Per-residue optimisation of protein structures: Rapid alternative to optimisation with constrained alpha carbons

Ondřej Schindler[1,2*], Tomáš Svoboda[1,2], Gabriela Bučeková[1,2], Radka Svobodová[1,2]

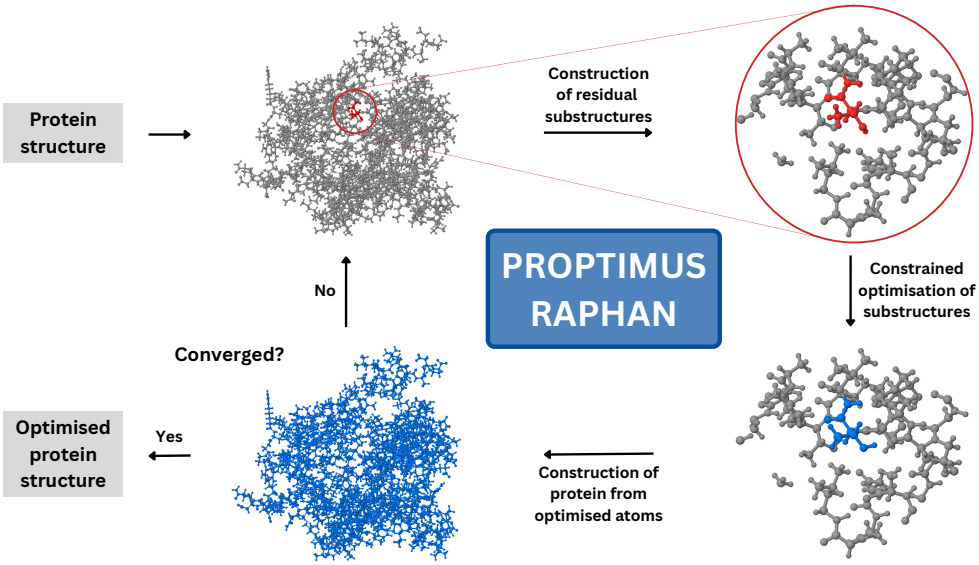
[1*]National Centre for Biomolecular Research, Masaryk University, Kamenice 753/5, Brno, 625 00, Czech Republic.
[2*]CEITEC – Central European Institute of Technology, Masaryk University, Kamenice 753/5, Brno, 625 00, Czech Republic.

*Corresponding author(s). E-mail(s): ondrej.schindler@mail.muni.cz; Contributing authors: tomsvo@mail.muni.cz; gabriela.bucekova@mail.muni.cz; radka.svobodova@mail.muni.cz;

Abstract

In recent years, the number of known protein structures has increased significantly. Predictive algorithms and experimental methods provide the positions of protein residues relative to each other with high accuracy. However, the local quality of the protein structure, including bond lengths, angles, and positions of individual atoms, often lacks the same level of precision. For this reason, protein structures are usually optimised by a force field prior to their application in further research sensitive to structural quality. Protein structure optimisation, however, is computationally challenging.

In this paper, we introduce a general method *Per-residue optimisation of protein structures: Rapid alternative to optimisation with constrained alpha carbons* (PROPTIMUS RAPHAN). Rather than optimising the entire protein structure at once, PROPTIMUS RAPHAN divides the structure into overlapping residual substructures and optimises each substructure individually. This approach results in computational time that scales linearly with the size of the structure. Additionally, we present PROPTIMUS RAPHAN$_{\text{GFN-FF}}$, a reference implementation of our method employing a generic, almost QM-accurate force field, GFN-FF.



We tested PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ on 461 AlphaFold DB structures and demonstrated that our approach achieves results comparable to the optimisation of the structure with constrained alpha carbons in significantly less time.
Scientific Contribution: The main contribution of this work is the PROPTIMUS RAPHAN method and its reference parallelisable implementation PROPTIMUS RAPHAN$_{\text{GFN-FF}}$. Because the time requirement increases linearly with the size of the structure, PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ optimises on average 5 000 atoms per hour and a common CPU. Therefore, prior to any research sensitive to protein structure quality, our method can be employed to obtain protein structures closer to QM-accuracy.

Keywords: protein, residue, structure optimisation, force fields, GFN-FF, AlphaFold DB

1 Introduction

Proteins are a fundamental part of all living organisms and are extensively studied for their crucial roles in many biological functions [1]. The 3D structure of a protein, the arrangement of atoms in space, is the crucial factor that determines the function of a protein [2]. In recent years, the number of predicted and experimentally determined

protein structures has increased significantly. The Protein Data Bank [3] currently contains more than 240 thousand experimentally determined protein structures. On the other hand, more than 200 million predicted structures are available in the AlphaFold DB [4, 5], and over 700 million structures are stored in the ESM Metagenomic Atlas [6]. Furthermore, several freely available web services are available to predict the structure of a protein sequence, with the latest ones being based on artificial intelligence [7–9].

Machine learning methods aim to accurately determine the positions of alpha carbons in a protein and thus the relative positions of individual amino acid residues. Current machine learning approaches can predict the positions of alpha carbons in a protein with accuracy comparable to experimental methods [10]. In contrast, detailed characteristics of amino acid side chains, such as bond lengths, bond angles and residue interactions, are not predicted with the same precision [11, 12]. A similar phenomenon is also observed in experimentally determined protein structures, where the local quality of the structure could be limited by the accuracy of the experiment (e.g., resolution in the case of X-ray crystallography).

The local quality of protein structure is essential for correct results in various applications of computational chemistry [13]. Applications sensitive to structure quality are, for example, docking [12], QSPR modelling [14], partial atomic charge calculations [14], or calculations using semiempirical or QM/MM methods [15]. Therefore, predicted protein structures are usually optimised before conducting the main research [13, 16].

