UNIWERSYTET
MIKOŁAJA KOPERNIKA
W TORUNIU

Chair of Biophysics
Institute of Physics
Head
Prof. Dr hab. Wiesław Nowak

Review of PhD Thesis
„Navigating Protein Conformational Landscapes : AN AI and Molecular Dynamics Approach"
by Daniel Wiczew

If one thinks about medicine and have some understanding of science, immediately proteins come to his/her mind. Proteins are present in every cell and cells form tissues and then organs. The wellbeing of a human body depends on orchestrated action of proteins. Proteins are complex biopolymers with (usually) very fine tuned 3D architectures. The stability and dynamics of a protein depend on its free energy landscape (FES). We still use Ansfinsen hypothesis that the native conformational state is closely related to the absolute minimum in this FES. Often proteins change their conformation to perform physiological function, for example ion channels present in cell membranes open or close depending on a local electric field (voltage gated channels). Since proteins easily have over 1000 atoms, the description requires some 3000 or more coordinates. Description of geometry changes in such spaces is very difficult or next to impossible. Therefore, for many years in structural biology and biophysics one of the main research task is a question how to reduce such dimensionality in a rational way. What coordinates describe changes pivotal for physiologically relevant conformations? [redacted] presenting in conceivable way all physics important for a proper physiology of a protein (aka: a cell, an organ, a body = good health). Mr **Daniel Wiczew, MSc in his PhD dissertation aimed at this long standing problem: How to [redacted] for a protein in an automatic way? He took relatively modern approach based on [redacted] and computer simulation of proteins dynamics. Two main problems were addressed: (1)** [redacted]

Better methods for theoretical investigation of proteins may lead (in a long term) to better medicine. **In my opinion the topic of this Thesis is very modern, it is linked to important scientific questions and such task for this doctoral project, albeit very ambitious, is well justified.**

The thesis has been prepared as a join doctorate (cotutelle) in Wroclaw University of Technology, Lab of Biomedical Engineering (Poland) and Universite de Lorraine (C2MP), Laboratoire de Physique et Chimie Theoriques (France). It has a form of a 150 pages long classical dissertation. The editorial style is not quite typical for Polish standards (wide margins, numerous side notes, for of references) but changes are acceptable, and to some extent it facilitates reading of this advanced material. The Thesis has four parts: State of the art (55pp), Methods (38 pp), Results (28 pp), Discussion (6pp). Each part has 3 chapters, divided into subchapters and sections. Such detailed chopping of the content has

Chair of Biophysics
Faculty of Physics, Astronomy and Informatics
Institute of Physics
NICHOLAS COPERNICUS UNIVERSITY IN TORUŃ
ul. Grudziądzka 5, 87-100 Toruń, Polska
tel. +48 56 611 32 04, NIP: 879-017-72-91, REGON: 000001324
e-mail: wiesiek@umk.pl, internet: https://www.ifiz.umk.pl/instytut/struktura/kb/

some advantages (it is easy to see the whole structure and content of the work) but also disadvantages (covered so many diverse topics that sometimes material is too condensed and I felt a lack of flow in many places). The bibliography has 161 positions. Typesetting is of very high quality, selection of references is excellent. I liked also a very useful "Glossary" of acronyms used in this Thesis (9pp). Without that perhaps everybody including a specialist would be lost due to so many (>65) acronym-coded methods/concepts etc. used in the text! A standard list of figures and a list of tables are also provided. English usage is acceptable, I have noticed quite a number of mistakes, and several typos.

Now I will present a content all parts and my assessment of that material.

State of the art

The author starts from **discussion of limitations of MD sampling**. Indeed, myoglobin has over 3000 close lying energy minima as has been shown in the famous paper by Elber and Karplus already in 1987 (ScienceVol. 235, No. 4786 „Multiple Conformational States of Proteins: A Molecular Dynamics Analysis of Myoglobin"). In more complicated proteins important minima are separated by large free energy barriers. So, MD sampling may never visit certain regions of energy landscape and that is a serious limitation of the MD methodology. One should remember that ergodicity in MD it is only a hypothesis. Thus many "enhanced sampling" techniques have been proposed to ensure visits of MD trajectories in less accessible regions (such as LES or metadynamics). Such sampling is obviously biased (statistics is not correct), so numerous attempts were exercised to extract accurate kinetics from such biased data. Mr Wiczew discusses in a greater detail Grisanov reweighting method, focused on extraction of [redacted] Slow conformational changes are particularly important in physiology and medicine context, but they are very difficult to obtain from the classical MD simulations. Protein folding, for example, may take seconds, while the most advanced simulations for medically relevant systems reach currently only microsecond MD timescale. The next part is about [redacted] This is quite justified since one of such methods -[redacted] served as a reference to check results of the new method developed by D. Wiczew and co-authors.

