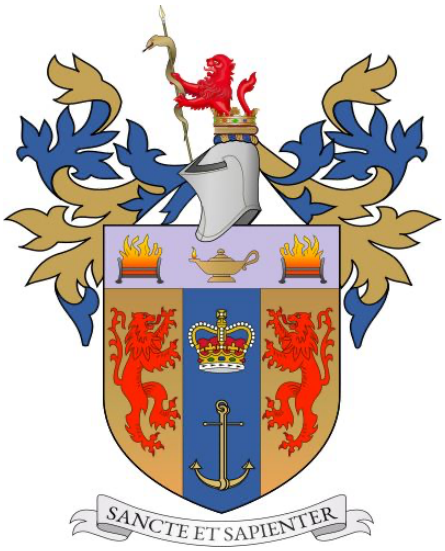
The role of missense mutations on the catalytic activity of enzymes: a study in pathological processes and for industrial applications

Mateeva, Teodora

Awarding institution:
King's College London

The role of missense mutations on the catalytic activity of enzymes: a study in pathological processes and for industrial applications



Teodora Mateeva

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
King's College London.

Department of Physics
King's College London

May 2024

Abstract

This thesis explores the relationships between enzyme mutations and their impact on catalytic function. This is considered from two angles: firstly, in cases where missense mutations lead to pathological processes in humans, and secondly, from a contrasting perspective where mutations confer benefits and are harnessed for the engineering and optimization of enzymes.

Through the integration of genomic and proteomic data, two enzymes emerged that are correlated with the toxicity of α-synuclein. The P5B-ATPase ATP13A2 and the phosphatase Synaptojanin-1 (Synj-1) were independently identified to be implicated in neurodegenerative diseases through various mutations.

In Chapter 3, I have modeled ATP13A2, focusing on elucidating details on the active site composition, conformation, and the role of specific amino acids in the catalytic reaction. This is needed to be able to quantitatively investigate the effect of mutations near the active site of the protein, during the different conformational states. I show the binding mode of the ATP substrate in the presence of one and two Mg^{2+} cations, in the E1 conformational state leading to E1P. The Molecular Dynamics simulations and QM/MM potential energy scans give strong evidence that ATP13A2 completes the autophosphorylation reaction with two Mg^{2+} ions in the active site. I show that without Arg686 the barrier height of the reaction is considerably higher while Lys859 is crucial for stabilizing the reactant state. Additionally, upon the analysis of the Molecular Dynamics trajectories, several binding pockets are identified, which is likely where the ATP13A2 cargo binds.

In Chapter 4, a method for the classification of enzyme variants is proposed, based on the predicted effect on the catalytic rate, coming from the mutations. This method is based on Molecular Dynamics simulations of the variants at/around the rate-limiting step and integration with Machine Learning algorithms. I look at variants that are similar to wild type Galactose Oxidase and variants with significant structural differences (> 10 mutations). Some of the variants are modeled with non-native substrates to create a model that can classify

variants that convert a diverse substrate range. This approach achieves excellent classification accuracy and high precision and recall with the current dataset.

In Chapter 5, structural exploration is conducted on the 5-phosphatase domain of Synaptojanin-1 (Synj-1). The 5-phosphatase domain is modeled embedded in a membrane, to gain insights into its substrate interaction. This modeling work can inform the design of inhibitors for disorders in which Synj-1 is overexpressed.

The thesis concludes by introducing a new method for calculating electron transfer rates. This method can be applied in the investigation of electron transfer in a biological context involving an enzyme mutation.

Overall, this thesis aims to contribute to a deeper understanding of the structural and functional implications of missense mutations in several specific cases, using traditional physics-based computational approaches and to also test the integration of these methods with Machine Learning, in the context of enzyme optimization, particularly when limited experimental data is available.

Publications

This thesis incorporates publications. The following manuscripts were published during my PhD, and I am either the first or one of the main authors:

Direct calculation of Electron Transfer Rates.
Z Benda, **T Mateeva***, E Rosta
The Journal of Physical Chemistry Letters, 10.1021/acs.jpclett.3c02624.

Structural dynamics and catalytic mechanism of ATP13A2 (PARK9) from simulations.
T Mateeva, M Klahn, E Rosta
The Journal of Physical Chemistry B 125 (43), 11835-11847

Combining data integration and molecular dynamics for target identification in α-synuclein aggregating neurodegenerative diseases: Structural insights on Synaptojanin-1 (Synj1).
K Jenkins, **T Mateeva**, I Szabó, et al.
Computational and structural biotechnology journal 18, 1032-1042

*Equal contributions.

Acknowledgments

First, I would like to thank Prof. Edina Rosta for giving me the opportunity to pursue a PhD. Next, I would like to thank my two other supervisors – Dr. Marco Klaehn and Dr. Hao Fan. Thank you, Marco, for explaining many concepts so patiently but also thank you for many wonderful chats about topics like history and travel which I greatly enjoyed. Many thanks are due to Dr. Hao Fan, who accepted me into his group and helped me integrate at the Bioinformatics Institute (BII) during a difficult period for me. I would also like to express my sincere thanks to Prof. Chris Lorenz who has kindly helped me with many administrative tasks.

Next, I would like to thank the many people in the Rosta group who have introduced me to various computational tools and have also been the source of many excellent discussions. I would always appreciate your kind help in the early days of my PhD. I will begin with Dr. Pedro Buigues who was the first PhD student I met at the Rosta group and deserves the first mention. Thank you for the warm welcome to the group! Of course, I would also like to thank Dr. Magd Badaui and Dr. Dénes Berta who were the first to help me learn about Molecular Dynamics and answered my endless questions. It must have been very tiring but for the most part, you didn't show it. Our political chats were always great fun at the end of the day, Dénes. I also want to thank you for taking the time to help me correct mistakes in writing and not only! I appreciate your help immensely. Special thanks are due to Dr. Tamás Földes who gave me advice and suggestions on more figures than I can count. Thanks are due to all other Rosta group members, past and present, with an honorary mention to Dr. Zsuzsana Koczor-Benda who has worked alongside me on a project I found quite challenging. Thanks for being so helpful in making this project move forward. And of course, thanks go out to Sam Martino who started his PhD at the same time as me and due to this circumstance has been unfortunate to have had the most chats with me. Thank you for the last-minute machine learning chats, Sam! Thanks, are also due to some of the people working in Hao's group, mainly Shreyas who provided constructive feedback for my ML models and Achal who gave me helpful hands-on advice on coding.

Thanks are due to my good friend Yoanna, who has listened to me complain about computations and coding more than a healthy amount and has always provided funny commentary and cheered me up. Thanks to Helen, my good friend on my Singapore journey who was with me during some rocky months and always provided a level-headed perspective. To my oldest friend in the world Janny, thanks for sticking around over 20 odd years. To the Mateevi family, who have experienced my good but also my not-so-good sides.

I dedicate this thesis to my two grandmothers – Dora (the first Dr. Mateeva) and Ivanka.

"…Nearly everything is really interesting if you go into it deeply enough."

– Richard Feynman

Contents

List of Abbreviations

Chapter 1

INTRODUCTION

1.1 Motivation

Enzymes are essential for life in all six kingdoms – Bacteria, Archaea, Protista, Fungi, Plantae, and Animalia. Whether we are referring to the smallest nitrogen-fixing cyanobacteria, the fungi which cause pathological processes in plants, or the entirety of *Homo sapiens*, all prokaryotic and eukaryotic species rely for their survival on the precise functioning and coordination of various enzymes.[1,2] In humans, the loss of catalytic function, caused by missense mutations, often results in a range of pathological processes.[3–6] Even in cases when the catalytic function of an enzyme is not lost, mutations can lead to a range of disorders by other mechanisms which will be discussed at length in this thesis.[7,8] The understanding of the exact mode in which a missense mutation impacts the catalytic activity is detrimental to developing and assigning the right therapy. At the same time, the field of enzyme engineering has harnessed the power of introducing mutations to create improved biocatalysts for biotechnology, biomedicine, and life sciences, capable of catalytic activity and substrate selectivity out of reach for native enzymes.[9–13] These days enzymes are routinely engineered to improve other properties as well, such as enantioselectivity, expressibility, solubility, and thermal stability.[14]

In this thesis, I aim to gain insight into the relationship between structural changes in the three-dimensional structure of proteins, caused by missense mutations, and the respective effect the structural change has on the catalytic function. In the 3rd chapter of this thesis, I explore the ATP13A2 enzyme, in which missense mutations are known to cause a range of neurodegenerative diseases,[1,15–22] despite the lack of clarity on how some of these mutations are implicated in the development of pathology. I use a combination of Molecular Dynamics (MD) simulations, Density Functional Theory (DFT), and Quantum Mechanics/Molecular Mechanics (QM/MM) calculations to investigate the catalytic mechanism of the wild type protein to elucidate details on the binding of ATP and Mg^{2+} in the active site. I then use these

findings to investigate how the substitution of amino acids in the active site affects the catalytic function of ATP13A2. This allows us to explain how some mutations may adversely affect the catalytic function of the enzyme. In the 4th chapter, I study the effect of missense mutations on Galactose Oxidase (GO) where beneficial mutations are utilized to achieve enhanced substrate selectivity and improved activity. I use a combination of MD and Machine Learning (ML) to predict the effect of a combination of mutations, in combination with non-native substrates, on the catalytic rate of the GO enzyme. The 5th chapter focuses on the integration of genomic and proteomic data to find proteins that are correlated with the toxicity of α-synuclein. The protein Synaptojanin-1 (Synj-1) was identified and the first fully atomistic model of the 5-phosphatase domain in a membrane-embedded setting was provided. Its binding to phosphatidylinositol-4,5-bisphosphate (PIP_2), an important lipid in membrane trafficking, was also probed in detail. In the final chapter of this thesis, I present a new method for the calculation of the rate of electron transfer. This method can be used when studying the catalytic mechanism of enzymes and might be of interest in cases where the transfer of an electron forms part of the catalytic mechanism. There are many examples of this, including in pathological processes. One of these examples is NADH Dehydrogenase (Complex I), which is a part of the mitochondrial electron transport chain. Pathological mutations of this enzyme and other enzymes that take part in the mitochondrial-encoded Electron Transport Chain likely disrupt the rate of electron transfer, leading to energy deficiency and mitochondrial dysfunction, contributing to several cancers.[23]

