DE-STRESS: A user-friendly web application for the evaluation of protein designs

Michael J. Stam[1] and Christopher W. Wood[2]*.

1. School of Informatics, University of Edinburgh, 10 Crichton St, Newington, Edinburgh EH8 9AB.
2. School of Biological Sciences, University of Edinburgh, Roger Land Building, Edinburgh EH9 3FF.

* To whom correspondence should be addressed.

Abstract

Motivation

It is becoming routine to design protein structures *de novo*, with many interesting and useful examples in the literature. However, most sequences of designed proteins could be classed as failures when characterised in the lab, usually as a result of low expression, misfolding, aggregation or lack of function. This high attrition rate makes protein design unreliable and costly. These limitations could potentially be addressed if it were quick and easy to generate a set of high-quality metrics and information regarding designs, which could be used to make reproducible and data-driven decisions about which designs to characterise experimentally.

Results

We present DE-STRESS (DEsigned STRucture Evaluation ServiceS), a web application for the evaluation of structural models of designed and engineered proteins. DE-STRESS has been designed to be simple, intuitive to use and responsive. It provides a wealth of information regarding designs, as well as tools to help contextualise the results and formally describe the properties that a design requires to be fit for purpose.

Availability

DE-STRESS is available for non-commercial use, without registration, through the following website: https://pragmaticproteindesign.bio.ed.ac.uk/de-stress/. Source code for the

application is available on GitHub: https://github.com/wells-wood-research/de-stress. The data used to generate reference sets is available through a GraphQL API, with the following URL: https://pragmaticproteindesign.bio.ed.ac.uk/big-structure/graphql.

Introduction

There has been rapid development in the field of *de novo* protein design over recent years, with more groups producing increasingly ambitious designs with complex behaviour, often applied in cellular environments (Ben-Sasson et al., 2021; Glasgow et al., 2019; Harrington et al., 2021; Herud-Sikimić et al., 2021; Pirro et al., 2020; Sesterhenn et al., 2020; VanDrisse et al., 2021).

Despite the great promise of *de novo* protein design, it remains the domain of highly specialist research groups, as there are significant barriers blocking broader adoption as a methodology. One major challenge is that only a fraction of designs adopt stable, folded structures when expressed (Huang et al., 2016), and it can be difficult to identify these models using the metrics calculated during the design process alone (Radom et al., 2018). This is especially challenging for designs with complex requirements that are needed for targeted applications.

Here we present DE-STRESS (DEsigned STRucture Evaluation ServiceS), a user-friendly web application for evaluating structural models of designed and engineered proteins. We aim to provide the user with as much information as possible about their designs before they select sequences to characterise experimentally.

