MLtopKin
Kinase conformation prediction using topological features and machine learning

Chatzitolios Vasileios-Klearchos, ID: 7115152400026
Konstantinidis Konstantinos, ID: 7115152400017

July 7, 2025

Algorithms in structural boinformatics — final project

Contents

1 Introduction

1.1 Kinases

Kinases are enzymes belonging to the transferase class, which catalyze the transfer of phosphorus-containing groups from a donor molecule to an acceptor molecule; they are grouped under the Enzyme Commision (EC) number 2.7[1]. They are responsible for the phosphorylation of proteins, which is a key post-translational modification that regulates various cellular processes, including cell growth, differentiation, and metabolism [1]. Due to their crucial role in cellular and molecular processes, kinases are often implicated in various diseases, most notably cancer [2] and neurodegeneration [3].

Like all proteins, kinases are dynamic molecules that can adopt multiple conformations. These conformations are essential for their function, as they determine the accessibility of their active sites and the binding of substrates and inhibitors. The conformational state of a kinase is controlled by the positions and orientations of several key structural elements, such as the DFG motif (Asp-Phe-Gly) and the A-loop (activation loop) [4]. There are several methods to discern the conformation state of a kinase, such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy (cryo-EM). These methods provide high-resolution structural information, but they require additional work to analyze the structures and determine the conformation state, such as manual inspection [4].

1.1.1 Kinase conformation classification

Some automated methods exist, like the one proposed by Natarajan Kannan et al. in 2017 [5]. This method uses the ϕ, ψ, $\chi1$, and pseudo-dihedral angles as features to classify the kinase structures into active and inactive states. Using this feature extraction method and a simple Random Forest model, the authors managed to achieve a classification accuracy of over 97%. However, this method requires an alignment step prior to the angle extraction step to locate the conserved domains necessary, which might lead to errors as it is sensitive to any alterations. Additionally, the method is limited to enzymes with highly conserved domains, as its feature generation step depends on this conservation.

A newer method was proposed by Simonson et al. in 2024 [6] which uses 78 geometric features like dihedral angles, distances, and solvent-accessible surface area (SASA) and an XGBoost model. The structures were clustered based on DFG geometry and then annotated manually for conformation (DFG-in, DFG-out, and other). They achieved over 99% classification accuracy. This method also relied on a common sequence alignment to the Pfam HMM for kinases, introducing the same potential issues.

[1] `https://enzyme.expasy.org/EC/2.7.-.-`

1.2 Topological descriptors as features

Based on the limitations outlined above, we hypothesize that topological descriptors may serve as a robust alternative for predicting the conformational state of kinases. Unlike angle-based or motif-dependent methods, topological descriptors are inherently alignment-free, eliminating the need for preprocessing steps such as structural alignment or conserved domain identification [7].This is particularly advantageous, as alignment steps can introduce errors, propagate biases, and limit generalizability across diverse kinase families.

Moreover, topological methods—especially those derived from persistent homology—can capture global and multi-scale geometric features of the protein structure, such as loops, voids, and cavities. These features may be particularly relevant in the context of kinase conformational transitions, which often involve subtle but distributed structural rearrangements [8].

Topological features also offer a degree of noise tolerance and robustness to missing or non-canonical data, making them suitable for applications beyond highly curated crystal structures [9]. As such, we believe this approach has the potential to generalize better to novel folds, cryptic conformations, or kinases with incomplete annotations.

Ultimately, by designing a simple, one-step, alignment-free pipeline based on topological and geometric features—specifically Betti numbers and solvent-accessible surface area (SASA)—we aim to provide a proof of concept that these descriptors alone are sufficient to discriminate kinase activation states. This work contributes to the growing effort to explore low-dimensional, interpretable, and generalizable representations in structural bioinformatics. The following sections describe the dataset selection, feature computation pipeline and classification framework used to assess the predictive power of these descriptors.

2 Methods

2.1 Data

The KinCore database[2] is a comprehensive web-based database and tool [10], which contains information on the structures of protein kinases and their conformational states. We downloaded the index file containing all available kinase structures, which includes the PDB ID, the conformation state, the gene name, and other relevant metadata. Based on that, we grouped the structures by gene, in order to download equal numbers of active and inactive structures for each gene. This step was necessary to ensure that the dataset is balanced and that the model does not favor one class over the other. We downloaded a total of 2608 structures, 1308 inactive and 1300 active (this small discrepancy is due to the fact that some structures failed to pass the quality control steps, but since the number of structures is large enough and the classes are balanced, we did not proceed to any further filtering).

