Integrating Machine Learning with Computational Chemistry to Predict Xylan Chain Properties at Scale

Maciej Staszak(✉)

Poznan University of Technology, 60-965 Poznań, Poland
maciej.staszak@put.poznan.pl

Abstract. This study demonstrates the integration of advanced computational frameworks, including machine learning (ML) and quantum chemical methods, to address challenges in molecular modeling of xylan chains. Trained ML model based on the SchNet architecture was validated against classical density functional theory (DFT) results for xylan homologs up to 15 units. The ML model is employed to investigate energetic trends in considerably longer chains, up to 80 units, which would be computationally challenging for DFT due to both memory and time constraints. The comparison demonstrate that the ML model accurately captures structural and energetic trends, thereby illustrating its potential as a viable tool for studying large polysaccharide systems. This work highlights the respective strengths and limitations of both approaches, thereby providing a foundation for further ML-based exploration of complex biopolymers and their interactions in biological and industrial perspectives, and contributes to the growing trend of leveraging machine learning frameworks to enhance system development and scalability in computational science.

Keywords: xylan · DFT · machine learning · SchNet · polysaccharide informatics

1 Introduction

As computational tools become increasingly essential in scientific research, the study of polysaccharides has highlighted both their structural complexity and computational challenges, as they often form long, branched chains and intricate networks that present significant topological-like problems. Polysaccharides play a crucial role in biological structures and processes due to their unique physical and chemical properties. Xylan, a primary hemicellulose in plant cell walls, has drawn considerable interest in research, particularly in fields such as biofuel production, bioinformatics, computational material science, and biochemical engineering [1]. Given its complex structure and the diverse environments in which it functions, understanding xylan's behavior at the molecular level is essential. Traditionally, method such as DFT, a cornerstone in computational chemistry, has been employed to predict the properties of xylan and similar polysaccharides, particularly due to their high accuracy in conformational and energetic analyses [2]. DFT

R. Silhavy and P. Silhavy (Eds.): CSOC 2025, LNNS 1561, pp. 66–75, 2025.
https://doi.org/10.1007/978-3-032-03406-9_4

effectively models mechanical and structural properties of materials, with recent studies applying it to molecular systems like xylobiose, xylan's simplest oligomer, to explore bond energies and stability at the quantum level. DFT research underlines its ability to detail molecular interactions, proving invaluable in understanding xylan's behavior both in isolation and during solvent or enzyme interactions [2, 3]. While DFT delivers high accuracy, it becomes computationally expensive as xylan chain length increases. Chains exceeding 15 units face time and memory constraints, limiting DFT's scalability for practical applications [4].

The development of ML models offers a complementary approach to molecular modeling, enabling rapid computations with a reasonable degree of accuracy. ML models like SchNet have demonstrated promising results in predicting equilibrium molecular properties with high precision, making them suitable for large, complex molecules like xylan chains [2, 5]. Unlike DFT, ML models can handle significantly larger molecular systems due to their scalable architecture [6]. Beyond ML and DFT methods, studying xylan's behavior in biochemical contexts often involves its interaction with enzymes like xylanase. This enzyme degrades xylan by cleaving β-1,4-glycosidic bonds, a key process in biofuel production and biotechnology [7]. Insights from ML models on the structure and behavior of xylan could enhance our understanding of enzyme-substrate interactions and facilitate the design of more efficient xylanase-based processes [8].

This study presents a computational comparison of DFT and ML in predicting xylan chain properties. By validating the ML model against DFT for short xylan chains, it establishes a baseline for accuracy. The validated ML model will then explore energetic trends in longer chains, emphasizing scalability and predictive power. This computational framework draws attention to the advantages and limitations of utilizing ML for the analysis of large biomolecular systems, with a particular focus on its applications in scalable simulations and predictive modelling. The findings provide an insight into the development of advanced software solutions and methodologies tailored to xylan research and other complex biopolymer systems.

