Artificial Intelligence in the Life Sciences 1 (2021) 100016

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

# Fiscore package: Effective protein structural data visualisation and exploration

Auste Kanapeckaite

*School of Pharmacy, University of Reading, Hopkins Building, Reading RG6 6UB, United Kingdom*

A B S T R A C T

The lack of bioinformatics tools to quickly assess protein conformational and topological features motivated to create an integrative and user-friendly R package. Moreover, the *Fiscore* package implements a pipeline for Gaussian mixture modelling making such machine learning methods readily accessible to non-experts. This is especially important since probabilistic machine learning techniques can help with a better interpretation of complex biological phenomena when it is necessary to elucidate various structural features that might play a role in protein function. Thus, *Fiscore* builds on the mathematical formulation of protein physicochemical properties that can aid in drug discovery, target evaluation, or relational database building. In addition, the package provides interactive environments to explore various features of interest. Finally, one of the goals of this package was to engage structural bioinformaticians and develop more robust and free R tools that could help researchers not necessarily specialising in this field. Package *Fiscore* (v.0.1.3) is distributed free of charge via CRAN and Github.

## 1. Introduction

*Fiscore* R package was developed to quickly take advantage of protein topology/conformational feature assessment and perform various analyses allowing a seamless integration into relational databases as well as machine learning pipelines [1]. The package builds on protein structure and topology studies which led to the derivation of the Fi-score equation capturing protein dihedral angle and B-factor influence on amino acid residues (Eqs. (1) and (2)) [1]. The introduced tools can be very beneficial in rational therapeutics development where successful engineering of biologics, such as antibodies, relies on the characterisation of potential binding or contact sites on target proteins [1,2]. Moreover, translating structural data into scores can help with target classification, target-ligand information storage, screening studies, or integration into machine learning pipelines [1,2]. As a result, Fi-score, a first-of-its-kind *in silico* protein fingerprinting approach, created a premise for the development of a specialised and freely distributed R package to assist with protein studies and new therapeutics development [1].

*Fiscore* package allows capturing dihedral angle and B-factor effects on protein topology and conformation. Since these physicochemical characteristics could help with the identification or characterisation of a binding pocket or any other therapeutically relevant site, it is important to extract and combine data from structural files to allow such information integration [1,3,4]. Protein dihedral angles were selected as they contain information on the local and global protein structural features where protein backbone conformation can be highly accurately recreated based on the associated dihedral angles [1,4]. Furthermore, since Ramachandran plot, which provides a visualisation for dihedral angle distributions, namely $\phi$ (phi) and $\psi$ (psi), allows only a holistic description of conformation and cannot be integrated with traditional parametric or non-parametric density estimation methods, a specific transformation was required to use this data. An additional parameter, specifically the oscillation amplitudes of the atoms around their equilibrium positions (B-factors) in the crystal structures, was also used. B-factors encompass a lot of information on the overall biomolecule structure; for example, these parameters depend on conformational disorder, thermal motion paths, and the rotameric state of amino acids side-chains. B-factors also show dependence on the three-dimensional structure as well as protein flexibility [1,4]. Normalised dihedral angles (standard deviation scaling to account for variability and distribution) and scaled B-factors (min-max scaling) (Eq. (1)) were integrated into the Fi-score equation (Eq. (2)). It is important to highlight that B-factors need to be scaled so that different structural files can be compared and that the dihedral angle normalisation transforms angular data into adjusted values based on the overall variability [1]. Thus, combining dihedral angle and B-factor values into a single parameter provides a way to extract information on individual residues, residue clusters, motifs, and structural features. This information can be efficiently transferred into machine learning to detect data characteristics not easily identifiable otherwise.

$$B_{i-norm} = \frac{B_i - B_{\min}}{B_{\max} - B_{\min}}$$

**Equation 1**. Min-max normalisation and scaling of B-factors where $B_{i-norm}$ is a scaled B-factor, $B_i$ - B-factor for a selected $C_\alpha$ atom in a chain, Bmax - the largest B-factor value for all $C_\alpha$ B-factors in a protein, Bmin - the smallest B-factor value for all $C_\alpha$ B-factors in a protein. B-factor normalisation is based on the full length protein.

$$Fi_{score} = \frac{1}{N} \sum_i \frac{\phi_i \psi_i}{\sigma_{\phi_i} \sigma_{\psi_i}} B_{i-norm}$$

**Equation 2**. Fi-score evaluation where N is the total number of atoms for which dihedral angle information is available, $\phi$ and $\psi$ values represent dihedral angles for a specific $C_\alpha$ atom, $\sigma_{\phi_i}$ and $\sigma_{\psi_i}$ represent cor-

*E-mail address:* auste.kan@algorithm379.com

https://doi.org/10.1016/j.ailsci.2021.100016
Received 16 November 2021; Accepted 18 November 2021
Available online 26 November 2021

responding standard deviations for the torsion angles and $B_{i-norm}$ is a normalised B- factor value for the $C_\alpha$ atom. B-factor, $\sigma_{\phi_i}$ and $\sigma_{\psi_i}$ normalisation is based on the full length protein.

In order to identify meaningful clusters based on the structural complexity, Gaussian mixture models (GMM) were selected as a primary machine learning classifier [1]. The strength of GMM lies in the probabilistic model nature since all data points are assumed to be derived from a mixture of a finite number of Gaussian distributions with unknown parameters [1,5]. Consequently, the soft classification of GMM where a data point has a probability of belonging to a cluster is much more suitable to assess biological parameters compared to other hard classification techniques in machine learning, such as k-means, which provide only a strict separation between classes. GMM pipeline offers a number of benefits to categorise protein structural features and the information can be used to explore amino acid grouping based on their physicochemical parameters. The designed GMM implementation takes care of the information criterion assessment to fine tune the number of clusters for modelling and predicts the best suited model for the expectation-maximisation (EM) algorithm to maximise the likelihood of data point assignments [1,5]. As a result, protein residues can be grouped based on their Fi-scores where this information can be used to identify emerging patterns in the protein conformation or topology.

Nur77 protein was used as a case example to demonstrate various package functionalities. Nuclear receptor subfamily 4 group A member 1 (NR4A1), also known as Nur77/TR3/NGFIB, is a member of the nuclear receptor superfamily and regulates the expression of multiple target genes [6]. This nuclear receptor is classified as an orphan receptor since there are no known endogenous ligands. Nur77 has the typical structure of a nuclear receptor which consists of N-terminal, DNA binding, and ligand-binding domains. This regulatory protein plays many potentially therapeutically relevant roles regulating cell proliferation and apoptosis [6]. Consequently, the Nur77 protein is an excellent example to highlight how in-depth structural analysis and classification could be used in better understanding protein functions and finding druggable binding sites or identifying ligands.

Based on the need to develop integratable and specialised tools for protein analyses, the *Fiscore* package was developed to assist with a wide spectrum of research questions ranging from exploratory analyses to therapeutic target assessment (Fig. 1). The introduced set of new tools provides an interactive exploration of targets with an easy integration into downstream analyses. Importantly, the package and associated tools are written to be easy to use and freely available facilitating analyses for non-specialists in structural biology or machine learning.