2.2 Feature extraction

The feature extraction pipeline consists of two main steps: the computation of the solvent-accessible surface area (SASA) and the computation of the persistence diagram. For the computation of the SASA, we used FreeSASA[3] which is a Python library with bindings to the C library of the same name [11]. To generate the persistence diagrams we used the Gudhi library[4] which is a library for topological data analysis, written in C++ with Python bindings [12]. To extract the persistence images from the persistence diagrams, we used the persim library[5] which is a Python library for computing persistence images and other topological features from persistence diagrams [13].

2.3 Model training and evaluation

For the Logistic Regression, Gaussian Naive Bayes, Linear Discriminant Analysis, Support Vector Machine and Random Forest Classifiers we used the scikit-learn library[6] which is a Python library for machine learning [14]. For the LightGBM model we used the LightGBM library[7] which is a gradient boosting framework that uses tree-based learning algorithms [15]. The baseline models were trained using the default hyperparameters, while the Random Forest model was tuned using the Optuna library[8] which is a hyperparameter optimization framework [16]. The baseline models were evaluated using 6 common metrics: Matthews Correlation Coefficient (MCC), F1-score, Accuracy, Precision, Recall and Specificity — as provided by the scikit-learn library.

[2] http://dunbrack.fccc.edu/kincore/home

[3] https://freesasa.github.io/python/

[4] https://gudhi.inria.fr/

[5] https://persim.scikit-tda.org/en/latest/index.html

[6] https://scikit-learn.org/stable/

[7] https://lightgbm.readthedocs.io/en/stable/

[8] https://optuna.org/

3 Results

We evaluated whether geometric and topological features, specifically Solvent-Accessible Surface Area (SASA) and Betti numbers could be used as descriptors in classifying kinase conformations as active or inactive. To this end, we tried tested various combinations of features types and normalization strategies: SASA-only features, Betti-only features, SASA + Betti (no scaling), SASA + Betti (scaled). All models were evaluated using 6 common metrics: Matthews Correlation Coefficient (MCC), F1-score, Accuracy, Precision, Recall and Specificity.

3.1 Classifier performance

We observed that both the Betti-only and SASA-only feature sets achieved moderate classification performance individually. However, when combined, the SASA + Betti feature set consistently outperformed across all classifiers and evaluation metrics. This pattern suggests that local geometric properties (captured by SASA) and global topological structure (quantified by Betti numbers) provide complementary information relevant to kinase activation. Specifically, topological features may capture changes in global structural organization such as loop formation or cavity closure, while SASA reflects local solvent exposure patterns that accompany conformational changes like activation loop movement or rearrangement of the ATP-binding site.

The below figure (Figure 1) shows the performance of each classifier using the best-setup (combined SASA + Betti features, with scaling) and no hyperparameter tuning. The classifiers were evaluated using 6 common metrics: Matthews Correlation Coefficient (MCC), F1-score, Accuracy, Precision, Recall and Specificity. The results indicate that tree-based models (Random Forest and LightGBM) and the Logistic Regression model achieved the highest performance, with accuracies between 73 and 74%.

3.2 Random Forest hyperparameter tuning

To further improve the performance of the Random Forest classifier, we performed a hyperparameter tuning step using the Optuna library [16]. The hyperparameters we tuned were the number of estimators, the maximum depth, the criterion for splitting, the minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node, and the maximum number of features to consider when looking for the best split.

When tested on unseen data, the classifier achieved an accuracy of 0.74. The corresponding confusion matrix is shown in Figure 2.