The main [redacted] in this Thesis is [redacted] This algorithm is **based on so called** [redacted] **group**. A clue is a [redacted] in a protein structure: we do not like conformations [redacted] – [redacted] The framework proposed by Mr Wiczew [redacted] possible physiological transitions in a protein under study. This is quite good idea, and may

Chair of Biophysics
Faculty of Physics, Astronomy and Informatics
Institute of Physics
NICHOLAS COPERNICUS UNIVERSITY IN TORUŃ
ul. Grudziądzka 5, 87-100 Toruń, Polska
tel. +48 56 611 32 04, NIP: 879-017-72-91, REGON: 000001324
e-mail: wiesiek@umk.pl, internet: https://www.ifiz.umk.pl/instytut/struktura/kb/

have serious advantages (in terms of efficiency in learning) with respect to the standard method. A part of the framework called [redacted] what protein conformation will be used for further modeling. Sure enough, the most promising are those conformations which have the [redacted] Presentation of the new method is illustrated by a few diagrams. It is a pity that all fonts here (and in many, many other figures in this Thesis, eg. 6.2, etc.) are so small. It is a technical flaw of this, otherwise nicely edited, Thesis.

Chapter 4 [redacted] is a kind of a review of methods based on [redacted] This is a very comprehensive subject, therefore only the most relevant information could be presented in this "concentrated" chapter. Discussed are features spaces, types of neural networks, limitations of DL approach, and biases in DL. This material is not new, it is based on original papers, however it should be included to prepare readers to understand better methods developed by the PhD candidate, for example [redacted] However, I have found also some new material: an interesting analysis of [redacted] with respect to standard challenges of [redacted] **In my opinion this chapter is a proof of very good knowledge of current literature related to DL and classical reaction coordinates studies.** The topic is highly mathematical but the quality of presentation in this chapter is not good. I saw numerous English language mistakes, and reading was further obstructed by numerous side notes inserted in the text in a ridiculous way (may be italics, location at the bottom of a page or a different font might facilitate presentation of side notes and could add to the clarity). I appreciated clear diagrams that help to understand some methods better.

In the Chapter 5 we have **a good presentation of Voltage Gated Ion Channels**. The topic is again huge, and presented data are very basic, but in my opinion all important for modeling and further presentation data are provided here. Figure 5.2 nicely presents changes in Voltage Sensor Domain and helps to understand that in such channels numerous physiological states are expected. The author underlines how kinetic model may facilitate understanding physiology of such channels (page 48). Here we have also a neat presentation of idea of Markov State Models (MSM), how they are extracted from MD and what we can learn from MSM. At the end a short account on Kv1.2 channel is given. This channel is large, so in real simulations presented in this Thesis only a part was used in MD. Obviously it a bias, in my opinion (we can discuss that during public defense) data from modeling can't be attributed to the whole channel and conclusions/observations should be taken with "a grain of salt". But for testing of a new theoretical method this simplification is reasonable and justified. Formula (5.2) contains a concept of "electrical distance" taken from the literature. This is a strange concept to me as a physicist, especially I can hardly understand the formula $\delta(r) = (\frac{\partial}{\partial V})\,\varphi(r)$. Is charge Q just a scalar or position dependent function?

Chair of Biophysics
Faculty of Physics, Astronomy and Informatics
Institute of Physics
NICHOLAS COPERNICUS UNIVERSITY IN TORUŃ
ul. Grudziądzka 5, 87-100 Toruń, Polska
tel. +48 56 611 32 04, NIP: 879-017-72-91, REGON: 000001324
e-mail: wiesiek@umk.pl, internet: https://www.ifiz.umk.pl/instytut/struktura/kb/

Methods (38 pp)

This is the main part of the Thesis, it presents computational methods developed during the PhD project and to large extent contains new and original material. Presentation addresses many methodological aspects and the present reviewer had some difficulty to judge what material was an adaptation of existing methods and what part was an original idea and contribution of the candidate.