1.2 Missense mutations and their implication in pathological processes

Mutations that happen because a single nucleotide substitution has occurred, and the amino acid-encoding codon has changed, are defined as “missense point mutations”. The result is that one amino acid in the enzyme sequence gets swapped for a different one (Figure 1.1). Mutations can occur naturally during cell division and as a result of extrinsic factors in which case the mutation is not inherited and is defined as a somatic mutation.[24] Some mutations do not occur randomly during the cell division but are rather passed through the progeny and are defined as germline mutations. Germline mutations can be neutral and not cause any pathogenic effects. However, in the cases when they do, such as in many cancers, the

contribution of germline mutations towards the progression and susceptibility of the respective cancer, has been quantified and is subject to ongoing research, such as in the infamous BRCA1 and 2 genes.[25–27] Variants frequently occurring in a population are termed polymorphisms and single nucleotide polymorphisms (SNPs) are common genetic variations among populations.[24] It is important to note that the pattern of inheritance is detrimental – dominant missense mutations are such that their presence in one allele is sufficient to cause a phenotypic expression of the disorder. Recessive missense mutations, on the other hand, require the mutation to be present in all alleles of the gene for the disorder to manifest. In this work, the pattern of inheritance is not discussed as the main point of interest is how a missense mutation affects the catalytic mechanism of the protein in situations when the phenotype resulting from the mutation is already present. In this thesis, the term which will be used for an amino acid swap in the protein sequence, resulting from a missense mutation, is going to be referred to as a protein mutation. This is usually denoted in the literature with p. before the mutation, for example, p.Thr512Ile would mean that at position 512 of the protein sequence, a threonine is swapped for isoleucine. This clarification needs to be made as some literature sources refer to amino acid swaps as “replacements” while others refer to the term “mutations”. Additionally, the terms “protein” and “enzyme” are used interchangeably in this thesis as all proteins I have studied are enzymes, but it should be clarified that there are cases this doesn’t hold.

The ways in which mutated proteins cause pathological processes can be divided into a few categories. The first category that I am going to discuss is the one of protein mutations that affect the protein’s thermodynamic stability and folding.[28,29] Thermodynamic stability is defined as the difference in folding free energy between the native and the denaturated state (ΔG_f)[30] and it can be quantified to calculate the difference in stability between a wild type enzyme and a mutated variant. Mutations that impact enzyme stability frequently result in accelerated degradation of the enzyme, causing a change in the enzyme's concentration at the steady state. For example, specific mutations occurring in the dystrophin protein, which is mainly found in muscle cells, lead to misfolding which reduces the presence of properly functioning dystrophin, ultimately giving rise to muscular dystrophy. Some of the mutations identified in patients with muscular dystrophy have been observed to play a pathological role

by causing the protein to not fold correctly in the N-terminal actin-binding domain causing dystrophin to aggregate in a cross-beta structure similar to that found in amyloid diseases.[31]

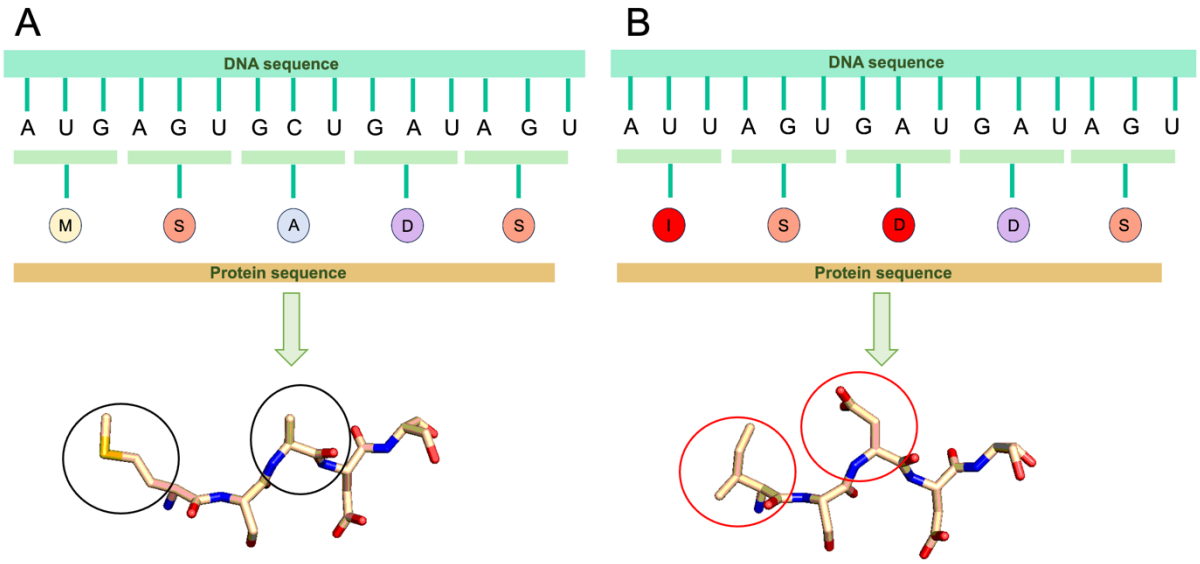


Figure 1.1. (A) A DNA sequence of nucleotides and the wild type protein sequence resulting from the respective DNA sequence. The three-dimensional structure resulting from the protein sequence is shown below. (B) A single nucleotide gets exchanged for a different nucleotide which in some cases causes an amino acid swap in the protein sequence. The three-dimensional representation below shows how this affects the protein structurally.

Similarly, it has been shown that protein mutations in the human mismatch protein 2 (MSH2) give rise to folding defects and subsequent proteasome-dependent speeded degradation. Since MSH2 is responsible for recognizing and binding to DNA mismatches occurring when the DNA strands are not correctly base-paired during replication, the increased degradation of the protein presents itself pathologically as Lynch syndrome, an inherited disorder that increases the risk of many types of cancer, in particular colon cancer.[32] Protein mutations affecting the structural integrity and folding can also contribute to altered protein-protein interactions of the mutated variant or affect the interaction with other signaling biomolecules or lipids.[33,34] This is the case for mutations in the CBS (cystathionine β-synthase) protein identified in patients with homocystinuria, a disorder that affects metabolism. Missense mutations in CBS change the structural and energetic features of the C-terminal regulatory

domain, such that it can no longer undergo conformational changes in response to S-adenosylmethionine, leaving it in a constantly open conformation.[33] These examples from pathological processes demonstrate the importance of being able to quantify and evaluate the thermodynamic stability of different mutated variants. Computationally, several methods exist for evaluating thermodynamic stability such as FoldX,[35] Rosetta-ddG,[36] and many others. FoldX employs an empirical force field and is designed for the prediction of stability upon a few mutations. Since it is not very computationally intensive to use, it can be a useful supplement to the design of stabilizing mutations.[35] Rosetta-ddG, while more computationally intensive, aims to predict the change in free energy upon mutation. Rosetta-ddG's advantage over a lot of other methods is that it employs sampling techniques to explore the conformational space. Its disadvantage is that it is slower and takes a longer time to evaluate many structures. The change in ΔG (or ΔΔG), when a point mutation is introduced, is a good indication of whether the mutations will be unfavorable in terms of protein stability. There are also many Deep Learning (DL) algorithms developed that predict thermodynamic stability based on protein sequences and even changes in just a few amino acids.[37–39]

The second category or mechanism by which missense mutations contribute to the presentation and progression of pathological processes is through affected expression and localization of the protein within the cell.[8,40–42] The effect arising from these types of mutations can be quite difficult to predict and evaluate with traditional computational methods due to the variability of missense mutations causing mislocalization and the immense conformational space that needs to be explored. Recently, DL algorithms such as Bidirectional Long Short-Term Memory Networks (LSTMs), which are used for processing sequential data such as protein sequences, have achieved great progress in predicting the localization of proteins from purely sequence information.[43] LSTMs have been successfully used to predict the site of expression for a range of protein families and have shown great promise in predicting the localization of the protein upon a few amino acid changes.[43–45]

One of the most common ways mutated proteins contribute to disease progression, however, is through the disruption of the protein's catalytic function.[3,46,47] In this thesis, these types of mutations are of more interest and will be subject to a more thorough discussion.

1.2.1 Subcategories of protein mutations disrupting the catalytic function of enzymes

Protein mutations that disrupt the catalytic function can be further divided into subcategories. The first subcategory consists of protein mutations that have an immediate role in the catalytic mechanism of the respective protein. For example, this could be situations in which an amino acid directly coordinates the ion cofactor in the active site or forms contacts with the substrate.[1,15,48] This subcategory also includes the cases where an amino acid directly participating in the catalytic mechanism gets mutated. This could be, for example, in situations when the amino acid performing the nucleophilic attack on the substrate, gets swapped for a different amino acid that can no longer serve as a nucleophile.[1,49] Another example for this subcategory is from cases when one of the residues in the active site which is involved in proton transfer gets mutated.[50] In some of these examples, the loss of the catalytic function cannot be rescued, which results in the severity of the disease being very pronounced.[1]

The second subcategory consists of mutations close to the active site that do not take part in the catalytic mechanism directly but either interact with other active site residues that are involved directly with the catalytic mechanism, change the active site conformation geometrically,[51] and/or affect substrate binding.[52] Most notably, when the electrostatic potential in the active site is different, as in a situation when a negatively or positively charged amino acid gets mutated to a neutral one, this can decrease the overall affinity of the mutated variant for the active site substrate.[53] The difference from subcategory I is that here the mutated amino acid is not needed to create a direct interaction with the substrate such as a stabilizing hydrogen bond but rather changes the overall affinity of the active site, i.e. the K_m constant is different. This is the case for a lot of ATPases where positively charged amino acids like Lys and/or Arg are needed for efficient ATP binding.[53]

A third subcategory can be considered which constitutes all mutations that are not spatially in the immediate active site and/or mutations that affect the binding of substrates in domains far from the active site[34] which has allosteric implications affecting the catalytic mechanism. Without changing the overall topology of the protein significantly, an allosteric signal can

transmit the effect of a perturbation to a different site in the protein structure.[54,55] This category also includes mutations that affect the flexibility of certain loops, which in turn can affect binding affinity to other important biomolecules, and again affect the catalytic mechanism indirectly.[56] This information is summarized in Table 1.1.

