

α Charges: partial atomic charges for AlphaFold structures in high quality

Ondřej Schindler^{1,2}, Karel Berka³, Alessio Cantara^{1,2}, Aleš Křenek⁴,
Dominik Tichý⁵, Tomáš Raček^{1,2,*} and Radka Svobodová^{1,2,*}

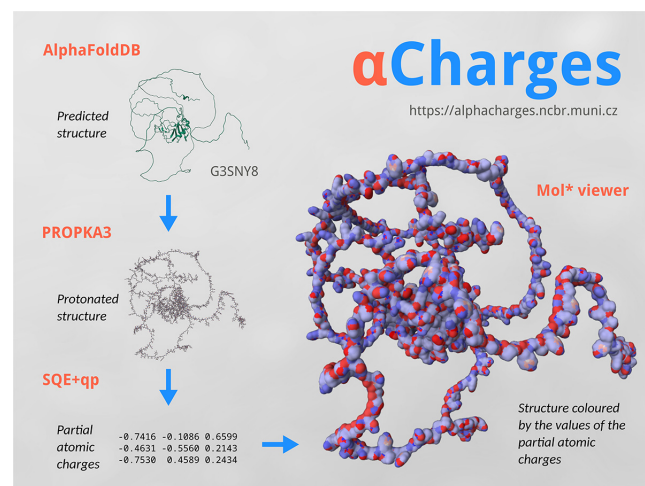
¹CEITEC – Central European Institute of Technology, Masaryk University, 625 00 Brno, Czech Republic, ²National Centre for Biomolecular Research, Faculty of Science, Masaryk University, 625 00 Brno, Czech Republic, ³Department of Physical Chemistry, Faculty of Science, Palacký University Olomouc, 779 00 Olomouc, Czech Republic, ⁴Institute of Computer Science, Masaryk University, 602 00 Brno, Czech Republic and ⁵Faculty of Informatics, Masaryk University, 602 00 Brno, Czech Republic

Received February 18, 2023; Revised March 17, 2023; Editorial Decision April 17, 2023; Accepted April 24, 2023

ABSTRACT

The AlphaFold2 prediction algorithm opened up the possibility of exploring proteins' structural space at an unprecedented scale. Currently, >200 million protein structures predicted by this approach are deposited in AlphaFoldDB, covering entire proteomes of multiple organisms, including humans. Predicted structures are, however, stored without detailed functional annotations describing their chemical behaviour. Partial atomic charges, which map electron distribution over a molecule and provide a clue to its chemical reactivity, are an important example of such data. We introduce the web application α Charges: a tool for the quick calculation of partial atomic charges for protein structures from AlphaFoldDB. The charges are calculated by the recent empirical method SQE+qp, parameterised for this class of molecules using robust quantum mechanics charges (B3LYP/6-31G*/NPA) on PROPKA3 protonated structures. The computed partial atomic charges can be downloaded in common data formats or visualised via the powerful Mol* viewer. The α Charges application is freely available at <https://alphacharges.ncbr.muni.cz> with no login requirement.

GRAPHICAL ABSTRACT



INTRODUCTION

The AlphaFold2 prediction algorithm (1) provided us with markedly greater insight into the structural space of proteins. Specifically, the set of 200 000 experimentally determined protein structures in the Protein Data Bank (2) is currently extended by >200 million predicted protein structures in AlphaFoldDB (3). This enables the research community to consider and hypothesise on scientific questions that could not have been touched until now.

It is beneficial to enrich the predicted protein structures with properties describing their chemical behaviour to support this research. A typical example of such data is partial atomic charges (4), which map the electron distribution over the molecule and provide a clue to its chemical reactivity. Partial atomic charges enable us to compare binding sites, including their polarity (5), identify hydrophobic membrane parts (4), or drive protein-protein

*To whom correspondence should be addressed. Tel: +420 54949 4860; Email: radka.svobodova@mail.muni.cz
Correspondence may also be addressed to Tomáš Raček. Tel: +420 54949 6568; Email: krablk@mail.muni.cz

interactions (6). They are also used as inputs for molecular docking (7), molecular dynamics (8), and other simulations (9,10).