Chapter 6 presents [redacted] At the heart of learning and making [redacted] which makes [redacted] landscape. In short, the Reinforcement Learning, adopted here, indicates what [redacted] The [redacted] is presented in the Thesis rather clearly and somewhat explained in Fig. 6.2 (too small!) The neural network architecture is discussed too (in Fig. 6.4). Technically the [redacted] was implemented using popular language [redacted] and other packages. [redacted] were used to generate statistics, NymPy library was also employed as well as [redacted] packages to run MD simulations (on graphical cards). PyEmma tool was used in subsequent analyzes. Observing the effective implementation of such diverse packages I infer that Mr Wiaczew has very good technical, programming and computational skills, very crucial in this field of science. During development he had to make many important decision, including selection of metrics used to asses performance his new method. 2D log-his and Fraction of New Microstates were selected. To have data for comparison, "ground truth" conformational spaces were generated (via MD) for three systems: WLALLS pentapeptide, RfaH_CDT and Chignolin miniproteins. Simulation times were measured in hundreds of microseconds. In the simulation protocol I didn't find important information about possible usage of Period Boundary Conditions (in explicit solvent cases). Two older methods were used to extract meaningful diverse states from those "real" conformational spaces, namely: [redacted]

Understanding kinetics means that main conformational states of a protein must be defined and then some rate constants for transitions between them should be determined. **In this Thesis Markov State Model has been exploited ("no memory") to extract long living states and a concept of Mean First Passage Time was used to describe kinetics. This part presents very high quality of implementation** of those rather old concepts, for example, measures were taken to avoid overfitting in determination of MSM via tICA method. Here I have two questions:

Chair of Biophysics
Faculty of Physics, Astronomy and Informatics
Institute of Physics
NICHOLAS COPERNICUS UNIVERSITY IN TORUŃ
ul. Grudziądzka 5, 87-100 Toruń, Polska
tel. +48 56 611 32 04, NIP: 879-017-72-91, REGON: 000001324
e-mail: wiesiek@umk.pl, internet: https://www.ifiz.umk.pl/instytut/struktura/kb/

(1) [redacted]To what extent parameters proposed in the Thesis, tuned here for a few systems, are general?
(2) Generation of states seems to be very time-consuming and needs a lot of attention of a researcher. [redacted]

In Chapters 7 and 8 a model of real ion channel is [redacted] method. This is a test of performance and presentation of the utility of this [redacted] approach. The starting model was taken from the previous work of Delemotte et al. (2017). Selection of features (numerous distances between amino acids) is a logical choice for me. The usage of large 4 fs time step is not so obvious, unless fixing of high-frequency motions by SHAKE or RATTLE is involved, here a variant of SHAKE - SETTLE was used. Boosting phase is desired, but I wonder what was the effect of 200K heating on the whole structure (? any RMSD plots). We do want to explore new minima, but we do not want to destroy a channel, I presume. The selection of the force field and MD simulations setting was correct and corresponds to state-of-art standards.

In the Chapter 8 non-standard procedure for finding [redacted] Here [redacted] are employed. The architecture of this new variant of [redacted] icely presented in Fig. 8.3, together with discussion on advantages of the proposed model. In short, the scheme consists of few layers, learning is done through presenting pairs of configurations. After optimization the [redacted] predicts [redacted] The whole idea is very appealing and promising, however I was a little disappointed that this algorithm is so sensitive to details of selected features (Page 90). It seems that D. Wiczew optimized this set using a trial-and-error procedure. Is it possible to develop a reasonable method that wouldn't require human knowledge and time? Fine-tuning of features and **parameters presented in the Chapter 8 worked well, but it seems to constitute just a clever heuristics. The Chapter gives solid presentation on many aspects of kinetics, including estimation of kinetic rates from Mean First Passage Times. It covers a lot of complex material** and perhaps that was the reason that a number of typos in this chapter is particularly large (I can show them in my copy of the Thesis). The most positive value of this part is presentation of new computational approach for protein dynamics studies, seeking help in modern neural network techniques. I rank this part high.