## 2. Methods

*Fiscore* package architecture is divided into exploratory and advanced functions (Fig. 1). Several key packages, such as ggplot2 [7], Bio3D [8], plotly [9], and mclust [10], are also employed to create an easy-to-use analytical environment where a user-friendly machine learning pipeline of GMM [1] allows for a robust structural analysis. GMM implementation is designed to include the optimal cluster number evaluation (Bayesian information criterion; BIC), automatic model fitting in the EM phase of clustering, model-based hierarchical clustering, density estimation, as well as discriminant analysis [1,11]. Researchers also have an option to perform advanced exploratory studies or integrate the package into their development pipelines. *Fiscore* also takes care of raw data pre-processing and evaluation with optional settings to adjust how the analyses are performed. The package was built using functional programming principles with several R S3 methods to create objects for PDB files [12]. *Fiscore* is accompanied by documentation and vignette files to help the users with their analyses [11]. Since PDB files are typically large, the documentation provides a compressed testing environment as well as a detailed tutorial. Additional visualisations were generated with PyMol [13]. Proteins were retrieved from the Protein Data Bank database [14]. Protein sequence alignments were performed with PSI-BLAST using default parameters and a single iteration [15]. Hydrophobicity plots for Nur77 functional analysis were generated with the following parameters: window = 15,weight = 25, model="exponential". Student's *t*-test (two-sided, unpaired, sig. level=95%) was performed in the R programming environment.

## 3. Results

### 3.1. Data preparation

The workflow begins with the PDB file pre-processing and preparation. The user should also generally assess if the structure is suitable for the analysis; that is, the crystallographic data provides a good resolution and there are no or a minimal number of breakages within the reported structure. Function *PDB_process* takes a PDB file name which can be expressed as 6KZ5.pdb or path/to/the/file/6KZ5.pdb. One of the function's dependencies is package Bio3D [8], this useful package provides several tools to begin any PDB file analysis. In addition, the *PDB_process* function can take a path parameter which can point to a directory where to split PDB files into separate chain files (necessary for the downstream analysis). If this option is left empty, a folder in the working directory will be created automatically. If the user splits multiple PDB files in a loop, they will be continuously added to the same folder. After the processing, the function *PDB_process* returns a list of split chain names. It is important to highlight that PDB files need to be split for the downstream processing so that separate chains can be analysed independently.

After a file or files are pre-processed the function *PDB_prepare* can be used to prepare a PDB file to generate Fi-score and normalised B-factor values as well as secondary structure designations. The function takes a PDB file name that was split into separate chains, e.g., 6KZ5_A.pdb, where a letter designates a split chain. The file is then cleaned and only the complete entries for amino acids are kept for the analysis, i.e., amino acids from the terminal residues that do not contain both dihedral angles are removed. The function returns a data frame with protein secondary structure information '*Type*', Fi-score values per residue '*Fi_score*', as well as normalised B-factor values for each amino acid $C_\alpha$ '*B_normalised*' (Fig. 2). Extracting protein secondary structure information, i.e., '*Type*', helps to prepare a data object so that the information about a target can be supplied into cheminformatics or other bioinformatics pipelines where structural features are important to assess protein sites and amino acid composition. These features are new and extend structural file exploration possibilities compared to, for example, other software packages, such as Bio3D [8].

Function calls are simple and user-friendly:

```
#General function for pre-processing raw PDB files

pdb_df<-PDB_process(pdb_path)

#Cleaning and preparation of PDB file

pdb_df<-PDB_prepare(pdb_path)

#Explore the output

head(pdb_df)

#The package allows to call test data directly
  for the Nur77 example file

pdb_path<- system.file("extdata", "6kz5.pdb", package="Fiscore")
```

### 3.2. Exploratory analyses

The scope of the exploratory analyses provides options to evaluate physicochemical parameters, such as dihedral angles, B-factors, or hydrophobicity scores, and visualise their distribution (Fig. 1).

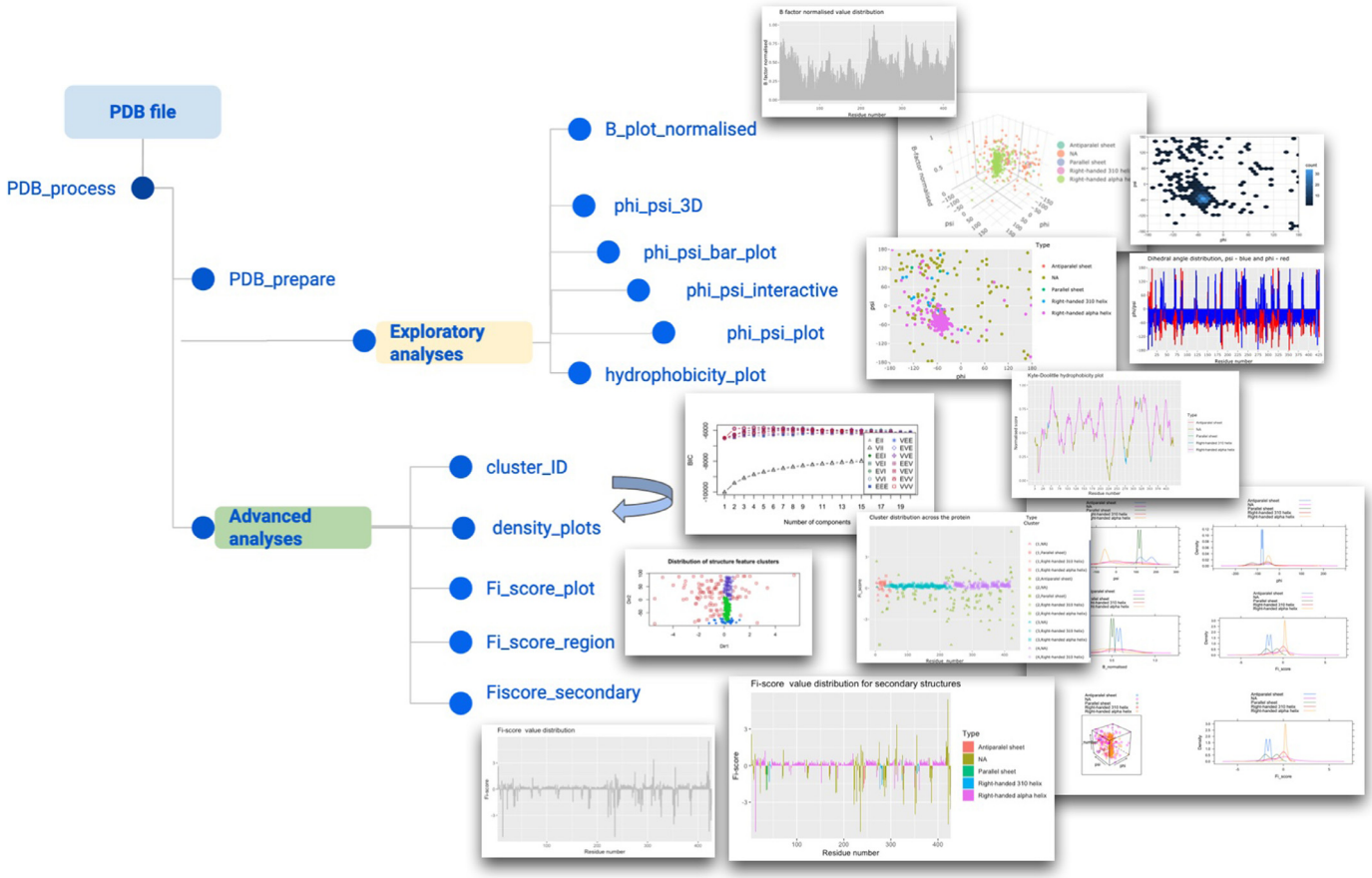


**Fig. 1.** Schematic visualisation of the package features.

```
##                phi       psi       chi1      chi2 chi3 chi4 chi5 df_resno
##  31.A.ALA  54.94701  53.80667         NA        NA   NA   NA   NA       31
##  32.A.ASN -60.91976 -18.01379 -157.93838 -76.10748   NA   NA   NA       32
##  33.A.LEU -64.18792 -45.94048  176.74103  46.17107   NA   NA   NA       33
##  34.A.LEU -65.44501 -38.46630  -88.61643 158.49518   NA   NA   NA       34
##  35.A.THR -67.68541 -43.43184   75.18019        NA   NA   NA   NA       35
##  36.A.SER -76.88914 -17.40880  157.45159        NA   NA   NA   NA       36
##           df_res B_factor B_normalised   Fi_score                   Type
##  31.A.ALA    ALA    30.07   0.35506724 0.37663226 Right-handed alpha helix
##  32.A.ASN    ASN    15.68   0.18227666 0.07176640 Right-handed alpha helix
##  33.A.LEU    LEU     7.79   0.08753602 0.09261096 Right-handed alpha helix
##  34.A.LEU    LEU     9.85   0.11227185 0.10140389 Right-handed alpha helix
##  35.A.THR    THR    20.79   0.24363593 0.25696344 Right-handed alpha helix
##  36.A.SER    SER    23.69   0.27845821 0.13372748 Right-handed alpha helix
```