2 Modeling

In the background of emerging trends in computational methodologies, this study integrates advanced cheminformatics tools and machine learning models to analyze xylan molecules. In the proposed calculations on the energy calculated by ML model, a number of parameters are analyzed, allowing characterization of the structure and molecular properties of xylan molecules. Here the RDKit cheminformatics package [9] to calculate the solvent accessible surface area (SASA) using the Lee-Richards algorithm [10]. The Lee-Richards method, also referred to as the "slice" algorithm, employs a two-dimensional representation of the molecule, with each atom represented by a point and the molecule divided into thin slices perpendicular to a rotation axis. For each slice, the algorithm calculates the accessible surface area by rolling a spherical probe (typically representing a water molecule) around the van der Waals surface of the atoms. Subsequently, the total SASA is calculated by integrating the accessible arcs over all slices.

The implementation in RDKit employs the FreeSASA library, which provides an efficient and robust calculation of SASA values. This approach was selected for its high accuracy, although it may be more computationally intensive than alternative methods, such as the Shrake-Rupley algorithm. The default probe radius of 1.4 Å was used to simulate water accessibility, and atoms were classified using the Protor classification scheme [11] for distinguishing between polar and non-polar surface areas.

2.1 The Surface Parameters

To obtain energy results relative to surface area, the total SASA was calculated to measure the molecule's interactive surface with its surroundings. The SASA per monosaccharide unit, the average SASA in glycosidic bond regions, the total glycosidic bond SASA and the glycosidic SASA fraction were calculated.

2.2 Energy Parameters

The total energy of the molecule is an indicator of the overall structural stability. The energy per monosaccharide unit provides an indication of the average energy associated with a single monomer, thus facilitating comparison between fragments of the structure. Furthermore, the analysis incorporates the energy density per unit surface area, which indicates the energy concentration per surface area of a single monosaccharide unit. In contrast, the glycosidic surface energy represents the energy per unit area within the glycosidic bond area. The local energy density index integrates data on energy, surface area, and number of monosaccharide units and atoms, providing a composite parameter that characterizes the local energy density in a molecule.

2.3 Embedding

In order to generate reliable three-dimensional structures for xylan homologs, the Enhanced Torsion Distance Geometry embedding algorithm [12], as implemented in the RDKit software, was employed. This approach combines distance geometry with empirical data about molecular geometry, to produce high-quality conformations. The algorithm uses a distance bounds matrix, which ensures the enforcement of realistic interatomic distances based on covalent radii and van der Waals interactions. These constraints ensured avoiding unrealistic atom overlaps and maintaining proper bonding distances.

The `firstMinimization` procedure was conducted, so that internal energy minimization was applied within the RDKit embedding framework. This process reduced steric hindrances and obeyed chirality and stereochemistry. Although this internal minimization step did not perform a full force field optimization, it employed a penalty-based approach to refine atomic distances towards the desired values. The resulting structure closely matched natural conformations, guided by RDKit's torsion angle preferences derived from crystallographic data, favoring experimentally observed angles for rotatable bonds. It is also interesting that ML model managed particularly well on such somewhat unbalanced structures, shown below.

2.4 Deep Learning Model

The ML model was trained on the QM9 dataset, which contains information about small organic molecules, using the advanced SchNet neural network architecture [13]. The SchNet model was designed by Schütt et al. [14] to efficiently represent molecular structure, thereby enabling the prediction of the properties of molecules based on their atomic structure. The model's architecture is based on three principal components: the calculation of distances between pairs of atoms, a basis of radial Gaussian functions, and a multilayer SchNet representation with a specified number of atomic interactions. The cutoff value for atomic interactions has been set to 5 Å, which allows the model to consider local atomic interactions over a sufficiently wide range. Optimization was conducted using the AdamW algorithm [15] with a selected learning rate of 1e-4. The data underwent a series of transformations, including the removal of energy offsets, in order to facilitate the optimal processing of large data sets.

3 Results

The accuracy of the SchNet model for xylan chains was validated by comparison with DFT calculations. A B3LYP function with the 6-311G(2df,p) basis set was employed using Psi4 library [16, 17], with 10 GB of memory and four computational threads allocated. Due to the constraints of the hardware, DFT calculations were conducted for chains comprising a maximum of 15 mer units. The longer structures would necessitate a significantly greater allocation of RAM resources.