Table 1.1 Subcategories of mutations that cause loss-of-catalytic function or affect the catalytic rate in enzymes. Examples from particular cases involved in disease are shown in the final row of the table.

Subcategory **I**	Subcategory **II**	Subcategory **III**
• mutated residue coordinates the central metal ion. • mutated residue coordinates the substrate – ATP, GTP, etc. • mutated residue is a nucleophilic base or proton/electron acceptor.	• mutated residue coordinates another amino acid which takes part in the catalytic mechanism. • mutated residue is in the immediate active site and interferes sterically with the catalytic mechanism. • mutated residue affects the charge distribution of the active site.	• mutated residue is not in the immediate active site but affects substrate binding in the active site allosterically/affects the ability of the active site to bring together the cofactor.
• H1069Q in ATP7B,[48] Wilson disease.	• G12C, G12D in RAS,[3] present in many cancers.	• V94M in UDP-galactose, 4-epimerase in type III galactosemia.[56]

To illustrate the discussed subcategories, three examples from human disease where mutations affect the catalytic rate of the enzyme, are shown in Fig. 1.2.

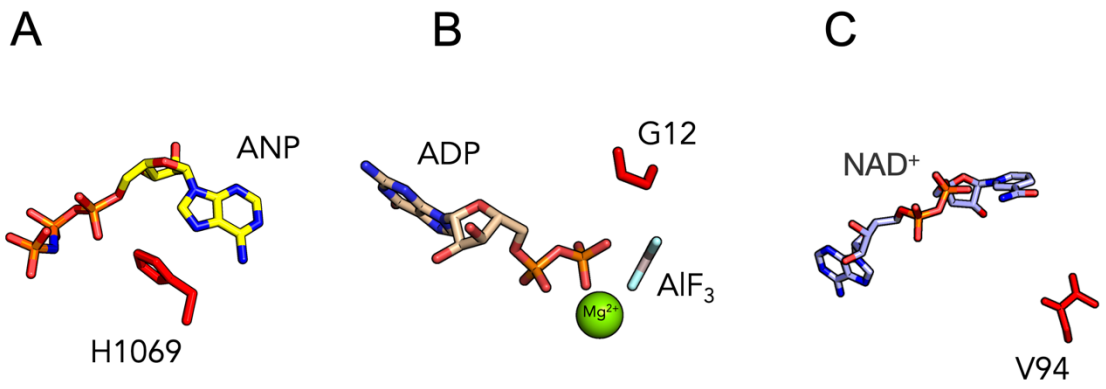


Figure 1.2. (A) An example of subcategory I where the amino acid that gets mutated (H1069) is directly involved in the substrate binding by forming a hydrogen bond to the β-phosphate of the substrate. (B) Example of subcategory II where the amino acid that gets mutated (G12) is not coordinated to the substrate but is in the immediate active site – pathological mutations of this residue are known to alter the charge distribution of the active site and introduce bigger amino acids that interfere sterically. (C) In this example for subcategory III the amino acid that gets mutated (V94) is not located in the immediate active site (it does not coordinate any of the catalytic residues or cofactors) and it does not participate in the catalytic mechanism directly but has an effect on the catalytic rate. Crystal structures used to illustrate the subcategories are the following: 8IOY[57] for the ATP7B protein, 1WQ1 for HRAS,[58] and 1EK5[59] for the human UDP-galactose.

In the traditional computational approaches, to probe how a protein mutation affects the catalytic rate, one needs to have a detailed description of the mechanism in the wild type enzyme. Once the catalytic mechanism is known/established, usually a free energy profile is obtained for the rate-limiting step. The free energy profile is then obtained for the mutated variant.[3,60] By comparing differences in properties like Gibbs free energy of activation ($\Delta G^{\ddagger}$), it is possible to observe the effect on the catalytic rate coming from the mutation – if the rate is slower, unchanged, or faster. Depending on how much the $\Delta G^{\ddagger}$ is affected, it is possible to rule out whether there is complete or partial loss-of-catalytic function.

It is also important to introduce the concept of second-site compensatory mutations. These types of mutations occur either very close or in a distant location from the original residue and can alleviate the negative effects of a primary mutation, remediating the fitness loss of the original mutation. Sometimes viruses use this mechanism to restore infectivity even when

drastic deleterious mutations at the capsid are present.[61] This type of mutation is not well-understood, despite the important implications arising from the phenomenon. Being able to predict this type of mutation is important in the field of enzyme engineering, where a mutation desired for one quality, such as enhanced substrate specificity, results in diminishing of a different quality, catalytic activity for example. This kind of problem is of interest in the work discussed in Chapter 4.

1.3 Missense mutations in enzyme optimization

1.3.1 Directed evolution

One of the strategies for protein engineering is directed evolution. For applications relevant to the pharmaceutical industry directed evolution is utilized to improve the substrate selectivity and catalytic activity of enzymes, as well as tuning enantio- and regioselectivity. The field has progressed considerably since its conceptualization in the 1960s to the point where enzymes are routinely engineered to be more active, regio- and enantioselective with non-native substrates. An example is the engineering of enzymes capable of biocatalytic oxidation. These types of enzymes are excellent choices for renewable oxidation which is not harmful to the environment and achieves excellent catalytic turnover without the need to use toxic or unsustainable inorganic oxidants.[9–13,62] A prominent example is that of Galactose oxidase (GO) where the wild type enzyme catalyzes only a narrow range of substrates (galactose and galactose-containing oligosaccharides) and does not oxidize secondary alcohols but has been successfully engineered to convert a wide range of secondary alcohols, including bulky benzylic alcohols, through the application of directed evolution.[13]

The most common techniques used experimentally to enable directed evolution are error-prone PCR (epPCR), DNA shuffling, and saturation mutagenesis (SM).[62] Error-prone PCR is a technique that introduces mutations in already existing protein sequences during the PCR amplification process. It is often applied when there is little information on the structure and function of an enzyme. It can be considered “random” and requires the screening of large protein combinatorial libraries. Error-prone PCR is also useful when there is a specific target gene in which diversity is to be introduced through random mutations.[62,63] DNA shuffling is a

recombination-based technique that tries to mimic the way natural evolution works to create beneficial enzymes by the recombination of existing useful genes. It involves the combination of DNA fragments from related sequences, such as from homologous genes. Saturation mutagenesis (SM) is a technique that substitutes a single codon or a set of codons with all possible amino acids at the codon position identified to be of interest. It is common for SM to introduce mutations at sites lining the enzyme binding pocket so binding affinity to different substrates can be evaluated with different amino acid substitutions. Iterative Saturation Mutagenesis (ISM) builds on SM as it factors in the "best" mutant in a library at a given site and this mutant is used as the template for SM-based randomization at another site.[64] For improving selectivity and activity, SM generally achieves better results over error-prone PCR and DNA shuffling. All of the techniques described here are performed in a few consecutive experimental steps: library creation with new variants, library expression, and library screening.[62,65] Some methods can generate libraries through solid-phase-based gene synthesis or by utilizing the CRISPR gene editing system.[66] Rational design is based on structural analysis and in-depth computational modeling of enzymes by accounting for the physicochemical properties of amino acids. Generally, directed evolution is often complemented by rational design.

One of the main setbacks of directed evolution is that very vast combinatorial space needs to be explored – even when mutations can be introduced in a small region of interest such as an enzyme active site, the combination of possibilities of amino acids is huge. For instance, the randomization of 4 amino acids in an active site to all possible amino acid combinations yields 160,000 enzyme variants that need to be screened. It needs to be pointed out that statistically very few missense mutations achieve improved catalytic properties. About 70% of missense mutations are estimated to be neutral, 30-50% deleterious, and less than ~1% cause improvement of the catalytic properties. This makes it very difficult to identify mutations that are beneficial catalytically from such a small structural change and even more difficult to predict the effect from a combination of several mutations. In the next section, I am going to discuss how Machine Learning (ML) algorithms can be harnessed to learn about the structure-function relationship in enzymes from the currently existing data. I am going to outline some approaches and pitfalls, also discussing the algorithms as a function of the training data available.

1.3.2 Machine Learning in protein engineering

Machine Learning (ML) has recently become a very popular tool in enzyme engineering, largely due to the advancements in the processing power of computers and the wide availability of powerful GPUs (Graphical Processing Units). The other main factor is the availability of training data, with more than 251 600 000 sequence entries publicly available on Uniprot,[67,68] and more than 162 000 three-dimensional protein structures deposited in the Protein Data Bank (PDB) as of December 2023.[69] AlphaFold's success in predicting the three-dimensional structure of proteins from just sequence information,[70,71] vastly fueled the upheaval of Deep Learning (DL) algorithms aiming to predict various properties from protein sequences alone.