The most reliable way to compute partial atomic charges is using quantum mechanics (QM) (11), which can provide electron density distribution among molecular orbitals. Afterwards, its mapping to atoms can be calculated by population analyses, e.g. Natural population analysis (NPA) (12,13). Unfortunately, QM methods are very time-consuming, and therefore only applicable to small molecules.

A faster alternative is the usage of empirical charge calculation methods (4), which mimic QM approaches and are based on significantly less complex empirical equation systems describing the charge distribution. These empirical methods are parameterised using QM charges calculated by a certain combination of a QM method and a population analysis. Charges computed by the empirical method mimic QM charges used in the parameterisation process. Even though the empirical methods are fast, it is challenging to apply them to proteins. One reason is that proteins are very large systems, and the empirical methods must be optimised to handle them. Moreover, proteins are homogeneous (i.e. they contain atoms with similar neighbourhoods), complicating the parameterisation of empirical methods. For these reasons, empirical charge calculation methods applicable to proteins have been developed in the last several years. We developed the SQE+qp method (14), which is currently the best-performing approach for proteins. We are also the authors of ACC II (4), a web server that provides charge calculation for almost any structure using 20 empirical methods. ACC II also includes the SQE+qp approach.

A limitation of ACC II is that it requires a protonated protein structure. Additionally, it implicitly deals with electroneutral molecules (i.e. molecules that have a zero total charge). The total charge is computed as a sum of the formal charges of individual atoms specified in the input file. A structure with a non-zero total charge can be provided. However, the user has to compute it by him or herself and include it in the input file. Moreover, the empirical methods in ACC II are only parameterised for electroneutral molecules. On the other hand, protein structures produced by the AlphaFold2 algorithm and contained in AlphaFoldDB do not include hydrogens, and after protonation, most of them are not electroneutral.

For this reason, we introduce α Charges, which solves all these challenges and provides the following functionality:

- The input protein structure from AlphaFoldDB is protonated using PROPKA3 (15,16) based on the pH value provided by the user (physiological pH is used as the default value).
- After protonation, the total charge of the protein is computed and used as an input for the SQE+qp method.
- The SQE+qp method is parameterised directly for protein-like structures.
- α Charges integrates the Mol* viewer (17), a cutting-edge molecular visualiser also used in AlphaFoldDB (3), PDBe (2) and RCSB PDB (18).

Therefore, α Charges provides high-quality partial atomic charges for AlphaFoldDB structures in one click. The user

only has to provide the UniProt ID of the protein, and no additional knowledge or preprocessing is required.

DESCRIPTION OF THE WEB SERVER

The application back end is written in Python, using the Flask web framework (<https://flask.palletsprojects.com/>). The front end is built with the Bootstrap library (<https://getbootstrap.com/>).

SQE+qp method

The computation of partial atomic charges is carried out according to the SQE+qp model (14), which belongs to the family of electronegativity equalisation methods, like the popular EEM (19), QEq (8) or SQE (20).

As SQE added bond information to the EEM formalism, SQE+qp further improved it by adding a parameterised initial charge for each atomic type. In this context, an atomic type represents a class of atoms with similar characteristics (e.g. all carbon atoms with only single bonds or all hydrogen atoms bound to oxygen atoms).

These initial charges and other model parameters have to be chosen appropriately for the given class of molecules (e.g. peptides, drug-like molecules, nucleic acids) to compute high-quality partial atomic charges.

α Charges workflow

One of the major design decisions for α Charges was to provide the user with an intuitive interface that would facilitate the calculation of partial atomic charges without needing to configure all the parameters of the process. The workflow is depicted in Figure 1 and can be described in several steps:

1. *Downloading the structure from AlphaFoldDB API:* The first step is to enter the UniProt ID (or AlphaFoldDB ID) of a selected protein structure into the textbox on the landing page. The structure is downloaded in PDB format if the ID is valid. Currently, structures are downloaded in the latest version (i.e. AlphaFoldDB version 4, as of February 2023). However, if a user clicks the *Setup calculation* button, any prior version can be selected.
2. *Addition of hydrogens by PDB2PQR/PROPKA3:* Once the structure is downloaded, adding hydrogen atoms is the next step. For this purpose, α Charges uses PDB2PQR (21) with the PROPKA3 library (15,16). By default, the protonation is done using the physiological pH of 7.2. However, any value between 0 and 14 can be specified on the *Setup* page.
3. *Calculation of charges by SQE+qp:* If the protonation succeeds, partial atomic charges can be computed. Since the input structure might be large (i.e. tens of thousands of atoms), even quite effective empirical methods such as SQE+qp might be too memory-demanding and slow, since they have a cubic computational complexity with respect to the number of bonds in the molecule.