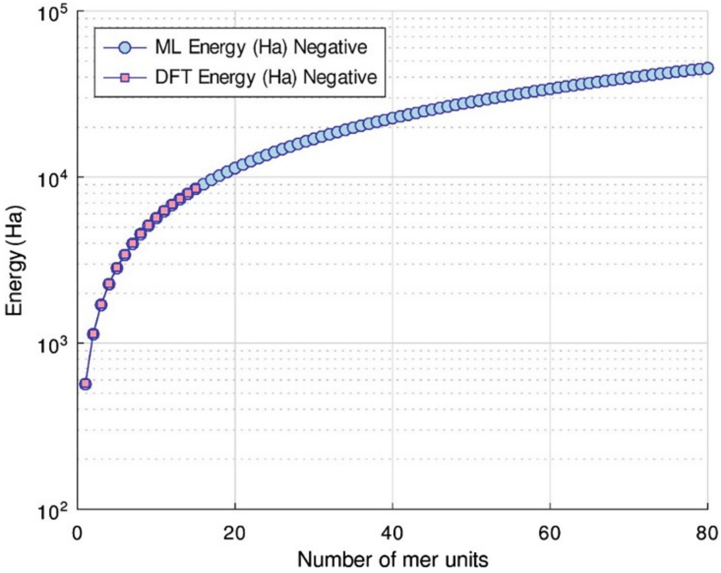

Fig. 1. Validation of ML model, comparison vs DFT

A comparison of the DFT and ML results (Fig. 1) demonstrated a high degree of correlation between the two methods across the 3 to 15 range of chain lengths investigated. The energy values predicted by SchNet were in close agreement with the reference DFT values. This validates that the ML model has effectively learned energy patterns from small molecules and can reliably predict the energies of much larger structures (Fig. 2), for which DFT calculations would be impractical or infeasible.

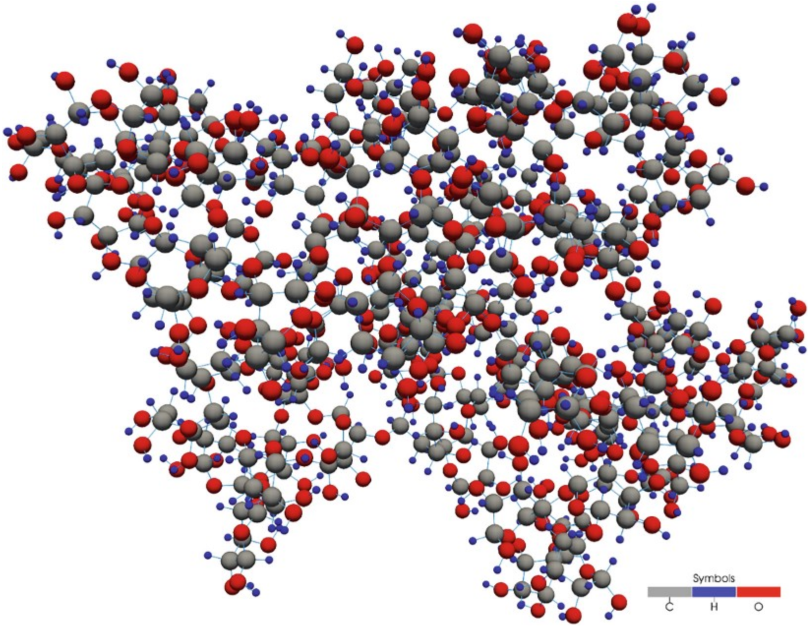

Fig. 2. Large xylan structure, $C_{400}H_{642}O_{400}$, 80 monosaccharide units

The substantial time savings achieved through machine learning (ML) techniques are evident (Fig. 3). As more monosaccharide groups are incorporated into the ML model, the computational power requirements exhibit only a slight increase, with this trend gradually stabilising. In contrast, the classical model based on DFT methods demonstrates a pronounced growth in computational demands, aligning with the anticipated computational complexity of $O(n^3)$.

The relationship between the local energy density index and molecule size reveals several significant patterns that facilitate a deeper comprehension of the way how energy density evolves with molecule size (Fig. 4).

