
TangledFeatures: Robust Feature Selection in Highly Correlated Spaces

Allen Daniel Sunny
University of Maryland, College Park
allens@umd.edu

Abstract

Feature selection is a fundamental step in model development, shaping both predictive performance and interpretability. Yet, most widely used methods focus on predictive accuracy, and their performance degrades in the presence of correlated predictors. To address this gap, we introduce TangledFeatures, a framework for feature selection in correlated feature spaces. It identifies representative features from groups of entangled predictors, reducing redundancy while retaining explanatory power. The resulting feature subset can be directly applied in downstream models, offering a more interpretable and stable basis for analysis compared to traditional selection techniques. We demonstrate the effectiveness of TangledFeatures on Alanine Dipeptide, applying it to the prediction of backbone torsional angles ϕ and ψ , and show that the selected features correspond to structurally meaningful intra-atomic distances that explain variation in these angles.

1 Introduction

Predictive modeling in structural biology has advanced rapidly, moving from early statistical methods to deep learning systems such as AlphaFold [1], which achieves remarkable accuracy in protein structure prediction. Yet, there is a gap in understanding the drivers behind these predictions: specific residues, motifs, or the correct structural interactions that determine outcomes [2].

For features identified as drivers to be meaningful, they must meet two goals. First, they should be biologically interpretable, mapping back to known structural or functional elements [3]. Second, they should be reproducible, consistently emerging across analyses rather than reflecting noise or model instability [4]. Meeting these goals is essential not only for scientific trust, but also for actionable insights, enabling mutational studies, protein design, and functional discovery [5].

Approaches to interpretability in structural biology fall under two categories: post-hoc and pre-hoc. Post-hoc methods, such as SHAP [6], Integrated Gradients [7], and Global Importance Analysis [8] apply explanatory frameworks after training. However, explanations often unstable in the presence of highly correlated structural features and may fail to yield biologically meaningful drivers [9]. Pre-hoc approaches have attempted to mitigate this by designing interpretable feature spaces before modeling. For example, clustering residue–residue contacts from molecular dynamics prior to Random Forest classification [10], or using sparse autoencoders to extract latent motifs from protein language models [11]. While these strategies reduce redundancy, they still exhibit instability. For example, different runs frequently produce divergent feature sets, leaving reproducibility an open challenge. Unlike earlier clustering-based pipelines, our method explicitly evaluates feature stability across runs, ensuring that identified drivers are not only interpretable but also reproducible.

To address the instability of existing interpretability methods, we place feature stability at the center of our approach. Our objectives are: (i) to evaluate the reproducibility of structural drivers across multiple analyses; (ii) to distinguish robust biological determinants from spurious correlations; and (iii) to provide a systematic framework for actionable biochemical insight.

To ground our study, we focus on the canonical case of *Alanine dipeptide*, a widely used benchmark in structural biology [12]. Our prediction targets are the backbone torsional angles ϕ (C–N–C $_{\alpha}$ –C) and ψ (N–C $_{\alpha}$ –C–N), with intra-atomic distances serving as the input features. This setup allows us to rigorously test whether our framework can identify stable and biologically meaningful drivers of conformational variation [13] (see Fig. 1).

Our contributions are threefold. First, we introduce TangledFeatures, a stability-based pipeline for feature selection in correlated spaces. Second, we validate it quantitatively through predictive accuracy and stability analyses. Third, we demonstrate its interpretability by showing that selected features correspond to meaningful intra-atomic distances in Alanine Dipeptide. Our results reveal TangledFeatures successfully maps cause and effect in complex correlated systems, presenting a novel and principled approach to feature selection, supporting reproducible research in structural biology.