To overcome the issue of computation complexity, α Charges uses the Cutoff/Cover approach (22), which first divides the molecule into smaller regions (spheres centred around individual residues). Then, the computation is carried out independently for each region. In the next step, the partial atomic charges are corrected so that

their sum equals the desired total molecular charge. First, the correction is computed for each small region. Then, the charges of the whole structure are revised in the same way.

The quality of the parameters for the empirical charge calculation method heavily influences the quality of the resulting charges. α Charges uses parameters fitted to reproduce B3LYP/6-31G*/NPA charges originally parameterised on the set of peptides and later optimised for proteins. Charges computed using these parameters were shown to achieve an excellent agreement with the reference charges (14).

4. *Visualisation of the charges with Mol**: After calculating partial atomic charges, the user is redirected to the Results page, whose central part is Mol* viewer (17). Mol* was extended to support the visualisation of partial atomic charges at several levels of protein structure depending on the chosen view. The default *cartoon* representation colours each residue by the sum of the charges of its atoms. The two other visualisation modes are *ball & stick* and *surface*.

The values of charges are mapped onto a linear blue-white-red gradient, where blue means a positive charge, white signifies zero charge, and red equals a negative charge. Two main colouring types are available in the application. In the default *relative colouring*, the gradient is mapped to the charges such that the highest and lowest charges are represented by the most saturated shades of blue and red, respectively. If a user wants to generate several images of various structures to compare the charge distribution, *absolute colouring* can be used. In this case, the maximum value of positive and negative charge can be configured manually.

5. *Download of computed charges*: Although the visual representation of the charges is undoubtedly valuable for humans, the charges can also be downloaded if further processing is required. α Charges includes files of several formats in a downloaded ZIP file for user convenience. These include:

- PDB file with added hydrogens
- PQR file with added hydrogens and calculated charges
- mmCIF file with added hydrogens and calculated charges
- a simple TXT file with calculated charges

Limitations

α Charges cannot compute partial atomic charges for proteins that contain structural errors because of an incorrect AlphaFold2 prediction. Furthermore, PROPKA3 is unable to protonate some structures correctly. In these cases, an appropriate error message is given to the user. The message lists all atoms that were identified as problematic. After clicking on their names, they are highlighted in Mol* so the user can visually inspect the issues.

RESULTS AND DISCUSSION

We put the application through extensive testing to confirm that it is stable and reliable. Details of this testing can be found in the Supplementary Data.

