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

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

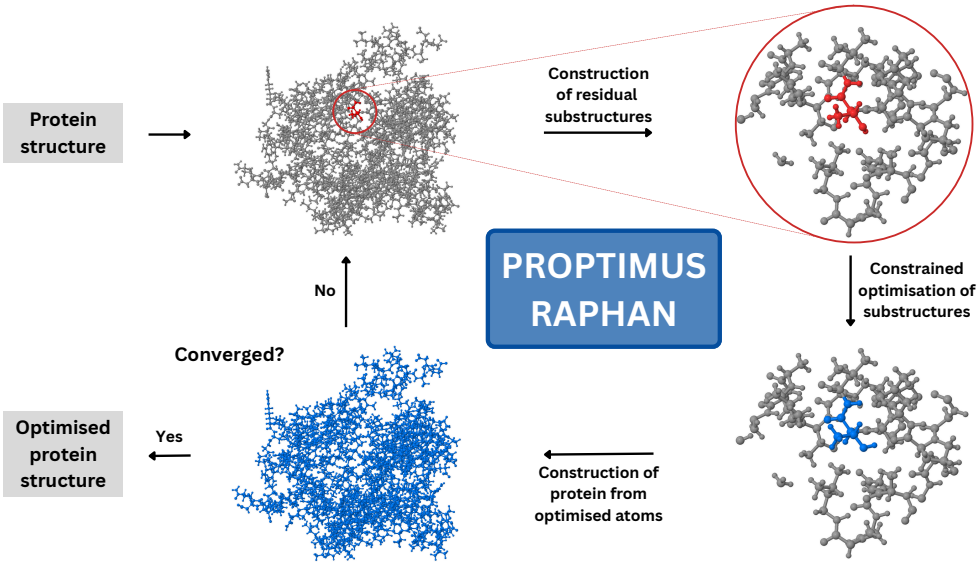
[1*]National Centre for Biomolecular Research, Masaryk University, Kamenice 753/5, Brno, 625 00, Czech Republic.
[2*]CEITEC – Central European Institute of Technology, Masaryk University, Kamenice 753/5, Brno, 625 00, Czech Republic.

*Corresponding author(s). E-mail(s): ondrej.schindler@mail.muni.cz; Contributing authors: gabriela.bucekova@mail.muni.cz; tomsvo@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, 2]. The 3D structure of a protein, the arrangement of atoms in space, is the crucial factor that determines the function of a protein [3]. In recent years, the number of predicted and experimentally determined protein structures has increased significantly. The Protein Data Bank [4]

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 [5, 6], and over 700 million structures are stored in the ESM Metagenomic Atlas [7]. 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 [8–10].

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 [11]. 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 [12, 13]. 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 [14]. Applications sensitive to structure quality are, for example, docking [13], QSPR modelling [15], partial atomic charge calculations [15], or calculations using semiempirical or QM/MM methods [16]. Therefore, predicted protein structures are usually optimised before conducting the main research [14, 17].

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 [18, 19]. However, even with the use of a force field, optimisation of proteins with more than ten thousand atoms remains computationally challenging [18]. 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 [5, 14, 17]. 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 [18].

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$_{\text{GFN-FF}}$) [18].

2 Methods

The PROPTIMUS RAPHAN method is developed based on a Cover approach, a divide-and-conquer algorithm, which has already been successfully utilised multiple times to calculate partial atomic charges [20–22]. 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 linear computational complexity with respect to the number of atoms. Moreover, the use of the divide-and-conquer algorithm ensures high parallelisability.

Before the PROPTIMUS RAPHAN computation begins, 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 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 [22]. 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.
2. **Constrained optimisation of substructures:** All residual substructures undergo optimisation. During this process, all *constrained atoms* are constrained and therefore their coordinates cannot be altered within optimisation. *Flexible atoms* are not constrained during optimisation, but are not used in the construction of the partially optimised structure in the following step. The number of optimisation steps is limited to prevent extensive changes in atom coordinates. Specifically, the number of optimisation steps is equal to the number of *optimised atoms* and the number of iterations.
3. **Construction of protein from optimised atoms:** *Optimised atoms* are taken from each optimised substructure and used to construct a new partially optimised protein structure.