**Fig. 2.** PDB file processing output.

Basic analyses are accessed through simple function calls to explore how dihedral angles and B-factors are distributed. These analyses offer interactive and easy visualisations of key parameters that are currently not offered in any other package. For example, while Bio3D [8] has many useful functionalities for the exploration of PDB files, '*Fi_score*' extends exploratory analyses by allowing a simplified and in-depth look into the key physicochemical parameters, such as B-factor value visualisation or generation of Ramachandran plots. Similarly, other freely available tools (distributed as an online service), such as Expasy ProtScale [16,17], provide only one dimensional assessment without incorporating structural features and do not process PDB files. '*Fi_score*', however, combines sequence, structural, and physicochemical analyses in simple function calls to quickly explore the user data.

```
#Calling a Ramachandran plot function

phi_psi_plot(pdb_df)
#Visualisation of dihedral angle juxtaposed distributions

phi_psi_bar_plot(pdb_df)

#B plot value visualisation

B_plot_normalised(pdb_df)

#Interactive plot to map amino acids via 2D distribution
```

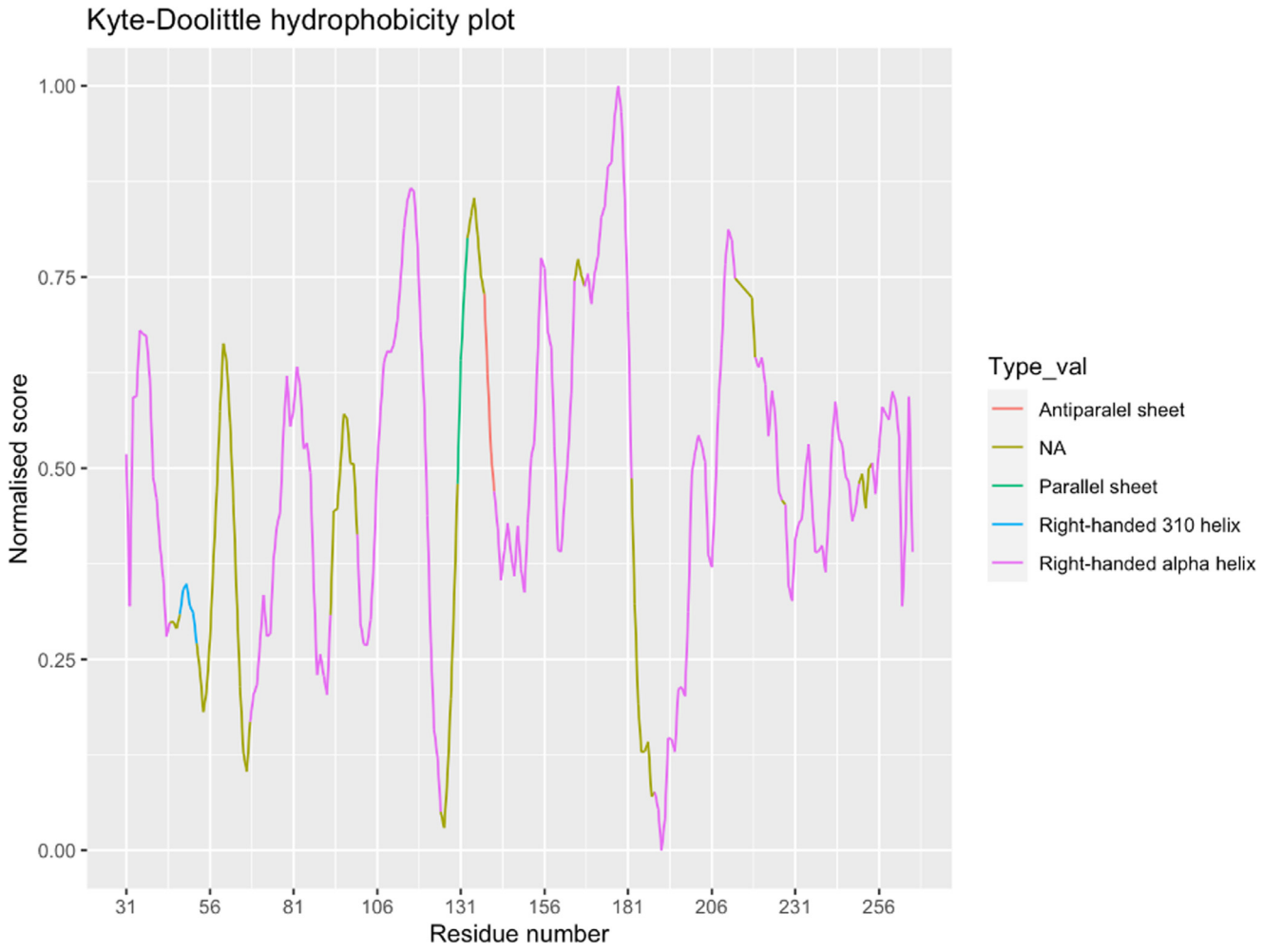


**Fig. 3.** Hydrophobicity plot with secondary structure superimposition.

```
#to precisely see what parameters an individual amino acid has

phi_psi_interactive(pdb_df)

#3D visualisation of dihedral angles and B-factor values

phi_psi_3D(pdb_df)
```

An especially useful functionality is the hydrophobicity visualisation with the superimposed secondary structure elements. To the author's knowledge, there are currently no tools implementing such a visualisation (Fig. 3). In contrast to Expasy ProtScale [16,17], it is possible to visualise hydrophobicity values and their corresponding secondary structure elements as extracted from the PDB file. Such an assessment provides a direct way to compare structural features based on their affinity to water. This can be very helpful in evaluating or predicting potential binding sites as well as bioengineering new proteins.

The package provides an easy to use wrapper:

```
#Alternatively an exponential model can be selected

hydrophobicity_plot(pdb_df,window = 9,weight = 25,model = "exponential")
```

The nuclear receptor was assessed to provide a case example for the introduced hydrophobicity analysis. The evaluation revealed an overall dynamic profile for the protein. Moreover, Nur77 evidently contains a relatively large number of right-handed alpha helices with the majority showing a hydrophobic profile, i.e., the larger the score, the more hydrophobic the region. Some likely disordered regions can be seen spanning 50–70 amino acids (Fig. 4). Another interesting region is around 126–136 amino acids since these amino acids undergo significant shifts in their hydrophilicity and hydrophobicity. Similarly, the region around 180–210 amino acids appears to be actively changing preferences from

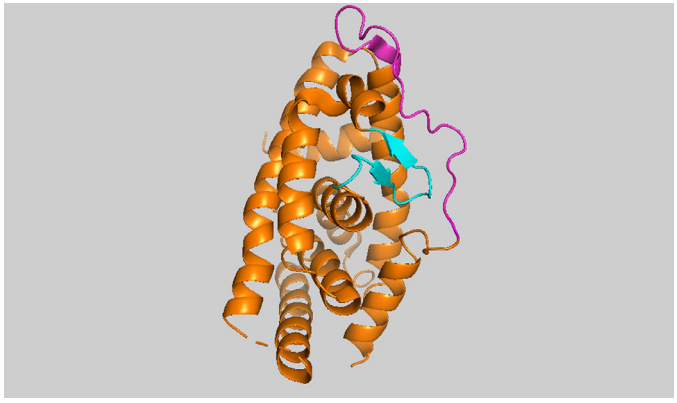

**Fig. 4.** Nur77 protein where magenta highlights are used to define a likely disordered region between 50 and 70 amino acids and the cyan color indicates a region between 127 and 140 amino acids.