The initial trend is a general decrease in index values, which is particularly pronounced for small molecules up to approximately 15–20 units in size. In this initial phase, the index values decrease rapidly before gradually stabilizing and transitioning to a smoother, more uniform decline. Three characteristic regions can be distinguished: an

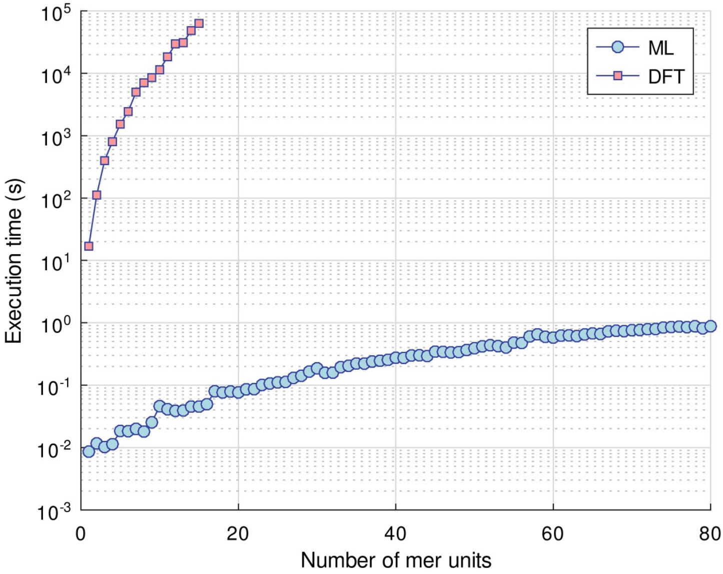

Fig. 3. Execution times for DFT and ML models at 11th Gen Intel(R) Core(TM) i7-11700K @ 3.60GHz.

initial period of rapid decline (for molecules with a size of 3–15 monosaccharide units), a transition region (around 15–25 units) and a plateau region (above 25 units) in which the index values remain stable at around -590 to -600 kJ/(mol•Å2).

These trends suggest that smaller polysaccharides (3–15 units) have a higher local energy density, indicating a more compact structure and stronger intramolecular interactions, resulting in higher energy concentration per unit area. In contrast, larger polymers (above 25 units) achieve a stable index level, which may indicate that their spatial organization obtains more uniform energy distribution, which is characteristic of structurally repetitive polymers. The fluctuations in the index values within the plateau region are attributed to local conformational variations of the chain, which do not markedly influence the overall energy stability. The local energy density index can be used to determine the critical length of a molecule's chain, after which the energy properties stabilize and further chain growth does not significantly alter the energy distribution.

The dependence of local energy density index on total energy (Fig. 5) indicate regions of higher negative energy density between −600 and −620 kJ/(mol-Å2). Such structures can be more challenging for enzymes due to their stronger intermolecular interactions and more compact spatial organization. Substrate accessibility is hindered and reorganization of the structure for enzyme action is associated with a higher energy cost.

For higher values of local energy density index, closer to -500 kJ/(mol-Å2), structures become more accessible to enzymes. The easier accessibility of the glycosidic bonds implies a lower remodeling cost and thus a higher efficiency of xylanases. This shows

