Molecular Dynamics and Quantum Machine Learning Model to Classify Polymorphisms in Multiple Sclerosis

Diego Rueda[1,2], Levy Bueno Alves[1], Milton Faria Junior[2], Silvana Giuliatti[1]

[1]Faculdade de Medicina de Ribeirão Preto – Universidade Federal de São Paulo (USP)
Avenida Bandeirantes, 3900 – 14049-900 – São Paulo – SP – Brazil

[2]Departamento de Exatas - Universidade de Ribeirão Preto (UNAERP)
Avenida Costábile Romano, 2201– 14096-900 – São Paulo - SP – Brazil

drueda@usp.br, levybuenoalves@usp.br, mfaria@unaerp.br,
silvana@fmrp.usp.br

Abstract. *Multiple Sclerosis is an autoimmune disorder associated with specific HLA-DRB1 alleles however, the structural mechanisms underlying this association remains poorly understood. We developed a model integrating molecular dynamics (MD), machine learning (ML), and quantum neural networks (QNN) to classify alleles as either risk-related or protective. MD simulations of HLA-peptide complexes were conducted, and distance matrices were used as input for a ML-QNN classifier. The combined complex protein (chains A and B) and peptide input resulted in the highest accuracy (0.990) and most stable convergence. This study shows the potential of quantum-classical models in identifying subtle structural patterns linked to autoimmune susceptibility.*

1. Multiple Sclerosis

Multiple Sclerosis (MS) is a chronic and progressive autoimmune disease that affects the central nervous system (CNS), resulting in a wide range of deficits, including sensory, motor, autonomic, and neurocognitive impairments in affected individuals [Adiele and Adiele 2019]. To date, the exact causes underlying the development of MS remain incompletely understood, although both genetic and environmental factors appear to contribute [Nourbakhsh and Mowry 2019]. The strongest genetic association with MS lies within the human leukocyte antigen (HLA) class II genes, particularly the HLA-DRB1*15:01 allele located on chromosome 6p21, which links adaptive immune function to MS susceptibility [Hollenbach and Oksenberg 2015].

Certain HLA-DRB1 alleles have been associated with either susceptibility to or protection against MS. For example, while the HLA-DRB115:01 and HLA-DRB115:03 polymorphisms are linked to increased risk, alleles such as HLA-DRB101:01, HLA-DRB111:01, HLA-DRB115:02, and HLA-DRB116:01 are considered protective factors [Hollenbach and Oksenberg 2015, Silvestri et al. 2019]. However, the reasons why some alleles confer a higher risk while others provide protection remain unclear.

HLA class II proteins are essential surface glycoproteins of the immune system responsible for presenting peptides to T cells, enabling the detection and response to foreign substances [Rock, Reits and Neefjes 2016]. One of the prevailing hypotheses used

to explain the development of MS is molecular mimicry [Chastain and Smiller 2012]. According to this hypothesis, peptide sequences from pathogens or metabolites mimic self-proteins, leading to autoimmune responses. In the context of MS, cross-reactivity can occur when T cells recognize antigens via HLA class II molecules, particularly when the EBNA-1 peptide from the Epstein-Barr virus (EBV) mimics myelin basic protein (MBP).

It is well established that polymorphic residues within the peptide-binding groove of HLA-DRB1 molecules are critical for determining peptide specificity and T-cell receptor (TCR) recognition. Alterations in these residues modify the biochemical characteristics of the so-called binding pockets (P1 to P9) within the class II HLA groove, where peptide side chains are accommodated. Peptides that bind with high affinity to a specific HLA class II allele often share conserved amino acid motifs, with relatively strict preferences in pockets P1, P4, P6, P7, and P9 [Karnaukhov et al. 2022].

Based on this knowledge, we hypothesize that MS-associated HLA-DRB1 alleles, when in a receptive conformational state, may adopt structural arrangements that enhance interaction with disease-specific antigens. This facilitated conformation may, in turn, promote the activation of autoreactive T cells by increasing the likelihood of HLA-antigen binding events related to MS.

Despite advances in our understanding, the specific contributions of key HLA residues to subtle differences in antigen interactions in MS remain poorly elucidated. Molecular dynamics (MD) simulations provide a powerful approach to investigate the dynamic interactions between class II HLA molecules and antigenic peptides. However, interpreting the complex data generated by MD simulations presents a significant challenge due to the high dimensionality and volume of information produced.

