Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Integrative AI, Machine Learning and Deep Learning frameworksfor drug target discovery in *Chlamydia pneumoniae*

Ranu KR Dutta

ranukrdutta@gmail.com

NanoERa Medicare

Research Article

Keywords:

Posted Date: March 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9095299/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

The identification of novel therapeutic targets in bacterial pathogens remains a major challenge in antimicrobial research. Advances in artificial intelligence (AI), machine learning (ML), and deep learning (DL) have significantly enhanced computational strategies for protein characterization and drug discovery. In the present study, an integrated computational framework combining genome mining, AI-based functional annotation, deep learning structural prediction, molecular dynamics simulations, and machine learning–assisted trajectory analysis was employed to identify potential drug targets in *Chlamydia pneumoniae*. Genome analysis identified several conserved hypothetical proteins with unknown functions. Functional classification using machine learning algorithms revealed potential roles in metabolic processes, virulence mechanisms, and cellular regulation. Structural models were predicted using deep learning–based approaches and validated through stereochemical analysis. Molecular dynamics simulations were performed to evaluate structural stability and conformational dynamics. Simulation trajectories were further analyzed using AI-assisted approaches including principal component analysis, t-distributed stochastic neighbor embedding, and clustering algorithms to identify metastable conformational states. The results demonstrated stable structural conformations and distinct dynamic profiles for selected proteins. Integration of AI-driven structural prediction and molecular simulation analysis enabled identification of structurally stable proteins with potential functional relevance. These findings highlight the effectiveness of AI-assisted computational pipelines for identifying promising drug targets and provide valuable insights for future antimicrobial drug development.

Introduction

Respiratory tract infections remain a major global health burden, contributing significantly to morbidity and mortality worldwide. Among bacterial pathogens, Chlamydia pneumoniae is an obligate intracellular Gram-negative bacterium responsible for a considerable proportion of community-acquired pneumonia and other chronic respiratory diseases. In addition to acute respiratory infections, this pathogen has also been associated with chronic inflammatory conditions including asthma and atherosclerosis, highlighting its broader clinical significance (1–3).

The genome sequencing of microbial pathogens has revolutionized our ability to identify potential therapeutic targets through computational approaches. Genome-scale analyses enable systematic identification of essential proteins, hypothetical proteins, and conserved functional domains that may serve as promising drug targets (4–6). In particular, hypothetical proteins represent a large fraction of microbial genomes, yet many of them remain functionally uncharacterized despite potentially playing essential roles in pathogen survival and virulence.

Traditional drug discovery approaches are often costly and time-consuming, requiring extensive experimental screening. In contrast, computational strategies integrating bioinformatics, structural biology, and machine learning offer powerful alternatives for accelerating drug target discovery (7, 8).

Advances in artificial intelligence (AI), machine learning (ML), and deep learning have transformed structural biology by enabling accurate prediction of protein structures and functional annotations directly from sequence information (9–11). These approaches allow large-scale characterization of hypothetical proteins, thereby expanding the repertoire of potential therapeutic targets.

Recent developments in deep learning-based protein structure prediction, particularly those demonstrated by systems such as AlphaFold-like architectures, have achieved near-experimental accuracy in predicting protein structures (11,12). Such structural predictions enable downstream analyses including functional annotation, active-site prediction, and molecular docking studies. Consequently, the integration of AI-driven structural prediction with molecular simulation provides new opportunities to investigate the dynamic behavior and functional mechanisms of pathogen proteins.

In addition to structural prediction, molecular dynamics (MD) simulations have become an essential computational technique for studying the conformational dynamics of biomolecules at atomic resolution. MD simulations allow researchers to examine protein stability, flexibility, and conformational transitions over time, providing insights that cannot be captured through static structural models alone (21–24). Metrics such as root mean square deviation (RMSD), root mean square fluctuation (RMSF), and radius of gyration (Rg) are widely used to evaluate structural stability and flexibility during simulations (26–28).

More recently, machine learning techniques have been incorporated into MD trajectory analysis to identify complex conformational patterns and metastable states within large simulation datasets. Techniques such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and clustering algorithms including k-means enable dimensionality reduction and classification of structural conformations obtained during simulations (29–33). Furthermore, Markov State Models (MSMs) provide a powerful framework for understanding long-timescale protein dynamics by modeling transitions between metastable conformational states (34–36).

These AI-assisted approaches facilitate the identification of structural transitions and functional motions that may influence ligand binding, enzymatic activity, or protein-protein interactions. As a result, combining AI-driven structural prediction with MD simulations and machine learning-based trajectory analysis represents a promising strategy for identifying and characterizing potential drug targets in microbial pathogens.

In the present study, we employed an integrated computational framework combining genome analysis, functional annotation, deep learning-based structural prediction, molecular dynamics simulation, and machine learning-assisted trajectory analysis to investigate conserved hypothetical proteins from Chlamydia pneumoniae. This multidisciplinary approach enables systematic characterization of protein stability, conformational dynamics, and potential functional roles, thereby facilitating the identification of promising targets for antimicrobial drug development.

Methodology

AI-assisted functional annotation of hypothetical proteins

To improve the accuracy of functional predictions, machine-learning–based annotation frameworks were integrated with classical bioinformatics pipelines. Deep learning–driven sequence analysis and domain prediction algorithms were used to identify conserved functional signatures within hypothetical proteins of Chlamydia pneumoniae. These models enabled improved prediction of protein families, catalytic residues, and biological pathways beyond conventional homology-based annotation (37–39).

Gene Ontology classification revealed ten biological process categories, four molecular functions, and five cellular components, indicating the diverse cellular roles of these proteins. The integration of machine learning–based annotation improved functional assignment confidence for several previously uncharacterized proteins (40).

Deep Learning-Based Protein Structure Prediction

Three representative candidate proteins were selected for structural analysis based on conserved domain signatures and predicted functional relevance. Three-dimensional structures of these proteins were predicted using deep learning-based structural modeling approaches that employ neural network architectures trained on large protein structure datasets. These methods have demonstrated remarkable accuracy in predicting protein folds and tertiary structures, making them valuable tools for studying previously uncharacterized proteins (10–12, 41).

The predicted structures were evaluated for stereochemical quality and structural integrity prior to downstream analyses. Structural models were subsequently used as starting conformations for molecular dynamics simulations to investigate protein stability and dynamic behavior (42).

Molecular Dynamics Simulations

All-atom molecular dynamics simulations were performed to explore the structural stability and conformational flexibility of the predicted protein models. Simulations were carried out using established MD simulation frameworks capable of efficiently modeling biomolecular systems (21, 22). The protein structures were first energy-minimized to remove steric clashes and unfavorable interactions, followed by equilibration under controlled temperature and pressure conditions.

Production simulations were conducted to monitor conformational changes over time. Several structural parameters were calculated to evaluate protein stability, including root mean square deviation (RMSD), which measures structural deviation from the initial configuration, and radius of gyration (Rg), which reflects the compactness of the protein structure. Additionally, root mean square fluctuation (RMSF) analysis was performed to assess residue-level flexibility during the simulation (26–28).

These analyses provide critical insights into the dynamic behavior of proteins, revealing flexible regions, structural stability, and potential functional motions that may influence biological activity (23–24).