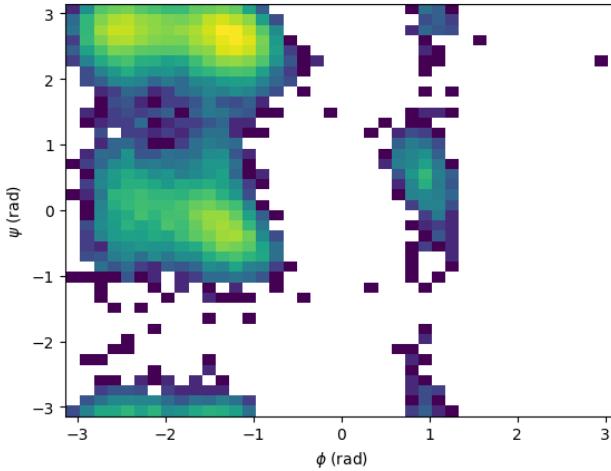


Figure 1: Illustration of the Ramachandran plot, showing allowed conformational regions for the backbone torsional angles ϕ and ψ . In this work, ϕ (C–N–C $_{\alpha}$ –C) and ψ (N–C $_{\alpha}$ –C–N) serve as prediction targets, while intra-atomic distances are used as input features.

2 Methods

2.1 Overview

Our objective is to develop a generalized feature selection algorithm that balances two central goals: preserving predictive accuracy while maintaining feature interpretability. As a proof-of-concept, we apply the algorithm to Alanine Dipeptide (Ace–Ala–Nme), a minimal peptide with a central alanine residue flanked by an acetyl group and an N-methylamide group. Its backbone conformation is defined by two torsional angles, ϕ (C–N–C $_{\alpha}$ –C) and ψ (N–C $_{\alpha}$ –C–N), which serve as the prediction targets. Each conformation is represented by a vector of intra-atomic distances among heavy atoms, yielding a feature space $d \in \mathbb{R}^m$, where m the number of intra-atomic distances. The task is to map d to torsional angles (ϕ, ψ) and to evaluate whether the algorithm can isolate a compact subset of distances that captures torsional variability while preserving predictive performance.

In this section, we focus on three components of the TangledFeatures pipeline. The correlation module c_{α} computes pairwise correlations and constructs a graph G , where nodes represent distances and edges denote correlations above a threshold. The selection module s_{β} operates on each connected component of G to extract a representative feature, thereby eliminating redundancy. The refinement module r_{γ} applies random forest feature selection to the set of representatives, filtering them further by predictive importance to produce the final subset of variables.

The overall process can be summarized as:

$$d \xrightarrow{c_{\alpha}} d_c \xrightarrow{s_{\beta}} d_s \xrightarrow{r_{\gamma}} d' \xrightarrow{f} (\phi, \psi).$$

2.2 Clustering

The first stage of TangledFeatures is to identify groups of highly correlated features while ensuring the prediction targets are treated consistently. Given the feature matrix $D \in \mathbb{R}^{n \times m}$, where n is the number of conformations and m the number of intra-atomic distances, we compute the pairwise Pearson correlation matrix $\Sigma \in \mathbb{R}^{m \times m}$. An undirected graph $G = (V, E)$ is then constructed with vertices V corresponding to features and edges

$$E = \{(i, j) : |\Sigma_{ij}| \geq \tau\},$$

where τ is a user-specified threshold. Connected components of G define clusters of features that convey largely redundant information. To account for the periodicity of torsional angles, the prediction targets (ϕ, ψ) are represented in cosine–sine form $(\cos \phi, \sin \phi, \cos \psi, \sin \psi)$, which removes discontinuities at angular boundaries and ensures stable downstream analysis.

2.3 Selection Module

For each correlated cluster C_k , the selection module s_β identifies a single representative distance using an ensemble-based stability selection procedure. Following the principle of stability selection [14], and ensemble feature selection methods such as Boruta [15], and Random Forest variable importance measures [16], we train multiple random forests where, in each run, one candidate distance is sampled from each cluster and evaluated alongside all uncorrelated variables when predicting the torsional targets $(\cos \phi, \sin \phi, \cos \psi, \sin \psi)$. We average feature importance scores across runs,

$$\hat{I}(d_{ij}) = \frac{1}{R} \sum_{r=1}^R I_r(d_{ij}),$$

and the distance with the highest \hat{I} within each cluster is retained as the representative. This ensemble approach reduces variance and prevents arbitrary choices among correlated distances, ensuring that the final representatives consistently capture the structural drivers of torsional variability.

2.4 Refinement Module

After cluster representatives are identified, the refinement module r_γ applies a Random Forest feature selection step based on cumulative importance coverage. The representatives are ranked by their Random Forest importance scores, and features are retained in descending order until the cumulative importance reaches 0.99 [16]. This threshold ensures that the final subset d' captures at least 99% of the predictive signal while discarding variables that contribute minimally. By combining local stability within clusters and global coverage across the model, the refinement module yields a compact set of distances that preserves predictive accuracy without sacrificing interpretability.