The main advantage ML has over the experimental techniques mentioned is that once trained to have high accuracy, ML algorithms should ideally generalize well on unseen data and can make predictions about the effect of unseen mutations. As already mentioned, statistically less than ~1% of point mutations achieve improved catalytic properties. This makes the deployment of directed evolution and rational design a rather slow and cumbersome process that usually takes many months to identify a beneficial mutation. ML does not remove the need for directed evolution but can rather make use of already existing data. One of the drawbacks of many current ML models is that while the goal is to generalize well on unseen data, models are usually only successful when applied to similar proteins to the ones in the original training set. Most commonly applied ML models based on standard Convolutional Neural Networks (CNNs) generalize poorly on protein predictions for very distinct subfamilies from the ones used in the training set. Therefore, the success of the model and the resultant predictions rely on the size, type, and quality of the training data. It is not uncommon for supervised models to generate negative examples by random association which can be an issue when training a binary classifier. For many prediction tasks, the format of the biological data is available only from positive examples.[72] This is also relevant to my work in Chapter 4, where variants of Galactose Oxidase with non-native substrates do not maintain the same catalytic rate as the WT GO enzyme. This means that the model does not see a particular class of substrates and their interactions with the protein (non-native substrates which achieve faster rate of conversion than the WT substrate, for example). Transformer-based

unsupervised language models can overcome some of these issues but can also suffer from the availability of training data to sample the probability space properly. Due to the high dimensionality of many biological problems, relative to the large unseen biological diversity, the prediction task can become very challenging. An example can be given from the field of computational immunology. A current challenge for Deep Learning models is the task of predicting the immunogenicity of an antigen, concerning a particular T-cell receptor response. Several components need to be considered by the model – the large variation of antigens, the polymorphic nature of human MHC I molecules, and the diversity of T-cell receptor structures. Any DL model needs to encode the sequences of all four components but also relies on the availability of experimental data, which is very scarce compared to the number of all possible pMHC-TCR binding combinations (a conservative estimate predicts $>3.6 \times 10^{15}$).[72–74] This is an example of where CNN-based and traditional DL classification models are particularly ill-equipped to deal with the complexity of the biological data.

What type of ML algorithm is deployed depends on the problem at hand. There is no single model that is going to outperform the others in all cases – one needs to consider various aspects not limited to but including time limits, application, data size, and dimensionality of the dataset.

1.3.2.1 Ensemble algorithms

The first point of consideration is the size of the data – for a small number of data points and features, commonly utilized are “tree-based” methods which use an ensemble of decision trees, such as the Random Forest (RF) or Gradient Boosted Decision Trees (GBDT) algorithms. A Random Forest is quick to train, optimize, and evaluate and a good go-to option for experimental datasets of limited size, such as the one used in this work.

Decision trees usually work by recursively splitting the dataset into subsets based on the calculation of the Gini index, thus selecting the most informative features. The Gini index is used to evaluate the quality of a split at each node. Each split is chosen to maximize the separation between classes or minimize the variance of the target variable. At each node of the tree, a decision is made based on a specific feature and a threshold value (see Figure 1.3).

This process continues until a stopping criterion is met, such as a maximum depth or a minimum number of samples per leaf.[75,76]

RFs incorporate a user-defined or default number of decision trees and instead of outputting the result of a single decision tree, an RF uses the majority vote to improve the predictive accuracy and control over-fitting. Since an RF is built from a bootstrapped sample of the training data, and at each split, a random subset of features is considered this introduces randomness by ensuring that different data points go into each of the trees. The GBDT algorithm, while also an ensemble algorithm, works differently by aiming to create one strong learner from the previously weaker learning trees rather than taking the majority vote of the decision trees. The RF and GBDT algorithms are utilized and discussed again in Chapters 2 and 4.

A

variant	x_0	x_1	y
1	4.0	2.0	1
2	3.0	5.0	0
3	3.1	5.3	0
4	4.5	1.9	1
5	4.6	1.7	1

B

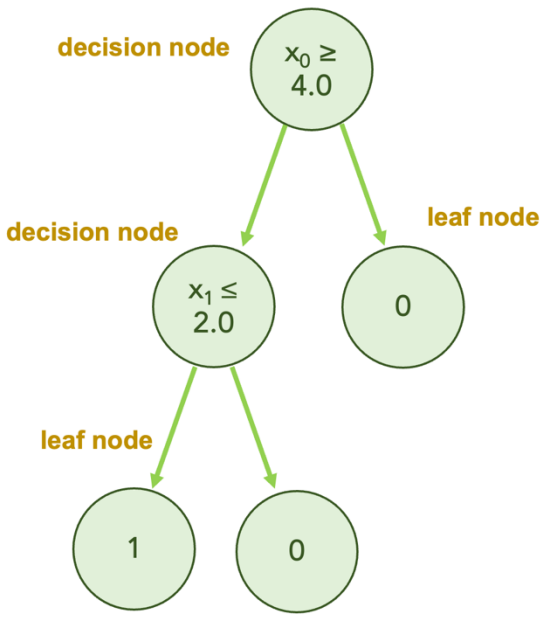


Figure 1.3. (A) Example dataset with 5 variants and two features x_0 and x_1. The target variable y takes binary values 0 and 1, which is an example of a classification task. The two features in this dataset are average distances from 3 replicas of a Molecular Dynamics simulation. (B) Example of a single decision tree. If a variant has a certain distance x_0 bigger than or equal to 4.0 Å, and x_1 distance is smaller or equal to 2 Å, the variant gets classified in class 1. If x_0 is smaller than 4.0 Å, it gets classified in class 0. This is a model example where the features are ideal, which is very rarely the case in real-life datasets,

hence the need for multiple decision trees. If a condition in a decision node is satisfied, the move is to the left and if not satisfied, the move is to the right.

1.3.2.2 Deep Learning approaches

Predicting the structure of a protein is only one of the many aspects of enzyme engineering. Predicting properties of the protein that arise from and depend on the dynamics of the folded protein and its interaction with other moieties, is a multi-dimensional and very complex problem.

To apply any DL model, one generally needs > thousand or at least several thousand data points, in the context of protein engineering, a thousand protein sequences or three-dimensional structures as input. Based on whether one is interested in properties such as binding affinity to a peptide or a ligand, information on and encoding of the peptide also needs to be available.[73,77] It is possible to apply DL models on less data but one has to be aware of overfitting. The second point of consideration is whether one is interested in predicting a discrete value such as whether a variant is active or inactive towards a certain substrate type, or continuous, such as by exactly what value a mutation would affect the activation free energy of a catalytic reaction or by how many degrees the thermostability of a protein will increase or decrease. Additionally, it is important to consider whether one is interested in predicting the effect of a single mutation or a combination of mutation positions. What the approach should be also depends on the format of the data available. An RF and other tree-based algorithms require all data points to have the same number of features, which is not the case for situations when one would be required to encode variant and peptide sequences as different variants would have a different number of amino acids or atoms, respectively. Where the sequential data is of varying length LSTMs have shown to provide useful solutions.[43,44] Most often, the chosen input representation must be adapted to proteins of variable lengths and be able to encode the relational information of the protein structure.[77] Ideally, the protein structure representation should account for the properties to make training efficient.

It is difficult to decide what the best protein encoding approach is. Generally, a protein sequence can be encoded in two ways – by its amino acid sequence or by the physical properties of the amino acids. Since amino acids have specific properties like hydrophobicity and charge, it is possible to encode an amino acid sequence as a combination of those properties. One of the common descriptors is the Identity descriptor which is a one-hot-encoding binary vector of the 20 natural amino acids. The zScales protein sequence descriptor, which uses physicochemical properties calculated from NMR and thin-layer chromatography (TLC) data, is represented by a five-dimensional vector descriptor for each amino acid. There are numerous ways to encode protein sequences and each might be better suited than another, based on what the purpose of the DL model is.[78] The BLOSUM62 matrix has also been used successfully to encode protein sequences for the prediction of segments in sequences, such as which part of the protein sequence belongs to a signaling peptide, etc.[79]

Currently, many DL algorithms exist that claim to be able to predict with high accuracy properties like thermal stability[80] and solubility,[81] as well as protein binding interfaces[82] and protein-protein interactions.[83] However, predicting catalytic activity, and more specifically how activity is influenced by the presence of one or more missense mutations, is very challenging. The mapping from sequence to function is tremendously complex because it involves thousands of molecular interactions that are coupled over multiple lengths and timescales. To the best of my knowledge, there is currently no existing DL architecture that can accurately predict the rate of catalysis in mutated proteins, across multiple protein subfamilies, relative to the WT enzyme. As already discussed, due to the size of the dataset used in my work, DL models were not utilized in Chapter 4. Therefore, a more in-depth discussion of possible model architectures will not be presented. This thesis utilizes traditional computational approaches to study the effect of missense mutations and integrates those with tree-based ML algorithms.

Chapter 2

METHODS

2.1 Molecular Mechanics

Molecular Mechanics (MM) methods, also known as force field methods, are applied for systems with many atoms, for example, a protein solvated in water. Since electrons are vastly lighter compared to nuclei, they move ultrafast, and it is assumed that the motions of electrons average out over the timescale of nuclear motion. Within the Born-Oppenheimer approximation and the framework of MM methods, the electronic motion is ignored. Electrons are not present in standard atomistic MM and most atoms are treated as point particles. Therefore, the microscopic state of the studied system is described as a function of only the position and momenta of the respective point particles. Needless to say, MM-based methods cannot calculate accurately properties that depend on the movement of electrons.

2.1.1 Force Field

A force field (FF) is essential for running classical MM Molecular Dynamics (MD) simulations as it provides the equations and parameters necessary to describe the potential energy of the modeled system. The potential energy of the system consists of the sum of all bonded and non-bonded interactions between the particles.

2.1.1.2 Bonded interactions

Bonded interactions encompass interactions between atoms that are connected by covalent bonds. These can be bond stretching, angle bending, dihedral or torsional interactions and improper torsions.

A harmonic potential with a force constant, k_{bond} , represents the stretching of covalent bonds, with the magnitude of k_{bond} representing the type and order of the bond. The

equilibrium bond length between the two particles is defined as r_0 . The less r_{ij} deviates from the equilibrium bond length r_0, the closer the potential will be to zero. The bond potential is described with the following term:

$$U_{bond} = \frac{1}{2}\sum_{ij} k_{bond}\left(r_{ij} - r_0\right)^2 \quad (2.1)$$

Angle bending is also described with a harmonic potential with the following term:

$$U_{angle} = \frac{1}{2}\sum_{ijk} k_{angle}\left(\theta_{ijk} - \theta_0\right)^2 \quad (2.2)$$

θ_0 is the equilibrium angle and the strength of k_{angle} depends on the atoms in the angle.