Force field methods are commonly used for protein optimisation. These methods replace the electronic structure of the molecule used in quantum chemistry methods with classical mechanical interatomic interaction potentials, resulting in a fast calculation [17, 18]. However, even with the use of a force field, optimisation of proteins with more than ten thousand atoms remains computationally challenging [17]. Therefore, various constraints can be introduced during protein structure optimisation to reduce the number of degrees of freedom and accelerate computation. Constraining

alpha carbons at a specific location in the space during optimisation is one of the most frequently used techniques [4, 13, 16]. This approach is particularly suitable for structures predicted by machine learning methods, as alpha carbons are the most accurately predicted atoms. However, even with these constraints, force field methods still retain their quadratic computational complexity in relation to the number of atoms for single-point energy calculations and, consequently, for the overall protein structure optimisation [17].

In this work, we introduce a general iterative method Per-residue optimisation of protein structures: Rapid alternative to optimisation with constrained alpha carbons (PROPTIMUS RAPHAN). In every iteration of the PROPTIMUS RAPHAN method, a substructure is created for each protein residue, which includes the residue and its neighbouring residues. These substructures are then optimised separately. Then, the optimised substructures are used to reconstruct a partially optimised protein structure. This approach allows for linear computational dependence on protein size. The method is applicable to any force field. In our work, we also establish a reference implementation of PROPTIMUS RAPHAN method employing a force field, GFN-FF (PROPTIMUS RAPHAN$_{\text{GFN-FF}}$) [17].

2 Methods and implementation

The PROPTIMUS RAPHAN method is developed based on the Cover approach, a divide-and-conquer algorithm, which has already been successfully utilised multiple times to calculate partial atomic charges [19–21]. The Cover approach is an approximation that neglects long-range weak interactions, thereby allowing the protein to be partitioned into smaller fragments. The PROPTIMUS RAPHAN divides the protein into overlapping residual substructures for each residue, and substructures are then optimised separately. The time required to optimise one substructure is approximately constant, as the substructures are similar in size. This leads to approximate linear

computational complexity with respect to the number of atoms. Moreover, the use of the divide-and-conquer algorithm ensures high parallelisability.

We developed a reference implementation of PROPTIMUS RAPHAN in Python. Before the PROPTIMUS RAPHAN computation begins, the protein structure is loaded using the BioPython library [22]. We refer to this BioPython object in the following text as the *target structure*, because it tracks and stores all intermediate results during the optimisation. Moreover, a set of *optimised atoms* is defined for each protein residue. *Optimised atoms* are all atoms of the residue except for the alpha carbon and the N-H peptide bond atoms. Instead, the N-H peptide bond atoms from the following residue are included in the *optimised atoms*.

PROPTIMUS RAPHAN is an iterative method. Each iteration consists of the following steps (see Figure 1):

1. **Construction of residual substructures:** A residual substructure is created for each non-converged *target structure* residue. The residual substructure includes all atoms that are within 6 Å to at least one *optimised atom* of this residue. Moreover, a minimum set of other atoms is included to ensure that only bonds between two carbon atoms are cut [21]. Atoms of each residual substructure that do not belong to the *optimised atoms* of the specific residue are divided into two groups:

   - *Flexible atoms* are all atoms that are closer than 4 Å to at least one *optimised atom* and are not alpha carbons.
   - *Constrained atoms* are the remaining atoms of the residual substructure that belong to neither the *optimised* nor the *flexible* atoms.

   We implemented this step using the BioPython [22] and RDKIT [23] libraries.

2. **Constrained optimisation of residual substructures:** All residual substructures undergo their own optimisation. During optimisation, the coordinates of *constrained atoms* cannot be altered and therefore only positions of *flexible atoms*

and *optimised atoms* are optimised. Residual substructures are optimised by GFN-FF [17] provided within the xtb software [24]. For that reason, we refer to our implementation as PROPTIMUS RAPHAN$_{\text{GFN-FF}}$. Solvation is included in the calculation by using the implicit solvent model ALPB [25]. The number of GFN-FF optimisation steps is limited to prevent extensive changes in atom coordinates. Specifically, the maximum number of GFN-FF optimisation steps is determined by the sum of *optimised atoms* of the residual substructure and the current PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ iteration index.

3. **Construction of protein from optimised atoms:** The *target structure* coordinates are updated based on the corresponding optimised atomic positions.

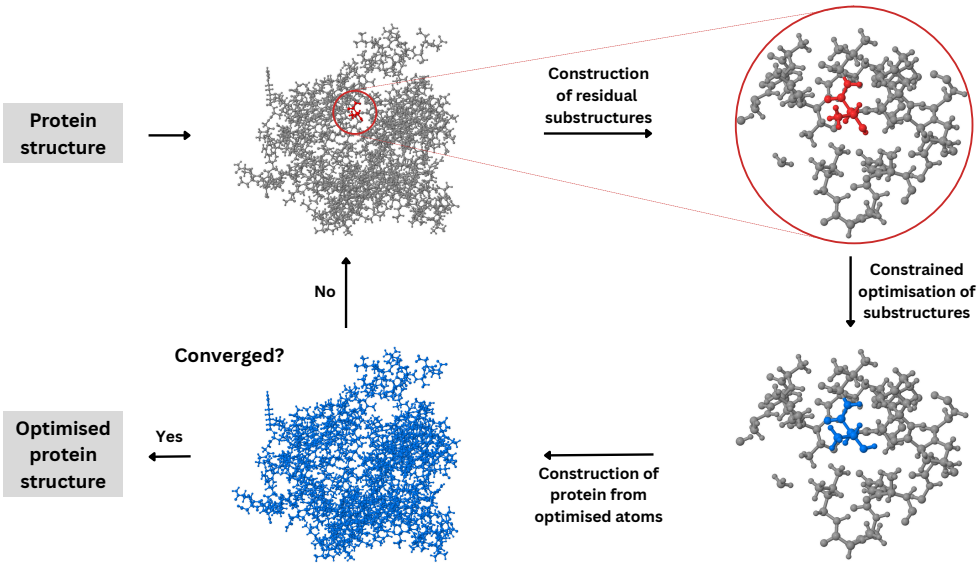


Fig. 1 Scheme of the PROPTIMUS RAPHAN method.

An advantage of PROPTIMUS RAPHAN is the ability to continuously exclude already converged residues (i.e., residues whose position no longer changes during optimisation) from the iterative process. This feature significantly speeds up the entire calculation.