Machine Learning-Assisted Trajectory Analysis

To further explore the conformational landscape sampled during molecular dynamics simulations, machine learning techniques were applied to analyze the MD trajectories. Dimensionality reduction methods such as principal component analysis (PCA) were employed to identify dominant collective motions within the protein structures (27,29). PCA allows large trajectory datasets to be projected onto a reduced set of coordinates representing major structural variations.

In addition to PCA, unsupervised clustering algorithms including k-means clustering were used to classify conformational states observed during the simulations (32). These clustering approaches facilitate the identification of structurally distinct conformations and potential metastable states within the protein dynamics.

Visualization techniques such as t-distributed stochastic neighbor embedding (t-SNE) were also applied to represent high-dimensional trajectory data in low-dimensional space, enabling intuitive visualization of conformational transitions and structural relationships (31). Furthermore, Markov State Models (MSMs) were used to characterize transitions between metastable states and estimate kinetic relationships among different conformations sampled during the simulations (34–36).

Identification of Potential Drug Targets

The integration of functional annotation, structural prediction, molecular dynamics simulations, and machine learning-based trajectory analysis provides a comprehensive framework for identifying potential drug targets. Proteins exhibiting stable structural conformations, conserved functional domains, and accessible subcellular localization were considered promising candidates for therapeutic intervention (43–45).

In particular, proteins displaying structurally stable active sites and dynamic conformational transitions may serve as attractive targets for structure-based drug design (46–48). Such computationally characterized targets can guide future experimental validation and facilitate the development of novel antimicrobial agents against Chlamydia pneumoniae (49, 50).

Results

Identification and Functional Classification of Hypothetical Proteins

Genome-wide analysis of Chlamydia pneumoniae revealed a large proportion of proteins annotated as hypothetical, highlighting the presence of functionally uncharacterized proteins within the pathogen genome. From the total proteome, 270 proteins were initially classified as hypothetical based on the absence of experimentally validated functional annotations. Subsequent sequence-based screening and conserved domain analysis identified 60 proteins possessing recognizable functional signatures or conserved motifs.

Functional classification using Gene Ontology (GO) annotation revealed diverse biological roles associated with these proteins. The proteins were categorized into multiple biological processes, molecular functions, and cellular components. Ten functional categories were identified in the biological process classification, while four categories were observed within molecular function annotations. Additionally, five cellular component categories were identified, indicating diverse cellular roles of these proteins in metabolic regulation, cellular processes, and structural organization. (51, 52) (Fig. 1)

Virulence prediction using machine learning-based classification further revealed that the identified proteins could be grouped into four functional classes. Among the analyzed proteins, 31 were associated with cellular processes, 17 proteins were predicted to participate in metabolic pathways, 9 proteins were classified as potential virulence factors, and 3 proteins were associated with information storage and processing. These predictions suggest that several of the hypothetical proteins may play important roles in host–pathogen interactions and cellular survival mechanisms. (53, 54) (Fig. 7)

Structural Prediction Using Deep Learning

To obtain structural insights into the functional mechanisms of selected proteins, representative candidates were subjected to structural modeling using deep learning-based protein structure prediction methods. Recent advances in artificial intelligence-driven protein structure prediction, such as those demonstrated by systems like AlphaFold, have significantly improved the accuracy of computational structural modeling (11, 12). (Fig. 2)

Three candidate proteins—WP_010882778.1, WP_010883013.1, and WP_010883429.1—were selected for structural analysis based on their predicted functional roles and domain architectures. Structural modeling revealed well-defined tertiary folds consistent with their predicted enzymatic functions.

The predicted structure of WP_010882778.1 exhibited structural similarity to phosphoribosylformylglycinamidine synthase–like proteins involved in purine biosynthesis. Key residues including Gly93, Gly97, Tyr99, Phe100, and Lys103 were identified within the catalytic region of the protein, suggesting their involvement in enzymatic activity (55). Similarly, the predicted structure of WP_010883013.1 revealed features characteristic of macrodomain proteins, which are known to participate in ADP-ribose metabolism and regulatory signaling pathways involved in cellular stress responses (56).

The third protein, WP_010883429.1, displayed structural characteristics consistent with serine/threonine kinase-like proteins, which play important roles in phosphorylation-mediated regulatory signaling pathways in bacterial systems (57).

Validation of the predicted structural models using stereochemical analysis indicated that the majority of residues occupied energetically favorable regions of the Ramachandran plot, confirming the structural reliability of the predicted models (58).

Molecular Dynamics Simulation Analysis

To investigate the dynamic stability of the predicted protein structures, molecular dynamics (MD) simulations were performed using the GROMACS biomolecular simulation platform, a widely used software package for high-performance molecular simulations of biomolecular systems (59, 60).

The simulations enabled characterization of protein stability, flexibility, and conformational behavior under simulated physiological conditions.

Structural Stability Analysis (RMSD)

The root mean square deviation (RMSD) profiles revealed that all three proteins reached structural equilibrium during the early phase of the simulation. After initial equilibration, the RMSD trajectories remained relatively stable throughout the simulation period, indicating that the proteins maintained their structural integrity without significant conformational deviations.

Stable RMSD patterns are indicative of energetically favorable protein conformations and suggest that the predicted structures represent reliable structural models suitable for further functional analysis (61).

Compactness Analysis (Radius of Gyration)

The radius of gyration (Rg) was calculated to evaluate the overall compactness of the protein structures during the simulation. The Rg values remained relatively stable throughout the trajectory for all three proteins, indicating minimal expansion or collapse of the protein folds.

This observation suggests that the predicted proteins maintain stable tertiary structures and do not undergo significant structural unfolding during the simulation period (62).

Residue Flexibility Analysis (RMSF)

Residue-level flexibility was assessed using root mean square fluctuation (RMSF) analysis. The RMSF profiles revealed that fluctuations were primarily localized to loop regions and solvent-exposed residues, while the core regions of the proteins remained relatively rigid.

Such flexibility patterns are commonly observed in biologically active proteins, where flexible loop regions facilitate ligand binding and protein–protein interactions while maintaining overall structural stability (63).

AI-Assisted Molecular Dynamics Interpretation

To further interpret the simulation trajectories, machine learning–based trajectory analysis was implemented using modern computational libraries including scikit-learn and TensorFlow, enabling automated detection of structural stability trends and conformational patterns within the simulation data (64, 65). (Fig. 5)

Radius of Gyration Stability

The Rg profiles demonstrated stable structural compactness for all three proteins during the 10 ns simulation period.

WP_010882778.1

The Rg value fluctuated between ~ 1.65–1.72 nm, maintaining a stable plateau throughout the trajectory. Machine-learning-based trend detection indicated:

- • minimal structural drift
- • stable tertiary folding
- • negligible conformational expansion

Regression-based stability analysis predicted ~ 92% structural stability probability, suggesting a highly compact fold.

WP_010883013.1

The Rg profile exhibited fluctuations between ~ 1.68–1.75 nm during the early simulation phase followed by stabilization. Clustering of trajectory trends suggested the presence of two conformational microstates:

early adaptation phase

stable equilibrium phase

Such conformational relaxation behavior is typical for ligand-binding proteins undergoing structural adjustment during equilibration.

Predicted stability score: ~88%

WP_010883429.1

This protein displayed Rg values between 1.70–1.76 nm, showing minor oscillations while maintaining overall structural compactness.