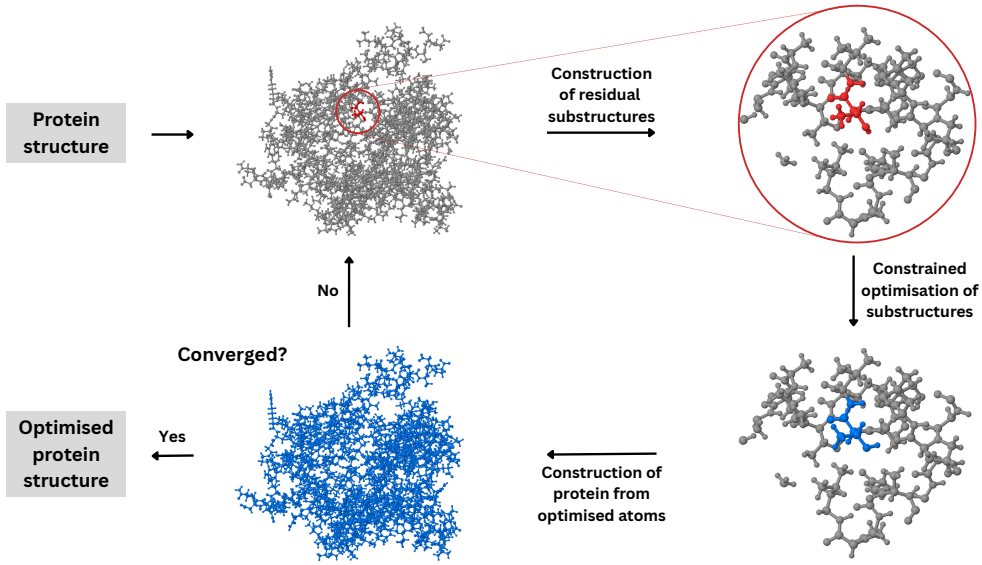


Fig. 1 Scheme of the PROPTIMUS RAPHAN method.

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.

2.1 Implementation

Our methodology can be used with any force fields. For our testing purposes, we decided to implement the PROPTIMUS RAPHAN method employing a force field, GFN-FF (PROPTIMUS RAPHAN$_{\text{GFN-FF}}$). These force fields are characterised by a near QM-accuracy and are suitable for a wide range of molecules [18, 23]. Solvation is

included by using the implicit solvent model ALPB [24]. For the purposes of testing and reproducibility, we also included the option to directly run the GFN-FF optimisation with constrained alpha carbons (GFN-FF$_{C\alpha}$). Method was implemented in Python, utilising the Biopython and RDKit libraries to manipulate protein structures [25, 26].

3 Results and Discussion

To compare the accuracy and speed of PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ with those of the original method GFN-FF$_{C\alpha}$, 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 [6]. In particular, the sample came from the fourth version (v4) of predictions for Swiss-Prot sequences [27]. 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 [28] with the PROPKA [29, 30] 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 SET$_{\text{ORIG}}$. Having the SET$_{\text{ORIG}}$ prepared, three different optimisation procedures were applied to it (see Figure 2). The following procedures took place:

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

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

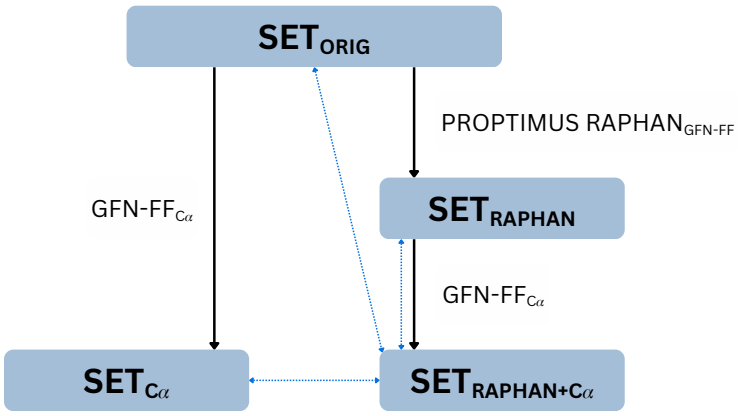


Fig. 2 Schematic diagram describing the relationships between datasets. Black lines represent optimisation. Blue dotted lines represent which datasets were compared in the following Chapter 3.2.