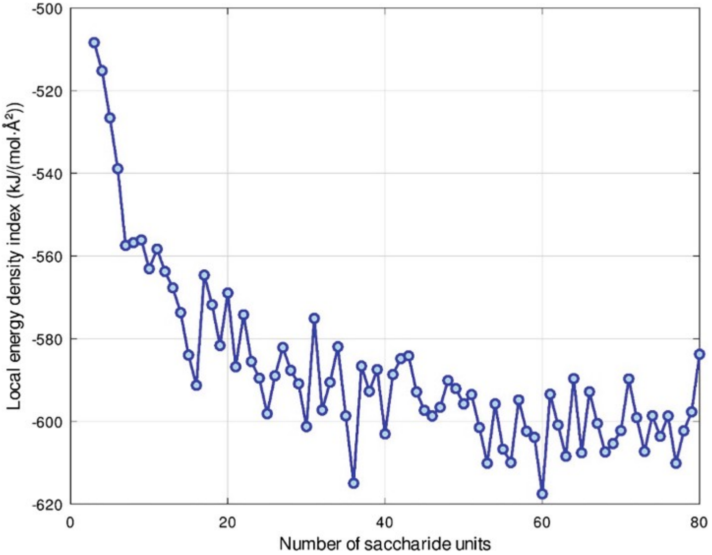

Fig. 4. Local energy density index vs polysaccharide size

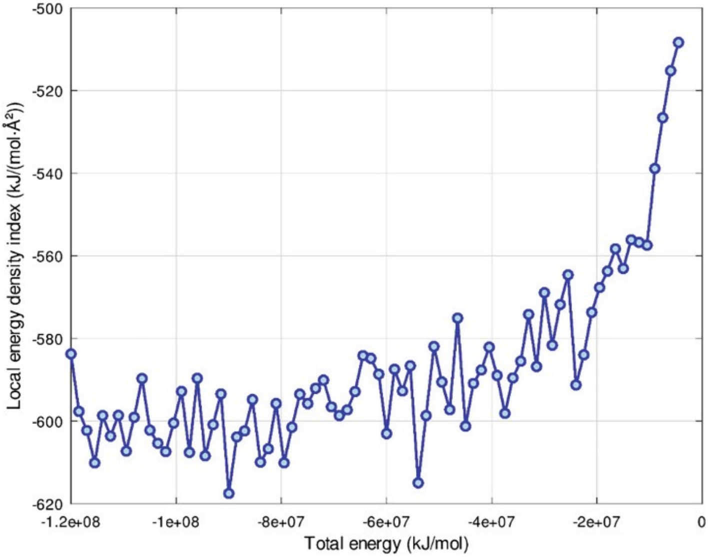

Fig. 5. Local energy density index vs total energy.

preferable by enzymes regions with higher local energy density index values, where bonds are more exposed.

In the field of enzyme kinetics, structures with higher local energy density index, located on the right side of the graph, can have higher values of turnover number, lower values of affinity for substrate and higher overall catalytic efficiency [18]. These properties indicate favorable conditions for efficient enzyme performance, which is particularly relevant in the context of biorefineries [19], where optimization of enzymatic processes may involve favoring substrates with a suitable Local Energy Density Index range or pre-treating the substrate to reduce its energy density.

4 Conclusions

This study demonstrates the potential of computational methodologies in advancing molecular modeling, with the SchNet model exhibiting high accuracy in predicting the energy profiles of xylan chains, effectively capturing the energy distribution trends observed in DFT calculations. This validation against DFT confirms the model's reliability for larger structures, showing it as an efficient alternative to resource-intensive quantum methods.

The analysis reveals the existence of distinct regions in the Local Energy Density Index curve, which highlight the varying structural characteristics of xylan chains as their length increases. The local energy densities of shorter chains (3–15 units) are higher, indicative of compactness and strong intramolecular interactions. In contrast, longer chains (25 units and above) reach a stable energy distribution, suggesting an optimal spatial organization typical of extended polymers.

From a biotechnological perspective, these insights have practical implications, particularly in the context of enzyme-catalyzed degradation processes. Xylanase enzymes, which are responsible for the breakdown of xylan chains, may operate more efficiently on regions with higher Local Energy Density Index values due to the increased accessibility of glycosidic bonds. It can therefore be surmised that modifying the xylan structure in order to optimize these regions, or alternatively, to target them selectively, could result in an enhancement of the catalytic performance.

Furthermore, the stabilizing trend in glycosidic bond surface energy with increasing chain length is consistent with a shift towards polymer-like characteristics, indicating the potential for an effective enzyme interaction threshold. Insights from this study highlight SchNet's usefulness in capturing complex molecular energy profiles, opening ways for targeted biotechnological applications in enzyme-catalyzed processes with polysaccharides like xylan, and also demonstrating the transformative role of advanced computational tools in system development and biotechnological innovation.

Acknowledgements. The project is supported by the Circular Bio-based Europe Joint Undertaking and its members. Funded by the European Union, grant agreement N° 101157779. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CBE-JU. Neither the European Union nor the granting authority can be held responsible for them.