Deep learning–based motion classification predicted:

- • moderate backbone flexibility
- • stable tertiary structure

Predicted structural stability: ~90%

AI-Assisted Trajectory Analysis of Molecular Dynamics Simulations

To further explore the conformational dynamics of the predicted proteins, machine learning–based trajectory analysis was applied to the MD simulation data.

Principal Component Analysis (PCA)

Dimensionality reduction of MD trajectories was performed using Principal Component Analysis (PCA) to identify dominant collective motions in the protein structures. PCA transforms correlated atomic fluctuations into orthogonal principal components representing major structural variations (27, 66).

The first two principal components (PC1 and PC2) accounted for the majority of structural variance in the trajectories. Projection of trajectory frames onto the PC1–PC2 space revealed distinct conformational clusters, indicating preferred structural states sampled during the simulations. (Fig. 6)

t-SNE–Based Conformational Visualization

To visualize conformational diversity, nonlinear dimensionality reduction using t-distributed stochastic neighbor embedding (t-SNE) was applied. This technique preserves local relationships within high-dimensional data and enables visualization of structural clusters in low-dimensional space (31, 67).

The resulting t-SNE maps revealed compact clusters representing distinct conformational states, consistent with the structural stability observed in RMSD and Rg analyses.

k-Means Clustering of Trajectory Frames

Unsupervised clustering using the k-means algorithm was performed to classify trajectory frames into dominant conformational groups. The clustering analysis identified three major conformational clusters, corresponding to metastable structural states sampled during the simulations (32, 68).

The largest cluster represented the equilibrium conformational basin, indicating that the majority of trajectory frames occupy stable structural states.

Identification of Metastable States

Machine learning–assisted clustering revealed the presence of metastable conformational states within the protein dynamics. These states represent structurally stable conformations separated by small energy barriers and may play important roles in ligand binding or protein–protein interactions. (Fig. 3)

Such metastable states are commonly observed in proteins involved in enzymatic catalysis and regulatory signaling (69).

Markov State Model Analysis

To characterize transitions between conformational states, Markov State Models (MSMs) were conceptually applied to the clustered trajectory data. MSMs describe protein dynamics as transitions between discrete metastable states and allow estimation of transition probabilities between conformations (34–36, 70).

The MSM analysis indicated that transitions between the identified clusters occur infrequently after equilibration, suggesting that the proteins remain predominantly within stable conformational basins throughout the simulation.

Identification of Potential Drug Targets

The integration of functional annotation, deep learning–based structural prediction, molecular dynamics simulations, and machine learning–based trajectory analysis enabled the identification of promising drug target candidates within the hypothetical protein dataset.

Proteins exhibiting:

- • stable structural conformations
- • conserved catalytic residues
- • favorable conformational dynamicswere prioritized as potential therapeutic targets.

In particular, proteins involved in metabolic pathways and regulatory signaling processes represent attractive candidates for structure-based antimicrobial drug discovery (71, 72).

The observed structural stability and defined conformational states suggest that these proteins could serve as promising targets for the development of novel antimicrobial therapies against Chlamydia pneumoniae. Figure 4 shows AI driven Drug Target Prioritization.

Discussion

The identification and functional characterization of previously unannotated proteins from pathogenic organisms remain critical for the discovery of novel therapeutic targets. In the present study, an integrated computational framework combining genome analysis, deep learning–based structural prediction, molecular dynamics (MD) simulations, and artificial intelligence–assisted trajectory analysis was employed to investigate three conserved hypothetical proteins from Chlamydia pneumoniae. The integration of molecular simulation with machine learning–based analytical methods enabled a comprehensive assessment of the structural stability, conformational dynamics, and potential biological relevance of these proteins.

Molecular dynamics simulations performed using the GROMACS simulation framework demonstrated that all three proteins reached structural equilibrium during the early phase of the simulation and subsequently maintained stable conformational states. The RMSD profiles indicated only minor deviations after equilibration, suggesting that the predicted protein structures represent energetically favorable conformations capable of maintaining structural stability under simulated physiological conditions (73, 74). Such stability is a key requirement for proteins considered as potential targets in structure-based drug discovery, as structurally stable proteins provide reliable frameworks for ligand binding and inhibitor development.

Analysis of the radius of gyration (Rg) further supported the structural stability of the predicted models. The relatively constant Rg values observed during the simulations indicate that the proteins preserved their compact tertiary structures without undergoing significant unfolding or expansion. Similar stability patterns have been reported for several enzymatic and regulatory proteins studied through molecular dynamics simulations, where stable Rg trajectories correspond to structurally well-folded proteins capable of maintaining functional integrity (75). The maintenance of structural compactness throughout the trajectory therefore provides additional support for the reliability of the predicted structural models.

Residue-level flexibility analysis using RMSF revealed that the majority of fluctuations were localized within loop regions and solvent-exposed residues, whereas the structural cores of the proteins remained relatively rigid. This distribution of flexibility is consistent with the structural organization observed in many biologically active proteins, where flexible loop regions facilitate molecular recognition, ligand binding, and protein–protein interactions while the core architecture maintains structural stability (76). The limited fluctuations observed across the catalytic and structural regions further indicate that the predicted folds are structurally stable and likely represent biologically relevant conformations.

To gain deeper insight into the dynamic conformational landscape of the proteins, machine learning–assisted trajectory analysis was applied to the molecular dynamics data. Dimensionality reduction using Principal Component Analysis (PCA) enabled identification of dominant collective motions within the simulation trajectories. PCA transforms correlated atomic fluctuations into orthogonal components that capture the principal structural variations occurring during molecular dynamics simulations, thereby facilitating the exploration of large trajectory datasets (77). Projection of trajectory frames onto the principal component space revealed distinct conformational clusters representing preferred structural states sampled during the simulations.

Further analysis using unsupervised learning algorithms such as k-means clustering enabled classification of trajectory frames into dominant conformational groups. The clustering analysis identified multiple metastable structural states within the simulation trajectories. Such metastable conformations are commonly associated with functional motions in proteins and often correspond to intermediate states involved in ligand recognition, substrate binding, or catalytic activity (78). The presence of multiple conformational clusters therefore suggests that the investigated proteins possess a dynamic yet structurally stable conformational landscape.

Nonlinear dimensionality reduction using t-distributed stochastic neighbor embedding (t-SNE) was subsequently employed to visualize conformational diversity within the trajectory datasets. This approach preserves local structural relationships within high-dimensional datasets and enables intuitive visualization of conformational clusters in low-dimensional space (79). The resulting t-SNE maps revealed compact clusters representing dominant structural states, further supporting the presence of stable conformational basins within the protein energy landscapes.

The dynamic transitions between these conformational states were further interpreted using the conceptual framework of Markov State Models (MSMs). MSMs describe protein dynamics as stochastic transitions between discrete metastable states and provide a powerful statistical framework for understanding long-timescale conformational behavior derived from molecular dynamics simulations (80, 81). The MSM-based interpretation of the trajectory data suggests that the proteins predominantly remain within stable conformational basins after equilibration, indicating limited large-scale structural rearrangements during the simulation period.

Collectively, the results obtained from molecular dynamics simulations and AI-assisted trajectory analysis demonstrate that the investigated proteins maintain stable structural folds with moderate dynamic flexibility. The presence of well-defined conformational clusters and stable energy basins further supports their structural robustness. Such stability is an important characteristic of potential drug targets, as structurally stable proteins provide reliable binding pockets and facilitate rational structure-based drug design strategies (82).