To address this challenge, we implemented a computational framework that integrates machine learning (ML) techniques and quantum neural network (QNN) to classify the different HLA-DRB1 alleles found in the population. To the best of our knowledge, this is the first study to combine MD simulations, ML, and QNN methodologies to investigate structural and dynamic features that may subtly influence MS susceptibility across different HLA-DRB1 polymorphisms.

3. Goals

This study aims to investigate whether conformational and energetic differences in the peptide binding groove of different HLA-DRB1 alleles influence susceptibility to MS. We employ a combined machine learning and quantum neural network approach using MD trajectories as input, aiming to classify alleles as either risk-associated or protective.

4. Methods

This study was designed to combine MD, ML and QNN models into a framework (Figure 1). Machine learning, including convolutional neural networks (CNN) and linear neural networks, was employed to capture independent features. CNN were further utilized to model feature interactions and QNN were applied to model advanced nonlinear dynamics and explore higher-order patterns from complex feature interactions.

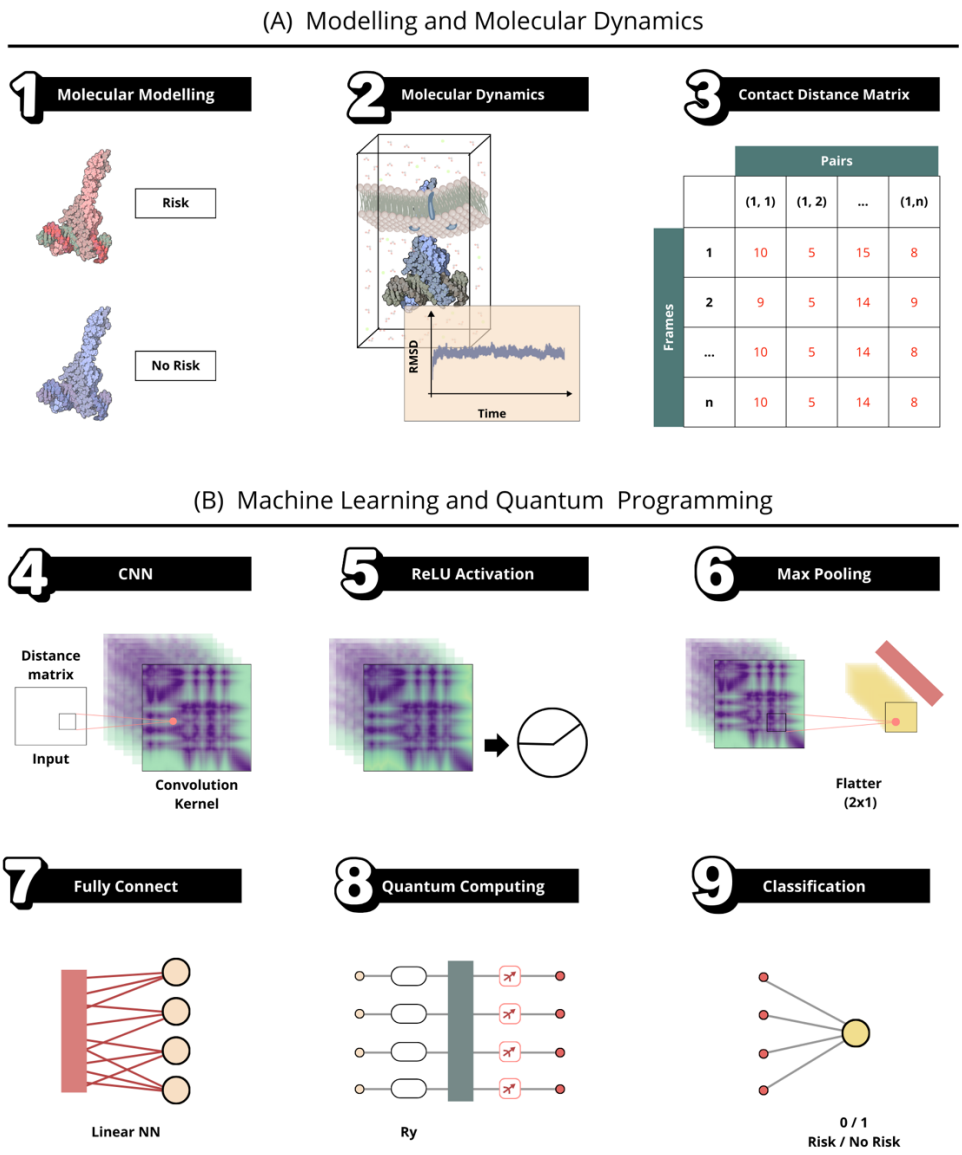

Figure 1. Overview of DM-MLWorkflow of the modeling and predictive analysis process. (A) Molecular modeling and molecular dynamics simulations: (1) protein structures of risk and non-risk variants are modeled; (2) molecular dynamics simulations are performed to assess temporal behavior; (3) contact distance matrices are extracted from the trajectories. (B) Machine learning and quantum computing: (4) distance matrices are used as input for a convolutional neural network (CNN); (5) ReLU activation is applied; (6) max pooling layer reduces dimensionality; (7) fully connected layer processes the features; (8) quantum computing is used for advanced processing; (9) final classification distinguishes between risk and non-risk variants.

4.1. Modeling and Molecular Dynamics Simulations

To obtain a complete model of the HLA-DRB115:01 heterodimer, crystallographic coordinates from the Protein Data Bank (PDB ID: 1YMM) were used as a starting structure in MODELLER [Sali and Blundell 1993]. Structural gaps were complemented using AlphaFold models AF-P01903-F1 and AF-P01911-F1. This model was used as a template to construct the DRB501:01 allele (used as a control) and four additional polymorphic HLA-DRB1 alleles: DRB115:01 (risk), DRB115:03 (risk), DRB101:01 (no risk) and DRB115:02 (no risk). The four heterodimers were embedded into heterogeneous lipid bilayers using CHARMM-GUI [Wu et al. 2014] and subjected to 500ns of MD simulations with GROMACS [Van Der Spoel et al. 2005]. The resulting trajectories were analyzed using MDAnalysis [Michaud-Agrawal et al. 2011].

4.2. Machine Learning and Quantum Neural Network