The harmonic Urey-Bradley potential is included in some force fields to account for the interdependence between bond stretching and angle bending. It defines an equilibrium distance u_0, between the 1,3 atoms in a bond angle.

$$U_{Urey\ Bradley} = \frac{1}{2}\sum_{ijk} k_u\left(u_{ik} - u_0\right)^2 \quad (2.3)$$

The cosine potential which describes dihedral angles of interconnected atoms, $ijkl$, can be expressed in a few ways but one of the most widely used expressions is with the following term:

$$U_{dihedrals} = \sum_{ijkl}\sum_{n=1}^{N} k_{\varphi,n}\left[1 + \cos\left(n\varphi_{ijkl} - \delta_n\right)\right] \quad (2.4)$$

Here $ijkl$ is a set of four connected atoms. Each triplet of atoms ijk or jkl defines a half-plane and the angle of intersection, φ_{ijkl}, of these half-planes is the dihedral angle. Bond dihedrals can be described by a sum of cosine potentials, with Nminima each at a phase-shift of δ_n, with a force constant $k_{\varphi,n}$.

The harmonic potential used to model improper angles between atoms $ijkl$ is described with the following term:

$$U_{impropers} = \sum_{ijkl} k_{\omega} \left(\omega_{ijkl} - \omega_0 \right)^2 \quad (2.5)$$

The sum of these terms makes up the bonded contributions of the FF which need to be calculated to obtain the bonded interactions which contribute to the potential energy of the system.

Values in the force field that are used for bonded interactions, such as force constants and equilibrium values, for example, are specific to the force field used. The work in this thesis is done utilizing the all-atom CHARMM36 force field[84] and in some cases, the Amber96[85] FF and generally uses the readily available force constants. Each particle in these FFs gets assigned an atom type, rather than simply using the element. This allows to differentiate the environment around particles of the same element. For example, a carbon atom bonded to oxygen will be assigned a different atom type compared to a carbon atom bonded to another carbon. In some cases, wild card parameters are defined. Usually, new organic molecules modeled need careful parametrization which is done by quantum mechanics-level methods such as DFT calculations and integration of experimental data, for instance, from NMR.

2.1.1.3 Non-bonded Interactions

The Leonard-Jones potential provides a way to calculate the force between two atoms continuously with their separation. This way of calculating the force between atoms was an advancement over the original hard sphere potential in which the force between two atoms was not calculated unless a collision between the said atoms occurred.

The 12-6 Lennard Jones potential is used to describe the van der Waals term of non-bonded interactions as it gives rise to both attractive interactions at the medium to long distances whilst still providing repulsive potential at short distances.

The Lennard-Jones potential is expressed as:

$$U_{LennardJones} = \sum_{nonbonded\ pairs} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \quad (2.6)$$

where the parameter ε_{ij} represents the strength of interaction and σ_{ij} the distance between particles i and j. Other functional forms also exist (e.g. 10-6 L-J, Buckingham,[86] Morse potentials).

The long-range electrostatic interactions can be described using the Coulomb electric potential with the following term:

$$U_{Coulomb} = \sum_{non-bonded\ pairs} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_{rel} r_{ij}} \quad (2.7)$$

Where q_i and q_j are the electric charge of the particles i and j; ε_0 is the permittivity of free space, ε_{rel} is the relative permittivity of the environment that the particles are in and r_{ij} is the distance between the two particles i and j.

2.2 Phase Space

To be able to discuss computational simulations of proteins, in particular Molecular Dynamics (MD) simulations, it is useful to introduce the concept of Phase Space. The concept comes from classical mechanics and is particularly useful in the study of dynamic systems, such as molecular systems undergoing thermal motion. For a system containing *N* atoms, 6*N* values are needed to define the state of the respective system in phase space (3 coordinates per atom and 3 components of the momentum). Since there are multiple dimensions corresponding to each particle's position and momentum, the overall phase space is high-dimensional.

To illustrate this with an example molecule, if one takes the simple hydrocarbon methane, which has 5 atoms, there would be 30 values needed to describe the state of methane in phase space - three coordinates (x, y, z) describe the position of each of the 5 atoms and three

momenta each define a point in the 6N-dimensional phase space. There are cases where this does not hold, such as if you have an isotropic potential when the system's energy will be invariant to rotations and translations, therefore, you have 6 fewer degrees of freedom.

Having introduced this concept, one can think of an MD trajectory as a sequence of points in phase space that are connected in time as each new configuration is calculated from the previous one before it.

2.3 Exploring the Potential Energy Surface

The potential energy of a system is a multi-dimensional function of the coordinates of the said system. In most general cases, the potential energy of an enzyme consisting of 2500 atoms will be a function of 7500 Cartesian coordinates. The relationship between the energy of a protein and its coordinates is usually explained through a potential energy surface (PES). Running MD simulations over 1 μs is not always possible due to the computational cost of running long simulations. Due to this, many biological events which happen over longer timescales, are not going to be observed (for example, protein unfolding). Generally, the probability of reaching a state of higher energy decreases exponentially according to the Boltzmann factor. For this reason, if we are interested in modeling any rare event, the standard approach is to introduce a biasing potential along a chosen reaction coordinate (RC). This allows to sample regions of the PES that would otherwise remain unexplored. How this biasing potential can be removed to obtain an unbiased surface will be discussed in section 2.4.4 which summarizes some of the most common unbiasing methods.

A minimum stationary position on the PES represents a structure of the modeled protein in which the net inter-atomic force on each atom is close to zero. A minimum on the PES doesn't necessarily mean that the net inter-atomic force on each atom is exactly zero. Mathematically, a "perfect" stationary point is one at which the first derivative of the potential energy with respect to each geometric parameter is zero. In practice, an energy minimization of the starting protein structure is performed before the start of any conformational sampling to minimize any forces resulting from residue clashes or poor three-

dimensional structural prediction. The user defines the number of steps for the minimization algorithm or the minimum force to be reached in the instruction file.

2.4 Molecular Dynamics

Molecular Dynamics simulations provide a way to 'observe' the dynamics of a system of interest, from which one can calculate various properties of the respective system. One can derive atomic positions in a time sequence by applying Newton's equations of motion. This way of observing the dynamics is deterministic because a new state is calculated from the previous state in which the system is found. A trajectory arising from the dynamics of the system is obtained by solving the differential equation coming from Newton's second law:

$$\frac{d^2x_i}{dt^2} = \frac{Fx_i}{m_i} \quad (2.8)$$

Here Fx_i is the force acting on a particle with mass m_i along one coordinate x_i.

2.4.1 Integration algorithms

When simulating an enzyme comprising of 1000 amino acids, each containing at least 10 atoms, the force acting on each of the >10 000 particles depends on the position of each individual particle with respect to the rest of the simulated particles. Under the influence of a continuous potential, the motions of all particles are coupled together, giving rise to a many-body problem that cannot be solved analytically. In this case, the equations of motion are integrated using *finite difference method*s. The idea behind finite difference methods is that the integration stage has to be broken down into small time steps separated by a fixed time dt, which is typically between 1 and 2 femtoseconds in standard protein MD simulations. This is done so that at each step, the total force acting on the individual particle is computed as a vector sum of its interactions with the other particles in the simulated system. The acceleration of the particles is then calculated from the force and combined with positions and velocities at time t to generate new positions and velocities at a short time ahead $t + 2t$ During the chosen time interval, the force is assumed to stay constant. The atoms then get

moved to new positions, and an updated set of forces acting on each atom is re-calculated in an iterative procedure until a user-defined time limit is reached. The time a simulation has been run can be calculated by multiplying the time step by the number of steps the user has originally pre-defined in the instruction file. The final output is a trajectory in which one can observe the dynamics of the system from a starting position, over the course of the selected time. The trajectory shows how the dynamic variables change with time.

The Verlet algorithm

Originally developed in 1967, The Verlet algorithm[87] laid the foundation for the most popular algorithms for integrating the equations of motion used in MD simulations. It uses the positions and accelerations at a time *t*, and the positions from a previous step, $\mathbf{r}(t-t)$ to calculate new positions at a time $(t+t)$, $\mathbf{r}(t+t)$. The relationship between the positions, and the velocities at a time *t* can be expressed with the following equations:

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t\mathbf{v}(t) + \frac{1}{2}\delta t^2\mathbf{a}(t) + \ldots \qquad (2.9)$$

$$\mathbf{r}(t-\delta t) = \mathbf{r}(t) - \delta t\mathbf{v}(t) + \frac{1}{2}\delta t^2\mathbf{a}(t) - \ldots \qquad (2.10)$$

Adding the two equations together results in:

$$\mathbf{r}(t+\delta t) = 2\mathbf{r}(t) - \mathbf{r}(t-\delta t) + \delta t^2\mathbf{a}(t) \qquad (2.11)$$

The Verlet algorithm[87] suffers from several shortcomings. One of those is that the velocities do not appear explicitly in the algorithm. This means that velocities are only available in the next step once the positions have been updated. This obviously can result in loss of precision, as well as the fact that the contribution of the kinetic energy to the total energy as a function of a specific position cannot be computed exactly. An additional shortcoming is that at the beginning of the MD simulation, there is only one set of starting positions for all simulated entities, however, the Verlet algorithm requires positions from a previous step also, which requires to employ some additional methods to calculate the positions at a time step $t - t$. In this sense, the Verlet algorithm is not self-sufficient.

The Leap-frog algorithm

An improvement upon the Verlet algorithm[87] is the leap-frog algorithm[88] which explicitly includes the velocity. The name of the algorithm comes from the fact that the velocities are updated before the positions, or "leap" before the positions. The positions are then calculated and 'leap' before the velocities as a result of the following relationships:

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t\,\mathbf{v}\left(t + \frac{1}{2}\delta t\right) \qquad (2.12)$$

$$\mathbf{v}\left(t + \frac{1}{2}\delta t\right) = \mathbf{v}\left(t - \frac{1}{2}\delta t\right) + \delta t\,\mathbf{a}(t) \qquad (2.13)$$

First, the velocities are calculated from the velocities at a time and the acceleration at a time *t*.

$$\mathbf{v}(t) = \frac{1}{2}\left[\mathbf{v}\left(t + \frac{1}{2}\delta t\right) + \mathbf{v}\left(t - \frac{1}{2}\delta t\right)\right] \qquad (2.14)$$

The leap-frog algorithm does not solve the problem existing originally with the Verlet – the positions and the velocities are not updated simultaneously. Therefore, it does not solve the issue with the kinetic energy mentioned earlier.