To increase accuracy, the entire iterative process is repeated twice. In the second run, 8 Å is used instead of 6 Å when creating substructures. In addition, all atoms of the residue except for the alpha carbon are considered to be *optimised atoms* of the residue.

For the purposes of testing and reproducibility, we also included in the PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ implementation the option to directly run the GFN-FF optimisation with constrained alpha carbons (GFN-FF$_{\text{C}\alpha}$). In addition, we provide Python scripts for comparing protein structures using the Biopython [22], RDKit [23], Biotite [26], and MDAnalysis [27] libraries, which were used in Section 3. The PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ implementation, alongside the scripts used for comparison, is provided in the supplementary material.

3 Results and Discussion

To compare the accuracy and speed of PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ with those of the original method GFN-FF$_{\text{C}\alpha}$, we designed a series of experiments and selected a test dataset.

3.1 Datasets

The testing dataset would contain selected protein structures from AlphaFold DB [5]. In particular, the sample came from the fourth version (v4) of predictions for Swiss-Prot sequences [28]. The selection was performed in such a way that exactly one random structure was selected from each interval of 10 on the size scale ranging from 200 to 5 000 heavy atoms. Consequently, the dataset contained 480 structures. Then it was necessary to furnish the structures with hydrogens, since no predictions from AlphaFold DB included them. For this, we employed the tool PB2PQR [29] with the PROPKA [30, 31] titration state method. For each structure, the pH was randomly selected from a range of 0 to 14. This was the original dataset, and hereafter we will

refer to it as $\mathrm{SET}_{\mathrm{ORIG}}$. Having the $\mathrm{SET}_{\mathrm{ORIG}}$ prepared, three different optimisation procedures were applied to it (see Figure 2). The following procedures took place:

- Optimisation using GFN-FF$_{\mathrm{C}\alpha}$ method. The subjected set of structures will be further denoted as $\mathrm{SET}_{\mathrm{C}\alpha}$.
- Optimisation using PROPTIMUS RAPHAN$_{\mathrm{GFN\text{-}FF}}$ method. The subjected set of structures will be further denoted as $\mathrm{SET}_{\mathrm{RAPHAN}}$.
- Optimisation using PROPTIMUS RAPHAN$_{\mathrm{GFN\text{-}FF}}$ method and a subsequent one using the GFN-FF$_{\mathrm{C}\alpha}$ method. The subjected set of structures will be further denoted as $\mathrm{SET}_{\mathrm{RAPHAN+C}\alpha}$.

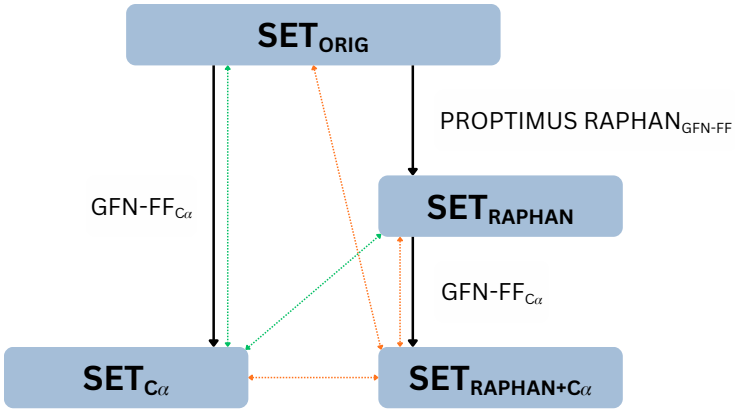


Fig. 2 Schematic diagram describing the relationships between datasets. Black lines represent optimisations. Dotted lines represent the pairs of datasets compared as described in Sections 4 (green lines) and 4.1 (orange lines).

All optimisations were launched within our implementation. All optimisations were performed on a single AMD EPYC 9454 CPU. The GFN-FF$_{\mathrm{C}\alpha}$ optimisation for 4 structures did not converge, and for 15 structures it exceeded our operating memory limit of 196 GB. Therefore, only the remaining 461 protein structures were used for further analysis. All four datasets, along with the scripts used, are available in the supplementary files.

4 Comparison of PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ and GFN-FF$_{\text{C}\alpha}$ optimised structures

GFN-FF$_{\text{C}\alpha}$ is a reference method to PROPTIMUS RAPHAN$_{\text{GFN-FF}}$. Therefore, in this Section, SET$_{\text{C}\alpha}$ with SET$_{\text{RAPHAN}}$ are compared. SET$_{\text{ORIG}}$ is also included in this comparison to evaluate differences between SET$_{\text{C}\alpha}$ and SET$_{\text{RAPHAN}}$ relative to the original structures (see Figure 2). The structures were compared in terms of the mean absolute deviation (MAD)s of atom positions, bond lengths, bond angles and dihedral angles. In addition, the MADs of canonical dihedral angles ϕ, ψ, ω and the side-chain dihedral angle χ_1, acting as soft descriptors for local displacements in the protein backbone, were calculated [17]. The comparative analysis results are summarised in Table 1. Percentiles and associated histograms of differences are available in the supplementary files.

Table 1 Comparison of SET$_{\text{C}\alpha}$ with sets of structures SET$_{\text{RAPHAN}}$ and SET$_{\text{ORIG}}$. Mean absolute deviation of atom positions, bond lengths, bond angles, dihedral angles and four characteristic dihedral angles are included.

	SET$_{\text{RAPHAN}}$ vs. SET$_{\text{C}\alpha}$	SET$_{\text{ORIG}}$ vs. SET$_{\text{C}\alpha}$
MAD atomic position [Å]	0.074	0.407
MAD bond length [pm]	0.075	4.851
MAD bond angle [°]	0.136	2.071
MAD dihedral angle [°]	1.611	11.124
MAD ϕ [°]	1.814	10.275
MAD ψ [°]	1.686	9.227
MAD ω [°]	1.156	5.770
MAD χ_1 [°]	1.307	7.981

The data in Table 1 show a high similarity between SET$_{\text{RAPHAN}}$ and SET$_{\text{C}\alpha}$. The structures exhibit the highest degree of similarity in bond lengths, with a MAD of only 0.075 pm, which is at the level of accuracy of the PDB format. The MAD values for atomic positions (0.074 Å) and bond angles (0.136°) also show good agreement. The

MADs for dihedral angles range around 1.5°. Relative to the MADs between SET_{ORIG} and $\text{SET}_{\text{C}\alpha}$, all these values are $5-65\times$ lower, with the most significant improvement occurring in bond lengths and the smallest in dihedral angles.