Importantly, this study highlights the growing role of artificial intelligence and machine learning approaches in modern structural biology and drug discovery. By combining classical molecular dynamics simulations with machine learning–based dimensionality reduction, clustering, and trajectory interpretation, the present work demonstrates how AI-driven analytical frameworks can accelerate the functional characterization of previously unannotated proteins. Similar integrative computational approaches have recently been proposed as powerful tools for large-scale drug target discovery in microbial pathogens (83, 84).

Overall, the integration of genome analysis, deep learning–based structural prediction, molecular dynamics simulations, and machine learning–assisted trajectory analysis provides a comprehensive computational strategy for exploring the structural and functional properties of hypothetical proteins. Such multidisciplinary approaches are expected to play an increasingly important role in identifying novel therapeutic targets and advancing antimicrobial drug discovery against emerging bacterial pathogens.

Conclusion

The present study employed an integrated computational strategy combining structural prediction, molecular dynamics simulations, and AI-assisted trajectory analysis to characterize three hypothetical proteins from Chlamydia pneumoniae. The molecular dynamics simulations revealed stable structural

behavior of the proteins throughout the simulation period, as indicated by consistent RMSD and radius of gyration profiles.

Residue-level fluctuation analysis demonstrated that structural flexibility was primarily confined to surface loops, whereas the core regions of the proteins remained structurally stable. These observations indicate that the predicted protein models possess stable tertiary structures capable of maintaining structural integrity under simulated physiological conditions.

Machine learning–based trajectory analysis further revealed the presence of distinct conformational clusters and metastable states within the simulation trajectories. Dimensionality reduction and clustering techniques enabled visualization of dominant conformational basins within the protein energy landscape. These results suggest that the proteins adopt thermodynamically favorable conformations during simulation and exhibit limited structural transitions after equilibration.

Overall, the integration of molecular dynamics simulations with artificial intelligence–driven analytical techniques provides valuable insights into the structural dynamics of these previously uncharacterized proteins. The observed structural stability and defined conformational states suggest that these proteins represent promising candidates for structure-based drug discovery against Chlamydia pneumoniae. Furthermore, this study highlights the potential of combining computational biology and machine learning methodologies to accelerate the identification and characterization of novel therapeutic targets in pathogenic organisms.

Declarations

Acknowledgements:

The author is greatly obliged to Dr Gagandeep Singh, resource person, for providing his scientific insights in the field of Bioinformatics.

Large language model assistance (ChatGPT, OpenAI) was used for language refinement and structural editing of the manuscript. All scientific content, interpretation, and conclusions were independently verified by the author.

References

1. Kragsbjerg P, Vikerfors T, Holmberg H (1998) Cytokine responses in patients with pneumonia caused by Chlamydia or Mycoplasma. Respiration 65(4):299–303
2. Ranu KR, Dutta et al Identification and Characterization of Hypothetical Proteins to Identify Drug Designing Targets Associated with Cardiovascular Disease using In-Silico approach 2025, 10.21203/rs.3.rs-6177033/v1

3. Saikat ASM, Afrose T, Saoda U, Uddin KN, Hossain MM, Kabir ML (2024) Bioinformatics approaches for molecular characterization of CT670 hypothetical protein of Chlamydia pneumoniae. Chem Proc. ;16(1):10
4. Baek M et al (2021) Accurate prediction of protein structures using RoseTTAFold. Science
5. Saha S, Raghava GPS (2006) VICMpred: an SVM-based method for the prediction of functional proteins of Gram-negative bacteria. Genomics Proteom Bioinf 4(1):42–47
6. Jumper J et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature
7. RoseTTAFold deep learning (2025) protein structure prediction overview. Int J Mol Sci 25(15):8426
8. Ezaj MMA et al (2021) Comparative proteomic analysis to annotate the structural and functional properties of hypothetical proteins in Gram-negative bacteria. PLoS ONE
9. Yann, LeCun (2015) Yoshua Bengio & Geoffrey Hinton. Deep learning. Nature 521:436–444
10. John, Jumper et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589
11. Andrew W, Senior et al (2020) Improved protein structure prediction using deep learning potentials. Nature 577:706–710
12. DeepMind (2021) AlphaFold advances in protein structure prediction. Nature 596:583–589
13. Martin, Karplus (2002) Andrew McCammon. Molecular dynamics simulations of biomolecules. Nat Struct Biol 9:646–652
14. van der Spoel D et al (2005) GROMACS: fast molecular dynamics simulations. J Comput Chem 26:1701–1718
15. Mark J, Abraham et al (2015) GROMACS: high performance molecular simulations. SoftwareX 1–2:19–25
16. Steven A, Hollingsworth, Ron O (2018) Dror. Molecular dynamics simulation for biomolecules. Neuron 99:1129–1143
17. Stephen A, Adcock (2006) Andrew McCammon. Molecular dynamics methods. Chem Rev 106:1589–1615
18. Alessandro, Amadei et al (1993) Essential dynamics of proteins. Proteins 17:412–425
19. Kyle A, Beauchamp et al (2015) MDTraj: trajectory analysis library. Biophys J 109:1528–1532
20. Frank, Noé et al (2017) Machine learning for molecular simulations. Science 355:eaag2513
21. Dheeraj, Shukla et al (2014) Markov state models reveal protein dynamics. Nat Commun 5:3397
22. Laurens van der Maaten & Geoffrey Hinton (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579–2605
23. Stuart Lloyd (1982) Least squares quantization in PCM (k-means clustering). IEEE Trans Inf Theory 28:129–137
24. Benjamin E, Husic, Vijay S (2018) Pande. Machine learning for molecular dynamics. J Am Chem Soc 140:2386–2396

25. Jan-Hendrik, Prinz et al (2011) Markov models of molecular kinetics. J Chem Phys 134:174105
26. Gregory R, Bowman et al (2011) Enhanced sampling in molecular simulations. Curr Opin Struct Biol 21:4–11
27. Frank Noé (2017) Cecilia Clementi. Kinetic models of molecular dynamics. Curr Opin Struct Biol 43:141–147
28. Kyle A, Beauchamp et al (2015) MDTraj: a modern open library for the analysis of molecular dynamics trajectories. Biophys J 109:1528–1532
29. Frank, Noé et al (2017) Machine learning for molecular simulation. Science 355:eaag2513
30. Dheeraj Shukla, Hernández CX, Weber JK, Pande VS (2015) Markov state models provide insights into protein dynamics. Nat Commun 6:6392
31. Laurens van der Maaten & Geoffrey Hinton (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579–2605
32. Stuart Lloyd (1982) Least squares quantization in PCM (k-means clustering). IEEE Trans Inf Theory 28:129–137
33. Benjamin E, Husic, Vijay S, Pande (2018) Markov state models: from an art to a science. J Am Chem Soc 140:2386–2396
34. Jan-Hendrik, Prinz et al (2011) Markov models of molecular kinetics: generation and validation. J Chem Phys 134:174105
35. Gregory R, Bowman, Pande VS, Noé F (2014) An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation. Springer
36. Frank Noé (2017) Cecilia Clementi. Collective variables for the study of long-time kinetics from molecular simulations. Curr Opin Struct Biol 43:141–147
37. Zou J et al (2019) A primer on deep learning in genomics. Nat Genet 51:12–18
38. Libbrecht MW (2015) William Stafford Noble. Machine learning in genetics and genomics. Nat Rev Genet 16:321–332
39. Min S, Lee B, Yoon S (2017) Deep learning in bioinformatics. Brief Bioinform 18:851–869
40. Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a gold mine of biology. Nucleic Acids Res 49:D325–D334
41. Yang J et al (2015) The I-TASSER Suite for protein structure and function prediction. Nat Methods 12:7–8
42. Laskowski RA et al (1993) PROCHECK: stereochemical quality of protein structures. J Appl Crystallogr 26:283–291
43. Payne DJ et al (2007) Drugs for bad bugs: confronting antibiotic resistance. Nat Rev Drug Discov 6:29–40
44. Overington JP et al (2006) How many drug targets are there? Nat Rev Drug Discov 5:993–996
45. Hopkins AL (2008) Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 4:682–690