Methods and Results

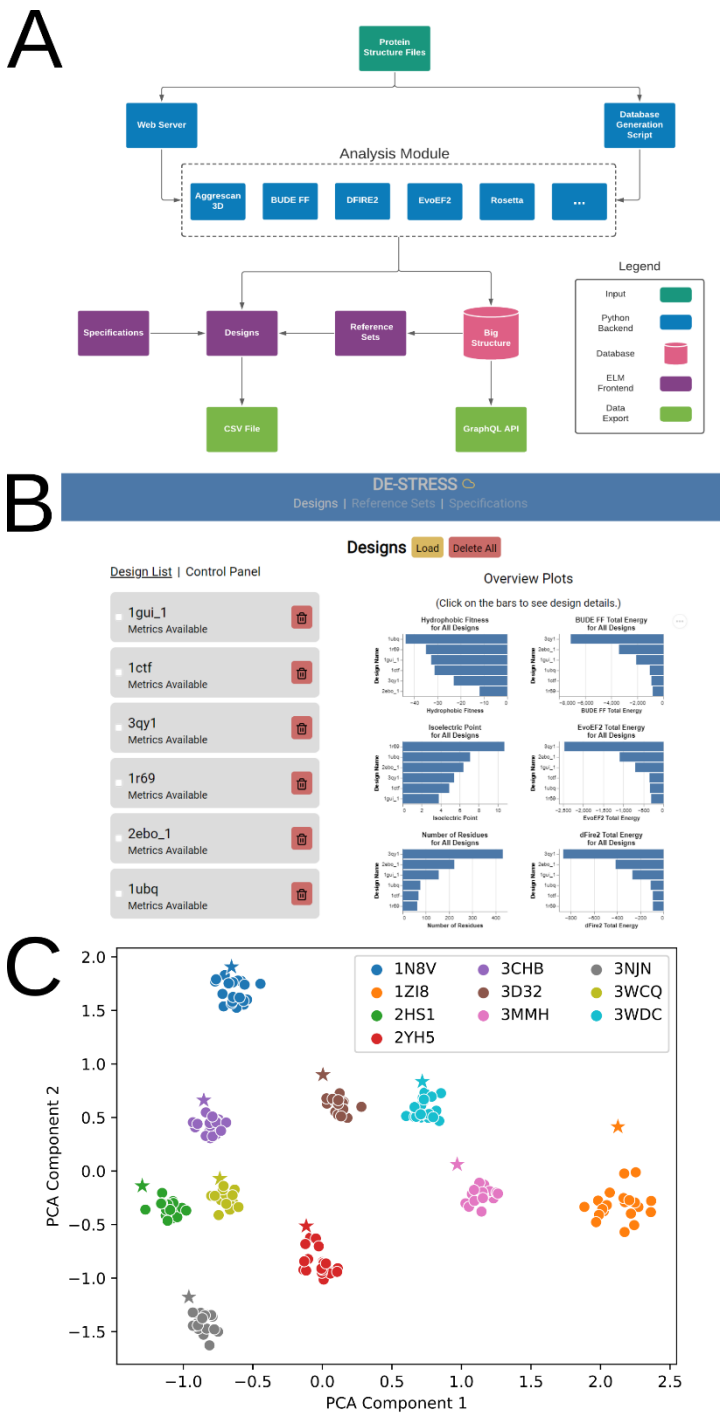

Figure 1: Overview of the DE-STRESS application. A) Architecture of the application. B) The "Designs" page. C) Principal component analysis of DE-STRESS metrics generated for experimentally-determined structures (stars) and folding decoys (circles).

The DE-STRESS application consists of a simple and intuitive user interface, written in Elm/JavaScript, and a backend web stack, consisting of

Gunicorn/Flask/GraphQL/PostgreSQL (figure 1A). The interface has three main sections that the user can explore: Designs, Reference Sets and Specifications.

On the Designs page, users can upload models of proteins (in PDB format) to the DE-STRESS server, where all the included metrics will be calculated for each design. Once the metrics have been calculated, an overview of the whole batch of designs is provided on the front page (figure 1B). Detailed information can be viewed for each design, as well as a comparison to the active Reference Set and Requirement Specification (*vide infra*).

On the Reference Sets page, users can define a set of known protein structures from the PDB (Berman et al., 2003), which can be used as a basis of comparison for their designs. We have precalculated the metrics included in DE-STRESS for the biological units of 82,010 protein structures, as defined by the PDBe (http://ftp.ebi.ac.uk/pub/databases/pdb/data/biounit/). The remaining structures in the PDB either did not contain protein, contained formatting errors in the PDB file or, in the case of large structures, failed to return results within a reasonable timeframe. Using these data, the user can define their own reference sets by submitting a list of PDB accession codes, enabling them to compare their designs to relevant structures. Additionally, two default reference sets are provided as an example, based on high-quality structures from Top500 (Hobohm and Sander, 1994) and Pisces (Wang and Dunbrack, 2003). Once a reference set has been defined, aggregated metrics are presented alongside the metrics for the user's designs. All the data used to generate the reference sets is available to search and download, programmatically and interactively, through a GraphQL API available at the following url: https://pragmaticproteindesign.bio.ed.ac.uk/big-structure/graphql.

Finally, the Specifications page allows the user to define "Requirement Specifications", which encapsulate the properties their designs should have in order to be fit for purpose. The user can define complex rules that can be used to filter designs, alongside associated metadata. We plan to expand the role of the specifications in the future, allowing the user to

capture more information about their design intent and export the specification to be used by other programmes.

A variety of external software packages are used by the DE-STRESS web server to calculate metrics for uploaded protein structures. Basic information about the protein structure is extracted using ISAMBARD (Wood et al., 2017), including information such as the isoelectric point and composition of the sequence, as well as implementations of a few metrics from the literature, such as packing density (Weiss, 2007) and hydrophobic fitness (Huang et al., 1995). In addition to these metrics, DE-STRESS applies a range of scoring functions that are well established in the protein-design field, such as BUDE (McIntosh-Smith et al., 2012, 2015), EvoEF2 (Huang et al., 2020), Rosetta (Alford et al., 2017) and DFIRE2 (Yang and Zhou, 2008). Finally, Aggrescan3D (Kuriata et al., 2019) calculates an aggregation propensity score for protein structures. Additional metrics will be incorporated into DE-STRESS in future releases. These metrics are presented on the Design Details page, alongside a visualisation of the model, using the NGL JavaScript library (Rose and Hildebrand, 2015; Rose et al., 2016), and other information such as secondary structure assignment using DSSP (Kabsch and Sander, 1983; Touw et al., 2015).