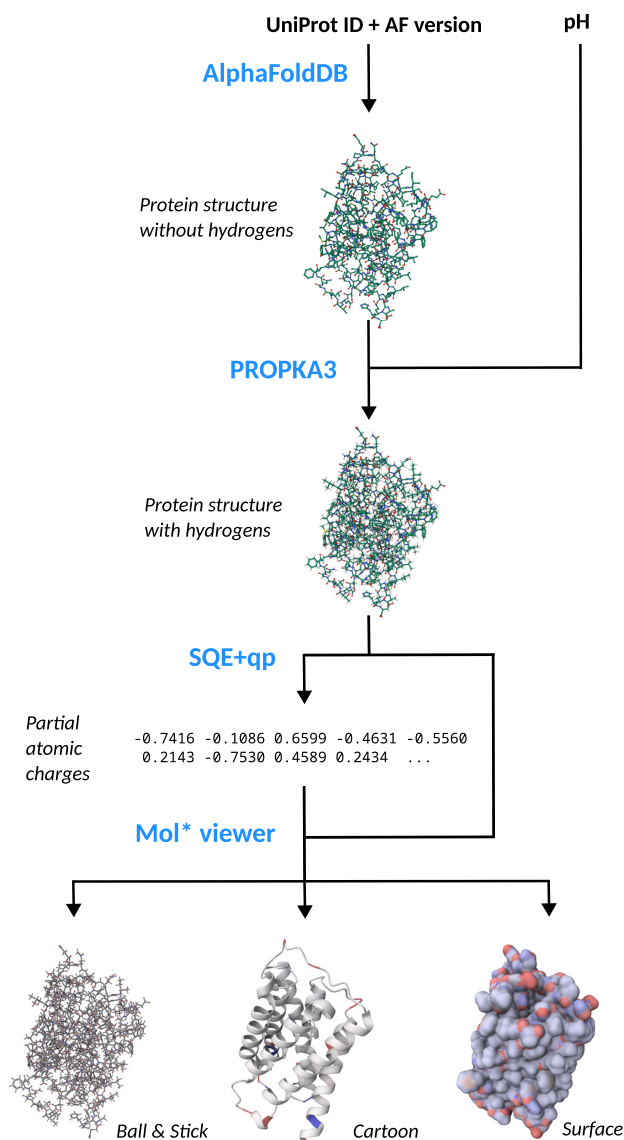


Figure 1. Schematic depiction of α Charges workflow. The schema includes a structure of porin (UniProt ID: A0A2E51BK2) originating from *Euryarchaeota archaeon*.

Furthermore, we provide three use cases demonstrating various applications of α Charges. They are presented interactively on the α Charges webpage.

Example I: P-glycoprotein

P-glycoprotein is one of the ABC transporter proteins that decrease drug accumulation in cancer cells (23). It is a 170kDa protein consisting of a nucleotide-binding domain and a transmembrane domain (24). We used a P-glycoprotein structure model from *Caenorhabditis elegans* (UniProt ID P34712). Visualising the charges (Figure 2) computed by α Charges demonstrates the differences in charge distribution between the transmembrane parts and extracellular/intracellular ones.

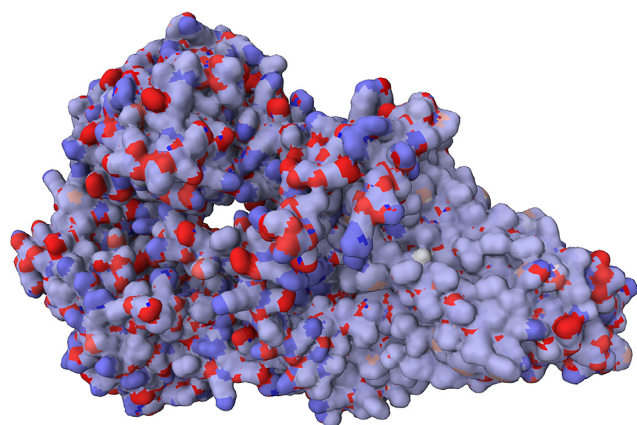


Figure 2. Structure of P-glycoprotein from *Caenorhabditis elegans* (UniProt ID: P34712). Partial atomic charges calculated by α Charges are visualised on the protein's surface. They show distinct areas of the protein: The nonpolar transmembrane part (mostly very light blue due to the charge being around zero on the surface) and the polar surface of extracellular and cytoplasmic parts (with a mosaic of blue positive and red negative charges).

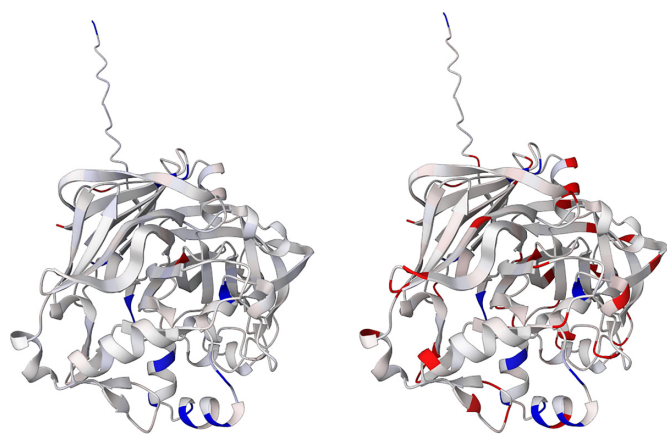


Figure 3. Structure of pepsin (UniProt ID: P00791), partial atomic charges calculated by α Charges are visualised on the cartoon model of the protein showing significant charge differences between an active form (left image, pH 2) and an inactive form (right image, pH 7).