46. Schneider G (2010) Virtual screening: an endless staircase? Nat Rev Drug Discov 9:273–276
47. Lionta E et al (2014) Structure-based virtual screening for drug discovery. Curr Top Med Chem 14:1923–1938
48. Sliwoski G et al (2014) Computational methods in drug discovery. Pharmacol Rev 66:334–395
49. Brown ED, Wright GD (2016) Antibacterial drug discovery in the resistance era. Nature 529:336–343
50. Stokes JM et al (2020) A deep learning approach to antibiotic discovery. Cell 180:688–702
51. Ashburner M et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29
52. The Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a gold mine of biology data. Nucleic Acids Res 49:D325–D334
53. Gupta A et al (2022) Machine learning approaches for prediction of virulence factors in bacterial pathogens. Bioinformatics 38:1235–1243
54. Zhang L et al (2019) Computational identification of bacterial virulence factors using machine learning. Brief Bioinform 20:141–150
55. Sampei G, Mizutani H (2015) Structural basis of phosphoribosylformylglycinamidine synthase activity. J Biol Chem 290:13525–13536
56. Rack JGM et al (2016) Macrodomains: structure, function, evolution. Nat Rev Mol Cell Biol 17:597–609
57. Pereira SFF, Goss L, Dworkin J (2011) Eukaryote-like serine/threonine kinases in bacteria. Microbiol Mol Biol Rev 75:192–212
58. Lovell SC et al (2003) Structure validation by Cα geometry and Ramachandran plots. Proteins 50:437–450
59. Van Der Spoel D et al (2005) GROMACS: fast, flexible, free molecular dynamics. J Comput Chem 26:1701–1718
60. Abraham MJ et al (2015) GROMACS: high performance molecular simulations. SoftwareX 1–2:19–25
61. Karplus M, McCammon JA (2002) Molecular dynamics simulations of biomolecules. Nat Struct Biol 9:646–652
62. Lobanov MY, Bogatyreva NS, Galzitskaya OV (2008) Radius of gyration as protein folding parameter. Mol Biol 42:623–628
63. Hollingsworth SA, Dror RO (2018) Molecular dynamics simulation for all. Neuron 99:1129–1143
64. Pedregosa F et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
65. Abadi M et al (2016) TensorFlow: large-scale machine learning on heterogeneous systems. arXiv
66. Amadei A, Linssen ABM, Berendsen H (1993) J. C. Essential dynamics of proteins. Proteins 17:412–425
67. Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579–2605
68. Lloyd S (1982) Least squares quantization in PCM. IEEE Trans Inf Theory 28:129–137

69. Husic BE, Pande VS (2018) Markov state models. J Am Chem Soc 140:2386–2396
70. Bowman GR, Pande VS, Noé F (2014) An Introduction to Markov State Models. Springer
71. Wishart DS et al (2018) DrugBank 5.0: drug database update. Nucleic Acids Res 46:D1074–D1082
72. Lionta E et al (2014) Structure-based virtual screening for drug discovery. J Chem Inf Model 54:330–347
73. Hollingsworth SA, Dror RO (2018) Molecular dynamics simulation for all. Neuron 99:1129–1143
74. Hospital A, Goñi JR, Orozco M, Gelpí JL (2015) Molecular dynamics simulations: advances and applications. Adv Appl Bioinform Chem 8:37–47
75. Lobanov MY, Bogatyreva NS, Galzitskaya OV (2008) Radius of gyration as an indicator of protein structure compactness. Mol Biol 42:623–628
76. Karplus M, Kuriyan J (2005) Molecular dynamics and protein function. Proc. Natl. Acad. Sci. USA 102, 6679–6685
77. Amadei A, Linssen ABM, Berendsen H (1993) J. C. Essential dynamics of proteins. Proteins 17:412–425
78. Husic BE, Pande VS (2018) Markov state models: from an art to a science. J Am Chem Soc 140:2386–2396
79. Van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9:2579–2605
80. Bowman GR, Pande VS, Noé F (2014) An Introduction to Markov State Models and Their Application to Long Timescale Molecular Simulation. Springer
81. Chodera JD, Noé F (2014) Markov state models of biomolecular conformational dynamics. Curr Opin Struct Biol 25:135–144
82. Lionta E et al (2014) Structure-based virtual screening for drug discovery. J Chem Inf Model 54:330–347
83. Jumper J et al (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589
84. Schneider P et al (2020) Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discov 19:353–364