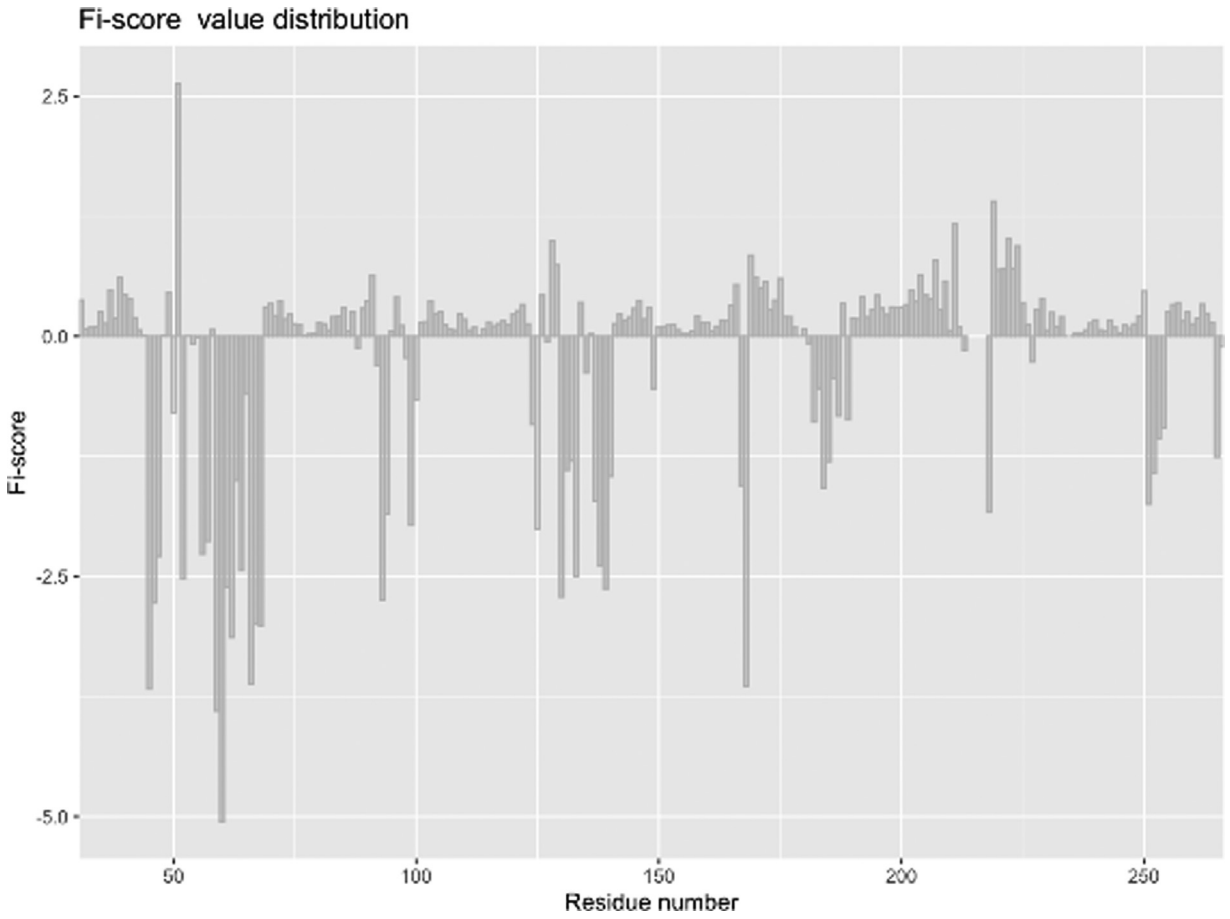


**Fig. 5.** Fi-score distribution for Nur77.

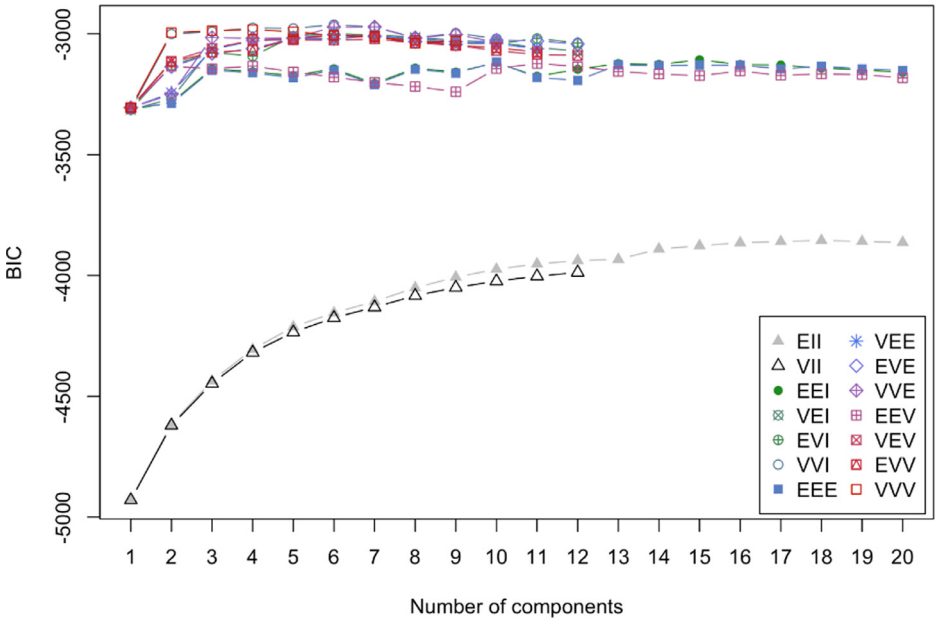


**Fig. 6.** Gaussian mixture modelling output showing Bayesian information criterion evaluation.

little solvent to being solvent exposed. This might suggest that the site undergoes considerable movements or actively engages other proteins or the DNA sequence. The disordered elements in this sequence stretch also imply that the region has to likely accommodate various rearrangements. Thus, studying these sites could provide hints at functionally important protein domains or subdomains (Figs. 3 and 4). Finally, evaluating N and C terminal sites for the purpose of protein engineering, we can see that a histidine tag would not significantly disrupt the conformation of the molecule and the C-terminus is probably the best site for the tag.

It is worth commenting on the derivation of the hydrophobicity scoring since the algorithmic nature of the process provides several impor-

```
## Best BIC values:
##          VVI,6        VVE,7        VVE,6
## BIC      -2962 -2971.310677 -2971.756243
## BIC diff     0    -9.310508    -9.756073
## ------------------------------------------------------------------
## Dimension reduction for model-based clustering and classification
## ------------------------------------------------------------------
##
## Mixture model type: Mclust (VVI, 6)
##
## Clusters  n
##        1 15
##        2 16
##        3 41
##        4 87
##        5 39
##        6 34
##
## Estimated basis vectors:
##                        Dir1      Dir2
## Residue_number 0.00047634  0.056189
## Fi_score       0.99999989 -0.998420
##
##                 Dir1      Dir2
## Eigenvalues  0.74861   0.55583
## Cum. %      57.38945 100.00000
```

**Fig. 7.** Output table for Gaussian mixture modelling evaluation.

tant analytical angles. The function builds on the Kyte-Doolittle hydrophobicity scale [1,18] to detect hydrophobic regions in proteins. Regions with a positive value are hydrophobic and those with negative values are hydrophilic. This scale can be used to identify both surface-exposed as well as transmembrane regions, depending on the window size used. However, to make comparisons easier, the original scale is transformed from 0 to 1 (similar scaling is also implemented in Expasy ProtScale [16,17]). The function requires a PDB data frame generated by *PDB_prepare* and the user needs to specify a window parameter to determine the size of the window for hydrophobicity calculations. The selection must be any odd number between 3 and 21 with the default being 21. Another parameter is weight that needs to be supplied to the function to establish a relative weight of the window edges compared to the window center (%); the default setting is 100%. Finally, a model parameter provides an option for weight calculation; that is, the selection determines whether weights are calculated linearly ($y = k \cdot x + b$) or exponentially ($y = a \cdot b^x$); the default model is 'linear'. The function evaluates each amino acid in a selected window where a hydrophobic influence from the surrounding amino acids is calculated in. While the terminal amino acids cannot be included into the window for centering and weighing, they are assigned unweighted values based on the Kyte-Doolittle scale [18]. The plot values are all scaled from 0 to 1 so that different proteins can be compared without the need to convert.

