Dynamics of Structures, Interactions
of Macromolecules in Biology

Instituts thématiques Inserm
Institut national de la santé et de la recherche médicale
BIGR
UMR-S1134

Professor Catherine ETCHEBEST
Group Leader

Report on the manuscript PhD thesis presented by Daniel WICZEW
"Navigating Protein Conformational Landscapes : An AI and Molecular Dynamics Simulation »

The document presented by M. Wiczew is written in English and details the research conducted as part of a "co-tutelle" thesis between Université de Lorraine and Wroclaw University of Science and Technology.
His work involved developing new methods to comprehensively describe the protein [redacted] leading to the calculation of accurate kinetic properties at a lower computational cost. These methodologies were tested on various systems, ranging from small peptides and proteins to the most ambitious case: the Voltage Sensor Domain of the Kv1.2 potassium channel.
The dissertation is structured into five classical sections: a short introduction titled "How to Read the Thesis," which primarily summarizes the thesis, followed by a "State of the Art" section and then the "Method" chapter. The final two chapters focus on results and a discussion, with a few perspectives. Each section is further divided into chapters.
The "State of the Art" section contains three chapters. The first one (Chapter 3) begins with a detailed examination of the limitations in current methods for efficiently exploring the [redacted] and calculating kinetic properties. Three main aspects are discussed: (i) the quality of the force field, which favors the exploration of some regions at the expense of others, (ii) the use of biased simulations, which require reweighting methods to accurately estimate kinetic features, and (iii) the use of [redacted] approaches for defining Markov State Models, which require assumptions (e.g., discretization and state definition) to be effective in kinetic calculations.
While this chapter is thorough, I did not see a discussion about the definition of the reaction coordinate (based on collective variables, CVs), which is a key question. It is described in the next sections (Chapter 4), but for consistency, I suggest adding a small, specific paragraph in this chapter. For example, in the case of the protein folding process, the fraction of native contacts (Q) is frequently used, but it is not necessarily the most appropriate reaction coordinate (see, for instance: doi: 10.1021/acs.jpcb.3c06678).
This chapter concludes with a presentation of the [redacted], a [redacted] by M. Wiczew, which is intended to better explore [redacted] The key features of the method are well described and illustrated. Its limitations are only briefly discussed at the chapter's end, so I question why it is introduced in this "Limitations" chapter, as it will be further detailed in the "Method Section." This minor issue could easily be resolved by moving this paragraph to a new chapter, perhaps titled [redacted]An Alternative Algorithm for Discovering [redacted]

Email : catherine.etchebest@inserm.fr DSIMB, UMR-S1134 INSERM, Université Paris Cité, Hopital Necker Mobile: +33(6) 86 70 56 61
Bâtiment Lavoisier, 149 Rue de Sèvres PARIS 75015 France