2.5 Objective

Feature Stability: To assess the robustness of selected variables, we repeated the entire pipeline across multiple resampled training sets. We quantified feature importance using SHAP values for each run [17]. We then evaluated stability in two complementary ways. First, we computed the Spearman rank correlation [18] between runs to capture the consistency of feature importance rankings across resamples. Second, we applied the Kuncheva index [19], which compares the overlap of the top-k most important features across runs, adjusted for chance. Together, these measures assess whether both the relative ordering of features and the membership of the most important subsets remain stable under data perturbations [20]. A high stability score indicates that the algorithm consistently highlights the same core drivers of torsional variability [14].

Predictive Accuracy: To evaluate whether interpretability is preserved without sacrificing performance, we trained predictive models (OLS, Random Forest [21], XGBoost [22], and SVM [23]) on the selected subsets to predict torsional targets $(\cos \phi, \sin \phi, \cos \psi, \sin \psi)$. We measured accuracy on held-out conformations using root mean squared error (RMSE) and the coefficient of determination (R^2).

3 Experiments

We evaluated TangledFeatures against popular feature selection methods on the Alanine Dipeptide system: LASSO, Elastic Net Regularization (ENR), Random Forest Recursive Feature Evaluation (RFE), Boruta and a no-selection baseline. Each feature selection method was paired with four predictive algorithms: Ordinary Least Squares (OLS), Random Forest Regression, XGBoost, and Support Vector Machines (SVM) — yielding 24 pipelines in total (see Table. 1). For each pipeline, models were trained on 80% of the data and evaluated on the remaining 20%. To assess robustness, we repeat the entire training and evaluation process 10 times with independent bootstrap resamples of the training data. We evaluated along three axes: predictive accuracy, stability and interpretability.

3.1 Predictive Accuracy

Metrics: We evaluated predictive performance by training models to map intra-atomic distances to torsional targets in cosine–sine form ($\cos \phi, \sin \phi, \cos \psi, \sin \psi$). For each pipeline, we measured accuracy on a held-out test set using root mean squared error (RMSE) and the coefficient of determination (R^2). We computed metrics separately for ϕ and ψ .

Results: Results are summarized in Table 1. All models achieve high predictive accuracy ($R^2 > 0.9$ for most pipelines). The best performance for ϕ is achieved by SVM with Elastic Net (RMSE = 0.05, $R^2 = 0.99$), while the best performance for ψ is also obtained with SVM + Elastic Net (RMSE = 0.84, $R^2 = 0.89$). Higher accuracy for SVM arises from leveraging redundant features, whereas TangledFeatures trades a slight loss in accuracy for a more stable, non-redundant feature set.

	ϕ				ψ			
	OLS	RF	XGBoost	SVM	OLS	RF	XGBoost	SVM
No Feature (baseline)	0.19 / 0.93	0.29 / 0.83	0.08 / 0.98	0.06 / 0.99	0.65 / 0.71	0.69 / 0.80	0.67 / 0.80	0.83 / 0.69
LASSO	0.21 / 0.92	0.09 / 0.98	0.09 / 0.98	0.05 / 0.99	0.89 / 0.64	0.69 / 0.79	0.63 / 0.82	0.84 / 0.70
Elastic Net (ENR)	0.20 / 0.92	0.07 / 0.98	0.07 / 0.99	0.05 / 0.99	0.90 / 0.64	0.66 / 0.81	0.67 / 0.81	0.87 / 0.89
RF Recursive Eval (RFE)	0.30 / 0.83	0.10 / 0.97	0.10 / 0.97	0.05 / 0.99	0.65 / 0.81	0.67 / 0.81	0.65 / 0.82	0.89 / 0.67
Boruta	0.22 / 0.91	0.07 / 0.98	0.07 / 0.98	0.05 / 0.99	0.91 / 0.64	0.65 / 0.82	0.91 / 0.64	0.84 / 0.70
TangledFeatures	0.26 / 0.87	0.09 / 0.98	0.09 / 0.97	0.09 / 0.98	0.97 / 0.61	0.67 / 0.81	0.97 / 0.60	0.86 / 0.75

Table 1: Predictive accuracy for torsional angle prediction. Rows list feature selection methods, while columns show predictive algorithms (OLS, RF, XGBoost, SVM) for each torsional angle ϕ and ψ . We report results as RMSE / R^2 . The best overall results for ϕ and ψ are shown in bold. We highlight the TangledFeatures row for comparison. TangledFeatures remains competitive with other selection methods while pruning redundancy, demonstrating that compact, interpretable subsets can be obtained without substantial loss of predictive fidelity.

