acids. *Cell Chemical Biology*, 26(5):645–651.e3, 2019. (document)
61. M. T. Weinstock, M. T. Jacobsen, and M. S. Kay. Synthesis and folding of a mirror-image enzyme reveals ambidextrous chaperone activity. *Proceedings of the National Academy of Sciences of the United States of America*, 111(32):11679–11684, 2014. (document)
62. T. O. Yeates and S. B. Kent. Racemic Protein Crystallography. *Annual Review of Biophysics*, 41:41–61, jun 2012. (document)
63. L. Zhao and W. Lu. Mirror image proteins. *Current opinion in chemical biology*, 22:56–61, 2014. (document)
64. C. Zhou, Y. Zheng, and Y. Zhou. Structure prediction of membrane proteins. *Genomics, proteomics & bioinformatics / Beijing Genomics Institute*, 2(1):1–5, 2004. (document)

R-plot