4.1 Comparison of PROPTIMUS $\text{RAPHAN}_{\text{GFN-FF}}$ optimised structures with their corresponding $\text{GFN-FF}_{\text{C}\alpha}$ local minima

Based on the results in Table 1, it is not possible to determine to what extent the remaining differences between $\text{SET}_{\text{RAPHAN}}$ and $\text{SET}_{\text{C}\alpha}$ are due to the approximative nature of PROPTIMUS $\text{RAPHAN}_{\text{GFN-FF}}$ (i.e., the neglect of long-range interactions) and how much they stem from PROPTIMUS $\text{RAPHAN}_{\text{GFN-FF}}$ converging to a different local minima on the $\text{GFN-FF}_{\text{C}\alpha}$ potential energy surface representing an alternative protein conformation. It is not at all unusual for various optimisation techniques to converge to more or less different minima in a complex space [32]. For this reason, three additional comparisons were performed (see Figure 2):

- The comparison of $\text{SET}_{\text{RAPHAN}}$ and $\text{SET}_{\text{RAPHAN+C}\alpha}$ evaluates the differences between structures optimised by PROPTIMUS $\text{RAPHAN}_{\text{GFN-FF}}$ and their corresponding local minima on the potential energy surface reached by applying $\text{GFN-FF}_{\text{C}\alpha}$. Therefore, this comparison assesses the error caused by the approximative nature of PROPTIMUS $\text{RAPHAN}_{\text{GFN-FF}}$ relative to the nearest reachable local minima on the $\text{GFN-FF}_{\text{C}\alpha}$ potential energy surface.
- The comparison of $\text{SET}_{\text{C}\alpha}$ and $\text{SET}_{\text{RAPHAN+C}\alpha}$ expresses the differences between local minima on the $\text{GFN-FF}_{\text{C}\alpha}$ potential energy surface reached by applying $\text{GFN-FF}_{\text{C}\alpha}$ on the original structures versus on structures optimised by the PROPTIMUS $\text{RAPHAN}_{\text{GFN-FF}}$ method. This comparison detects whether PROPTIMUS $\text{RAPHAN}_{\text{GFN-FF}}$ converges to a different local minimum on the $\text{GFN-FF}_{\text{C}\alpha}$ potential energy surface than the original $\text{GFN-FF}_{\text{C}\alpha}$ method does.

- The comparison of $\mathrm{SET_{ORIG}}$ and $\mathrm{SET_{RAPHAN+C\alpha}}$ allows an evaluation of the differences between $\mathrm{SET_{RAPHAN}}$ and $\mathrm{SET_{RAPHAN+C\alpha}}$ relative to the original structures.

The sets were compared using the same methodology as in Section 4. The comparative analysis results are summarised in Table 2. Percentiles and associated histograms of differences are available in the supplementary files.

Table 2 Comparison of $\mathrm{SET_{RAPHAN+C\alpha}}$ with other sets of structures $\mathrm{SET_{C\alpha}}$, $\mathrm{SET_{RAPHAN}}$ and $\mathrm{SET_{ORIG}}$. Mean absolute deviation of atom positions, bond lengths, bond angles, dihedral angles and four characteristic dihedral angles are included.

	$\mathrm{SET_{RAPHAN}}$ vs. $\mathrm{SET_{RAPHAN+C\alpha}}$	$\mathrm{SET_{C\alpha}}$ vs. $\mathrm{SET_{RAPHAN+C\alpha}}$	$\mathrm{SET_{ORIG}}$ vs. $\mathrm{SET_{RAPHAN+C\alpha}}$
MAD atomic position [Å]	0.033	0.056	0.412
MAD bond length [pm]	0.063	0.061	4.851
MAD bond angle [°]	0.080	0.105	2.071
MAD dihedral angle [°]	0.664	1.257	11.143
MAD ϕ [°]	0.881	1.339	10.302
MAD ψ [°]	0.816	1.246	9.247
MAD ω [°]	0.596	0.839	5.778
MAD χ_1 [°]	0.590	0.993	8.018

A comparison reveals that, for all tested metrics except bond lengths, the structures from $\mathrm{SET_{RAPHAN}}$ and $\mathrm{SET_{RAPHAN+C\alpha}}$ exhibit approximately half the difference compared to the structures from $\mathrm{SET_{RAPHAN}}$ and $\mathrm{SET_{C\alpha}}$ from Section 4. In addition, the comparison between $\mathrm{SET_{C\alpha}}$ and $\mathrm{SET_{RAPHAN+C\alpha}}$ proves the difference of these two sets of local minima on the $\mathrm{GFN\text{-}FF_{C\alpha}}$ potential energy surface. This confirms that the PROPTIMUS $\mathrm{RAPHAN_{GFN\text{-}FF}}$ and $\mathrm{GFN\text{-}FF_{C\alpha}}$ methods converge to different local minima on the $\mathrm{GFN\text{-}FF_{C\alpha}}$ potential energy surface. This fact is further demonstrated by the histogram of atomic position differences shown in Figure 3a). The histogram also highlights that the methods converge to alternative conformations,

particularly for residues whose side chains form few or no hydrogen bonds (i.e., non-polar residues and residues on the structure's surface). These regions are naturally flexible: the side chains switch between several energetically similar conformations.