A privacy first approach has been taken when implementing DE-STRESS. No login is required to use the application and no data regarding the user, or their designs, are stored on our server. Designs are submitted directly to an in-memory job queue, with no associated metadata, and the results are returned directly to the user. All data regarding the user's designs are stored locally on the device used to access the website and can be exported to a CSV file for further analysis. With this architecture, we aim to give the user confidence in submitting their designs to the server. However, if they would like to take further steps to ensure that no one could access their data, they can run a local instance of the web application, which we have made as simple as possible by containerising the application.

We envisage that DE-STRESS will be useful for generating descriptive information and statistics that could be manually examined by users to choose designs that meet the needs

of their application. Beyond this, the datasets that DE-STRESS creates could be useful for automatic identification of high-quality designs using data-driven methods. As a simple example of this, we attempted to identify folding decoys from experimentally-determined structures. Using the DE-STRESS web application, we generated and exported metrics for a random sample of 10 experimentally-determined structures, along with 200 decoys (20 per structure) generated by 3DRobot (Deng et al., 2016). The metrics were normalised, using min-max scaling, and principal component analysis was performed. After this, the first two principal components were plotted against each other, and the experimentally-determined structure, with their associated decoys, formed neat clusters (figure 1C). Furthermore, the experimentally-determined structures were close to, but distinct from, the main cluster, indicating that the metrics included in DE-STRESS could be used to automatically identify high-quality models using machine learning. The dataset and the associated scripts for performing this analysis are available on GitHub: https://github.com/wells-wood-research/stam-m-wood-c-de-stress-2021.

Conclusions

DE-STRESS enables both non-experts and seasoned protein designers to rapidly evaluate their designs, providing a framework for making reproducible, data-driven decisions about which design to take forward for experimental characterisation. While some protocols and applications have been developed to address some of the same challenges as DE-STRESS (Bernhofer et al., 2021; Guffy et al., 2018; Yallapragada et al., 2020), none of them have the same breadth of metrics and tools, all packaged in a user-friendly web application.

It is our aim that using DE-STRESS will reduce the failure rate of designs taken into the lab, thus increasing the efficiency of protein design, making it more accessible and reliable as a technique.

Acknowledgements

The authors would like to thank Lynne Regan and Dek Woolfson for feedback on the manuscript and application, the UoE School of Biological Sciences IT department for infrastructure support, and the developers of the software used to generate many of the metrics included in DE-STRESS.

Funding

CWW is supported by an Engineering and Physical Sciences Research Council Fellowship (EP/S003002/1). Michael Stam is supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. This work was supported by the Wellcome Trust-University of Edinburgh Institutional Strategic Support Fund (ISSF3).

References

Alford, R.F., Leaver-Fay, A., Jeliazkov, J.R., O'Meara, M.J., DiMaio, F.P., Park, H., Shapovalov, M.V., Renfrew, P.D., Mulligan, V.K., Kappel, K., et al. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J. Chem. Theory Comput. *13*, 3031–3048.

Ben-Sasson, A.J., Watson, J.L., Sheffler, W., Johnson, M.C., Bittleston, A., Somasundaram, L., Decarreau, J., Jiao, F., Chen, J., Mela, I., et al. (2021). Design of biologically active binary protein 2D materials. Nature *589*, 468–473.

Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. Nature Structural & Molecular Biology *10*, 980–980.

Bernhofer, M., Dallago, C., Karl, T., Satagopam, V., Heinzinger, M., Littmann, M., Olenyi, T., Qiu, J., Schütze, K., Yachdav, G., et al. (2021). PredictProtein – Predicting Protein Structure and Function for 29 Years. BioRxiv 2021.02.23.432527.

Deng, H., Jia, Y., and Zhang, Y. (2016). 3DRobot: automated generation of diverse and well-packed protein structure decoys. Bioinformatics *32*, 378–387.

Glasgow, A.A., Huang, Y.-M., Mandell, D.J., Thompson, M., Ritterson, R., Loshbaugh, A.L., Pellegrino, J., Krivacic, C., Pache, R.A., Barlow, K.A., et al. (2019). Computational design of a modular protein sense-response system. Science *366*, 1024–1028.

Guffy, S.L., Teets, F.D., Langlois, M.I., and Kuhlman, B. (2018). Protocols for Requirement-Driven Protein Design in the Rosetta Modeling Program. J. Chem. Inf. Model. *58*, 895–901.

Harrington, L., Fletcher, J.M., Heermann, T., Woolfson, D.N., and Schwille, P. (2021). De novo design of a reversible phosphorylation-dependent switch for membrane targeting. Nature Communications *12*, 1472.