Example II: pepsin

Pepsin is an enzyme that plays a major role in protein digestion in the stomach. It is secreted as a zymogen and activated by the acidic pH created by the stomach parietal cells. Pepsin is most effective at a pH of approximately 1.5 to 2, and becomes inactive when the pH rises above 6 (25). The enzyme denatures at a pH 8 (26). Differences in the charge distribution between its active form (pH 2) and inactive form (pH 7) can be seen in this use case (Figure 3). The alkaline environment causes an increase in negative charges, which contributes to the structural instability of the pepsin, causing the shift to an inactive form of the protein (27,28).

Example III: PIN proteins

The PIN family proteins control plant growth by regulating auxin export from the cytosol to the extracellular space (29). Eight types of PIN transport proteins are known (PIN1

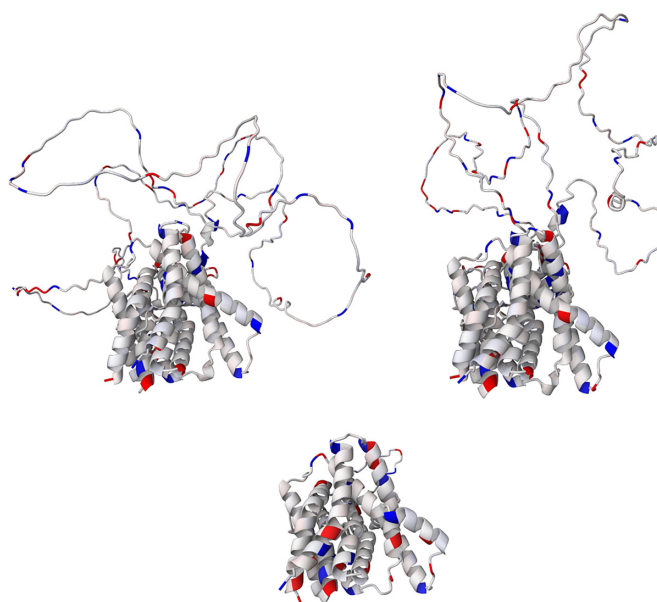


Figure 4. Structure of PIN3 (top left image, UniProt ID: Q9S7Z8), PIN7 (top right, UniProt ID: Q940Y5) and PIN5 (bottom image, UniProt ID: Q9FFD0) from *A. thaliana*. Partial atomic charges calculated by α Charges are visualised in a cartoon model.

to PIN8) and they are divided into two classes: canonical PINs (PIN1-4 and PIN7) and non-canonical PINs (PIN5, PIN6 and PIN8) (29). This year, structures of three of them were discovered and published (PIN1 (30), PIN3 (31) and PIN8 (29)), while structures of the remaining ones are waiting to be revealed. But thanks to AlphaFoldDB, we can see their predicted structure and compare their charge distribution.

Charge distribution of canonical PINs is very similar, even though sequence and structure strongly differ. It can be seen especially when we compare PIN3 and PIN7 (see Figure 4, top images). On the other hand, charge distribution of non-canonical PINs differs from each other and also from canonical PINs (see PIN5 in Figure 4, bottom image).

CONCLUSION

In this article, we presented α Charges, a novel web application for calculating partial atomic charges on protein structures produced by the AlphaFold2 algorithm and available in AlphaFoldDB. α Charges utilises the SQE+qp empirical charge calculation method, parameterised using B3LYP/6-31G*/NPA quantum mechanical charges. α Charges allows users to download charges (in PQR, mmCIF, or plaintext formats) or visualise them via three main structure visualisation models (cartoon, ball & stick, and surface). The web application is easy to use and is platform-independent. Documentation explaining the usage of the tool is provided on the webpage.