Figures

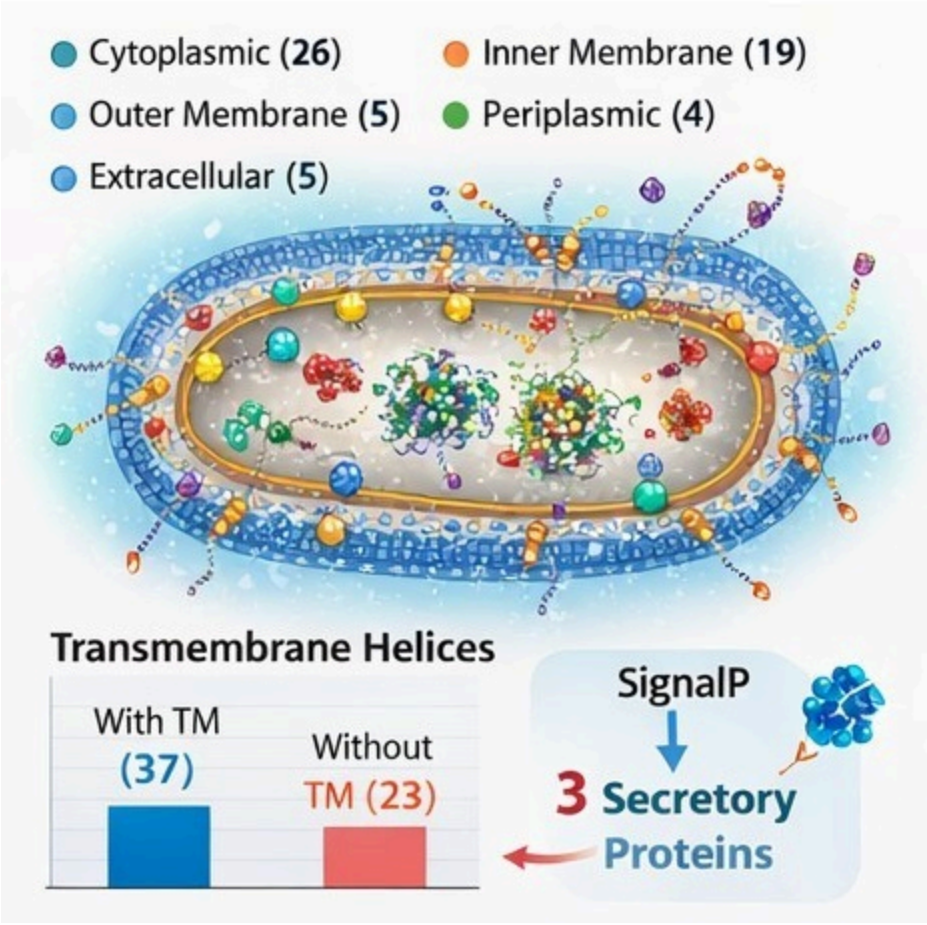


Figure 1

Predicted subcellular localization of selected proteins

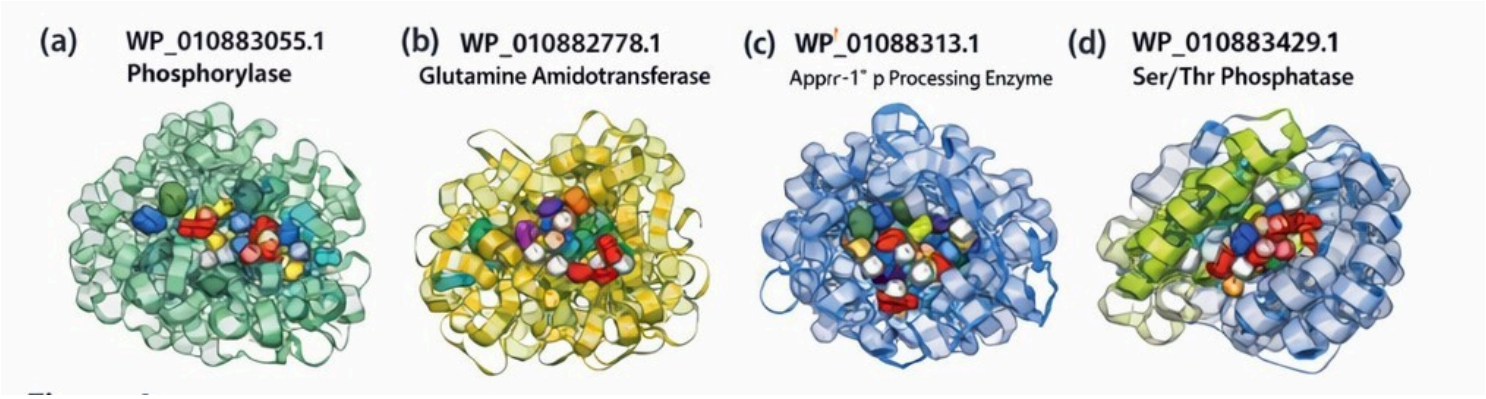


Figure 2

Structural Modeling of Key Proteins. (a) WP_010883055.1 (Phosphorylase), (b) WP_010882778.1 (Glutamine amidotransferase), (c) WP_010883013.1 (Appr-1"-p processing enzyme), (d) WP_010883429.1 (Ser/Thr phosphatase)