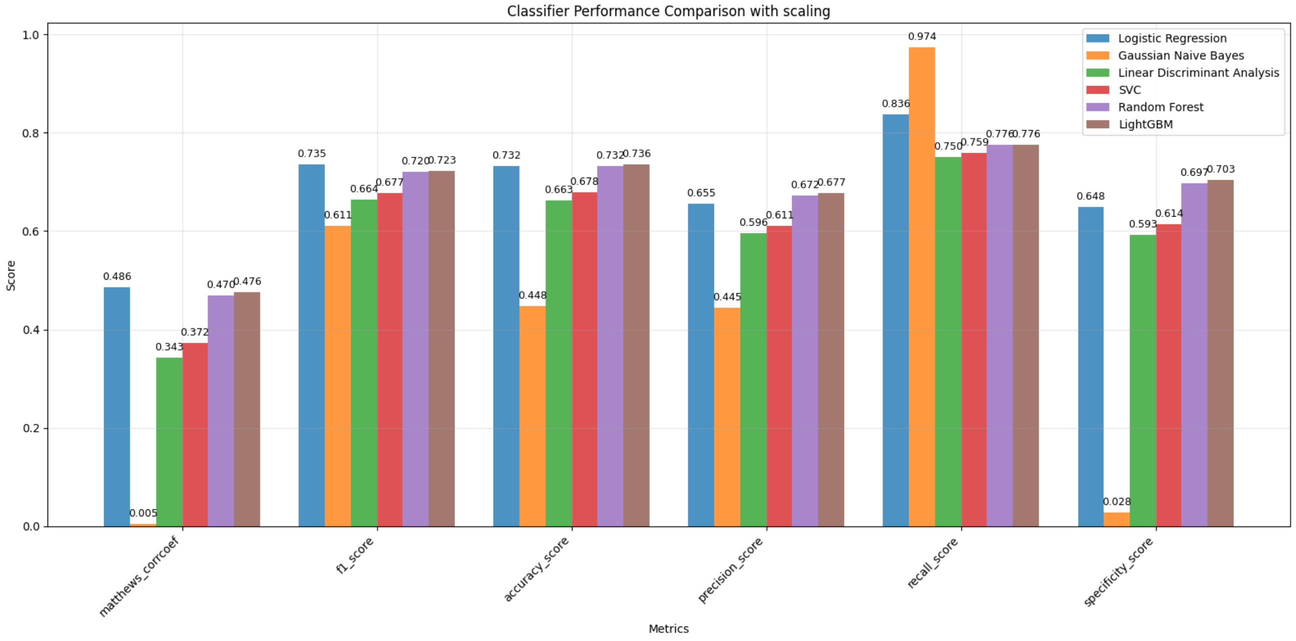

Figure 1: Different classification models performance using the best feature set (SASA + Betti, scaled). The models are evaluated using 6 common metrics: Matthews Correlation Coefficient (MCC), F1-score, Accuracy, Precision, Recall and Specificity.

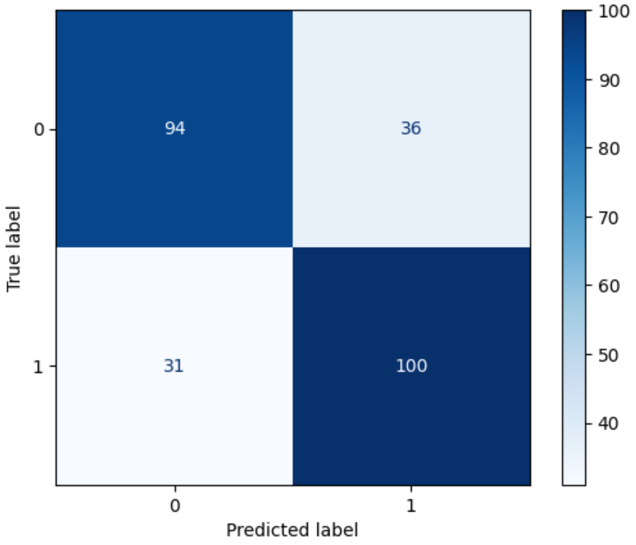

Figure 2: Confusion matrix of the Random Forest classifier after hyperparameter tuning. The model achieved an accuracy of 0.74 on unseen data.

4 Conclusion

This study explored the potential of using solvent-accessible surface area (SASA) and topological descriptors (Betti numbers) as features for classifying kinase conformational states. Our results show that these two types of descriptors, when used together, provide sufficient discriminatory power to achieve solid classification performance across multiple machine learning models. The best-performing configuration, using both feature types with normalization and a Random Forest classifier demonstrated that these compact and interpretable features can meaningfully distinguish between active and inactive kinase conformations.

An important insight from this work is the complementarity between SASA and Betti numbers. While SASA captures local geometric properties such as surface exposure of key residues, Betti numbers encode global topological information like the presence of cavities, loops, and structural components. This suggests that conformational shifts in kinases manifest both locally and globally, and both types of features are necessary to represent the full structural signal.