Results (28 pp),

Chapter 9 starts with presentation of a very useful and clever study – [redacted] 2D simple analytical potential energy surface. It is shown that during set simulation time (1 microsecond) [redacted] performs better than the standard MD. Better sampling of critical regions is achieved by usage of [redacted] part, it means that the algorithm works as planned. A comparison of [redacted] and [redacted] gave also interesting observations: for a small pentapeptide all methods have similar performance, but for a bigger RfaH-CTD system [redacted] leads. This

Chair of Biophysics
Faculty of Physics, Astronomy and Informatics
Institute of Physics
NICHOLAS COPERNICUS UNIVERSITY IN TORUŃ
ul. Grudziądzka 5, 87-100 Toruń, Polska
tel. +48 56 611 32 04, NIP: 879-017-72-91, REGON: 000001324
e-mail: wiesiek@umk.pl, internet: https://www.ifiz.umk.pl/instytut/struktura/kb/

refers also to sampling efficiency (Fig. 9.4). Impressive outcome was achieved in studies of Chignolin folding/unfolding processes: CS gave kinetic data in time a factor of 10 shorter than methods reported earlier. The analysis performed for Chignolin is quite interesting, however in Fig. 9.9 way too small symbols were used for "red dots" – plots are barely understandable.

The most ambitious system was studied in Chapters 10 and 11 – the voltage sensor domain of Kv1.2 channel. The set of analyzed data seems to be huge: 4068 trajectories 100 ns each. I didn't notice information who generated those trajectories. Also one part of preprocessing protocol (page 112) seems to be tricky: how centering of the membrane in the XY plane was done – just a shift by a certain vector or rotations were involved? What about protein part coordinates? On page 113 a discussion of "gating charge" Q is presented. Frankly, I had difficulty to understand how this gating charge was calculated. The idea is based on (unpublished) work by L. Delamotte et al (2017, biorxiv..). Individual $\delta(z_i)$ parameters were estimated from position based formula, with only one spacial coordinate involved. Some two mysterious parameters are used in the formula presented on page 113. In the text we have discussion about histograms of Q gathered from MD trajectories, but in Fig. 10.1 free energy is presented. In older papers from K. Schulten group (Biophys J.) gating charges of the order 10-12e are discussed for voltage gated ion channels, here maximum Q value is 2….or 3.2 (in Fig. 10.2). I hope to discuss definition and calculations of Q during public defense of this Thesis. The picture presented in Fig. 10.3 shows several states (In 2D IC1 /[redacted]But information what color corresponds to particular states (E, Δ, Γ) is missing.

In the Chapter 11 detailed analysis of kinetics of Kv1.2 domain is continued. **A valuable result has been obtained by checking sensitivity of performance [redacted] selector of features.** Nice data show that within a reasonable rage of features [redacted]re almost identical (Fig. 11.2). Another useful result is determination of the valuable number of slow modes to be used in interpretation of data (3 or 6 modes are optimum here). Another very important and practical issue is selection of microstates (clusters) in Markov State Model. The same refers to studies of optimum lag time in[redacted]Adopted value is rather small (10 ns) but allows for test of Markov property (lack of memory) in a rather short (100 ns) dynamical process. Simple and limited Chapman –Kolmogorow test results indicate that the studied model was Markovian, indeed.

After all that have work the final analysis come: search for macrostates in voltage dependent domain: three macrostates were found (Fig. 11.9) and calculation of transition rates between them. This is something that biologist and experimentalist want the most. Data are presented in Tab. 11.1, we see that rates estimated as inverse of MFPT are in the range of 10^4 - 10^5 sec^{-1}. Calculated transitions are 1-2 orders of magnitude faster than in another published experimental work. In my opinion such discrepancy (given theoretical character of the Thesis and so many assumptions made) is acceptable.

Chair of Biophysics
Faculty of Physics, Astronomy and Informatics
Institute of Physics
NICHOLAS COPERNICUS UNIVERSITY IN TORUŃ
ul. Grudziądzka 5, 87-100 Toruń, Polska
tel. +48 56 611 32 04, NIP: 879-017-72-91, REGON: 000001324
e-mail: wiesiek@umk.pl, internet: https://www.ifiz.umk.pl/instytut/struktura/kb/

UNIWERSYTET
MIKOŁAJA KOPERNIKA
W TORUNIU

Chair of Biophysics
Institute of Physics
Head
Prof. Dr hab. Wiesław Nowak

Discussion (6pp).