3.2 Stability

Metrics: We quantified stability using the Kuncheva index and Spearman rank correlation, as described in Section 2.5. We computed metrics separately for torsional angles ϕ and ψ .

Results: Figures 2a–b report stability for the ϕ torsional angle, based on SHAP-derived feature importance. Under the Kuncheva index (panel a), TangledFeatures achieves consistently higher overlap of the top- k SHAP features across bootstrap runs than ENR or RFE. ENR shows sharp drops in overlap when correlated predictors are present, while RFE often retains different redundant distances, leading to low reproducibility. In contrast, TangledFeatures curves remain flat and close to the maximum, indicating nearly identical feature subsets across resamples. The Spearman correlation results (panel b) reveal the same trend at the level of feature rankings. ENR and RFE correlations fluctuate strongly across runs, reflecting instability in the relative importance assigned to correlated predictors. TangledFeatures achieves near-perfect rank correlation across resamples, showing that both the selected subset and the relative ordering of features are highly reproducible. For a more detailed analysis, including results for ψ , we refer the reader to the Appendix 4i, 5i.

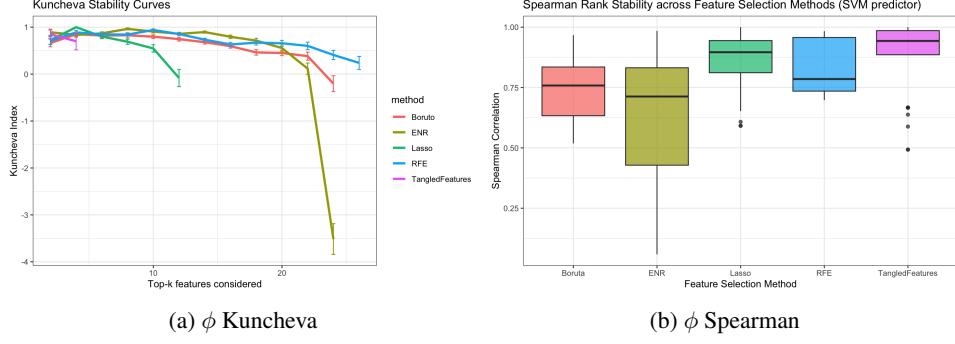


Figure 2: Stability for the ϕ torsional angle. (a) Kuncheva index, (b) Spearman rank correlation.

3.3 Interpretability

To assess interpretability, we projected the features selected by TangledFeatures onto the Alanine Dipeptide structure. The retained distances consistently aligned with backbone and near-backbone interactions—such as N–C, C–C, and C–N that are well-established determinants of ϕ and ψ torsional variability [24] [25] [26]. Figure 3 highlights these distances on the atomic diagram of Alanine Dipeptide. As an additional comparison, we also visualize the subset obtained from LASSO, the feature selection method that retained the second-fewest features among baselines. Unlike Tangled-Features, LASSO often selects redundant or chemically less meaningful distances, limiting interpretability despite its compactness. Taken together, these results demonstrate that TangledFeatures produces subsets that are not only stable and predictive, but also consistent with structural biology knowledge of peptide backbone flexibility. A full list of selected distances can be found in the Appendix B.