The next chapter (Chapter 4) introduces [redacted] methods that are now used to discover kinetics. By using these methods, M. Wiczew aims to reduce [redacted] in the Feature Engineering process that models phenomena, while proposing an integrated strategy for [redacted]
This chapter is very pedagogical and explains the key aspects of various algorithms. I greatly appreciate the excellent balance between illustrations and theoretical details. It demonstrates that M. Wiczew has genuinely developed expertise in this field. The sections on limitations and biases are highly informative, and the key role of data and its efficient use is well discussed. Both [redacted] learning ones (PCA, tICA, VAMP) for discovering reaction [redacted] are clearly detailed, with their limitations and advantages briefly compared. A semantic question arises: should we consider these last [redacted] learning methods as non-learning methods at all? This raises a small matter of debate! The final section of this chapter revisits the [redacted] providing a different perspective compared to its earlier mention, focusing on how it can address various challenges. It also revisits [redacted] It's unclear to me why these two types of methods are still mentioned here, but I am confident that M. Wiczew will clarify their relevance during the defense.
Chapter 5 describes the biological system explored by M. Wiczew—the Voltage-Gated Ion Channel—and the challenges of elucidating the kinetics of its [redacted] While this chapter includes specific data on the system, it also revisits the problem of [redacted] Although this reiteration emphasizes important points, it overlaps slightly with previous sections. Additionally, this chapter finally provides a detailed explanation of the Markov State Model methodology, which has been mentioned since Chapter 3. It may be more appropriate to move this section (or at least part of it) to an earlier chapter, as it applies more generally than just to the biological system in the chapter's title. Nevertheless, the presentation is very clear and comprehensive, so this suggestion is merely for better organization. The chapter concludes with a return to the challenges of kinetics discovery for the Kv1.2 channel.
The next three chapters describe the methodological strategies developed by M. Wiczew. Firstly, the [redacted] algorithm is detailed, with all choices clearly justified. This section is quite accessible, even for readers without deep expertise. Performance metrics are also provided. M. Wiczew compares various [redacted] strategies, including classical molecular dynamics simulations and [redacted] The test systems—2D potential environments and three small systems [redacted]—are described in detail. Information on the MD simulations is provided, but it would be helpful to include a structural representation, at least for [redacted] The results for [redacted] show that although most of the improved [redacted] methods outperformed classical molecular dynamics simulations, their performance was comparable (Fig 6.6). However, the efficiency of [redacted] is impressive compared to the other methods. This result raises several questions that will be discussed during the defense.
I have a few comments on Figures 6.7 and 6.8, which show PCA for [redacted] CTD; they could be better described and discussed (e.g., what variables were analyzed, what variance is explained by each PC?). Regarding the [redacted] system, the features used to describe the folded state are a set of distances computed from the X-ray structure (if I

understood correctly). It would be helpful to explain how the unfolded state is defined with this set of distances. I mean is there a threshold to define the unfolded state

In Chapter 7, the application of [redacted] to the VSD Kv1.2 is discussed. The chapter begins with a description of Feature Selection, a critical step that is, in this case, guided by accumulated knowledge of the system. The data production protocol includes some adjustments, such as using a [redacted] that involves performing [redacted] simulations [redacted]. During the [redacted], the integration timestep was reduced. Why was this necessary? This choice, which increases computational time, should be briefly justified. I also have a question about hydrogen mass repartition. This strategy allows for an increased timestep, but it has been shown to introduce artifacts in the calculation of protein-ligand recognition kinetics (Volume 122, Issue 5, 7 March 2023, Pages 802-816). I would like to hear M. Wiczew's opinion on this point.

Chapter 8 covers all stages of [redacted] for VSD Kv1.2. It starts with an explanation of the [redacted] procedure (a deep learning model), including its general principles, architecture, and strategy for avoiding overfitting. Key steps—network training, hyperparameter optimization, cross-validation, and model selection—are carefully explained. The work is conducted rigorously, and M. Wiczew honestly acknowledges the limitations of the strategy and the risks of artifacts that may occur at different stages, along with solutions to minimize these risks. I still have some questions to address during the defense, such as how the number of microstates is selected. If I understand correctly, the [redacted] approach helps in this choice, and K-means clustering is further used for classification. However, is there an objective measure to determine this number? (I may have missed it, but I am not sure I found it in the manuscript.) Could alternative clustering strategies avoid an *a priori* choice of the number of clusters (the “K” in K-means)? This is a recurring question when clustering is involved, but it doesn't detract from my excellent opinion of M. Wiczew's strategy.

Chapters 9 to 11 are dedicated to the results, and they are relatively short. The first chapter describes the application of [redacted] to the toy systems: [redacted] The method appears highly effective for [redacted] compared to alternative approaches. Is there an explanation for why the performance gap is larger for [redacted] I’m also not sure if the [redacted] procedure was used with [redacted] if so, could that explain the improvement? Overall, the results are intriguing, but the figures could be better explained to assist the reader. For example, Figure 9.1E would benefit from additional clarification. Similarly, I am not entirely clear on what each of the different colored curves in Figures 9.5 represents. As a minor point, Figure 9.9 is cited in the manuscript before Figure 9.8—was there a reason for this order? Thus, the results section suffers slightly from a lack of detail, in contrast to the highly detailed and well-presented Method section.