All optimisations were launched within our implementation. All optimisations were performed on a single AMD EPYC 9454 CPU. The GFN-FF$_{\text{C}\alpha}$ optimisation for 4 structures did not converge, and for 15 structures it exceeded our memory limit of 196GB of operating memory. 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.

3.2 Comparison of protein structures

As we explain in detail later on, we compared SET$_{\text{RAPHAN+C}\alpha}$ with other sets of structures SET$_{\text{C}\alpha}$, SET$_{\text{RAPHAN}}$ and SET$_{\text{ORIG}}$ (see Figure 2) to evaluate the accuracy of PROPTIMUS RAPHAN$_{\text{GFN-FF}}$. 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 respective

protein backbone, were calculated [18]. In this analysis, RDKit, Biopython, Biotite and MDAnalysis libraries were used to perform the comparisons [25, 26, 31, 32]. Python scripts used to compare optimised structures are available in the supplementary files. The results of comparative analysis are shown in Table 1.

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

	$\mathrm{SET}_{\mathrm{C\alpha}}$ vs. $\mathrm{SET}_{\mathrm{RAPHAN+C\alpha}}$	$\mathrm{SET}_{\mathrm{RAPHAN}}$ vs. $\mathrm{SET}_{\mathrm{RAPHAN+C\alpha}}$	$\mathrm{SET}_{\mathrm{ORIG}}$ vs. $\mathrm{SET}_{\mathrm{RAPHAN+C\alpha}}$
MAD atomic position [Å]	0.056	0.033	0.412
MAD bond length [pm]	0.113	0.064	4.893
MAD angle [°]	0.154	0.081	2.102
MAD dihedral angle [°]	2.174	0.735	12.526
MAD ϕ [°]	1.339	0.881	10.302
MAD ψ [°]	1.246	0.816	9.247
MAD ω [°]	0.839	0.596	5.778
MAD χ_1 [°]	0.993	0.590	8.018

It is not at all unusual for various optimisation techniques to converge to more or less different minima in complex space [33]. Therefore, to evaluate the extent to which the minimum approached by PROPTIMUS $\mathrm{RAPHAN}_{\mathrm{GFN\text{-}FF}}$ diverges from the one attained by $\mathrm{GFN\text{-}FF}_{\mathrm{C\alpha}}$, we compare $\mathrm{SET}_{\mathrm{RAPHAN+C\alpha}}$ with $\mathrm{SET}_{\mathrm{C\alpha}}$. The extra step with $\mathrm{GFN\text{-}FF}_{\mathrm{C\alpha}}$ applied in addition to PROPTIMUS $\mathrm{RAPHAN}_{\mathrm{GFN\text{-}FF}}$ is to eliminate the error induced by the approximative nature of PROPTIMUS $\mathrm{RAPHAN}_{\mathrm{GFN\text{-}FF}}$ and to compare in terms of actual local minima on the potential energy hypersurface of the $\mathrm{GFN\text{-}FF}_{\mathrm{C\alpha}}$ method. The fact that structures from $\mathrm{SET}_{\mathrm{C\alpha}}$ and $\mathrm{SET}_{\mathrm{RAPHAN+C\alpha}}$ are slightly different demonstrates that PROPTIMUS $\mathrm{RAPHAN}_{\mathrm{GFN\text{-}FF}}$ can indeed converge to a slightly different local minimum on the potential energy hypersurface than $\mathrm{GFN\text{-}FF}_{\mathrm{C\alpha}}$ does.