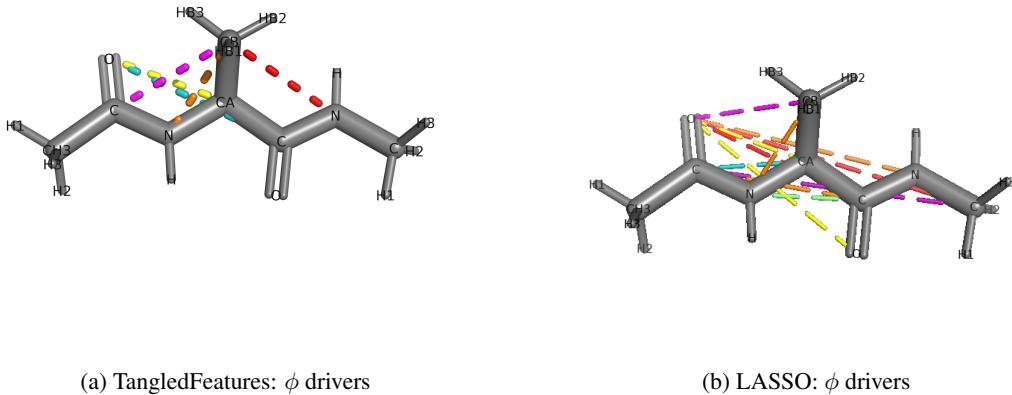


Figure 3: Comparison of selected feature subsets for the ϕ torsional angle.

4 Conclusion

These experiments demonstrate that TangledFeatures excels when datasets contain structured correlations among predictors, a setting where classical methods may either oversimplify (e.g., PCA [27]) or retain redundant features (e.g., LASSO, Boruta). Nevertheless, in cases where predictive accuracy is the sole objective and interpretability is less critical, PCA or LASSO may remain competitive alternatives. For the community at large, TangledFeatures provides a principled way to reconcile interpretability with robustness, ensuring that feature selection reflects the true informational diversity of the data.

Acknowledgments

The author gratefully thanks Xinyu Gu and Dr. Pratyush Tiwary (University of Maryland, College Park) for providing early access to the dataset and an accompanying image used in this work, as well as for helpful preliminary feedback during the development of *TangledFeatures*.

References

- [1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishabh Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.
- [2] David Medina-Ortiz, Ashkan Khalifeh, Hoda Anvari-Kazemabad, and Mehdi D. Davari. Interpretable and explainable predictive machine learning models for data-driven protein engineering. *Biotechnology Advances*, 79:108495, March 2025.
- [3] K. D. Ahlquist, Lauren A. Sugden, and Sohini Ramachandran. Enabling interpretable machine learning for biological data with reliability scores. *PLOS Computational Biology*, 19(5):e1011175, May 2023.
- [4] Anna Braghetto, Enzo Orlandini, and Marco Baiesi. Interpretable Machine Learning of Amino Acid Patterns in Proteins: A Statistical Ensemble Approach. *Journal of Chemical Theory and Computation*, 19(17):6011–6022, September 2023.
- [5] Valerie Chen, Muyu Yang, Wenbo Cui, Joon Sik Kim, Ameet Talwalkar, and Jian Ma. Applying interpretable machine learning in computational biology — pitfalls, recommendations and opportunities for new developments. *Nature methods*, 21(8):1454–1461, August 2024.
- [6] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions, November 2017. arXiv:1705.07874 [cs].
- [7] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks, June 2017. arXiv:1703.01365 [cs].
- [8] Peter K Koo, Matthew Ploenzke, Praveen Anand, and Steffan B Paul. Global Importance Analysis: A Method to Quantify Importance of Genomic Features in Deep Neural Networks.
- [9] Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A. Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl. General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models, August 2021. arXiv:2007.04131 [stat].
- [10] Tiffany J Callahan, Jie Shi, Kevin J Cheng, Michael A Sauer, Taras V Pogorelov, and Sara Capponi. Interpretable Machine Learning Uncovers Structural Determinants of Wnt-Wls Binding Specificity from Extended Atomistic Simulations.
- [11] James Lyons, Abdollah Dehzangi, Rhys Heffernan, Alok Sharma, Kuldip Paliwal, Abdul Sattar, Yaoqi Zhou, and Yuedong Yang. Predicting backbone $C\alpha$ angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *Journal of Computational Chemistry*, 35(28):2040–2046, October 2014.
- [12] Michael Feig. Is Alanine Dipeptide a Good Model for Representing the Torsional Preferences of Protein Backbones? *Journal of Chemical Theory and Computation*, 4(9):1555–1564, September 2008.