Thus, the hydrophobicity analysis can be especially useful when preparing to engineer proteins for various expression systems as the superimposition of structural features and hydrophobicity scores can help deciding if a protein region or domain is likely to be solvent exposed or prefer hydrophobic environments. For example, assessing the hydrophobicity and structural milieu of the N or C terminal amino acids can help selecting which terminal site should be tagged (as was demonstrated with Nur77). Moreover, this tool could be broadly applied in drug discovery studies involving the assessment of protein-protein interactions, protein-nucleic acid interactions, and membrane association events based on physicochemical characteristics.


**Fig. 8.** The Nur77 protein cluster identification with secondary structure elements.


**Fig. 9.** Nur77 cluster identification.

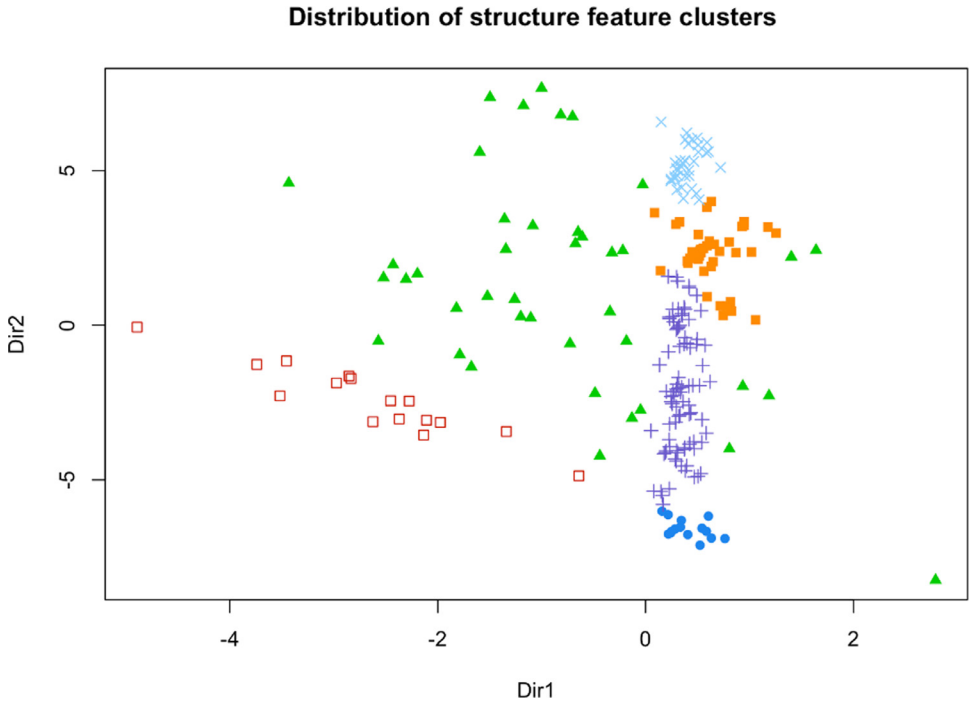


**Fig. 10.** Dimension reduction plot for the identified clusters.

### 3.3. Advanced analyses

Advanced analyses provide an opportunity to evaluate Fi-score distributions and take advantage of a streamlined GMM pipeline (Fig. 1). The main impetus for the development of this pipeline was the need for functions and data modelling tools that could be made freely accessible to non-experts. By contrast, commercial solutions, namely Schrȕdinger chemical simulation software [19], or non-commercial/semi-commercial solutions, including PSIPRED, AutoDock, MGLtools, and Expasy [16,20–22], lack a simple software platform to summarise and assess protein structural data that can be integrated into machine learning. While the mentioned software suites or online workbenches

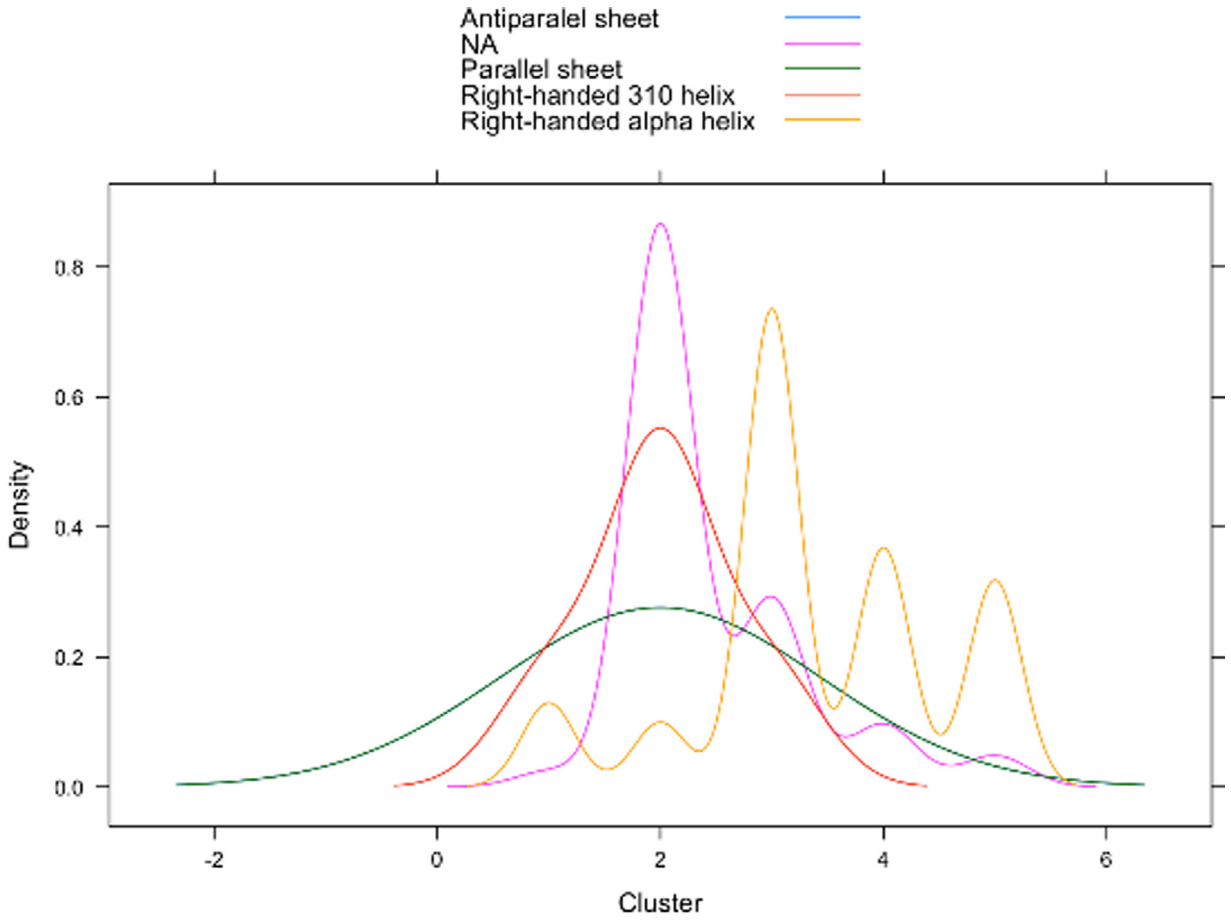


**Fig. 11.** Protein cluster density plots.

provide many useful functionalities, there is no one solution to use machine learning based inferences on the user's structural data. '*Fi_score*' enables researchers to quickly extract, assess, and summarise key features of their data and incorporate that information into downstream analyses or custom pipelines. That is, more advanced users are also given opportunity to supply custom parameters for the GMM workflow and extract probabilities from the output to use scores in other analyses or integrate the values in their own discovery pipelines.