For that reason, the accuracy of PROPTIMUS $\mathrm{RAPHAN}_{\mathrm{GFN\text{-}FF}}$ is expressed in the comparison of $\mathrm{SET}_{\mathrm{RAPHAN}}$ with $\mathrm{SET}_{\mathrm{RAPHAN+C\alpha}}$. The comparison describes how

similar the structure optimised by the PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ is to the nearest local minimum on the potential energy hypersurface of GFN-FF$_{\text{C}\alpha}$. The values in Table 1 show that the structures of SET$_{\text{RAPHAN+C}\alpha}$ and SET$_{\text{RAPHAN}}$ are almost identical on average. The MAD of atomic positions is 0.033 Å. The MAD of bond lengths is 0.064 pm, which is at the level of accuracy of the PDB format. The MAD of bond angles is smaller than 0.1° and the one of dihedral angles is smaller than 1°. This shows that the approximative nature of PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ produces only a minimum deviation.

Comparing SET$_{\text{RAPHAN+C}\alpha}$ with SET$_{\text{ORIG}}$ allows us to evaluate the contribution of PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ in relation to the difference between the non-optimised state and the local minimum on the hypersurface of GFN-FF$_{\text{C}\alpha}$ to which PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ approximates, as already explained. The contribution PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ can be calculated as the difference between SET$_{\text{ORIG}}$ vs. SET$_{\text{RAPHAN+C}\alpha}$ and SET$_{\text{RAPHAN}}$ vs. SET$_{\text{RAPHAN+C}\alpha}$. We can see that the most significant advance achieved by the PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ optimisations was in bond lengths and angles, where the MAD decreased by 77× and 26× respectively. The MADs for other similarity metrics decreased by $10 - 15\times$.

For visual comparison, the structure with the worst MAD of atomic positions between structures from SET$_{\text{RAPHAN}}$ and SET$_{\text{RAPHAN+C}\alpha}$ with less than a thousand atoms is shown in Figure 3. While most atoms are in good agreement, the atoms of residues CYS9 and GLY8 differ. To analyse such outliers between SET$_{\text{RAPHAN}}$ and SET$_{\text{RAPHAN+C}\alpha}$ in detail, a histogram of the atom position deviations distribution is visualised in Figure 4. The histogram shows that on rare occasions, the deviations caused by the approximative nature of PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ can be very significant. The histogram also shows colours on the bars corresponding 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. These regions are naturally flexible: the side chains switch between several energetically similar conformations.

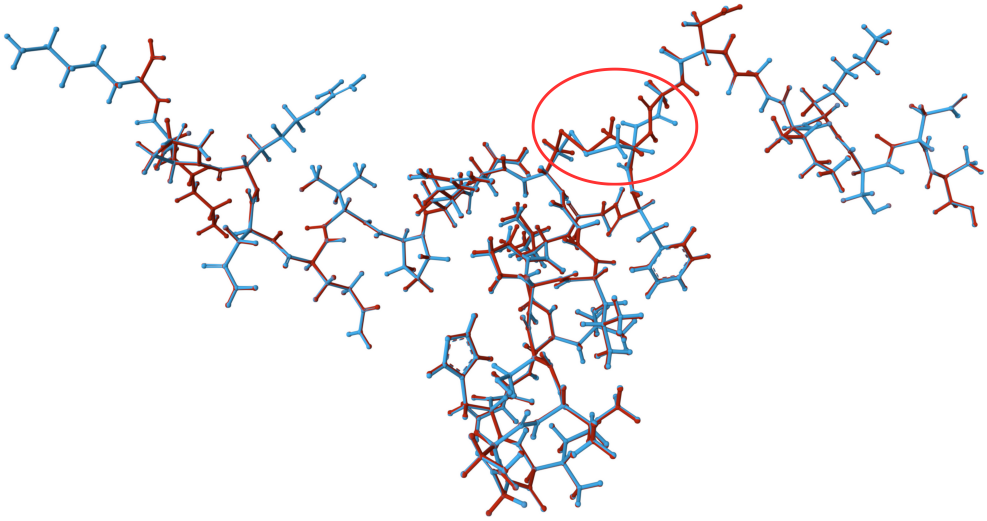


Fig. 3 The structure with the worst MAD of atomic positions with less than a thousand atoms from the testing dataset with UniProtKB AC P83224 optimised by PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ (blue) and PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ + GFN-FF$_{\text{C}\alpha}$ (red). The MAD of atomic positions for 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.