- [13] Takuma Kikutsuji, Yusuke Mori, Kei-ichi Okazaki, Toshifumi Mori, Kang Kim, and Nobuyuki Matubayasi. Explaining reaction coordinates of alanine dipeptide isomerization obtained from deep neural networks using Explainable Artificial Intelligence (XAI). *The Journal of Chemical Physics*, 156(15):154108, April 2022.
- [14] Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems*, 12(1):95–116, May 2007.
- [15] Miron B. Kursa, Aleksander Jankowski, and Witold R. Rudnicki. Boruta – A System for Feature Selection. *Fundamenta Informaticae*, 101(4):271–285, July 2010.
- [16] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in forests of randomized trees.
- [17] Neil Jethani, Mukund Sudarshan, Ian Covert, Su-In Lee, and Rajesh Ranganath. FastSHAP: Real-Time Shapley Value Estimation, March 2022. arXiv:2107.07436 [stat].
- [18] Joost C. F. De Winter, Samuel D. Gosling, and Jeff Potter. Comparing the Pearson and Spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. *Psychological Methods*, 21(3):273–290, September 2016.
- [19] Ludmila I Kuncheva. A STABILITY INDEX FOR FEATURE SELECTION.
- [20] Anne-Claire Haury, Pierre Gestraud, and Jean-Philippe Vert. The Influence of Feature Selection Methods on Accuracy, Stability and Interpretability of Molecular Signatures. *PLoS ONE*, 6(12):e28210, December 2011.
- [21] Marvin N. Wright and Andreas Ziegler. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1), 2017. arXiv:1508.04409 [stat].
- [22] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, August 2016. arXiv:1603.02754 [cs].
- [23] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998.
- [24] Kyle A. Beauchamp, Yu-Shan Lin, Rhiju Das, and Vijay S. Pande. Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. *Journal of Chemical Theory and Computation*, 8(4):1409–1414, April 2012.
- [25] Stefano Piana, John L Klepeis, and David E Shaw. Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. *Current Opinion in Structural Biology*, 24:98–105, February 2014.
- [26] Olga Ksenofontova and Vasily Stefanov. Conformational flexibility of the pharmacologically important insulin analogues. *Advances in Biological Chemistry*, 03(05):512–517, 2013.
- [27] Felipe L. Gewers, Gustavo R. Ferreira, Henrique F. de Arruda, Filipi N. Silva, Cesar H. Comin, Diego R. Amancio, and Luciano da F. Costa. Principal Component Analysis: A Natural Approach to Data Exploration. *ACM Computing Surveys*, 54(4):1–34, May 2022. arXiv:1804.02502 [cs].

A Software Availability

This work also contributes a reproducible R package, `TangledFeatures`, available on GitHub and prepared for CRAN submission. The package implements the core functions for redundancy detection, pruning, and stability analysis, along with visualization utilities for correlation structures. Documentation includes example workflows across multiple domains, and extensions are underway to handle high-dimensional genomics, complex financial markets, and large-scale government datasets. By making the tool openly available, we aim to encourage adoption by both researchers and practitioners seeking interpretable and redundancy-aware feature selection methods.

B Selected Distances

Table 2 and Table 3 list the intra-atomic distances retained by `TangledFeatures` after correlation pruning and refinement. These represent the compact subsets ultimately selected for ϕ and ψ , respectively by `TangledFeatures`.