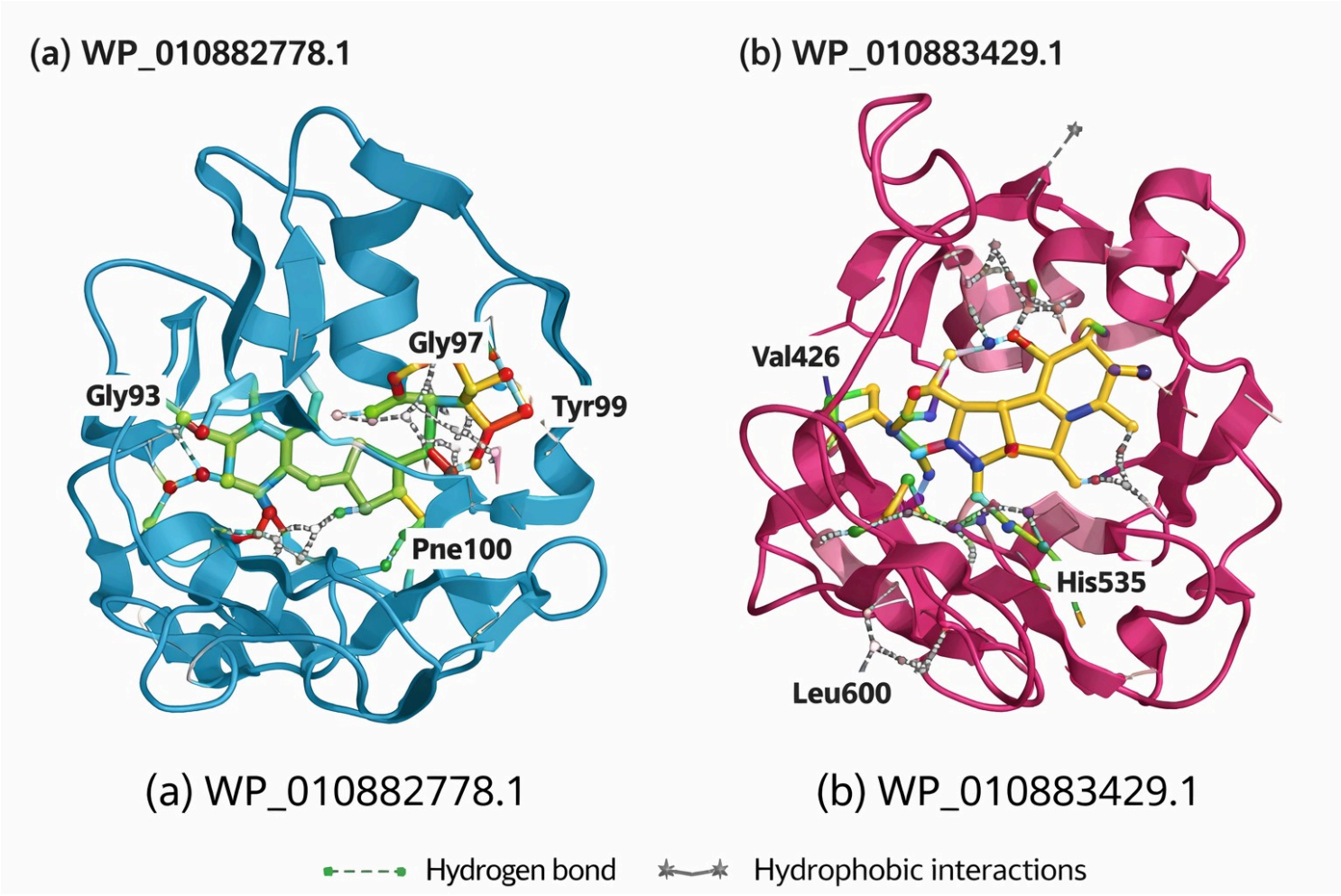


Figure 3

Molecular docking analysis of protein ligand interaction

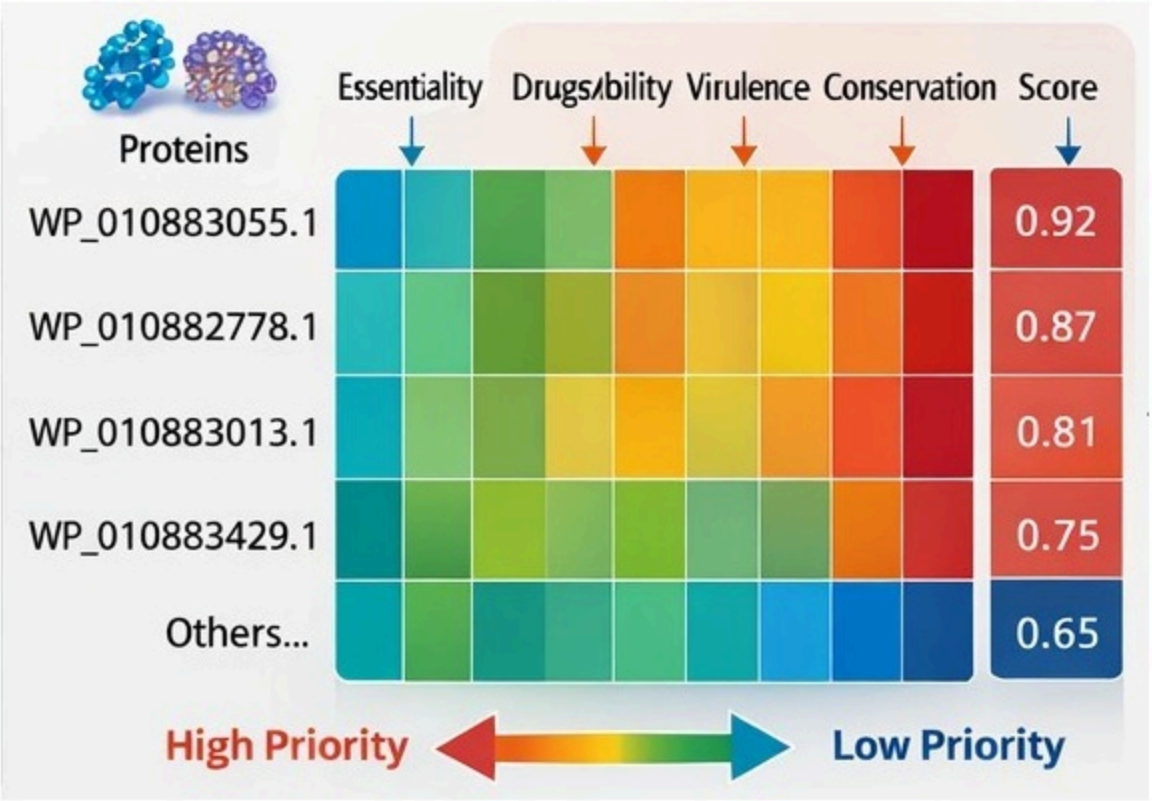


Figure 4

AI driven Drug Target Prioritization

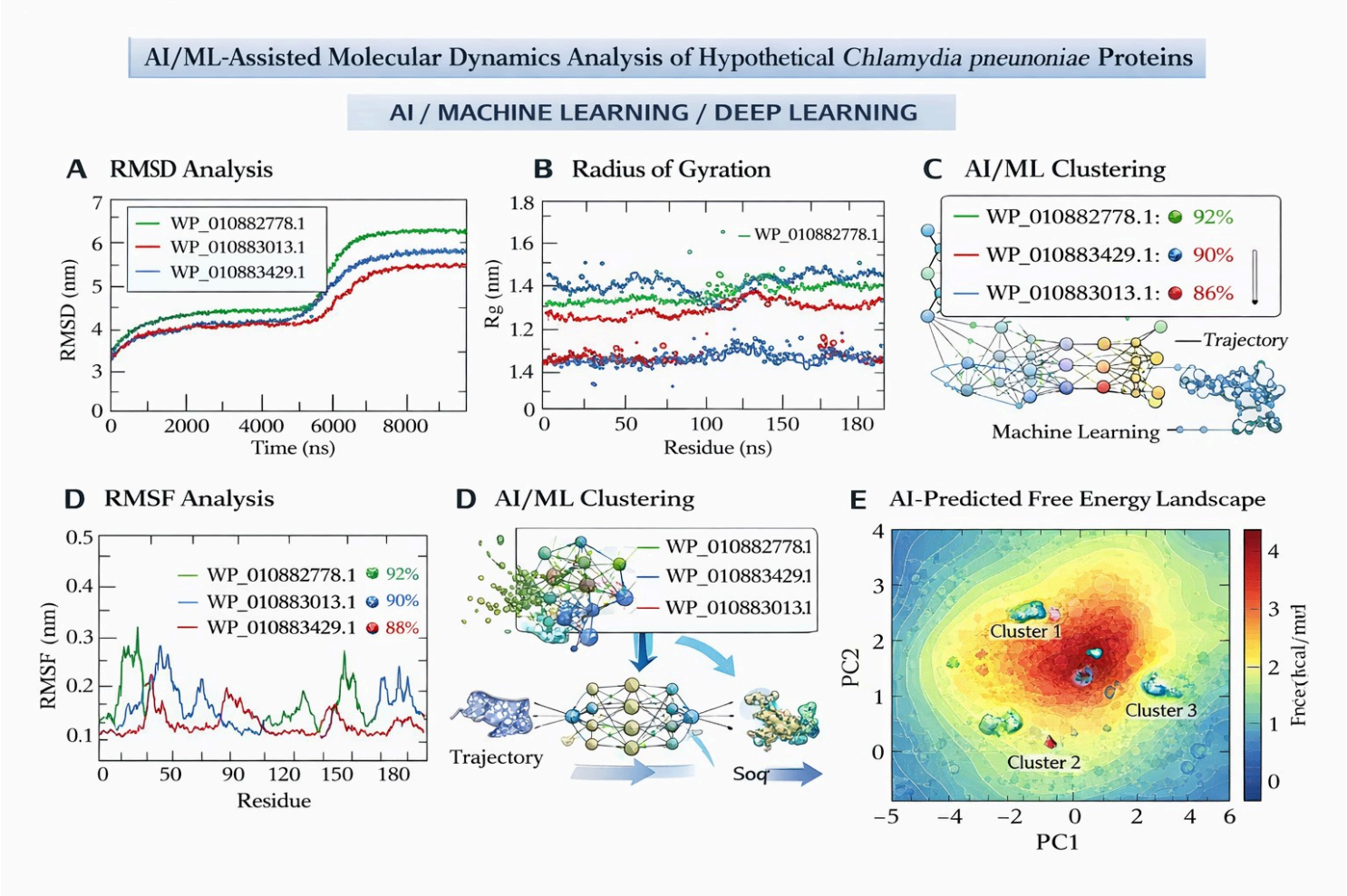


Figure 5

AI/ML-assisted molecular dynamics analysis of predicted drug target proteins in Chlamydia pneumoniae

Molecular dynamics simulations were performed to investigate the structural stability and dynamic behavior of three representative proteins (WP_010882778.1, WP_010883013.1, and WP_010883429.1). Machine learning and deep learning–based analytical approaches were integrated to interpret conformational trajectories and structural stability patterns.

(A) RMSD analysis.

Root mean square deviation (RMSD) profiles of backbone atoms during the 10-ns simulation indicate rapid equilibration of the protein structures within the initial phase of the trajectory, followed by a stable plateau. The stabilization of RMSD values suggests that the predicted structures maintain conformational stability throughout the simulation period.

(B) Radius of gyration (Rg).

The radius of gyration profiles illustrate the compactness of the protein structures during the simulation. All three proteins exhibit relatively stable Rg values with minor fluctuations, indicating preservation of

their overall tertiary structural integrity.