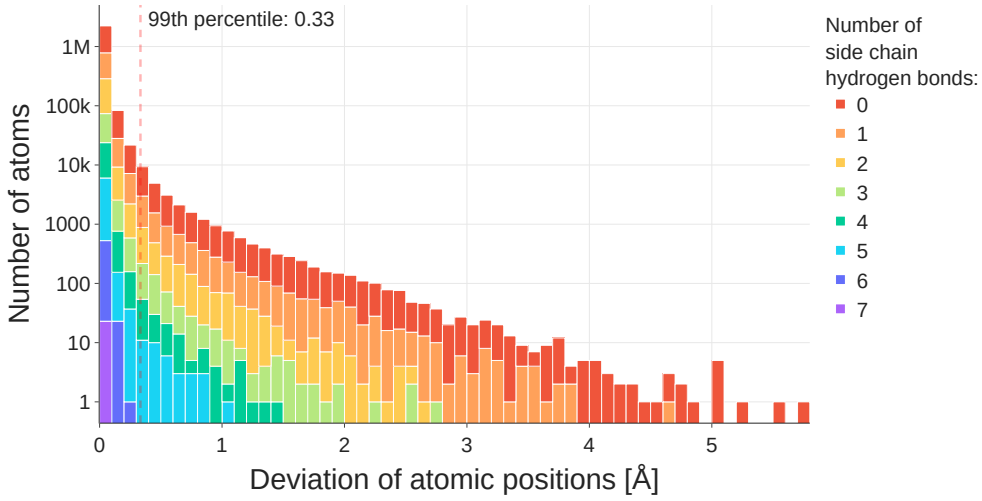


Fig. 4 Histogram of atom position deviations between $\mathrm{SET_{RAPHAN}}$ and $\mathrm{SET_{RAPHAN+C\alpha}}$. Bars are coloured according to the number of hydrogen bonds of the side chain residue to which they belong. The image was created using the Plotly tool [35].

3.3 Comparison of optimisation times

To evaluate the computational efficiency, the computational time of PROPTIMUS $\mathrm{RAPHAN_{GFN\text{-}FF}}$ and $\mathrm{GFN\text{-}FF_{C\alpha}}$ optimisations of original structures was compared. This comparison is illustrated in Figure 5. In the case of $\mathrm{GFN\text{-}FF_{C\alpha}}$, the computation time increases quadratically [18]. In contrast, the computation time of PROPTIMUS $\mathrm{RAPHAN_{GFN\text{-}FF}}$ increases linearly with respect to the number of atoms. The high variability in PROPTIMUS $\mathrm{RAPHAN_{GFN\text{-}FF}}$ calculation duration is due to the fact that the total calculation time is not linear with respect to the number of atoms, but rather to the number and size of residual substructures. For practical use, it can be said that the minimum speed of the method is 3 600 atoms per hour on the CPU,

which was used for testing purposes. The average speed, according to the testing, is 5 000 atoms per hour and the CPU used.