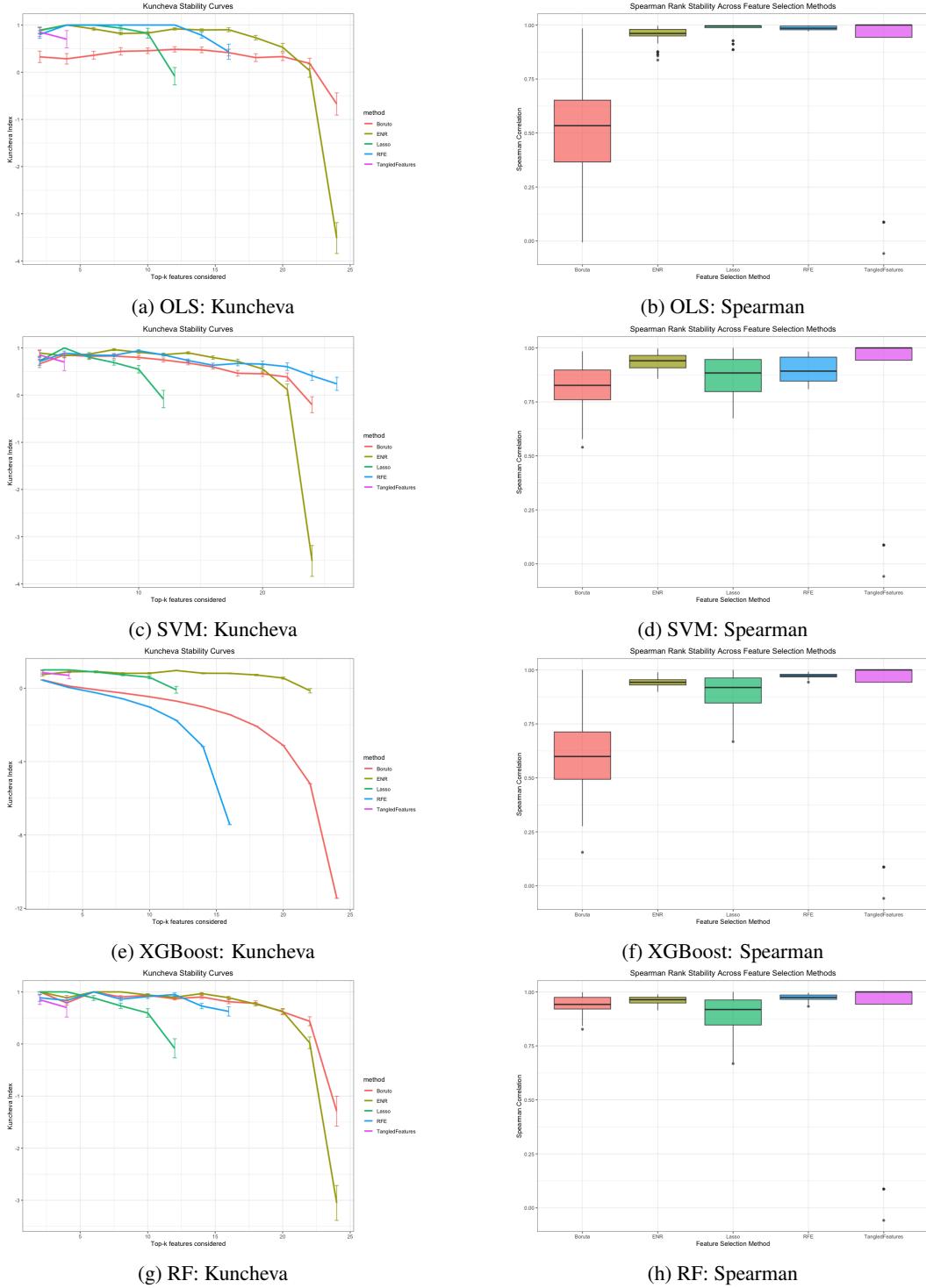
Atom Distance	Category
ACE1-CH3 \leftrightarrow ALA2-CB	Cap-related/Flexible
ACE1-C \leftrightarrow ALA2-C	Flexible
ACE1-CH3 \leftrightarrow ALA2-CA	Flexible
ACE1-C \leftrightarrow ACE1-O	Rigid
ACE1-CH3 \leftrightarrow NME3-N	Cap-related
ACE1-C \leftrightarrow NME3-N	Cap-related
ACE1-CH3 \leftrightarrow ALA2-C	Rigid

Table 2: Unique atom–atom distances selected by `TangledFeatures` for ϕ and their categorical assignments.

Atom Distance	Category
ACE1-CH3 \leftrightarrow ALA2-C	Rigid
ACE1-C \leftrightarrow NME3-N	Cap-related
ACE1-C \leftrightarrow ACE1-O	Rigid
ACE1-C \leftrightarrow ALA2-O	Rigid
ACE1-CH3 \leftrightarrow NME3-N	Cap-related
ACE1-C \leftrightarrow ALA2-C	Flexible