Distance matrices extracted from MD trajectories were used as input for ML-QNN framework. The method was employed to focus on residues that significantly contribute to differences in dynamic behavior among -DRB1 alleles. The input features are the contact distances between the alpha carbons of the residues in the DRA (res. 1-80) and DRB1 (res. 1-95) chains, which form the HLA interaction groove. A threshold of ≤ 10Å was set to capture the pairs that most contribute to the groove dynamics. From the 500 ns of MD simulations, frames after 2.5 ns, when the complexes had stabilized, were used. A total of 200,000 frames and 921 features were obtained for complexes between chain A of the allele and peptide, and 200,000 frames with 1,059 features for complexes between chain B and peptide. From these matrices, a subset of 1,000 frames (samples) was randomly selected. Eighty percent of the samples were designated as the training set, and twenty percent as the test set.

For this purpose, the libraries PennyLane (https://pennylane.ai), along with Scikit-Learn (https://scikit-learn.org/stable/), Torch (https://pytorch.org) were employed.

The CNN consisted of a one-dimensional convolutional layer with one input channel, eight output channels, and a kernel size of five, followed by a ReLU activation function. A one-dimensional max pooling layer with a kernel size of two was then applied. The output was flattened into a one-dimensional vector and passed through a linear neural network, which produced four output values used as input features for the quantum circuit.

The QNN was implemented using a quantum circuit with four qubits. Classical input features were encoded into quantum states via AngleEmbedding using RY rotations, and StronglyEntanglingLayers were applied to introduce entanglement and trainable rotations among the qubits. The circuit output corresponded to the expectation value of the Pauli-Z operator on each qubit, resulting in four real-valued outputs. These values were then passed through a linear neural network with a sigmoid activation function, producing a single real-valued output interpreted as 0 for risk and 1 for no risk associated alleles.

The hybrid ML-QNN model was trained using the Adam optimizer, an adaptive gradient descent method, with a learning rate of 0.001 for 100 epochs.

5. Results and Discussion

The hybrid CNN-QNN model was evaluated using three distinct input configurations derived from distance matrices involving two protein chains (A and B) and one peptide chain (C): (i) distances between protein chain A and the peptide (A/C), (ii) distances between protein chain B and the peptide (B/C), and (iii) combined distances from both chains A and B to the peptide (A/C_B/C). The objective was to assess the ability of the model to classify molecular configurations as either risk-associated (class 0) or non-risk (class 1).