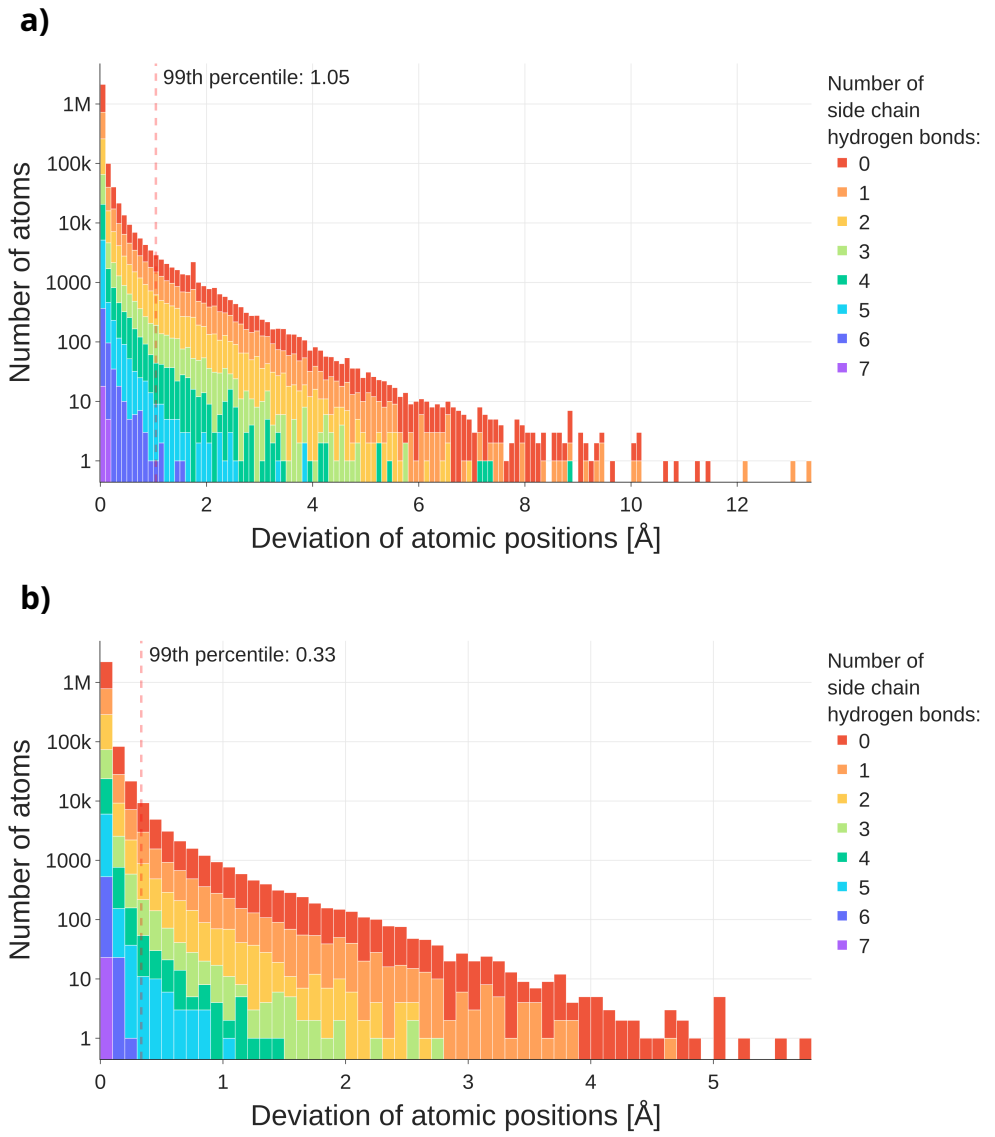


Fig. 3 Histograms of atom position deviations between a) $\mathrm{SET}_{\mathrm{C}\alpha}$ and $\mathrm{SET}_{\mathrm{RAPHAN+C}\alpha}$ and b) $\mathrm{SET}_{\mathrm{RAPHAN}}$ and $\mathrm{SET}_{\mathrm{RAPHAN+C}\alpha}$. Bars are coloured according to the number of hydrogen bonds of the side chain residue to which they belong. The image was created using the Plotly tool [33].

A comparison between the $\text{SET}_{\text{RAPHAN}}$ and $\text{SET}_{\text{RAPHAN+C}\alpha}$ sets in Table 2 shows that the atomic positions of protein structures optimised by PROPTIMUS $\text{RAPHAN}_{\text{GFN-FF}}$ deviate, on average, by 0.033 Å from their positions in the corresponding reachable local minimum on the $\text{GFN-FF}_{\text{C}\alpha}$ potential energy surface. The MAD of bond lengths is 0.063 pm. The MAD of bond angles is smaller than 0.1° and the one of dihedral angles is smaller than 1°. The most significant improvement relative to the original structures by the PROPTIMUS $\text{RAPHAN}_{\text{GFN-FF}}$ optimisations is achieved in bond lengths and angles, where the MADs decrease by 77× and 26× respectively. The MADs for other similarity metrics decrease by $10 - 17\times$.

For visual comparison, the structure with the worst MAD of atomic positions between structures from $\text{SET}_{\text{RAPHAN}}$ and $\text{SET}_{\text{RAPHAN+C}\alpha}$ with less than a thousand atoms is shown in Figure 4. While most atoms are in good agreement, the atoms of residues CYS9 and GLY8 differ. To analyse such outliers between $\text{SET}_{\text{RAPHAN}}$ and $\text{SET}_{\text{RAPHAN+C}\alpha}$ in detail, a histogram of the atom position deviations distribution is visualised in Figure 3b). The histogram shows that on rare occasions, the deviations caused by the approximative nature of PROPTIMUS $\text{RAPHAN}_{\text{GFN-FF}}$ can be very significant. Bars in the histogram are coloured according to the number of hydrogen bonds in the amino acid side chains. It can be seen, by examining the colours, that these rare significant deviations in atom positions are usually found within regions that are poor in hydrogen bonds, meaning nonpolar regions or regions on the surface of the structure. As previously mentioned, the side chains in these regions are naturally flexible and exhibit high conformational variability.