Chapter 10 addresses the results of [redacted] applied to the VSD Kv1.2 channel. Here too, the description of the results is somewhat brief, even though the findings are interesting. Since these results are compared to previous studies, I would suggest providing a more thorough introduction to those earlier findings. This could be done either in the Method section or in this chapter, but they need to be described in more detail. This is especially important because there are differences in the height of the energy barriers, for instance. What could be the cause of these discrepancies? M. Wiczew suggests an

explanation related to the protocol used by the other authors. However, [redacted] procedure also play a role? This point requires further discussion.
Chapter 11 focuses on the analysis of the kinetics itself and the role of the selected features. The results show that most sets of features perform equally well, as indicated in Figure 11.1. Figure 11.2 draws similar conclusions, but I am unsure about the differences in the analysis between Figures 11.1 and 11.2. Similarly, I would appreciate additional commentary on Figures 11.4 and 11.6, as I am uncertain whether I have interpreted them correctly. The final section describes the kinetic parameters obtained from the model derived from M. Wiczew's strategy (Table 11.1). Would it be possible to add standard deviations for the k_{from_to} constants in Table 11.1 for consistency? These values are compared with results obtained in previous studies, which were based on different approaches with other assumptions. These previous results should be explained in more detail to better understand and assess the relevance of the assumptions made in this study.
I may be mistaken, but from examining the original paper, it seems the kinetics data were obtained for mutants (R1R5(W) and R1K5(W)). In addition to the assumptions made by Tao et al. in designing the model, could the differences stem from the nature of the system itself? This point could be discussed further. Additionally, it would be useful to explore the importance of states B and A that are missed in the sampling. The reaction coordinate, i.e., the Q gating charge, is quite novel yet highly relevant. While the sensitivity of this measure might be discussed in the manuscript, it will certainly be addressed during the defense.
The final chapter of the thesis summarizes the results and introduces a discussion. It also proposes new directions to address unresolved problems, such as improving [redacted] considering better [redacted] A more general discussion, reflecting on the [redacted] the optimization step (which is crucial in most deep learning approaches), the quantity of data, and the delicate balance between features and data examples, among other topics, would be valuable. I would also appreciate hearing M. Wiczew's personal opinions on these points.
Overall, M. Wiczew's work is impressive and highly original. It has been conducted with great rigor and applied to a very important biological system. Although all the challenges in determining accurate kinetic models have not yet been solved, M. Wiczew offers new ways to tackle these questions. The document is well-written and includes a relevant bibliography. I particularly appreciate the glossary section, where not only are the acronyms explained, but additional information is also provided about each term. Regarding the form of the document, I have a few minor remarks: there are some typographical errors that need correction. The choice to use sidenotes is interesting but they should appear alongside the relevant paragraphs, as is the case with the references, rather than in the middle of the sections they comment on. At times, they interrupt the flow of sentences, making them difficult to follow. This is easy to fix. I would also suggest being consistent with the placement of references: sometimes they appear in the margins, which is very convenient, but other times they are in the bibliography section at the end of the document. While this is not a major issue, I must admit that if possible, I would prefer the first option (with references in the margins). Similarly, I would appreciate larger figures with higher resolution. But these are very minor points.
In conclusion, it is clear that M. Wiczew has produced an important and original piece of work. It has been submitted to a highly reputed journal, and the reviewers were quite

positive, although questions remain. Therefore, I believe that M. Wiczew should be authorized to defend his thesis in front of the jury appointed by the Doctoral School to obtain a PhD diploma of Université de Lorraine. I look forward to discussing the various points raised in this report at the defense.

Paris, September 30th 2024

Professor Catherine ETCHEBEST