For the A/C configuration, the model exhibited robust classification performance. The confusion matrix indicated 111 true negatives (class 1) and 84 true positives (class 0), with only 5 false negatives. These results suggest that the spatial features derived from the distances between chain A and the peptide show highly informative patterns relevant for classification. A slight tendency to underclassify the no risk category was observed; however, the overall test accuracy remained high at 0.9750 (Figure 2A).

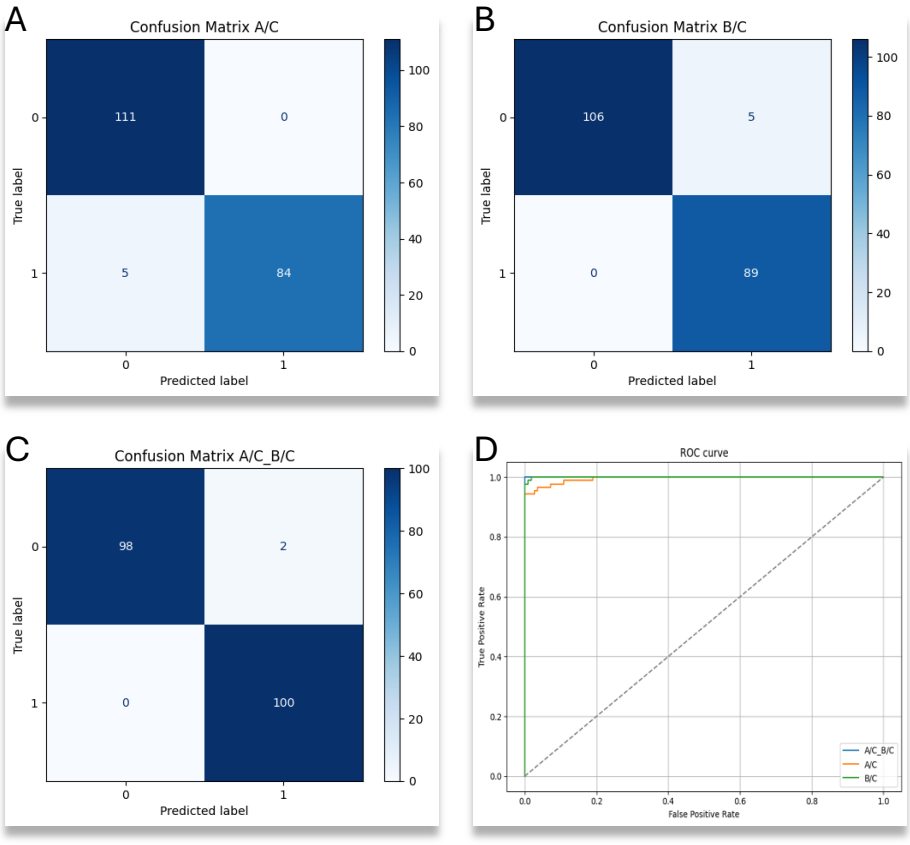

Figure 2. Confusion matrices and ROC curves showing the classification performance of the CNN-QNN model using three input configurations: A/C (A), B/C (B), and combined A/C_B/C (C). The combined configuration achieved the highest accuracy and AUC, highlighting the advantage of integrating spatial features from both protein chains (D).

In the B/C configuration, the model showed similarly strong performance, with 106 true negatives and 89 true positives, and only 5 false positives. This demonstrates that the spatial relationships between chain B and the peptide also carry discriminative structural information. The test accuracy for this configuration also reached 0.975, confirming the ability of the model to classify well with this input modality (Figure 2B).

When combining the A/C and B/C distance matrices, the model achieved its best overall performance, correctly classifying 98 instances of class 0 and all 100 instances of class 1. Only two misclassifications occurred, resulting in the highest test accuracy observed across all configurations (0.990) over 40 epochs. These findings show the advantage of integrating spatial features from both protein chains, allowing the model to capture a more complete representation of the interaction landscape between the dimer and the peptide (Figure 2C).

These classification outcomes were further supported by ROC curve analysis, which showed consistently high AUC values across all input configurations. Notably, the curve corresponding to the combined input (A/C_B/C) outperformed the individual configurations, confirming that the integration of complete spatial information enhances the discriminative capacity of the model. Overall, the hybrid model effectively combines classical convolutional feature extraction with quantum-based classification, demonstrating high predictive accuracy in distinguishing molecular states. These results reinforce the potential of quantum-classical hybrid models for analyzing complex spatial and structural biomolecular data (Figure 2D).