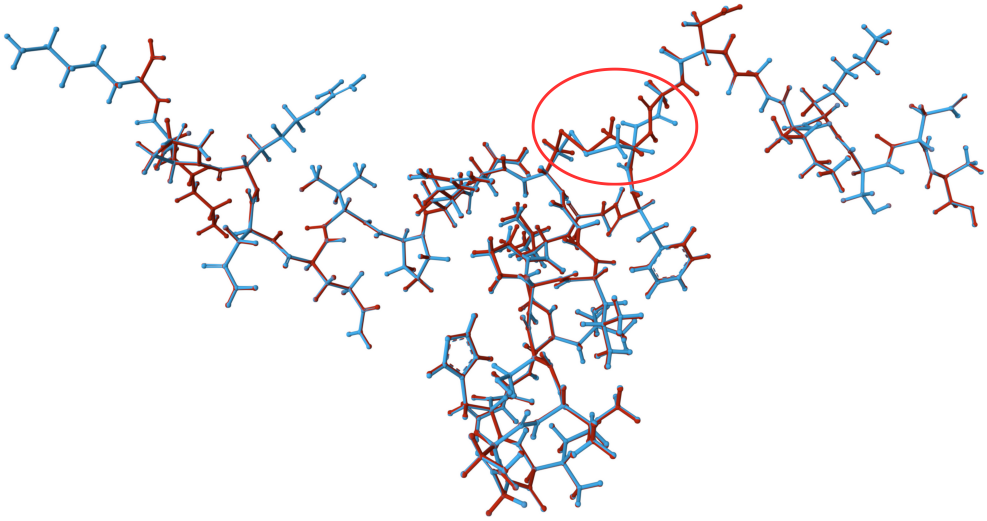


Fig. 4 The structure with the worst MAD of atomic positions with less than a thousand atoms from the testing dataset with UniProtKB AC P83224 optimised by PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ (blue) and PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ + GFN-FF$_{\text{C}\alpha}$ (red). The MAD of atomic positions for this structure is 0.075 Å. While most atoms are in good agreement, the atoms of residues CYS9 and GLY8 differ (red circle). The Mol* visualisation software [34] was used to create the figure.

4.2 Comparison of optimisation times and memory usage

To evaluate the computational efficiency, the computational time of PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ and GFN-FF$_{\text{C}\alpha}$ optimisations of original structures was compared. This comparison is illustrated in Figure 5. In the case of GFN-FF$_{\text{C}\alpha}$, the computation time increases quadratically [17]. In contrast, the computation time of PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ increases linearly with respect to the number of atoms. The high variability of PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ calculation duration stems from the fact that the total runtime is also influenced by the number of residual substructures, their respective atom counts, and the number of iterations required for PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ convergence. The method achieves a minimum throughput of 3,600

atoms per hour and an average speed of 5,000 atoms per hour on the CPU used for testing.

In addition to computational speed, memory efficiency is a crucial factor for the accessibility of the optimisation process. As previously noted in Section 3.1, the GFN-FF$_{C\alpha}$ optimisation of 15 structures exceeded the 196 GB memory limit. By contrast, PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ is significantly more memory-efficient. For the largest structure in our dataset (9,940 atoms), the optimisation required only 0.5 GB of RAM on a single CPU, and 3 GB when utilising 16 CPUs.

5 Conclusion

We developed the generic method PROPTIMUS RAPHAN as a fast alternative for the optimisation of proteins with constrained alpha carbons. Due to the use of the divide-and-conquer approach, the time complexity of PROPTIMUS RAPHAN is linear in relation to the size of the protein structure. We implemented our method as PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ employing the modern and almost QM-accurate force field GFN-FF and tested PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ on 461 structures from AlphaFold DB.

We found that the structures optimised by PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ are generally similar to those optimised by the GFN-FF$_{C\alpha}$ method. However, the results also demonstrate that PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ converges to different local minima on the GFN-FF$_{C\alpha}$ potential energy surface than the GFN-FF$_{C\alpha}$ method itself, particularly in flexible regions with a low number of side-chain hydrogen bonds. On the other hand, the structures optimised by PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ are comparable to their corresponding reachable local minima on the GFN-FF$_{C\alpha}$ potential energy surface, where the average atomic position deviation is 0.033 Å with outliers again