#Fi-score distribution plot to explore scores for corresponding amino acids

Fi_score_plot(pdb_df)

#Fi-score for a selected region

#this value for multiple sites can be stored in relational databases

Fi_score_region(pdb_df,50,70)

#Plot of Fi-score values with superimposed secondary structures

Fiscore_secondary(pdb_df)

For example, a Fi-score distribution plot captures several interesting regions in Nur77 around the 50, 130, and 180 amino acids (Fig. 5) that coincide with the Fi-score shifts and mirroring patterns. Some other regions are also picked up which should be studied in more detail based on the amino acid composition and 3D conformations. The uncovered characteristics can be juxtaposed to other similar sites to better understand interaction mechanisms. Such approaches are especially useful when comparing known structures with the newly identified or investigating potential structural outliers.

Extracted Fi-score values can be used in machine learning modelling and this is enabled through the function *cluster_ID*. This function groups structural features using the Fi-score and Gaussian mixture modelling where an optimal number of clusters and a model to be fitted during the EM phase of clustering for GMM are automatically selected (Fig. 6). The output of this analytical tool summarises cluster information and also provides plots to visualise the identified clusters based on the cluster number and BIC value (Fig. 7). These outputs can be used to better assess model performance for the select parameters if the users chose to customise their model building.

#User selected parameters

df<-cluster_ID(pdb_df,clusters = 5, modelNames = "VVI")

The users are advised to set seed for more reproducible results when initiating their projects. *cluster_ID* takes a data frame containing a processed PDB file with Fi-score values as well as a number of clusters to consider during model selection; by default 20 clusters ('max_range') are explored. In addition, a 'secondary_structures' parameter is needed to define whether the information on secondary structure elements from the PDB file needs to be included when plotting; the default value is TRUE. Researchers also have an option to select a cluster number to test 'clusters' together with 'modelNames'. However, it is important to stress that both optional entries need to be selected and defined, if the

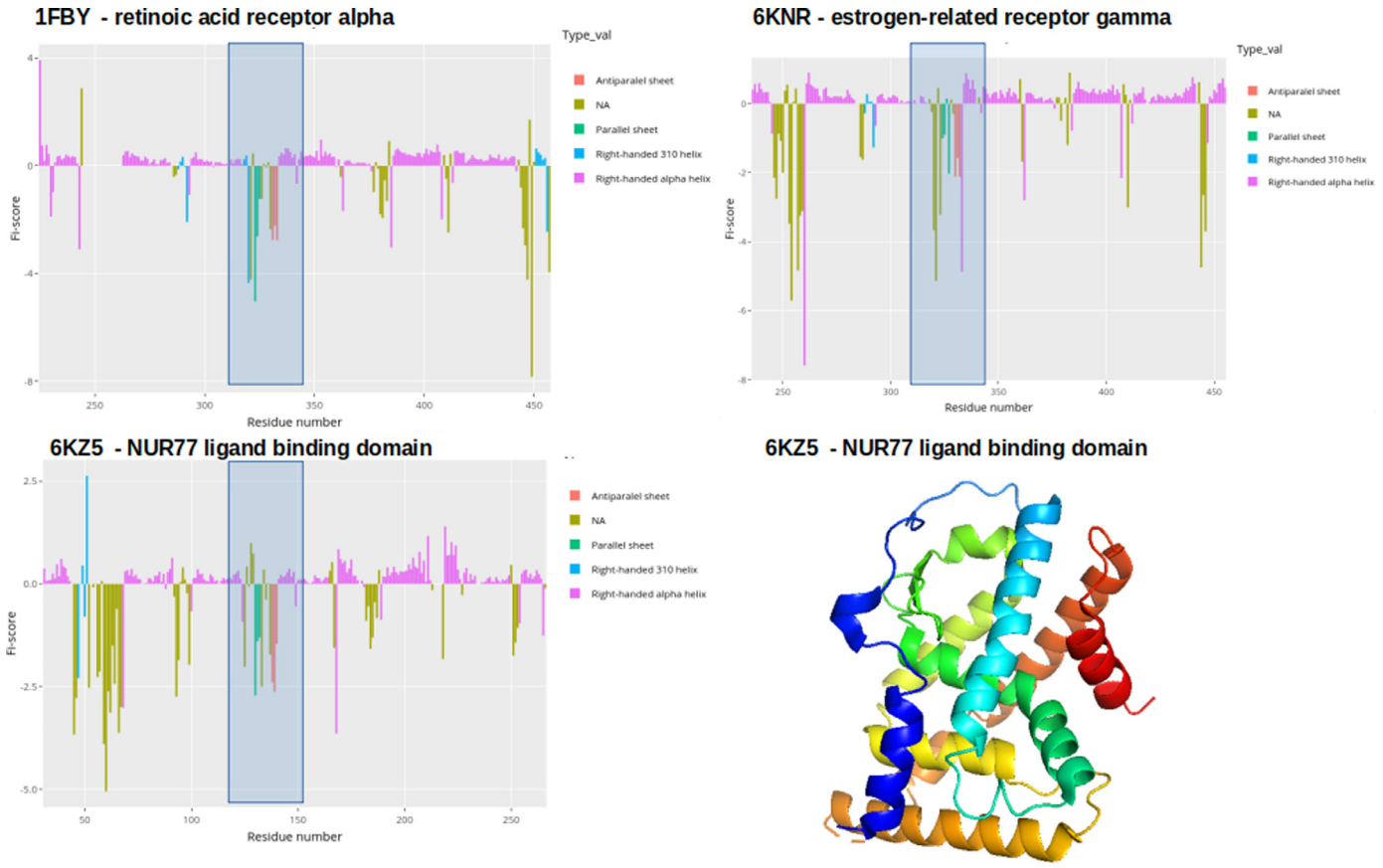


**Fig. 12.** Fi-score distribution plots with the Nur77 ligand binding domain (PDB ID:6KZ5), retinoic acid receptor alpha (PDB ID: 1FBY), and estrogen-related receptor gamma (PDB ID: 6KNR). Rainbow spectrum of the Nur77 structure allows to visualise the sequence from N-terminal (blue) to C-terminal (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

user wants to test other clustering options that were not provided by the automated BIC output. This is an advanced option and the user should assess the BIC output to decide which model and what cluster number he or she wants to try out. It is important to note that *cluster_ID* offers a user-friendly implementation of GMM where most technical decisions are already incorporated automatically.

A dimension reduction method for the visualisation of clustering is also automatically provided (Fig. 10). Dimension reduction is a useful technique to explore multi-dimensional biological data through key eigenvalues that define the largest information content of the explored features [10]. In other words, one can infer how well the explored characteristics define the data and if the classification is sufficient for downstream analyses. For example, in the case of Nur77 Fi-score clustering, this analysis allowed assessing how well the number of clusters separates data points based on their distribution features. Nur77 has six clusters which might indicate functionally and structurally distinct regions in the target protein. It appears that the data points are well separated into groups accounting for the different variability. The dimension reduction approach could also help deciding if a different number of clusters might better classify Fi-scores. One of the goals of building this software package was not only to provide accessible and easy-to-use functions but also to generate additional plots allowing to assess model performance and data point distribution.