The velocity Verlet algorithm

The velocity Verlet algorithm[87] developed by Swope *et. al*. in 1982 allows for the positions, velocities, and accelerations to be calculated at the same time using the following relationships:

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \delta t\,\mathbf{v}(t) + \frac{1}{2}\delta t^2\mathbf{a}(t) \qquad (2.15)$$

$$\mathbf{v}(t+\delta t) = \mathbf{v}(t) + \frac{1}{2}\delta t\left[\mathbf{a}(t) + \mathbf{a}(t+\delta t)\right] \qquad (2.16)$$

To obtain the velocities in the final step, the algorithm is implemented as a three-stage process. This is because the acceleration is required both at a time step t and $t + \delta t$. First, the positions at a time step $t + \delta t$ are calculated according to eq. (2.16) using the velocities and accelerations at a time t. The velocities at a time $t + \frac{1}{2}$ are then determined using:

$$\mathbf{v}\left(t + \frac{1}{2}\delta t\right) = \mathbf{v}(t) + \frac{1}{2}\delta t\mathbf{a}(t) \qquad (2.17)$$

This allows us to calculate the acceleration $\mathbf{a}(t + \delta t)$ from the new forces at the current positions and finally in the third stage the velocities at a time $t + \delta t$ can be obtained with the following equation:

$$\mathbf{v}(t+\delta t) = \mathbf{v}(t) + \frac{1}{2}\delta t\mathbf{a}(t) \qquad (2.18)$$

2.4.2 Thermostats and barostats

When conducting MD simulations of biological systems, for example of a protein solvated in water, the idea is to model the environment in a biological cell as closely as possible. For this reason, standard protein simulations are usually run under an almost constant temperature and pressure. Thermostat and barostat algorithms are introduced in MD simulations so that NVT and NPT ensembles can be sampled correctly. The equilibration stage of an MD simulation is a crucial step that is done before the production run or the stage of collecting data and making observations. Equilibration involves gradually adjusting the initial configuration of the system, typically by applying forces to the atoms or molecules within the simulation, until the system reaches a stable state where its properties no longer significantly change over time. Usually, the equilibration step of an MD simulation is completed in two steps. First, under the NVT ensemble, or canonical ensemble, where the number of particles N, the volume V, and the temperature T are constant. This is usually a fairly short step that

aims to get the temperature to a certain user-defined constant value or heat up the system. The next stage of the equilibration is conducted in the NPT or isobaric-isothermal ensemble where the number of particles N, pressure P, and temperature T are constant, and the idea is that the pressure gets quickly equilibrated to reach a plateau, while the volume is allowed to change. Before the production run of the simulation starts, it is useful to plot the temperature and pressure as a function of time to make sure those have reached a plateau. During the equilibration stage when the system is heated up, the Berendsen algorithm[89] is usually applied, which calculates the temperature $T'(t)$ at every integration step using the following term:

$$T'(t) = \frac{\sum_{i=1}^{N} m_i v_i^2}{N_f k_B} \qquad (2.19)$$

Where k_B is Boltzmann's constant, v_i is the velocity of particle i, m_i is the mass of particle i N is the number of particles and N_f is the number of degrees of freedom for the N particles.

Once $T'(t)$ is calculated, the atomic velocities are linearly rescaled by a factor λ:

$$\lambda = \sqrt{1 + \frac{\Delta t(T - T'(t))}{\tau T'(t)}} \qquad (2.20)$$

Barostat algorithms allow to couple pressure baths to MD simulations. The way these algorithms work in general is by resizing the size of the simulation box to account for the applied pressure. While many barostat algorithms exist, normally some of these are more appropriate for the equilibration stage of the simulation. With the Berendsen barostat[90,91] the pressure is being quickly equilibrated from a starting pressure, while other barostats are more appropriate for the production stage of the MD simulation where one assumes that pressure is almost constant (Nose-Hoover[92,93] or Parinello-Rahman).[94]

The Parinello-Rahman[94] barostat is typically used during the production run of the MD simulation as it can produce the correct isothermal-isobaric NPT ensemble. To be implemented, an extra term is added to the equations of motion:

$$\mathbf{a}_i(t) = \frac{F_i(t)}{m_i} - \mathbf{M}\frac{d\mathbf{r}_i(t)}{dt}$$
$$\mathbf{M} = \mathbf{b}^{-1}\left[\mathbf{b}\frac{d\mathbf{b}'}{dt} + \frac{d\mathbf{b}}{dt}\mathbf{b}'\right]\mathbf{b}'^{-1} \qquad (2.21)$$

Where b is a matrix representing the box vectors and its equation of motion can be expressed with the following term:

$$\frac{d^2\mathbf{b}}{dt^2} = V\mathbf{W}^{-1}\mathbf{b}'^{-1}\left(\mathrm{P} - \mathrm{P}_0\right) \qquad (2.22)$$

$$\left(\mathbf{W}^{-1}\right)ij = \frac{4\pi^2\beta_{ij}}{3\tau_p{}^2 L} \qquad (2.23)$$

Where P is the instantaneous pressure, P_0 is the reference pressure, V is the volume of the simulation box, W^{-1} is the inverse mass parameter matrix determining the strength of the pressure coupling. Prime notations indicate the variables of the extended system. β_{ij} is the isothermal compressibility, τ_p is the pressure time constant, and L is the largest box matrix element.

2.4.3 Periodic Boundary Conditions

Periodic boundary conditions (PBCs) are introduced in MD simulations to model large systems by using a unit cell. For example, a unit cell with water molecules is used to approximate water environment. A space-filling simulation box is introduced, images of which are repeated in the directions of the unit cell vectors. This is done in order for molecules to be uniformly affected by long range interactions regardless of their position. When one object, such as a water molecule, reaches the boundary of the box, it exits from one part of the unit cell, and appears on the opposite side with the same velocity. PBCs are often used in tandem with the

Ewald summation, where the Coulomb term is divided into a short-range component, treated in the real space U_{real}, a long-distance component $U_{reciprocal}$, treated as reciprocal, and a correction term $U_{correction}$ for when the particle is seeing its own image:

$$U_{el} = U_{real} + U_{reciprocal} + U_{correction} \qquad (2.24)$$

The reason most MD algorithms use the Ewald summation is the computational cost in computing U_{el} . The Ewald summation method uses a Fast Fourier Transform (FFT) called Particle Mesh Ewald to compute $U_{reciprocal}$, which considerably speeds up the calculation of the Coulomb term as a whole. Having said that, before a simulation starts one needs to set the parameters for the Coulomb term such as a cutoff point for the short-range component, etc. Simulations need to be overall neutral, otherwise the net electrostatic charge of the system will sum to an infinitely large charge, because of the applied PBC. A common practice to "neutralize" the simulated system is to add neutralizing ions such as sodium and chloride in appropriate concentrations.

2.4.4 Unbiasing methods

To sample rare events, different techniques for enhanced sampling have been developed in recent years. The one utilized in my work makes use of a bias potential but there are also other enhanced sampling methods that are not limited to the use of a bias potential. Examples are temperature-accelerated MD (TAMD),[95] Parallel Tampering (Replica Exchange),[96] and others.[97] Most techniques for exploring rare biological events rely on the identification of a collective variable (CV), representing a physical pathway, that allows the calculation of the free energy profile. Choosing the CVs has to be done very carefully and, in some cases, can be quite challenging.[97] This is, for example, when one is interested in an unbinding event,[97] such as a drug molecule leaving an active site, as opposed to a well-defined chemical reaction.

Weighted Histogram Analysis Method (WHAM)

Some of the methods for obtaining a free energy profile involve the prior generation of a Potential of Mean Force (PMF) which is a biased free energy profile. The PMF represents the free energy landscape as a function of the sampled Reaction Coordinate/s (RC/s).

One of the most common and widely used methods to unbias Umbrella Sampling (US)-type MD simulations, which reconstructs the free energy profile along one or more chosen RCs, is the Weighted Histogram Analysis Method (WHAM).[98]

Umbrella Sampling is one of the techniques developed to overcome the sampling problem, that of higher energy configurations being difficult to visit in unbiased MD simulations. It aims to overcome limited sampling at these configurations by restraining the system with added bias (typically harmonic potential). A set of N_w separate umbrella simulations or windows are carried out, with an umbrella potential being expressed as:

$$w_i(\xi) = K_i/2\left(\xi - \xi_i^c\right)^2 \qquad (2.25)$$

The potential restrains the system at the position ξ_i^c with a force constant K_i. From each of the umbrella windows N_w (the number of those can vary depending on how the RC is split), an umbrella histogram is recorded, representing the probability distribution along the RC biased by the umbrella potential. WHAM is then used to compute the PMF from the histograms. The main idea is that if one knows the probability distribution of the configurations with the bias potential, the probability distributions for the unbiased cases can also be obtained.