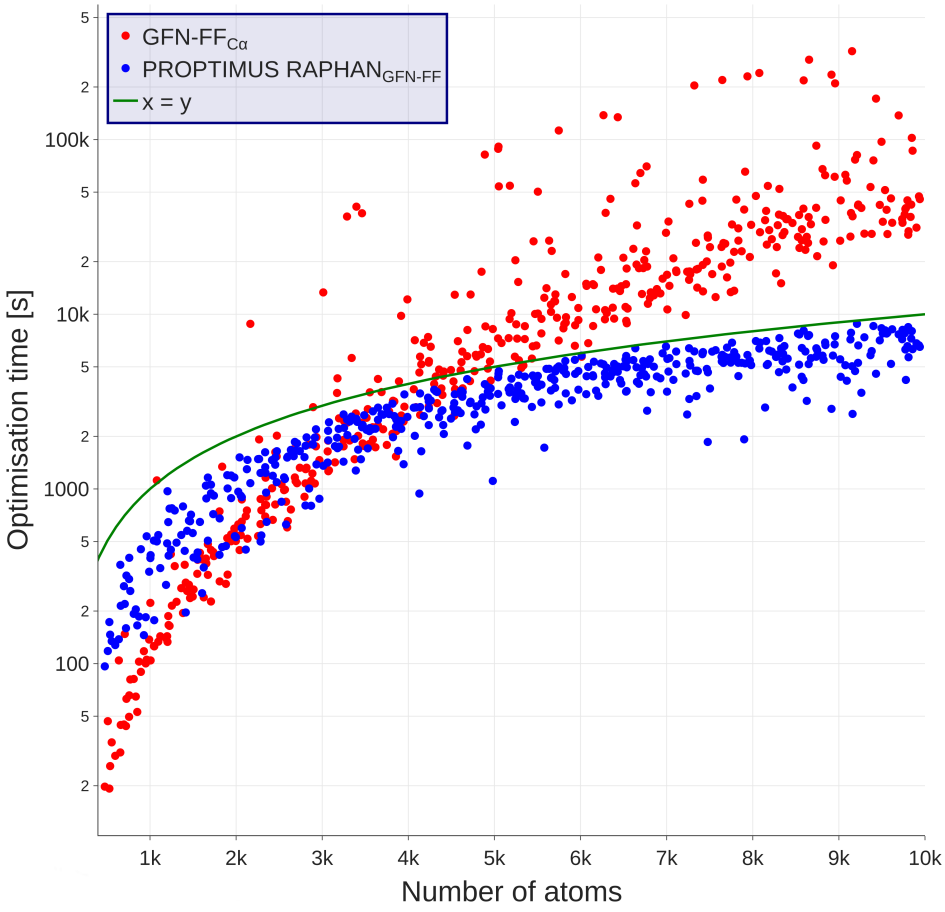


Fig. 5 Comparison of computational times of PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ and GFN-FF$_{C\alpha}$ optimisations. The image was created using the Plotly tool [33].

occurring in flexible regions. Thus, while PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ may converge to different conformations than the original GFN-FF$_{C\alpha}$ method, it approaches these alternative conformations with minimal error.

While the computation time for optimisation with constrained alpha carbons increases quadratically, the average speed of PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ is 5 000

atoms per hour on the CPU used for testing purposes. Another advantage of PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ is its low RAM usage, which further enhances its accessibility and practical utility. PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ is freely available, parallelisable, easy to install, and simple to use. Even the largest structures in AlphaFold DB could be highly optimised by PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ on a standard desktop computer in several hours.

Acronyms

GFN-FF$_{C\alpha}$ GFN-FF optimisation with constrained alpha carbons.

MAD mean absolute deviation.

PROPTIMUS RAPHAN Per-residue optimisation of protein structures: Rapid alternative to optimisation with constrained alpha carbons.

PROPTIMUS RAPHAN$_{GFN\text{-}FF}$ PROPTIMUS RAPHAN method employing a force field, GFN-FF.

PROPTIMUS RAPHAN$_{GFN\text{-}FF}$ + GFN-FF$_{C\alpha}$ PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ method, followed by optimisation by the GFN-FF$_{\text{C}\alpha}$ method.

6 Declarations

6.1 Availability of Data and Materials

The implementation of PROPTIMUS RAPHAN, including a description of its installation and use, is available at GitHub under the MIT licence at https://github.com/sb-ncbr/proptimus_raphan. The supplementary files supporting the conclusions of this article are available in the OneData repository at https://doi.org/10.58074/kap4-ka76 as well as archived implementation of PROPTIMUS RAPHAN.

6.2 Competing Interests

The authors declare that they have no competing interests.

6.3 Funding

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic (ELIXIR-CZ, Grant No. LM2023055). This work also received backing from the "Optimisation of protein structures predicted by artificial intelligence" grant (FR CESNET: 773R1/2025).

6.4 Authors' Contributions

Conceptualisation: OS. Data curation: OS, TS, GB. Formal analysis: OS, TS, GB. Funding acquisition: OS, RS. Investigation: OS, GB. Methodology: OS. Project administration: OS, GB, RS. Resources: OS, TS, RS. Software: OS, TS. Supervision: OS, RS. Validation: OS, TS, GB. Visualisation: OS, GB. Writing - original draft: OS, GB. Writing - review and editing: All authors.

6.5 Acknowledgements

Biological Data Management and Analysis Core Facility of CEITEC Masaryk University, funded by ELIXIR CZ research infrastructure (MEYS Grant No: LM2023055), is gratefully acknowledged for supporting the research presented in this paper. Computational resources were provided by the e-INFRA CZ (ID: 90254). The authors thank Lukáš Bohuš for valuable assistance with the manuscript review.

References

[1] Poluri, K.M., Gulati, K., Sarkar, S.: Structural and functional properties of proteins. In: Protein-Protein Interactions: Principles and Techniques: Volume

I, 1st edn., pp. 1–60. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-1594-8_1

[2] Scheeff, E.D., Fink, J.L.: Fundamentals of protein structure. Structural bioinformatics **44**, 15–39 (2003) https://doi.org/10.1002/0471721204.ch2

[3] Berman, H., Henrick, K., Nakamura, H.: Announcing the worldwide protein data bank. Nature structural & molecular biology **10**(12), 980–980 (2003) https://doi.org/10.1038/nsb1203-980

[4] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., *et al.*: Highly accurate protein structure prediction with alphafold. Nature **596**(7873), 583–589 (2021) https://doi.org/10.1038/s41586-021-03819-2

[5] Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., *et al.*: Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research **50**(D1), 439–444 (2022) https://doi.org/10.1093/nar/gkab1061

[6] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., Rives, A.: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**(6637), 1123–1130 (2023) https://doi.org/10.1126/science.ade2574

[7] Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., Beer, T.A.P., Rempfer, C., Bordoli, L., *et al.*: Swiss-model: homology modelling of protein structures and complexes. Nucleic Acids Research **46**(W1),

296–303 (2018) https://doi.org/10.1093/nar/gky427

[8] Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., Steinegger, M.: Colabfold: making protein folding accessible to all. Nature Methods **19**(6), 679–682 (2022) https://doi.org/10.1038/s41592-022-01488-1

[9] Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A.J., Bambrick, J., *et al.*: Accurate structure prediction of biomolecular interactions with alphafold 3. Nature **630**(8016), 1–3 (2024) https://doi.org/10.1038/s41586-024-07487-w

[10] Callaway, E.: 'it will change everything': Deepmind's ai makes gigantic leap in solving protein structures. Nature **588**(7837), 203–205 (2020) https://doi.org/10.1038/d41586-020-03348-4

[11] Terwilliger, T.C., Liebschner, D., Croll, T.I., Williams, C.J., McCoy, A.J., Poon, B.K., Afonine, P.V., Oeffner, R.D., Richardson, J.S., Read, R.J., *et al.*: Alphafold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. Nature Methods **21**(1), 110–116 (2024) https://doi.org/10.1038/s41592-023-02087-4