In addition, one of the most valuable features of this set of functions is to generate clusters with secondary structure information (Figs. 8 and 9).The produced interactive plots enable researchers to explore structural characteristics of a protein of interest (Figs. 8 and 9). Thus, the subdivision of a protein structure based on the physicochemical features offers a new way to detect and explore functional sites or structural elements. Figs. 9 and 10 clearly indicate that some structural elements in Nur77 are likely similar in their function and physicochemical characteristics. For example, different types of helices as well as beta sheets in some cases overlap in their Fi-score characteristics and the assigned cluster type. This detailed capture of structural elements can help evaluate conformational outliers or infer similarities for different motifs. Moreover, it can be clearly seen that the region around the 50 amino acid is set to be distinct from the other two sites around 130 and 180 amino acids which could suggest overall different motion and interaction profiles. These findings also correspond with the earlier observations for the hydrophobicity features (Fig. 3). A similar trend can be seen for N and C terminus clusters which form distinct groups and might indicate sites where the receptor mediates specific functions [6]. GMM guided analyses offer a novel way to extract patterns that might not be observable using other methods dependent on sequence based analytics, e.g., Expasy [16,17].

All previous analyses tie in with the function *density_plots* which provides a density plot set for $\phi/\psi$ angle distributions, Fi-scores, and normalised B-factors. 3D visualisation of dihedral angle distribution for every residue is also included. These plots can be used for a quick assessment of the overall parameters as well as to summarise the observations. Density plots are very useful when evaluating how well the selected features or scores separate protein structural elements and if a protein structure is of good quality (i.e., dihedral angles, B-factors, or Fi-scores provide reasonable separation between elements). The function also gives another reference point to establish if the selected number of clusters differentiates residues well based on the secondary structure

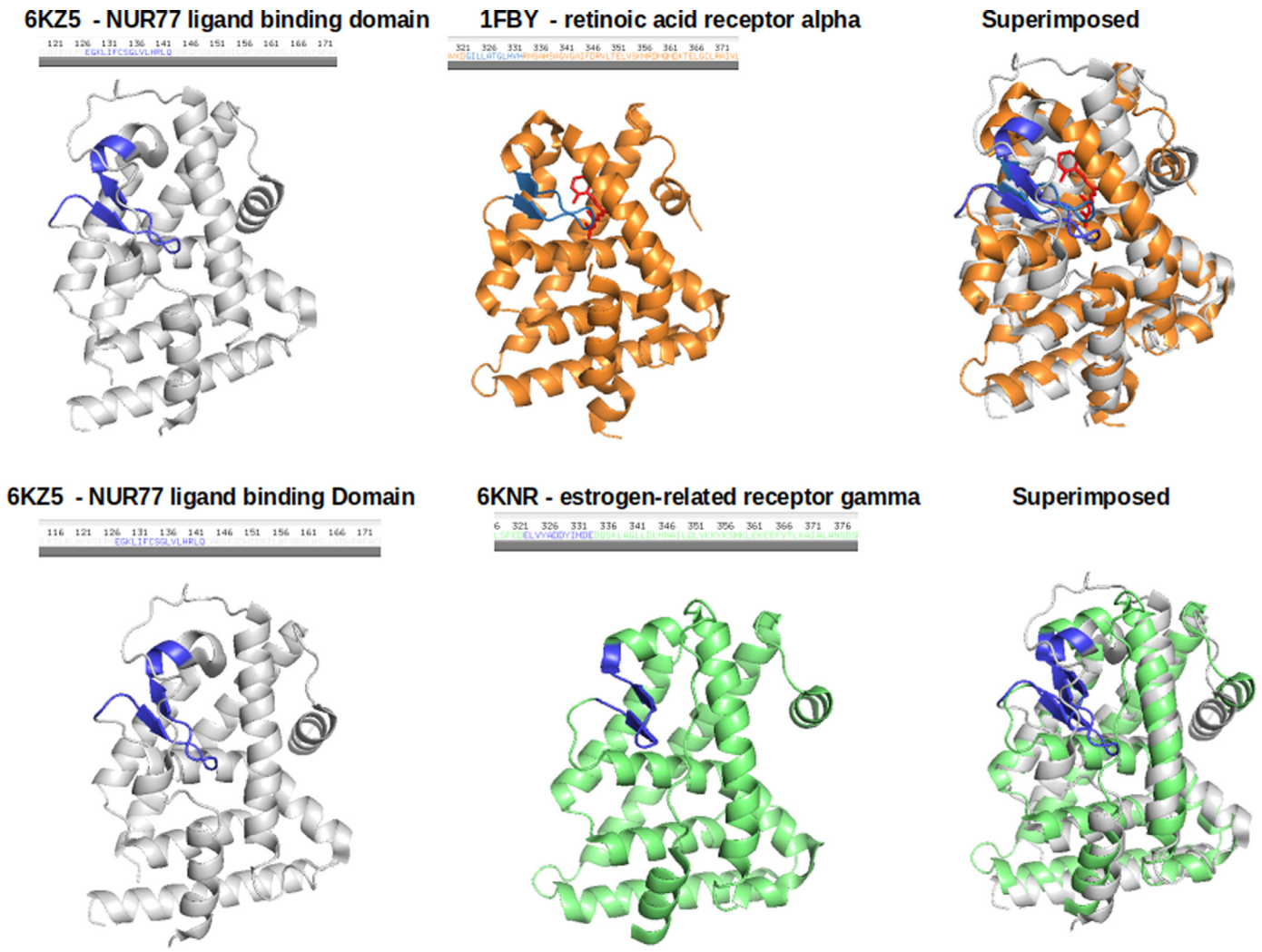


**Fig. 13.** PyMol generated plots to visualise protein structures where blue colors indicate the region identified through Fi-score patterns where cis-9 retinoic acid is highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

elements. In order to get this information, the user is only required to supply the output from the *cluster_ID* function (Fig. 11).

Data summary and evaluation

density_plots(pdb_df)

Data summary and evaluation including GMM outputs

cluster_IDs<-cluster_ID(pdb_df)

density_plots(cluster_IDs)

### 3.4. Case study: exploring potential ligands for the Nur77 orphan receptor

To demonstrate some of the *Fiscore* applications, potential ligands were searched for the Nur77 receptor which can be considered as a complex target since no endogenous ligands are known for this orphan receptor [6]. The first analysis step involved searching for other similar human proteins that did not belong to the Nuclear receptor subfamily 4. PSI-BLAST alignment analysis led to several candidate proteins, namely the retinoic acid receptor alpha (PDB ID: 1FBY) and estrogen-related receptor gamma (PDB ID: 6KNR) [15]. These proteins showed a good alignment to the Nur77 ligand binding domain sequence (average percent identity 31.68%; Suppl. Table 1) and were subsequently used for the structural and functional exploration. Comparing Nur77 Fi-scores with the retinoic acid receptor alpha and estrogen-related receptor gamma Fi-score distributions revealed several interesting patterns (Fig. 12). The shaded blue region highlights a matching distribution pattern for all the proteins and the Student's *t*-test confirmed that none of the distributions differed significantly (Fig. 12; Suppl. Table 2). Intriguingly, this region is involved in mediating interactions with retinoic acid in the retinoic acid receptor alpha (Fig. 13). Similarly, the estrogen-related receptor gamma (PDB ID: 6KNR) has a known inverse agonist binding to the same cavity created by paired alpha-helices, an anti-parallel beta sheet, and disordered stretches [23]. The inverse agonist exhibits several structural features, such as the scaffold size/orientation and key aromatic groups, that are similar to retinoic acid. Moreover, despite the different amino acid composition, the key physicochemical features are preserved in this site across the investigated proteins as can be seen from the superimposition studies (Fig. 13). These observations point to the fact that this region might be essential in accommodating binding events. Importantly, machine learning exploration (Fig. 8& 9) helped to classify Fi-scores around this region which revealed a repeating pattern very different from a surrounding N- and C-terminal regions. This further implies a site of special functional importance where data was grouped based on emerging probabilistic patterns in data point values. This case study suggests an interesting possibility that Nur77 with no known ligands might bind to chemical entities similar to retinoic

acid [6]. This is also supported by the alignment data and hydrophobicity plots (Suppl. Figs. 1–3) where Nur77 and the retinoic acid receptor alpha show substantial structural and physicochemical overlaps for this interactor site. Further molecular modelling and docking studies could aid in better understanding binding energetics and emerging interactions.