Among the three input configurations, the combined A/C_B/C representation led to the fastest and most stable convergence, achieving the highest classification accuracy and lowest loss. By epoch 40, the model reached an accuracy of 0.998 with a corresponding loss of 0.445, indicative of very high performance. This result shows the importance of integrating spatial features from both protein chains, allowing the model to more effectively capture the full complexity of the dimer–peptide interaction landscape (Figure 3).

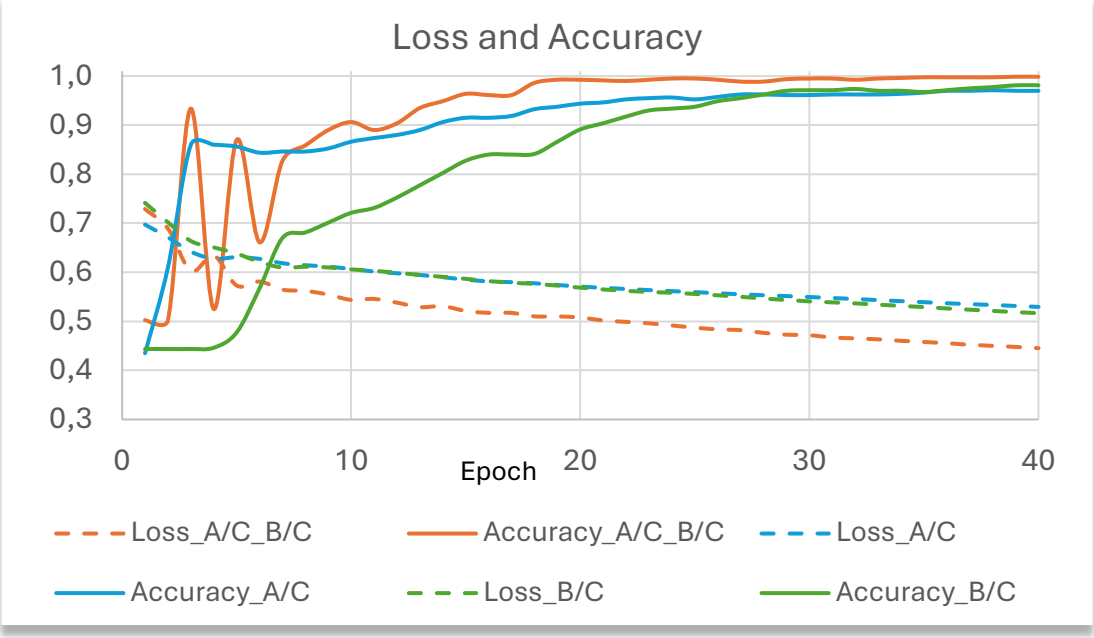

Figure 3. Training accuracy and loss curves over 40 epochs for the CNN-QNN model using A/C, B/C, and combined A/C_B/C inputs. The combined configuration showed the fastest convergence, highest accuracy, and lowest loss, highlighting the benefit of integrating spatial features from both protein chains.

In contrast, models trained exclusively on A/C or B/C distance matrices showed slightly lower performance, but still robust. The A/C configuration reached an accuracy of 0.970 with a final loss of 0.529, while the B/C configuration achieved 0.981 accuracy with a loss of 0.516. These findings suggest that although each complex independently contributes to discriminative structural information, their combination show a more comprehensive spatial representation.

Moreover, the learning curves for the combined A/C_B/C input show reduced oscillations and more rapid stabilization during early training epochs compared to the individual configurations. This observation further supports the hypothesis that a combined spatial feature space improves the ability of the model to learn, leading to better predictive performance. Therefore, our models were able to successfully classify risk and protective HLA-DRB1 alleles, demonstrating that the proposed approach can address the original biological question.

6. Conclusion

This study demonstrated the classification capabilities of a hybrid (MD-ML-QNN) model in distinguishing risk-associated from non-risk molecular complexes based on protein-peptide distance resulting from MD trajectories. Our evaluation across three distinct input configuration, individual protein chain interactions (A/C and B/C) and their combined representation (A/C_B/C), revealed the importance of comprehensive spatial molecular structures and their relationships.

By using 200,000 instances, a high number of features per dataset, and by sampling subsets of 1,000 frames across multiple iterations, we ensured both robustness and diversity in the training and testing processes.