[12] Scardino, V., Di Filippo, J.I., Cavasotto, C.N.: How good are alphafold models for docking-based virtual screening? Iscience **26**(1), 105920 (2023) https://doi.org/10.1016/j.isci.2022.105920

[13] Adiyaman, R., McGuffin, L.J.: Methods for the refinement of protein structure 3d models. International Journal of Molecular Sciences **20**(9), 2301 (2019) https://doi.org/10.3390/ijms20092301

[14] Geidl, S., Svobodová Vařeková, R., Bendová, V., Petrusek, L., Ionescu, C.-M.,

Jurka, Z., Abagyan, R., Koča, J.: How does the methodology of 3d structure preparation influence the quality of pka prediction? Journal of Chemical Information and Modeling **55**(6), 1088–1097 (2015) https://doi.org/10.1021/ci500758w

[15] Bursch, M., Mewes, J.-M., Hansen, A., Grimme, S.: Best-practice dft protocols for basic molecular computational chemistry. Angewandte Chemie International Edition **61**(42), 202205735 (2022) https://doi.org/10.1002/anie.202205735

[16] Read, R.J., Sammito, M.D., Kryshtafovych, A., Croll, T.I.: Evaluation of model refinement in casp13. Proteins: Structure, Function, and Bioinformatics **87**(12), 1249–1262 (2019) https://doi.org/10.1002/prot.25794

[17] Spicher, S., Grimme, S.: Robust atomistic modeling of materials, organometallic, and biochemical systems. Angewandte Chemie International Edition **59**(36), 15665–15673 (2020) https://doi.org/10.1002/anie.202004239

[18] Hornak, V., Abel, R., Okur, A., Strockbine, B., Roitberg, A., Simmerling, C.: Comparison of multiple amber force fields and development of improved protein backbone parameters. Proteins: Structure, Function, and Bioinformatics **65**(3), 712–725 (2006) https://doi.org/10.1002/prot.21123

[19] Raček, T., Schindler, O., Toušek, D., Horskỳ, V., Berka, K., Koča, J., Svobodová, R.: Atomic charge calculator ii: web-based tool for the calculation of partial atomic charges. Nucleic Acids Research **48**(W1), 591–596 (2020) https://doi.org/10.1093/nar/gkaa367

[20] Schindler, O., Berka, K., Cantara, A., Křenek, A., Tichỳ, D., Raček, T., Svobodová, R.: αcharges: partial atomic charges for alphafold structures in high quality. Nucleic Acids Research, 349 (2023) https://doi.org/10.1093/nar/gkad349

[21] Schindler, O., Svoboda, T., Rošinec, A., Raček, T., Bučeková, G., Tichỳ, D., Berka, K., Svobodová, R.: Pdbcharges: Quantum-mechanical partial atomic charges for pdb structures. Nucleic Acids Research **53**(W1), 401 (2025) https://doi.org/10.1093/nar/gkaf401

[22] Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., *et al.*: Biopython: freely available python tools for computational molecular biology and bioinformatics. Bioinformatics **25**(11), 1422 (2009) https://doi.org/10.1093/bioinformatics/btp163

[23] Rdkit: Open-source cheminformatics https://doi.org/10.5281/zenodo.591637

[24] Bannwarth, C., Caldeweyher, E., Ehlert, S., Hansen, A., Pracht, P., Seibert, J., Spicher, S., Grimme, S.: Extended tight-binding quantum chemistry methods. Wiley Interdisciplinary Reviews: Computational Molecular Science **11**(2), 1493 (2021) https://doi.org/10.1002/wcms.1493

[25] Ehlert, S., Stahn, M., Spicher, S., Grimme, S.: Robust and efficient implicit solvation model for fast semiempirical methods. Journal of Chemical Theory and Computation **17**(7), 4250–4261 (2021) https://doi.org/10.1021/acs.jctc.1c00471

[26] Kunzmann, P., Hamacher, K.: Biotite: a unifying open source computational biology framework in python. BMC bioinformatics **19**, 1–8 (2018) https://doi.org/10.1186/s12859-018-2367-z . [PubMed Central:PMC6167853] [doi:10.1186/s12859-018-2367-z]

[27] Michaud-Agrawal, N., Denning, E.J., Woolf, T.B., Beckstein, O.: Mdanalysis: a toolkit for the analysis of molecular dynamics simulations. Journal of Computational Chemistry **32**(10), 2319–2327 (2011) https://doi.org/10.1002/jcc.21787

[28] Consortium, U.: Uniprot: the universal protein knowledgebase in 2023. Nucleic Acids Research **51**(D1), 523–531 (2023) https://doi.org/10.1093/nar/gkac1052

[29] Dolinsky, T.J., Nielsen, J.E., McCammon, J.A., Baker, N.A.: Pdb2pqr: an automated pipeline for the setup of poisson–boltzmann electrostatics calculations. Nucleic Acids Research **32**(suppl_2), 665–667 (2004) https://doi.org/10.1093/nar/gkh381

[30] Søndergaard, C.R., Olsson, M.H., Rostkowski, M., Jensen, J.H.: Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pka values. Journal of Chemical Theory and Computation **7**(7), 2284–2295 (2011) https://doi.org/10.1021/ct200133y

[31] Olsson, M.H., Søndergaard, C.R., Rostkowski, M., Jensen, J.H.: Propka3: Consistent treatment of internal and surface residues in empirical pka predictions. Journal of Chemical Theory and Computation **7**(2), 525–537 (2011) https://doi.org/10.1021/ct100578z

[32] Schindler, O., Raček, T., Maršavelski, A., Koča, J., Berka, K., Svobodová, R.: Optimized sqe atomic charges for peptides accessible via a web application. Journal of Cheminformatics **13**(1), 45 (2021) https://doi.org/10.1186/s13321-021-00531-1

[33] Inc., P.T.: Collaborative Data Science. https://plot.ly

[34] Sehnal, D., Bittrich, S., Deshpande, M., Svobodová, R., Berka, K., Bazgier, V., Velankar, S., Burley, S.K., Koča, J., Rose, A.S.: Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Research **49**(W1), 431–437 (2021) https://doi.org/10.1093/nar/gkab314