Compared to traditional methods that rely on angle-based features or conserved domain alignment, our approach offers a lightweight, alignment-free alternative that is more robust to structural variability and applicable to diverse kinase families. However, there are still several limitations to address. First, the dataset size is relatively small, which may limit generalization. Second, topological features were extracted only from Cα atoms, which simplifies computation but may miss fine-grained details captured by all-atom topologies. Third, kinase conformational space is more continuous than binary; intermediate or ambiguous structures may challenge discrete classification schemes.

To further improve interpretability and broaden the scope of our approach, we propose the following extensions:

1. Identify the most informative topological features that drive model predictions, possibly through feature importance or ablation studies.
2. Analyze misclassified kinase structures to better understand failure cases and uncover potentially novel conformational patterns.
3. Evaluate cross-species performance by applying the trained models to kinases from different organisms.
4. Apply unsupervised clustering on the feature space to assess whether active and inactive states naturally form distinct clusters in topological or SASA-derived embeddings.

These steps aim to deepen our understanding of the structure-function relationship in kinases and guide future efforts to create generalizable, interpretable models for protein conformational analysis.

5 Code availability

The code used for this project is available on GitHub at the following link: `https://github.com/KonsKons26/MLtopKin`. The software is functional but is still under active development and we are open to suggestions and contributions from the community.

References

[1] S. K. Hanks, A. M. Quinn, and T. Hunter, “The protein kinase family: Conserved features and deduced phylogeny of the catalytic domains,” *Science*, vol. 241, no. 4861, pp. 42–52, 1988.

[2] J. Li, C. Gong, H. Zhou, J. Liu, X. Xia, W. Ha, Y. Jiang, Q. Liu, and H. Xiong, “Kinase inhibitors and kinase-targeted cancer therapies: Recent advances and future perspectives,” *International Journal of Molecular Sciences*, vol. 25, no. 5489, 2024.

[3] X. Wu, Z. Yang, J. Zou, H. Gao, Z. Shao, C. Li, and P. Lei, “Protein kinases in neurodegenerative diseases: current understandings and implications for drug discovery,” *Signal Transduction and Targeted Therapy*, vol. 10, p. 146, May 2025.

[4] N. R. Gough and C. G. Kalodimos, “Exploring the conformational landscape of protein kinases,” *Current Opinion in Structural Biology*, vol. 88, p. 102890, 2024.

[5] D. I. McSkimming, K. Rasheed, and N. Kannan, “Classifying kinase conformations using a machine learning approach,” *BMC Bioinformatics*, vol. 18, p. 86, Feb 2017.

[6] I. Reveguk and T. Simonson, “Classifying protein kinase conformations with machine learning,” *Protein Science*, vol. 33, no. 4, p. e4918, 2024.

[7] H. Adams, S. Chepushtanova, T. Emerson, E. Hanson, M. Kirby, F. Motta, R. Neville, C. Peterson, P. Shipman, and L. Ziegelmeier, “Persistence images: A stable vector representation of persistent homology,” 2016.

[8] I. Obayashi and Y. Hiraoka, “Persistence diagrams with linear machine learning models,” 2017.

[9] R. Gisolf, F. A. N. Santos, and F. Wierstra, “Beyond signal and noise: Unraveling scale invariance in neuroscience and financial networks with topological data analysis,” 2024.

[10] V. Modi and J. Dunbrack, Roland L, “Kincore: a web resource for structural classification of protein kinases and their inhibitors,” *Nucleic Acids Research*, vol. 50, pp. D654–D664, 10 2021.

[11] S. Mitternacht, “FreeSASA: An open source C library for solvent accessible surface area calculations,” *F1000Research*, vol. 5, p. 189, 2016. version 1; peer review: 2 approved.

[12] T. G. Project, *GUDHI User and Reference Manual*. GUDHI Editorial Board, 3.11.0 ed., 2025.

[13] D. Goldfarb and the scikit-tda developers, “Persim: Persistence diagram utilities.” `https://github.com/scikit-tda/persim`, 2019. Part of the scikit-tda project.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau,

M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “Lightgbm: a highly efficient gradient boosting decision tree,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, (Red Hook, NY, USA), pp. 3149–3157, Curran Associates Inc., 2017.

[16] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A next-generation hyperparameter optimization framework,” 2019.