DATA AVAILABILITY

α Charges application is freely available at <https://alphacharges.ncbr.muni.cz> (hosted by the highly available

Masaryk University computing cloud) with no login requirement. The user manual for the application is available at <https://github.com/sb-ncbr/AlphaCharges/wiki>, while the source code is accessible at GitHub under the MIT licence at <https://github.com/sb-ncbr/AlphaCharges/> and is also available in the Supplementary Data. The code is also available on Zenodo at <https://doi.org/10.5281/zenodo.7844436>.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Vladimír Horský for his valuable comments and proofreading.

Computational resources were supplied by the project ‘e-Infrastruktura CZ’ (e-INFRA CZ LM2023054) supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Core Facility Biological Data Management and Analysis of CEITEC Masaryk University, supported by ELIXIR CZ research infrastructure (MEYS Grant No: LM2023055) is gratefully acknowledged for the obtaining of the scientific data presented in this paper.

FUNDING

Ministry of Education, Youth and Sports of the Czech Republic: ELIXIR CZ [LM2023055]; IOCB Tech Foundation: The Martina Roeselová Memorial Fellowship [to O.S.]. Funding for open access charge: Masaryk University as per its Read and Publish agreement with the Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
- Varadi, M., Anyango, S., Appasamy, S.D., Armstrong, D., Bage, M., Berrisford, J., Choudhary, P., Bertoni, D., Deshpande, M., Leines, G.D. *et al.* (2022) PDBe and PDBe-KB: providing high-quality, up-to-date and integrated resources of macromolecular structures to support basic and applied research and education. *Protein Sci.*, **31**, e4439.
- Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A. *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
- Raček, T., Schindler, O., Toušek, D., Horský, V., Berka, K., Koča, J. and Svobodová, R. (2020) Atomic Charge Calculator II: web-based tool for the calculation of partial atomic charges. *Nucleic Acids Res.*, **48**, W591–W596.
- Kangas, E. and Tidor, B. (2001) Electrostatic complementarity at ligand binding sites: application to chorismate mutase. *J. Phys. Chem. B*, **105**, 880–888.
- Gitlin, I., Carbeck, J.D. and Whitesides, G.M. (2006) Why are proteins charged? Networks of charge-charge interactions in proteins measured by charge ladders and capillary electrophoresis. *Angew. Chem.*, **45**, 3022–3060.
- Park, H., Lee, J. and Lee, S. (2006) Critical assessment of the automated AutoDock as a new docking tool for virtual screening. *Proteins*, **65**, 549–554.
- Rappé, A.K. and Goddard III, W.A. (1991) Charge equilibration for molecular dynamics simulations. *J. Phys. Chem.*, **95**, 3358–3363.
- Nazarian, D., Camp, J.S. and Sholl, D.S. (2016) A comprehensive set of high-quality point charges for simulations of metal–organic frameworks. *Chem. Mater.*, **28**, 785–793.
- Boda, D., Gillespie, D., Nonner, W., Henderson, D. and Eisenberg, B. (2004) Computing induced charges in inhomogeneous dielectric media: application in a Monte Carlo simulation of complex ionic systems. *Phys. Rev. E*, **69**, 046702.
- Shankar, R. (2012) In: *Principles of Quantum Mechanics*. Springer, NY.
- Reed, A.E. and Weinhold, F. (1983) Natural bond orbital analysis of near-Hartree–Fock water dimer. *J. Chem. Phys.*, **78**, 4066–4073.
- Reed, A.E., Weinstock, R.B. and Weinhold, F. (1985) Natural population analysis. *J. Chem. Phys.*, **83**, 735–746.
- Schindler, O., Raček, T., Maršavelski, A., Koča, J., Berka, K. and Svobodová, R. (2021) Optimized SQE atomic charges for peptides accessible via a web application. *J. Cheminform.*, **13**, 45.