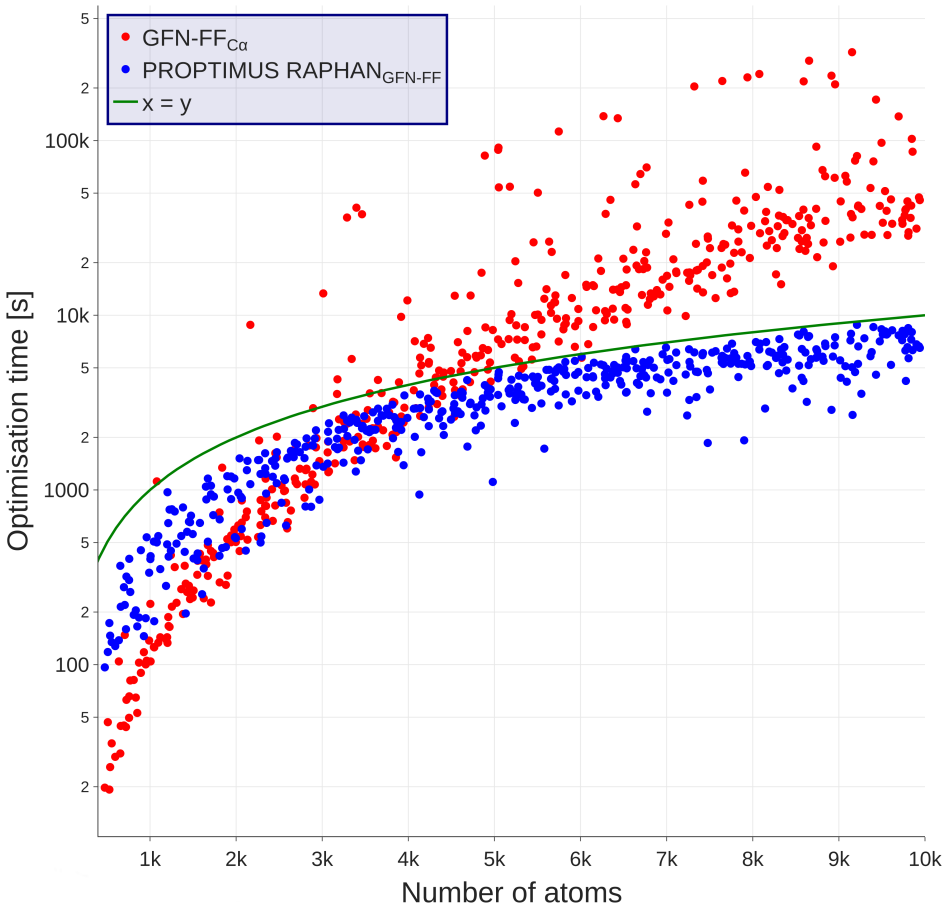


Fig. 5 Comparison of computational times of PROPTIMUS RAPHAN$_{GFN-FF}$ and GFN-FF$_{C\alpha}$ optimisations. The image was created using the Plotly tool [35].

4 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. We tested PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ on 461 structures from AlphaFold DB and demonstrated that the structures optimised by the PROPTIMUS RAPHAN$_{\text{GFN-FF}}$ are comparable to the structures optimised by the GFN-FF$_{\text{C}\alpha}$ method. Moreover, 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 and common CPU. PROPTIMUS RAPHAN 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.

5 Declarations

5.1 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.

5.2 Competing Interests

The authors declare that they have no competing interests.

5.3 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).

5.4 Authors' Contributions

Conceptualisation: OS. Data curation: OS, GB, TS. Formal analysis: OS, GB, TS. 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, GB, TS. Visualisation: OS, GB. Writing - original draft: OS, GB. Writing - review and editing: All authors.

5.5 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] Hille, B.: Ionic channels in nerve membranes, 50 years on. Progress in Biophysics and Molecular Biology **169**, 12–20 (2022) https://doi.org/10.1016/j.pbiomolbio.2021.11.003

[2] Raushel, F.M., Thoden, J.B., Holden, H.M.: Enzymes with molecular tunnels. Accounts of chemical research **36**(7), 539–548 (2003) https://doi.org/10.1021/ar020047k

[3] Thompson, M.C., Yeates, T.O., Rodriguez, J.A.: Advances in methods for atomic resolution macromolecular structure determination. F1000Research **9** (2020) https://doi.org/10.12688/f1000research.25097.1

[4] Burley, S.K., Bhatt, R., Bhikadiya, C., Bi, C., Biester, A., Biswas, P., Bittrich, S., Blaumann, S., Brown, R., Chao, H., *et al.*: Updated resources for exploring experimentally-determined pdb structures and computed structure models at the rcsb protein data bank. Nucleic Acids Research **53**(D1), 564–574 (2025) https://doi.org/10.1093/nar/gkae1091

[5] 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

[6] 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

[7] 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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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

[19] 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

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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

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

[26] 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

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

[28] 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

[29] 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

[30] 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

[31] 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]

[32] 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

[33] 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

[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

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