Overall, this example reveals that extracting patterns through scoring and machine learning could help identify proteins that have shared and functionally related features. Thus, *Fiscore* allows an easy implementation of protein structural data mining and classification without necessarily performing multiple visual inspections of the structures. These analytical principles can also be applied to explore other proteins of interest and their potential ligands.

## 4. Discussion

*Fiscore* package was developed to address the need for a simple-to-use, freely available, and adaptable set of tools for protein physicochemical feature exploration via machine learning. By contrast, other commercial, semi-commercial, or free software tools lack machine learning pipeline implementation to explore structural features and in most cases users need special knowledge to employ these pieces of software [8,16,19–22].

*Fiscore* package (Fig. 1) allows a user-friendly exploration of PDB structural data and integration with various machine learning methods. The package was benchmarked through several analytical stages that involved a diverse set of proteins (3352) to assess scoring principles [1] and package functionalities (1337 structures) [11]. With a number of helpful functions, including distribution analyses or hydrophobicity assessment in the context of structural elements, *Fiscore* enables the exploration of new target families and comprehensive data integration since the described fingerprinting captures protein sequence and physicochemical properties. Such analyses could be very helpful when exploring therapeutically relevant proteins. In addition, provided tutorials and documentation should guide researchers through their analysis and allow adapting the package based on individual project needs [1]. *Fiscore* could also aid in drug repurposing studies when a chemical compound needs to be juxtaposed to a number of potential targets. This was demonstrated during a native ligand search for Nur77 where a case study of a nuclear receptor revealed the usefulness of the introduced scoring and physicochemical data capturing via GMM. Furthermore, the novel scoring system as well as machine learning applications can lead to interesting insights about sites of structural and functional importance. The retrieved information could be used in comparative studies to search for other proteins that share similar features. For example, some of the shifts in Fi-score values coincide or precede post-translational modifications in Nur77 (Fig. 5) [24]. This information could be included in the future studies together with fingerprinting to better understand structural characteristics of this receptor.

Another important aspect of the *Fiscore* package is the simplification of complex analytical steps so that the researchers without an extensive background in structural bioinformatics or machine learning could still use the tools for their analyses, such as protein engineering, protein assessment, and data storage based on specific target sites. Thus, the interactive analytical and visualisation tools could become especially relevant in the pharmaceutical research and drug discovery studies as more complex targets and protein-protein interactions need to be assessed in a streamlined fashion. In other words, ability to translate structural data into parameters could accelerate target classification, target-ligand studies, or machine learning integration. Since target evaluation is paramount for rational therapeutics development, there is an undeniable need for specialised analytical tools and techniques that can be used in R&D or academic research. Implementing these novel approaches could significantly improve our ability to assess new targets and develop better therapeutics. As a result, the *Fiscore* package was developed to aid with therapeutic target assessment and make machine learning techniques free-to-use and more accessible to a wider scientific audience.

## 5. Funding

The package development was not supported by outside funding.

## Declaration of Competing Interest

The author reports no conflicts of interest.

## Acknowledgments

The author would like to thank the anonymous reviewers and code testers for helping to improve the package with their valuable suggestions and advice.

## Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.ailsci.2021.100016.

## References

[1] Kanapeckaitė A, Beaurivage C, Hancock M, Verschueren E. Fi-score: a novel approach to characterise protein topology and aid in drug discovery studies. J Biomol Struct Dyn 2021. doi:10.1080/07391102.2020.1854859.
[2] Du J, Guo J, Kand D, Li Z, Wang G, Wu J, et al. New techniques and strategies in drug discovery. Chin Chem Lett 2020;31. doi:10.1016/j.cclet.2020.03.028.
[3] Fauman E, Rai B, Huang E. Structure-based druggability assessment–identifying suitable targets for small molecule therapeutics. Curr Opin Chem Biol 2011;15. doi:10.1016/j.cbpa.2011.05.020.
[4] Faraggi E, Xue B, Zhou Y. Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. Proteins 2009;74. doi:10.1002/prot.22193.
[5] Reynolds D.. Gaussian mixture models2015; 10.1007/978-1-4899-7488-4
[6] Wu L, Chen L. Characteristics of nur77 and its ligands as potential anticancer compounds. Mol Med Rep 2018. doi:10.3892/mmr.2018.9515.
[7] Wickham H.. ggplot2: Elegant graphics for data analysis2016; https://ggplot2.tidyverse.org.
[8] Grant BJ, Rodrigues APC, ElSawy KM, McCammon JA, Caves LSD. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 2006;22:2695–6.
[9] Sievert C.. Interactive web-based data visualization with R, Plotly, and Shiny2020; https://plotly-r.com.
[10] Scrucca L, Fop M, Murphy TB, Raftery AE. Mclust 5: clustering, classification and density estimation using gaussian finite mixture models. R J 2016;8(1):289–317. doi:10.32614/RJ-2016-021.
[11] Kanapeckaite A.. Fiscore: effective protein structural data visualisation and exploration2021;R package version 0.1.3; https://github.com/AusteKan/Fiscore.
[12] Chambers J. Object-oriented programming, functional programming and R. Stat Sci 2014;29. doi:10.1214/13-STS452.
[13] DeLano W.. The PyMOL molecular graphics system, version 2.3.02021;.
[14] Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, et al. The protein data bank. Nucleic Acids Res 2000. doi:10.1093/nar/28.1.235.
[15] Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W, et al. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 1997. doi:10.1093/nar/25.17.3389.
[16] Gasteiger E., Hoogland C., Gattiker A., vaud S.D., ans Ron D. Appel M.R.W., Bairoch A.. Protscale2021;.
[17] Gasteiger E., Hoogland C., ans S'everine Du vaud A.G., ans Ron D. Appel M.R.W., Bairoch A.. Protein identification and analysis tools on the expasy server2005; 10.1385/1592598900
[18] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157(1):105–32. doi:10.1016/0022-2836(82)90515-0.
[19] Schrödinger. Schrödinger platform2021; https://www.schrodinger.com/platform.
[20] UCL. Predict secondary structure (psipred)2021; http://bioinf.cs.ucl.ac.uk/index.php?id=779.
[21] Institute T.S.R.. Autodock suite2021; https://autodock.scripps.edu/.
[22] the Sanner lab at the Center for Computational Structural Biology (CCRB). Mgltools software suite2021; https://ccsb.scripps.edu/mgltools/.
[23] Kimag J, Hwanga H, HeeseokYoona, Jae-EonLeed, Oh JM, An H, et al. An orally available inverse agonist of estrogen-related receptor gamma showed expanded efficacy for the radioiodine therapy of poorly differentiated thyroid cancer. Eur J Med Chem 2020. doi:10.1016/j.ejmech.2020.112501.
[24] Hornbeck P, Zhang B, Murray B, Kornhauser J, Latham V, Skrzypek E. Phosphositeplus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res 2015. doi:10.1093/nar/gku1267.