The main equations behind WHAM are:

$$P(\xi) = \frac{\sum_{i=1}^{N_w} g_i^{-1} h_i(\xi)}{\sum_{j=1}^{N_w} n_j g_j^{-1} exp\left[-\beta\left(w_j(\xi) - f_j\right)\right]} \qquad (2.26)$$

and

$$exp\left(-\beta f_j\right) = \int d\xi \; exp\left[-\beta w_j(\xi)\right] P(\xi) \qquad (2.27)$$

With β being the inverse temperature $1/k_B T$, k_B the Boltzmann constant and T the temperature, and n_j is the total number of datapoints in histogram h_j, and f_i is a free energy constant. The statistical insufficiency g_i is expressed by $g_i = 1 + 2\tau_i$ with the integration autocorrelation time τ_i of umbrella window i . $P(\xi)$ denotes the unbiased probability distribution that is related to the PMF via $\mathscr{W}(\xi) = -\beta^{-1}\ln\left[P(\xi)/P\left(\xi_0\right)\right]$. Here, ξ_0 is an arbitrary reference point where the PMF $\mathscr{W}\left(\xi_0\right)$ is zero.[99] WHAM is an iterative optimization process that aims to find the optimal unbiasing weights for each histogram. The optimized weights are available after convergence is achieved. To obtain the free energy profile, one needs to calculate the probability distribution first for the unbiased case and then the free energy function. One of the shortcomings of this method is that it assumes a proper equilibration sample was created which is often not the case. For example, the sampling in some biased runs may not be converged if the dynamics are slow and some high energy barrier events are not sampled.[100]WHAM also disregards the time sequence information within simulation trajectories and therefore kinetic information is lost.

Dynamic Histogram Analysis Method (DHAM)

The dynamic histogram analysis method (DHAM)[100] has several advantages over WHAM. Unlike WHAM, it does not disregard time sequence information. The goal of DHAM is to find the equilibrium free energy along a chosen reaction coordinate x in a way that considers dynamical information about the resulting time correlations. Unlike WHAM, DHAM is based on a global Markov state model (MSM) and uses a maximum likelihood estimate of a Markov transition matrix transition probabilities by using joint unbiasing of the transition counts from multiple US simulations along discretized RCs. The free energy profile can be obtained from the stationary distribution of the resulting Markov transition matrix.[101] Rosta and Hummer have developed an explicit approximation for this that does not require an iterative solution.[101]

The relation between biased and unbiased Markov transition probability matrices *M* can be expressed by solving the Smoluchowski diffusion equation[102] for transition probabilities $p(\, i \to j, \tau)$ from state i to j within a lag time τ:

$$\frac{M_{ji}^{k}}{M_{ji}^{0}} = \frac{p(\, i \to j, \tau)^{\,k}}{p(\, i \to j, \tau)^{\,0}}$$

$$= \frac{exp\left(-\left(\left(x_j - x_i\right) + \gamma\tau \, \frac{U_j^k - U_i^k + U_j^0 - U_j^0}{x_j - x_i}\right)^2 / 4D\tau\right)}{exp\left(-\left(\left(x_j - x_i\right) + \gamma\tau \, \frac{U_j^0 - U_j^0}{x_j - x_i}\right)^2 / 4D\tau\right)} \qquad (2.28)$$

with superscipt k denoting the biased simulation, 0 denoting the unbiased simulation. U is the potential energy along the reaction coordinate x, and $\gamma = D / k_B T$ is the mobility of the system. Expanding the squared terms in Equation (2.30) and omitting all τ^2 terms lead to the square root approximation at short lag times,

$$\frac{M_{ji}^{k}}{M_{ji}^{0}} \approx exp\left(-\left(U_j^k - U_i^k\right) / 2k_{\mathrm{B}}T\right) \qquad (2.29)$$

.

The unnormalized Markov matrix is defined as:

$$M_{ji} = \sum_{k=1}^{N} \frac{T_{ji}^{k}}{\sum_{l=1}^{N} n_i^l \, exp\left(-\left(u_j^l(c_j) - u_j^l(c_i)\right)\right) / 2k_{\mathrm{B}}T} \qquad (2.30)$$

where data is binned along x, and $T_{ji}^{k} = \sum_{t}^{L^k - \tau} \delta(x^k(t) \in i)\, \delta(x^k(t+\tau) \in j)$ gives the transition count from bin $i$to bin j in simulation window k, with data saved and analyzed at the frequency of the lag time τ from the overall length L^k of simulation k. $n_i^k = \sum_j T_{ji}^{k}$ is the number of transitions initiating from bin . The bias $u_i^l = U_i^l - U_i^0$ is evaluated at each bin center c_i assuming that the biasing is also done along x.

2.5 Quantum Mechanics-level based methods

To study bond breaking and bond formation, and chemical reactivity in general, one needs to be able to model the movement of electrons. Clearly, for this purpose, we need to move away from treating atoms as point particles but rather handle nuclei and electrons separately. Electrons taking part in chemical reactions, such as the catalytic reactions modeled in this work, need to be described quantum mechanically. All of the Quantum Mechanics (QM) and Quantum Mechanics/Molecular Mechanics (QM/MM) methods for calculating various properties in this thesis use Density Functional Theory (DFT) as their foundation so I am going to give a brief introduction to the main concepts behind DFT.

2.5.1 Density Functional Theory (DFT)

The central idea underpinning DFT is that the total electronic energy is a function of the overall electronic density. This concept was originally developed in the 1920s but in 1964 Hohenberg and Kohn were able to show that the ground state energy of a system and other properties are uniquely defined by the electron density of the said system.[103] Since every electron has three spatial coordinates and one spin coordinate, this makes the $4N$ dimensional electron wavefunction very complex. In contrast, the electron density depends only on three spatial coordinates in which the density ρ is defined, regardless of the size of the system. The aim of DFT is to express the electronic energy as a functional of the density:

$$E\left[\rho(\vec{r})\right] \qquad (2.31)$$

Constructing a functional $E[\rho]$ invokes some problems, because some of the contributions from the system are difficult to define and therefore unknown, coming from the many-body problem of interacting electrons. The great advancement came from Kohn-Sham's formalism which separated the interacting many-body problem into a set of non-interacting problems. They introduced fictitious non-interacting electrons with the same density as the real interacting system but with an effective potential that includes the effects of the electron-electron interaction. These non-interacting electrons are subject to an effective potential, which includes contributions from the external potential, the Hartree term (electron-electron

repulsion), and an exchange-correlation term. This eventually resulted in the Kohn-Sham formalism for the single determinant wavefunction expressed on a set of basis functions.

The summation over the occupied orbitals gives the electron density:

$$\rho = \sum_i \left|\phi_i\right|^2 \qquad (2.32)$$

The total energy functional is then defined as:

$$E^{KS} = T_s[\{\varphi i^{KS}\}] + E_{en}[\rho] + J[\rho] + E_{exc}[\rho] \qquad (2.33)$$

It is possible to compute every part of a Kohn-Sham DFT energy exactly apart from the last term, the exchange correlation energy $E_{ex}[\rho]$. The non-interacting kinetic energy can be calculated with the Kohn-Sham wavefunction, the Coulombic interactions can also be calculated and integrated over the density. Only the exact exchange correlation energy functional is not known. This term accounts for the quantum mechanical effects of electron exchange and correlation. The exchange part involves the antisymmetrization of the electron wave function, in order not to break the Pauli exclusion principle, and it is related to the fact that electrons are indistinguishable particles. The correlation part captures the quantum mechanical effects arising from the electron-electron interactions beyond what is accounted for by the mean-field approximation. The exchange-correlation functional must be approximated in practical calculations. There are many ways developed to do that through the years which will not be the subject of discussion here.

2.6 Quantum Mechanics/Molecular Mechanics

The hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) method is a simulation method where one part of the simulated system, usually where an important biochemical reaction is modeled, is treated quantum mechanically, while the rest of the system is simulated with a classical molecular mechanics-based method. The computational cost makes it virtually impossible to treat entire proteins quantum mechanically with the current

computing capabilities. A solution for this is to adopt QM/MM which allows to model biomolecular systems in an efficient way. QM/MM calculations combine the accuracy of ab initio methods with the speed of MM-based approaches thus allowing a big part of a protein to be simulated with an MM-based method while a much smaller region of interest, such as an active site, to be simulated with a QM-based method.

The total energy of the simulated system can be expressed in the following way:

$$E = E_{QM} + E_{MM} + E_{int} \qquad (2.34)$$

Where E_{QM} is the energy of the QM part, E_{MM} is the energy of the MM region, and E_{int} is the energy of interaction of the two regions. There are several ways to handle the electrostatic coupling which I am going to introduce briefly.

Embedding models which are applied to deal with the interaction energy E_{int} focus on geometry-based ways to split the whole region into individual parts where the QM/MM region in the case of proteins is usually split along individual chemical bonds (C-C bond in an amino acid, for example).[104] For every bond broken, there are two unpaired atoms that need to be somehow "capped". This is usually done through the introduction of linker atoms, usually hydrogens, to take up the free valence and not create a free radical unintentionally.

The simplest way to handle the electrostatic coupling between the QM and the MM regions is mechanical embedding. In the case of mechanical embedding, the QM/MM electrostatic interaction is treated as the electrostatics in the MM region – QM atoms are assigned the force field parameters of the force field used to describe the MM region and non-bonded terms are evaluated for pairs across the two regions. Mechanical embedding results in some oversimplifications – for example, when a chemical reaction is modeled in the QM region, this will result in a change in electron density. When the density changes, it would be expected to update the charges of the atoms. However, updating those charges would cause discontinuities in the PES. The MM charges assigned from the force field also do not

reproduce the true charge distribution of the inner region correctly. There are additional issues arising from this simplification, but those will not be discussed currently.[105]

Electrostatic embedding solves some of the problems introduced by mechanical embedding schemes. It defines the MM atoms as point charges in the QM input, thus allowing polarization of the electron density by the MM region. It is also possible to include a polarization effect or polarizable embedding on the MM atoms, introducing a need for a self-consistent iteration and a force field describing the MM region which can include polarization.[105]

2.7 Machine Learning

2.7.1 Unsupervised and Supervised Machine Learning

In unsupervised learning, the algorithm's aim is to find patterns or structures in data, without explicit labels. Unlike supervised learning, where the algorithm learns from labeled examples provided by a dataset, unsupervised learning operates on unlabeled data, relying solely on the inherent structure or relationships within the data. Common unsupervised learning methods include dimensionality reduction and clustering.