References

1. Mndlovu, H., et al.: Development of a fluid-absorptive alginate-chitosan bioplatform for potential application as a wound dressing. Carbohyd. Polym. **222**, 114988 (2019). https://doi.org/10.1016/j.carbpol.2019.114988
2. Ling, Z., Edwards, J.V., Nam, S., Xu, F., French, A.D.: Conformational analysis of xylobiose by DFT quantum mechanics. Cellulose **27**, 1207–1224 (2020). https://doi.org/10.1007/s10570-019-02874-3
3. Mohammed, A.S.A., Naveed, M., Jost, N.: Polysaccharides; classification, chemical properties, and future perspective applications in fields of pharmacology and biological medicine (a review of current applications and upcoming potentialities). J. Polym. Environ. **29**, 2359–2371 (2021). https://doi.org/10.1007/s10924-021-02052-2
4. Lin, Q., et al.: Molecular scale behavior of xylan during solvent-controlled extraction and precipitation. Phys. Chem. Chem. Phys. **25**, 28078–28085 (2023). https://doi.org/10.1039/D3CP01385E
5. Kiely, E., Zwane, R., Fox, R., Reilly, A.M., Guerin, S.: Density functional theory predictions of the mechanical properties of crystalline materials. CrystEngComm **23**, 5697–5710 (2021). https://doi.org/10.1039/D1CE00453K
6. BeMiller, J.N.: Polysaccharides: occurrence, significance, and properties. In: Fraser-Reid, B.O., Tatsuta, K., Thiem, J. (eds.) Glycoscience, pp. 1413–1435. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-30429-6_34
7. Bhardwaj, N., Kumar, B., Verma, P.: A detailed overview of xylanases: an emerging biomolecule for current and future prospective. Bioresour. Bioprocess. **6**, 40 (2019). https://doi.org/10.1186/s40643-019-0276-2
8. Sarangi, A., Thatoi, H.: Xylanase as a Promising Biocatalyst: A Review on Its Production, Purification and Biotechnological Applications. Proc. Natl. Acad. Sci., India, Sect. B Biol. Sci. (2024). https://doi.org/10.1007/s40011-024-01567-7
9. RDKit, https://www.rdkit.org/, last accessed 2024/12/04
10. Lee, B., Richards, F.M.: The interpretation of protein structures: Estimation of static accessibility. J. Molecul. Biol. **55**, 379-IN4 (1971). https://doi.org/10.1016/0022-2836(71)90324-X
11. Finney, J.L., Gellatly, B.J., Golton, I.C., Goodfellow, J.: Solvent effects and polar interactions in the structural stability and dynamics of globular proteins. Biophys. J . **32**, 17–33 (1980). https://doi.org/10.1016/S0006-3495(80)84913-7
12. Wang, S., Witek, J., Landrum, G.A., Riniker, S.: Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. J. Chem. Inf. Model. **60**, 2044–2058 (2020). https://doi.org/10.1021/acs.jcim.0c00025
13. Schütt, K.T., Sauceda, H.E., Kindermans, P.-J., Tkatchenko, A., Müller, K.-R.: SchNet – a deep learning architecture for molecules and materials. J. Chem. Phys. **148**, 241722 (2018). https://doi.org/10.1063/1.5019779
14. Schütt, K.T., Kindermans, P.-J., Sauceda, H.E., Chmiela, S., Tkatchenko, A., Müller, K.-R.: SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. (2017). https://doi.org/10.48550/ARXIV.1706.08566
15. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization (2019). http://arxiv.org/abs/1711.05101
16. Parrish, R.M., et al.: Psi4 1.1: an open-source electronic structure program emphasizing automation, advanced libraries, and interoperability. J. Chem. Theory Comput. **13**, 3185–3197 (2017). https://doi.org/10.1021/acs.jctc.7b00174

17. Smith, D.G.A., et al.: Psi4NumPy: an interactive quantum chemistry programming environment for reference implementations and rapid development. J. Chem. Theory Comput. **14**, 3504–3511 (2018). https://doi.org/10.1021/acs.jctc.8b00286
18. Jones, H.B.L., et al.: A complete thermodynamic analysis of enzyme turnover links the free energy landscape to enzyme catalysis. FEBS J. **284**, 2829–2842 (2017). https://doi.org/10.1111/febs.14152
19. Calvo-Flores, F.G., Martin-Martinez, F.J.: Biorefineries: achievements and challenges for a bio-based economy. Front. Chem. **10**, 973417 (2022). https://doi.org/10.3389/fchem.2022.973417