The model showed high performance across all configurations. However, the synergistic integration of distance features from both protein chains (A/C_B/C) proved to be superior, achieving the highest test accuracy of 0.990. This improved performance was further corroborated by ROC curve analysis, where the combined input configuration showed the highest AUC values. Furthermore, the results of combined A/C_B/C configuration show the faster and more stable convergence.

In summary, this research strongly supports the advantage of integrating complete spatial information in the analysis of complex biomolecular data. These findings show the way for future applications of quantum-classical hybrid models in diverse areas of biomolecular research and drug discovery.

Future work will focus on expanding the analysis to additional alleles, increasing the number of qubits in the quantum models, and integrating larger datasets.

7. References

Adiele RC, Adiele CA. (2019) “Metabolic defects in multiple sclerosis. Mitochondrion”. 2019 Jan; 44:7-14. doi: 10.1016/j.mito.2017.12.005. Epub 2017 Dec 13. PMID: 29246870.

Nourbakhsh B, Mowry EM. (2019) “Multiple Sclerosis Risk Factors and Pathogenesis”. Continuum (Minneap Minn). Jun;25(3):596-610. doi: 10.1212/CON.0000000000000725.

Hollenbach JA, Oksenberg JR. (2015) “The immunogenetics of multiple sclerosis: A comprehensive review”. J Autoimmun. Nov;64:13-25. doi: 10.1016/j.jaut.2015.06.010. Epub 2015 Jul 2. PMID: 26142251; PMCID: PMC4687745.

De Silvestri A, Capittini C, Mallucci G, Bergamaschi R, Rebuffi C, Pasi A, Martinetti M, Tinelli C. (2019) “The Involvement of HLA Class II Alleles in Multiple Sclerosis: A Systematic Review with Meta-analysis”. Dis Markers. Nov 6;2019:1409069. doi: 10.1155/2019/1409069. PMID: 31781296; PMCID: PMC6875418.

Rock KL, Reits E, Neefjes J. (2016) “Present Yourself! By MHC Class I and MHC Class II Molecules”. Trends Immunol. Nov;37(11):724-737. doi: 10.1016/j.it.2016.08.010. Epub 2016 Sep 7. PMID: 27614798; PMCID: PMC5159193.

Chastain EM, Miller SD. (2012) “Molecular mimicry as an inducing trigger for CNS autoimmune demyelinating disease”. Immunol Rev. Jan;245(1):227-38. doi: 10.1111/j.1600-065X.2011.01076.x. PMID: 22168423; PMCID: PMC3586283.

Karnaukhov V, Paes W, Woodhouse IB, Partridge T, Nicastri A, Brackenridge S, Shcherbinin D, Chudakov DM, Zvyagin IV, Ternette N, Koohy H, Borrow P, Shugay M. (2022) “HLA variants have different preferences to present proteins with specific molecular functions which are complemented in frequent haplotypes”. Front Immunol. Dec 20;13:1067463. doi: 10.3389/fimmu.2022.1067463. PMID: 36605212; PMCID: PMC9808399.

Sali A, Blundell TL. (1993) “Comparative protein modelling by satisfaction of spatial restraints”. J Mol Biol. Dec 5;234(3):779-815. doi: 10.1006/jmbi.1993.1626. PMID: 8254673.

Wu EL, Cheng X, Jo S, Rui H, Song KC, Dávila-Contreras EM, Qi Y, Lee J, Monje-Galvan V, Venable RM, Klauda JB, Im W. (2014) “CHARMM-GUI Membrane Builder toward realistic biological membrane simulations”. J Comput Chem. Oct 15;35(27):1997-2004. doi: 10.1002/jcc.23702. Epub 2014 Aug 7. PMID: 25130509; PMCID: PMC4165794.

D. Van Der Spoel, et al., (2005) “GROMACS: Fast, Flexible, and Free”, J Comput Chem, vol. 26, pp. 1701-1718, October.

Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJ. (2005) “ GROMACS: fast, flexible, and free”. J Comput Chem. Dec;26(16):1701-18. doi: 10.1002/jcc.20291. PMID: 16211538.

Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O. (2011) “MDAnalysis: a toolkit for the analysis of molecular dynamics simulations”. J Comput Chem. Jul 30;32(10):2319-27. doi: 10.1002/jcc.21787. Epub 2011 Apr 15. PMID: 21500218; PMCID: PMC3144279.