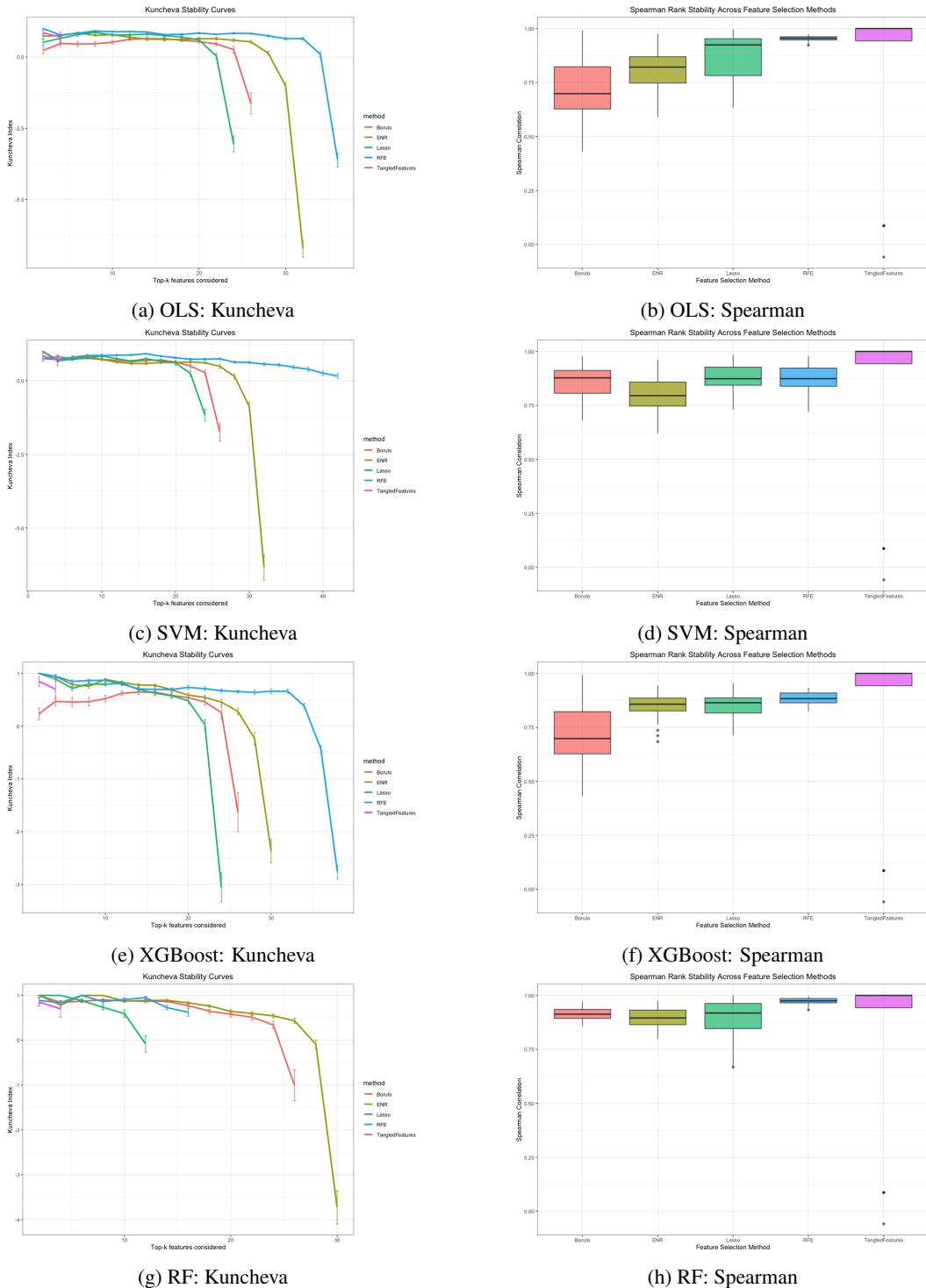
Table 3: Unique atom–atom distances selected by `TangledFeatures` for ψ and their categorical assignments.

Feature selection stability for ϕ torsional angle



(i) Stability curves (Kuncheva index, left column; Spearman rank correlation, right column) for ϕ torsional angle across predictive models (rows: OLS, SVM, XGBoost, RF).

Feature selection stability for ψ torsional angle



(i) Stability curves (Kuncheva index, left column; Spearman rank correlation, right column) for ψ torsional angle across predictive models (rows: OLS, SVM, XGBoost, RF).