Herud-Sikimić, O., Stiel, A.C., Kolb, M., Shanmugaratnam, S., Berendzen, K.W., Feldhaus, C., Höcker, B., and Jürgens, G. (2021). A biosensor for the direct visualization of auxin. Nature 1–5.

Hobohm, U., and Sander, C. (1994). Enlarged representative set of protein structures. Protein Science *3*, 522–524.

Huang, E.S., Subbiah, S., and Levitt, M. (1995). Recognizing native folds by the arrangement of hydrophobic and polar residues. J Mol Biol *252*, 709–720.

Huang, P.-S., Boyken, S.E., and Baker, D. (2016). The coming of age of de novo protein design. Nature *537*, 320–327.

Huang, X., Pearce, R., and Zhang, Y. (2020). EvoEF2: accurate and fast energy function for computational protein design. Bioinformatics *36*, 1135–1142.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577–2637.

Kuriata, A., Iglesias, V., Kurcinski, M., Ventura, S., and Kmiecik, S. (2019). Aggrescan3D standalone package for structure-based prediction of protein aggregation properties. Bioinformatics *35*, 3834–3835.

McIntosh-Smith, S., Wilson, T., Ibarra, A.Á., Crisp, J., and Sessions, R.B. (2012). Benchmarking Energy Efficiency, Power Costs and Carbon Emissions on Heterogeneous Systems. The Computer Journal *55*, 192–205.

McIntosh-Smith, S., Price, J., Sessions, R.B., and Ibarra, A.A. (2015). High performance in silico virtual drug screening on many-core processors. The International Journal of High Performance Computing Applications *29*, 119–134.

Pirro, F., Schmidt, N., Lincoff, J., Widel, Z.X., Polizzi, N.F., Liu, L., Therien, M.J., Grabe, M., Chino, M., Lombardi, A., et al. (2020). Allosteric cooperation in a de novo-designed two-domain protein. PNAS *117*, 33246–33253.

Radom, F., Plückthun, A., and Paci, E. (2018). Assessment of ab initio models of protein complexes by molecular dynamics. PLOS Computational Biology *14*, e1006182.

Rose, A.S., and Hildebrand, P.W. (2015). NGL Viewer: a web application for molecular visualization. Nucleic Acids Research *43*, W576–W579.

Rose, A.S., Bradley, A.R., Valasatava, Y., Duarte, J.M., Prlić, A., and Rose, P.W. (2016). Web-based molecular graphics for large complexes. In Proceedings of the 21st International Conference on Web3D Technology, (New York, NY, USA: Association for Computing Machinery), pp. 185–186.

Sesterhenn, F., Yang, C., Bonet, J., Cramer, J.T., Wen, X., Wang, Y., Chiang, C.-I., Abriata, L.A., Kucharska, I., Castoro, G., et al. (2020). De novo protein design enables the precise induction of RSV-neutralizing antibodies. Science *368*.

201 Touw, W.G., Baakman, C., Black, J., te Beek, T.A.H., Krieger, E., Joosten, R.P., and Vriend,
202 G. (2015). A series of PDB-related databanks for everyday needs. Nucleic Acids Research
203 *43*, D364–D368.

204 VanDrisse, C.M., Lipsh-Sokolik, R., Khersonsky, O., Fleishman, S.J., and Newman, D.K.
205 (2021). Computationally designed pyocyanin demethylase acts synergistically with
206 tobramycin to kill recalcitrant Pseudomonas aeruginosa biofilms. PNAS *118*.

207 Wang, G., and Dunbrack, R.L., Jr (2003). PISCES: a protein sequence culling server.
208 Bioinformatics *19*, 1589–1591.

209 Weiss, M.S. (2007). On the interrelationship between atomic displacement parameters
210 (ADPs) and coordinates in protein structures. Acta Crystallogr D Biol Crystallogr *63*, 1235–
211 1242.

212 Wood, C.W., Heal, J.W., Thomson, A.R., Bartlett, G.J., Ibarra, A.Á., Brady, R.L., Sessions,
213 R.B., and Woolfson, D.N. (2017). ISAMBARD: an open-source computational environment
214 for biomolecular analysis, modelling and design. Bioinformatics *33*, 3043–3050.

215 Yallapragada, V.V.B., Walker, S.P., Devoy, C., Buckley, S., Flores, Y., and Tangney, M.
216 (2020). Function2Form Bridge—Toward synthetic protein holistic performance prediction.
217 Proteins: Structure, Function, and Bioinformatics *88*, 462–475.

218 Yang, Y., and Zhou, Y. (2008). Ab initio folding of terminal segments with secondary
219 structures reveals the fine difference between two closely related all-atom statistical energy
220 functions. Protein Science *17*, 1212–1219.

221