Discussion of the results is a relatively short part of this Thesis. One should remember that many elements of discussion were scattered through previous Chapters. The author summarizes main achievements, some pitfalls (a lack of detection of A and B states in Kv1.2) and provides some suggestion for future developments and improvements. Presented material demonstrates that the PhD candidate is critical and see limitations of the method developed. I think that the last part should contain a concise section "Conclusions", they are not clearly delineated in this part.

General remarks:

(1) **In my opinion this Thesis is quite innovative: it is not just dull MD study of some systems, but offers new methodological advancement, in line with current trends of using AI to support science**. It required a great deal of computer work and programming and very good understanding of current trends in enhanced sampling and deep learning methods.

(2) I noticed quite many grammatical and typo errors, but this doesn't' t affect my high opinion about this Thesis.

(3) The Thesis **work is complete**, giving a background for further developments

(4) Mr Daniel Wiczew **has developed competences** in:
a) using classical and modern methods of computer modeling of proteins
b) developing new modeling methods and programming advanced software systems based on deep learning/neural networks
c) effective analysis of performance of complex computer codes
d) understanding of sampling problems and feature selections in multidimensional data
e) understanding structural dynamical aspects of exemplary ion channel important in medicine and drug design.

Conclusion:

Mr. Daniel Wiczew presented an excellent doctoral dissertation with a strong interdisciplinary character. He developed [redacted] with increasing complexity with existing ones. Tests were designed very well and the analysis was deep. The method works well and have a number of advantages. New data have been collected on a model of medically important voltage depended potassium

Chair of Biophysics
Faculty of Physics, Astronomy and Informatics
Institute of Physics
NICHOLAS COPERNICUS UNIVERSITY IN TORUŃ
ul. Grudziądzka 5, 87-100 Toruń, Polska
tel. +48 56 611 32 04, NIP: 879-017-72-91, REGON: 000001324
e-mail: wiesiek@umk.pl, internet: https://www.ifiz.umk.pl/instytut/struktura/kb/

UNIWERSYTET
MIKOŁAJA KOPERNIKA
W TORUNIU

Chair of Biophysics
Institute of Physics
Head
Prof. Dr hab. Wiesław Nowak

channel. The communication has been submitted to prestigious journal from Nature Group. I see potential of this method both for improvement and many applications.

The presented dissertation proves that the doctoral candidate has mastered the workshop of scientific research in computer science and biomedical engineering disciplines. Hopefully, the results will be published soon in international journals and (probably) presented at several major conferences.

I declare that the doctoral dissertation of Daniel Wiczew submitted to me for evaluation is an **original solution of a scientific problem** related to computer science and biomedical engineering as well as biochemistry and to less extent to medicine. The dissertation proves that the PhD candidate has general knowledge in the field of computer science, biomedicine and modeling of proteins and **is able to independently conduct scientific research** using computational biophysics methods.

For all the reasons presented above, I give my favorable opinion for the defense of the PhD thesis by Mr Daniel Wiczew at Universite de Lorraine (France) and/or Wroclaw University of Technology (Poland).

The reviewed work meets the statutory (i.e. the Act - Act of July 20, 2018 - Law on Higher Education and Science (Journal of Laws of 2018, item 1668, as amended) requirements for doctoral dissertations. **I apply for Mr D. Wiczew's admission to further stages of the procedure leading to obtaining a doctoral degree**.

Since the dissertation is innovative in many aspects [redacted] the results will be perhaps published in a specialist indexed journal distributed worldwide, the doctoral student mastered a number of advanced computational methods, and the way of presenting the results proves his deep expertise in the ML/AI methods used in MD simulations, I request the Scientific Councils of both Universities (UdL, PW) to award this doctorate with distinction, according to local rules.

W. Nowak

Wieslaw Nowak, prof. zw.

Torun, September 20, 2024

Chair of Biophysics
Faculty of Physics, Astronomy and Informatics
Institute of Physics
NICHOLAS COPERNICUS UNIVERSITY IN TORUŃ
ul. Grudziądzka 5, 87-100 Toruń, Polska
tel. +48 56 611 32 04, NIP: 879-017-72-91, REGON: 000001324
e-mail: wiesiek@umk.pl, internet: https://www.ifiz.umk.pl/instytut/struktura/kb/