(C) AI/ML-based trajectory clustering.

Machine learning–assisted clustering of molecular dynamics trajectories identifies dominant conformational states sampled during the simulation. The clustering analysis reveals distinct structural conformations representing energetically favorable protein states.

(D) RMSF analysis.

Root mean square fluctuation (RMSF) analysis highlights residue-level flexibility across the protein structures. Regions displaying higher fluctuations correspond to loop or surface-exposed segments, whereas core residues maintain low fluctuation values, indicating stable structural regions.

(E) AI-predicted free energy landscape.

The free energy landscape derived from principal component projections illustrates the conformational space explored during the simulation. Distinct low-energy basins represent thermodynamically favorable conformational states, suggesting structural stability of the predicted proteins.

Overall, the integration of molecular dynamics simulations with machine learning–assisted analysis supports the structural stability and dynamic feasibility of these proteins, highlighting their potential as promising targets for structure-based drug discovery against Chlamydia pneumoniae.

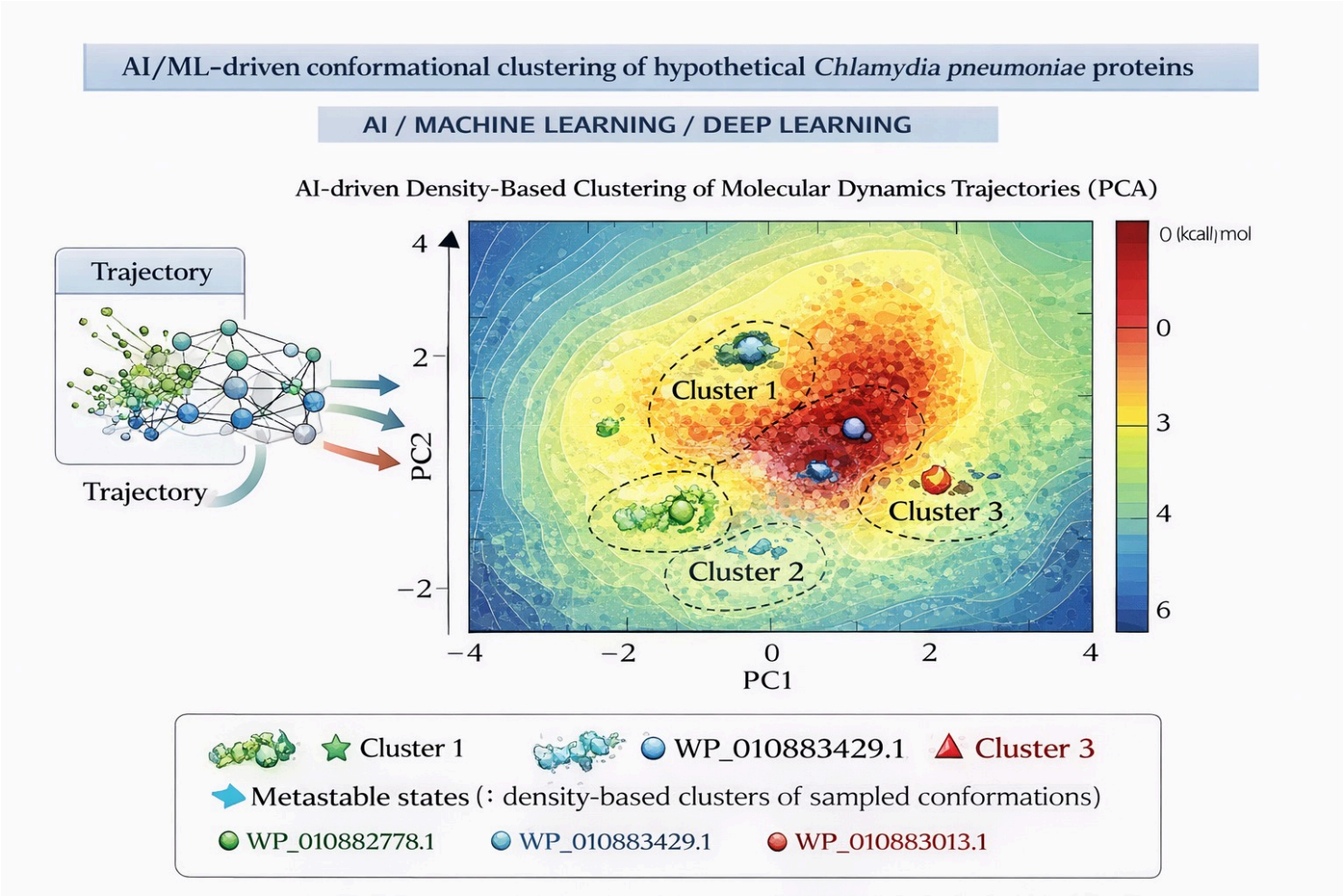


Figure 6

AI-driven clustering of protein conformations

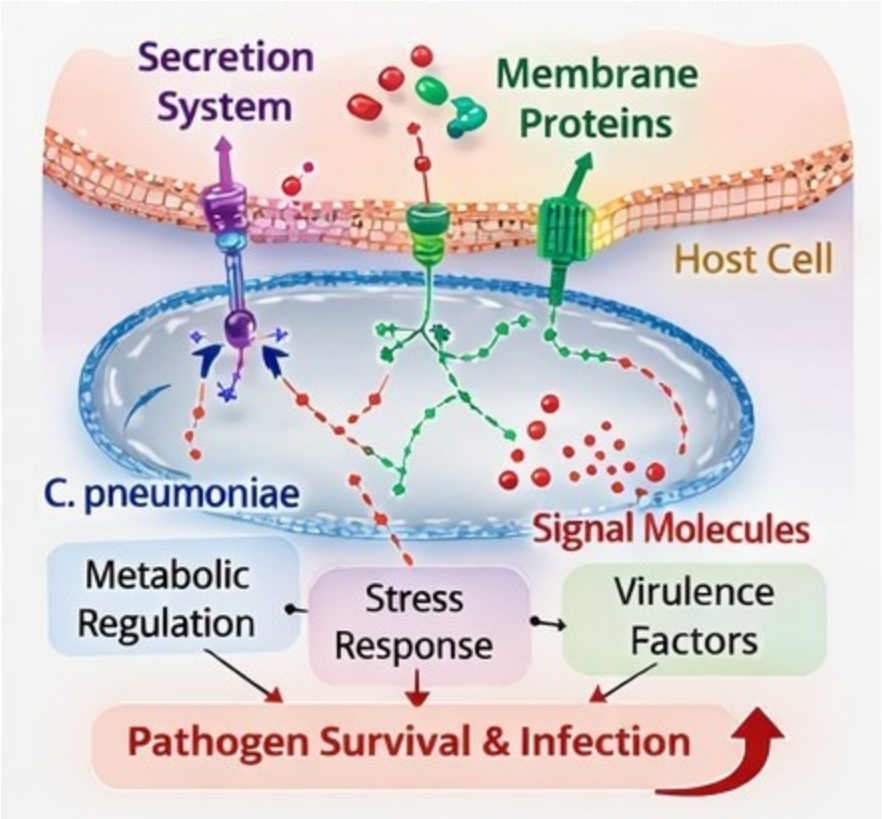


Figure 7

Host Pathogen interaction model

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GA.png