- Søndergaard, C.R., Olsson, M.H., Rostkowski, M. and Jensen, J.H. (2011) Improved treatment of ligands and coupling effects in empirical calculation and rationalization of pKa values. *J. Chem. Theory Comput.*, **7**, 2284–2295.
- Olsson, M.H., Søndergaard, C.R., Rostkowski, M. and Jensen, J.H. (2011) PROPKA3: consistent treatment of internal and surface residues in empirical pKa predictions. *J. Chem. Theory Comput.*, **7**, 525–537.
- Sehna, D., Bittrich, S., Deshpande, M., Svobodová, R., Berka, K., Bazgier, V., Velankar, S., Burley, S.K., Koča, J. and Rose, A.S. (2021) Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Res.*, **49**, W431–W437.
- Burley, S.K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G.V., Christie, C.H., Dalenberg, K., Di Costanzo, L., Duarte, J.M. *et al.* (2021) RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res.*, **49**, D437–D451.
- Mortier, W.J., Ghosh, S.K. and Shankar, S. (1986) Electronegativity equalization method for the calculation of atomic charges in molecules. *J. Am. Chem. Soc.*, **108**, 4315–4320.
- Nistor, R.A., Polihronov, J.G., Müser, M.H. and Mosey, N.J. (2006) A generalization of the charge equilibration method for nonmetallic materials. *J. Chem. Phys.*, **125**, 094108.
- Dolinsky, T.J., Nielsen, J.E., McCammon, J.A. and Baker, N.A. (2004) PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Res.*, **32**, W665–W667.
- Ionescu, C.-M., Sehna, D., Falginella, F.L., Pant, P., Pravda, L., Bouchal, T., Svobodová Vařeková, R., Geidl, S. and Koča, J. (2015) AtomicChargeCalculator: interactive web-based calculation of atomic charges in large biomolecular complexes and drug-like molecules. *J. Cheminform.*, **7**, 50.
- Leslie, E.M., Deeley, R.G. and Cole, S.P. (2005) Multidrug resistance proteins: role of P-glycoprotein, MRP1, MRP2, and BCRP (ABCG2) in tissue defense. *Toxicol. Appl. Pharmacol.*, **204**, 216–237.
- Ward, A.B., Szewczyk, P., Grimard, V., Lee, C.-W., Martinez, L., Doshi, R., Caya, A., Villaluz, M., Pardon, E., Cregger, C. *et al.* (2013) Structures of P-glycoprotein reveal its conformational flexibility and an epitope on the nucleotide-binding domain. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 13386–13391.
- Heda, R., Toro, F. and Tombazzi, C.R. (2022) Physiology, pepsin. In: *StatPearls*. StatPearls Publishing, Treasure Island, FL.
- Stanforth, K.J., Wilcox, M.D., Chater, P.I., Brownlee, I.A., Zakhour, M.I., Banecki, K. and Pearson, J.P. (2022) Pepsin properties, structure, and its accurate measurement: a narrative review. *Ann. Esophagus*, **5**, 31.
- Tanaka, T. and Yada, R.Y. (2001) N-terminal portion acts as an initiator of the inactivation of pepsin at neutral pH. *Protein Eng. Des. Sel.*, **14**, 669–674.
- Grahame, D.A., Dupuis, J.H., Bryksa, B.C., Tanaka, T. and Yada, R.Y. (2021) Improving the alkaline stability of pepsin through rational protein design using renin, an alkaline-stable aspartic protease, as a structural and functional reference. *Enzyme Microb. Technol.*, **150**, 109871.

29. Ung,K.L., Winkler,M., Schulz,L., Kolb,M., Janacek,D.P., Dedic,E., Stokes,D.L., Hammes,U.Z. and Pedersen,B.P. (2022) Structures and mechanism of the plant PIN-FORMED auxin transporter. *Nature*, **609**, 605–610.
30. Yang,Z., Xia,J., Hong,J., Zhang,C., Wei,H., Ying,W., Sun,C., Sun,L., Mao,Y., Gao,Y. *et al.* (2022) Structural insights into auxin recognition and efflux by Arabidopsis PIN1. *Nature*, **609**, 611–615.
31. Su,N., Zhu,A., Tao,X., Ding,Z.J., Chang,S., Ye,F., Zhang,Y., Zhao,C., Chen,Q., Wang,J. *et al.* (2022) Structures and mechanisms of the Arabidopsis auxin transporter PIN3. *Nature*, **609**, 616–621.