

HABILITATION THESIS REVIEWER'S REPORT

Masaryk University

Applicant

Stanislav Mazurenko, PhD

Habilitation thesis

Advanced Data Modelling for Protein Engineering

Reviewer

Prof. Armita Nourmohammad

Reviewer's home unit, institution

Yale Centre for Systems and Engineering Immunology (CSEI),
Departments of Immunobiology, Biomedical Engineering, and Physics
Yale University

Review of the Habilitation Thesis of Dr. Stanislav Mazurenko

Dr. Stanislav Mazurenko works broadly in computational biology, with a focus on protein science. His expertise covers both physics-based modeling of proteins and machine learning approaches to protein engineering. He has a strong publication record in this area, has obtained research funding, and has supervised several graduate students. He also teaches at the Faculty of Science on topics at the interface of artificial intelligence, biology, chemistry, and bioengineering. These courses fit well with his research profile and prepare students for a field that increasingly requires training across these disciplines. Taken together, Dr. Mazurenko's research, supervision, and teaching record fully meet the expectations for habilitation.

Summary and assessment of the thesis

The thesis presents two complementary approaches to the study of proteins. The first consists of physics-guided, bottom-up models that are directly connected to experimental data. The second consists of top-down statistical inference and machine learning methods, which learn patterns from data and may implicitly capture aspects of the underlying physics.

Dr. Mazurenko discusses both approaches appropriately and gives a balanced view of their respective strengths. He argues that physics-guided models are particularly useful in data-poor regimes, where the number of measurements is too small to train flexible statistical models, and where prior physical knowledge constrains the space of plausible explanations. He also points out that the interpretation of many experimental measurements requires a physical model. Protein thermal denaturation is a clear example. Quantities of interest for protein unfolding, such as the free energy difference between the folded and unfolded states or the melting temperature, are not measured directly. They must be inferred, for example from the heat capacity profile recorded by differential scanning calorimetry, through a

thermodynamic model of the unfolding process. The choice of model, including assumptions about reversibility and the number of intermediate states, directly affects the inferred values.

In Chapter 2, Dr. Mazurenko presents the models and computational tools he developed for the analysis of protein thermal denaturation. The main contributions are (i) the development of the CalFitter web server for fitting data from different types of thermal denaturation experiments within a common framework, (ii) the development of mathematical models for reheating experiments, which make it possible to separate reversible and irreversible components of unfolding, and (iii) the incorporation of these models into CalFitter. Making these models available through a web server lowers the barrier for experimental groups to apply consistent and physically grounded analyses to their data. These tools are useful to the protein engineering and biophysics communities, and Dr. Mazurenko's contribution to their development is substantial.

The thesis then turns to top-down machine learning approaches. Dr. Mazurenko first provides the relevant background and then discusses his own contributions. He describes two supervised learning applications: classifying enzyme pockets as buried or surface pockets, and classifying mutations in human cancer-related proteins as oncogenic or benign. He then discusses how he made use of unlabeled data through unsupervised and self-supervised training, applied to three tasks: (i) identifying evolutionary trends in protein phylogenetic trees to propose promising mutations, (ii) using the geometry of the latent space of a neural network to propose ancestor-like protein sequences, and (iii) comparing the dynamics of intrinsically disordered peptides in the presence and absence of a potential drug candidate using CoVAMPNets. Disordered peptides are a difficult target for conventional structure-based analysis, and this last application illustrates how learned representations of dynamics can address systems that are hard to describe with a small number of predefined coordinates.

In all of these examples, Dr. Mazurenko benchmarks the methods carefully, provides an informative analysis of the underlying biology, and clearly states the limitations of the models. This attention to validation and to the limits of applicability is important in machine learning applications to protein science, where performance on standard benchmarks does not always translate into reliable predictions for new systems.

In summary, the work presented in the thesis is of high quality, and Dr. Mazurenko demonstrates clear expertise in the topics he covers. He combines a solid understanding of protein biology with strong competence in computational and machine learning methods for protein engineering. This combination is necessary for applying machine learning to protein science in a meaningful way, as it allows the right questions to be asked and the results to be interpreted in light of the biology. The thesis is well written, with a clear logical structure and well-reasoned arguments.

Reviewer's questions for the habilitation thesis defence (number of questions up to the reviewer)

I have one question. The two parts of the thesis lay out the strengths and limitations of physics-based and machine learning models for protein engineering, and they make the case

that the two approaches are complementary. I would like the candidate to discuss how the two could be combined to retain the advantages of both: the data efficiency and interpretability offered by physical models, and the ability of machine learning methods to extract structure from large and heterogeneous datasets. Specifically:

- a) What forms of integration does the candidate consider most promising, for example physical constraints built into the model architecture or loss function, physics-based quantities used as features or priors, or machine learning models used to correct or parameterize physical models?
- b) In which data regimes would such hybrid approaches be most advantageous compared to either approach alone?
- c) Could the two be combined in an iterative scheme, for instance where physics-based analysis of experimental data provides training labels for machine learning models, and the models in turn guide which experiments or variants to measure next?

Conclusion

The habilitation thesis entitled *Advanced Data Modelling for Protein Engineering* by Dr. Stanislav Mazurenko **fulfils** requirements expected of a habilitation thesis in the field of Protein Engineering.

Date: 4/10/2026

Signature: Armita Nourmohammad