M U N I

Annex No. 10 to the MU Directive on Habilitation Procedures and Professor Appointment Procedures

HABILITATION THESIS REVIEWER'S REPORT

Masaryk University	
Applicant	Stanislav Mazurenko
Habilitation thesis	Advanced Data Modelling for Protein Engineering
Reviewer	Prof. Fabrizio Pucci
Reviewer's home unit, institution	Computational Biology and Bioinformatics, Université Libre de Bruxelles, Bruxelles, Belgium

Dr. Stanislav Mazurenko is a computational biologist developing computational methods, annotation tools, databases, and predictors for protein modelling and engineering. In this dissertation, the candidate presents 11 scientific publications to which he contributed substantially. The papers cover four main topics: (i) computational modelling of protein thermal stability and calorimetry data, including CalFitter and CalFitter 2.0 algorithms (papers 1, 2, 4); (ii) databases and benchmarking tools for protein stability and solubility, including FireProtDB, SoluProtMutDB, and BenchStab (papers 3, 5, 8); (iii) computational analysis of mutations, biomolecular dynamics, and structural features, with applications in precision oncology, drug effects, and enzyme function (papers 6, 7, 9); and (iv) AI-based protein engineering and evolution, including generative models for enzyme design and prediction of protein sequence evolution (papers 10, 11).

In his work, the candidate presents a dual approach to protein engineering. On one side, he develops more biophysics-oriented approaches, while on the other side he uses machine-learning-based methods. Although the presented models are substantially different, not only in methodology but also in the type of data they use, they address complementary aspects of protein engineering. The dissertation is well presented and technically sound, while remaining accessible to a relatively broad scientific audience. The main concepts and motivations are clearly introduced, whereas the individual publications provide detailed methodological and technical information. The selected works also show a coherent scientific progression, from the development of analytical and database resources toward more recent applications of machine learning and generative approaches to protein engineering.

The work of Dr. Mazurenko is well cited and appreciated within the scientific community. I myself have also used and analysed data from SoluProtMutDB and FireProtDB, which represent important and useful resources for researchers working on protein engineering. These databases are valuable because they provide curated and structured experimental information that can be used for method development, benchmarking, and validation by the entire community.

In summary, I consider the scientific work of Dr. Mazurenko and the habilitation thesis summarising some of his most relevant contributions to the field to fully warrant promotion to the position of Associate Professor. His work demonstrates scientific independence and the ability to develop computational resources with practical value for the community. I am confident that he will continue to bring interesting ideas, useful computational methods, and new approaches to the protein engineering community.

Reviewer's questions for the habilitation thesis defence

Q1. One aspect that I consider particularly important in the application of AI and machine-learning methods is the issue of generalization, including the risks of bias and overfitting. Models can sometimes achieve very strong performance on the data used for training or internal validation, while their performance may decrease substantially when applied to truly independent datasets.
In the model presented to predict missense mutations in oncogenic proteins, the candidate reports very high predictive performance, including an AUC of approximately 0.97 and a predictive accuracy close to 0.99 for oncogenic/benign classification. These are impressive results, but given how challenging this prediction task is, they also naturally raise questions about how model generalization was assessed and how potential overfitting or dataset-specific biases were controlled.
Could you comment on how these issues were addressed in that study, why they were not discussed in greater detail, and, more generally, how you currently approach the problems of generalization, bias, and overfitting in your present line of research?

Q2. One of the approaches that is rapidly becoming state of the art for predicting the effects of genetic variants are protein language models which learn high-dimensional latent representations of protein sequences and can capture structural, functional, and evolutionary information without relying exclusively on task-specific features.
While the candidate makes use of deep-learning approaches, protein language models do not appear to be discussed in much detail in the thesis or in the candidate's broader body of work. I would therefore like to ask how the candidate views the role of protein language models in this field. In particular, what do you consider to be their main advantages and limitations compared with more conventional deep-learning strategies?

Q3. The combination of physics-based and AI-based methodologies represents a valuable direction in protein modelling and engineering. As well described by the candidate in the thesis, physics-based approaches offer interpretability and direct links to molecular quantities, while machine-learning methods enable the exploitation of large datasets and a broader exploration of sequence and structural space.
Could the candidate elaborate on how these two domains could be integrated more closely? In particular, do you see AI-based methods as a way to improve our description and understanding of physical phenomena, rather than being used only as purely data-driven predictive tools? For example, in CoVAMPnet, molecular dynamics simulations are first performed and the AI model is then trained to identify and compare metastable conformational states. Do you think that, in the future, AI-based generative methods could partially or even completely replace molecular dynamics by directly generating physically meaningful conformational ensembles? Given that static protein structure prediction is almost solved, how far are we from achieving a comparable level of accuracy in predicting conformational dynamics and ensembles? Testo

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: 06/09/2026 Signature: Fabrizio Pucci