The most common dimensionality reduction techniques are the Principal Component Analysis (PCA) or t-distributed stochastic neighbor embedding (t-SNE), which aim to represent high-dimensional data in a lower-dimensional space while preserving the essential structure and relationships between data points. The signal-to-noise ratio often improves after dimensionality reduction, as the reduced-dimensional representation focuses on the most informative aspects of the data, leading to clearer patterns and structures.[106,107]

Clustering algorithms, such as k-means or hierarchical clustering, group similar data points together based on their characteristics. Center-based clustering like the k-means algorithm, partitions the data into clusters around central points or "centroids".[107] Each data point is assigned to the nearest centroid, resulting in clusters that are compact and well-separated. However, center-based clustering methods often struggle with non-spherical clusters and are

sensitive to initialization bias.[107] Hierarchical clustering, on the other hand, builds a tree-like hierarchy of clusters, either by agglomerative or divisive approaches. It does not require specifying the number of clusters beforehand and can capture clusters of varying shapes and sizes. Agglomerative hierarchical clustering starts with each data point as a separate cluster and iteratively merges them based on a similarity measure, resulting in a dendrogram that illustrates the nested clusters at different levels of granularity.

In supervised learning, one or several target properties, such as the enzyme thermal stability or solvent accessible surface area, for example, are predicted based on labeled training data. The goal is to engineer a predictor that will return labels/predictions for unseen data points on the basis of their descriptors or 'features'. Generally, it can be said that what sets apart unsupervised from supervised ML is the presence of labels in the training data set. When I discuss a predictor or a model, this refers to the mathematical structure by which the prediction y_i is made from the input data x_i. For a linear model, the prediction is based on a linear combination of weighted input features.

Principal Component Analysis (PCA)

The most common approach for linear dimensionality reduction is the Principal Component Analysis (PCA). It is used as a first-to-go approach in cases when one has many features and would like to reduce the dimensions to two or three principal components. The idea behind PCA is to define a set of orthogonal components through the eigendecomposition of the covariance matrix of the input data. There are N components in total, where N is the dimensionality of the input space. The component with the largest eigenvalue will be the one that maximizes the variance when the data is projected on it. By projecting onto the n components with the largest eigenvalues, the input data can be transformed into an n-dimensional representation in which the variance amongst data points is maximized. One can define the number of principal components that the model can then use for predictions, instead of the full set of original features, for example.

Clustering algorithms

There are many approaches to clustering that can be taken, depending on the type of data. Centroid-based clustering, which was used in this thesis, organizes the input data into non-hierarchical clusters. The most widely used centroid-based algorithm is k-means.[107] It is effective for a varying range of data types but as mentioned, it is sensitive to initial conditions and outliers. In the example in Figure 2.1, a PCA is performed on a high-dimensional set of features, and the first two principal components are used to check what the clusters will look like. This is an example of a very simple way to attempt finding clusters based on datasets with n-dimensional data set.

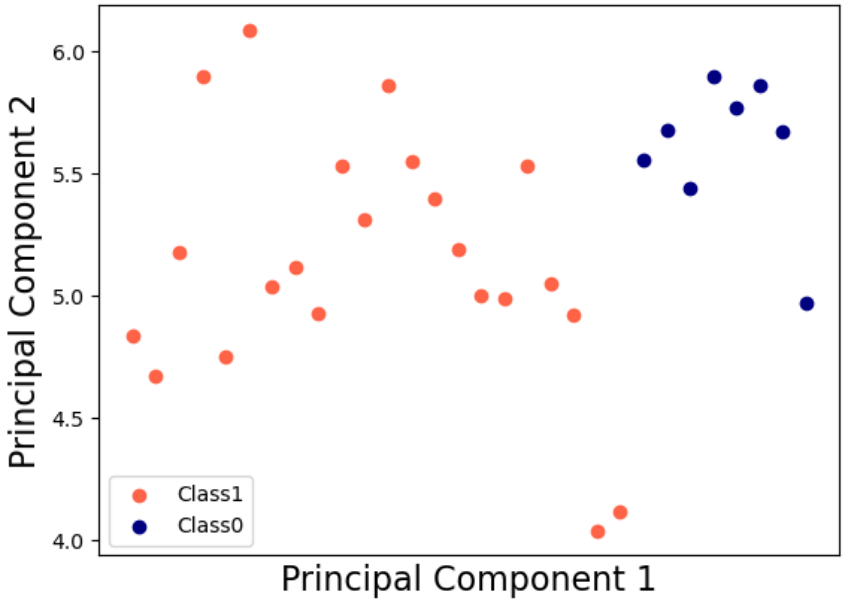


Figure 2.1. Example of non-hierarchical clustering, based on an initial dimensionality reduction by PCA. The two principal components are used to cluster the data using the k-means clustering algorithm.

Gradient Boosted Decision Trees

Gradient Boosting is an algorithm that combines many weak learners, such as decision trees, into a single strong learner, or model, by iteratively attempting to learn from the prediction error of the previous weaker learner.

The goal is to make a model F to learn to predict values in the form $\hat{y} = F(x)$ by minimizing the mean squared error $\frac{1}{2}\sum_i \left(\hat{y}_i - y_i\right)^2$, where i indexes over the training set of size n of values of the output variable y with: $\hat{y}_i$ being the predicted value, y_i the observed value, and n is the number of samples y.

The gradient boosting algorithm is built in the following way:

$$F_{m+1}(x_i) = F_m(x_i) + h_m(x_i) = y_i \qquad (2.35)$$

This is equivalent to $h_m(x_i) = y_i - F_m(x_i)$ (2.36)

So the algorithm will fit h_m to $y_i - F_m(x_i)$. The model F_{m+1} then attempts to improve the prediction based on the error of the previous model F_m. This iterative process repeats until a stopping criterion is met, such as a maximum number of iterations or if the (stronger) model begins to overfit.[108,109]

2.7.2 Splitting the Data

Any pre-processed, curated dataset, prepared to be used by a model, is typically split into subsets, to assess how effective the trained model will be on unseen data. This is achieved by allocating some of the data to training and testing sets, respectively. In this thesis, this was done with the scikit-learn implementation of the module train_test_split.[110]

Training Set

The training dataset is the part of the data that the model sees and learns from, to predict an outcome. The more diverse and representative the training data is, the more likely for the model to be able to generalize well on unseen data.

Testing Set

The testing set is the unseen part of the data used for evaluating the model. It is independent of the training set and should have a similar type of probability distribution of classes as the training set. Typically, 20 or 30% of the data is left out for testing but the data could also be split in different ways.

Validation Set

The validation set is used to fine-tune the hyperparameters of the model and provide an objective unbiased evaluation of the model.

2.7.3 Generalization capability, Overfitting and Underfitting

One of the most important concepts in Machine learning (ML) is the generalization capability of a model. This refers to the ability of models to predict/classify data samples never encountered before. Usually, two reasons prevent this – overfitting and underfitting. Overfitting refers to situations when the model learns too well on the training data. The model uses a combination of features that result in learning characteristics of the training set that allow it to predict well on the current testing set. However, these features and/or their combination are not necessarily suitable to describe similar unseen data. This commonly happens when too many features are selected on a dataset with a small sample size (50 features on a dataset with N=100 samples, for example). Underfitting results from the inability of a model to learn enough from the training dataset. Such a model will have poor performance in predicting the target variable of unseen data, just as well as in the case of overfitting, but for different reasons. To illustrate both concepts, I have generated synthetic data with a sine function and added some noise to it. I then fit polynomial regression models with different degrees (1, 4, and 15) to the data. As can be seen in Figure 2.2A, the model is underfitting and it cannot capture the data. In Figure 2.2B, the degree of the polynomial captures the trends in the data. As the degree of the polynomial increases, the model becomes more complex and fits the training data more closely. However, this increased complexity leads to overfitting, as the model starts capturing noise in the data (Figure 2.2C) rather than the underlying pattern.

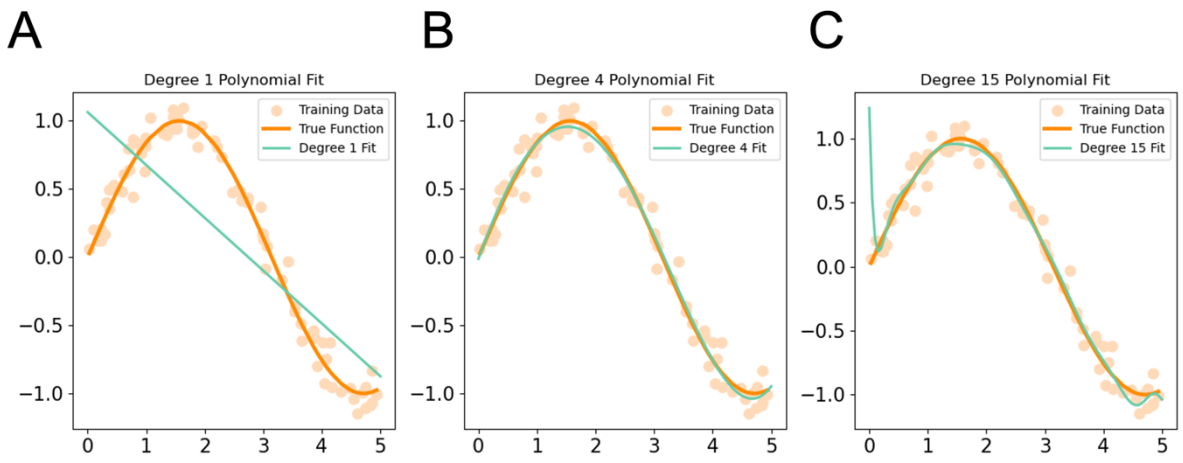


Figure 2.2 (A) A case of underfitting, the model cannot capture the pattern of the data. (B) The data is sampled well. (C) The degree of the polynomial is too big and it starts capturing the noise rather than real trends in the data, resulting in overfitting.

Chapter 3

Structural Dynamics and Catalytic Mechanism of ATP13A2 (PARK 9) from Simulations

This Chapter was published in The Journal of Physical Chemistry virtual special issue “Dave Thirumalai Festschrift” in 2021 and is reproduced here with permission from: Teodora Mateeva, Marco Klähn, and Edina Rosta, ‘Structural Dynamics and Catalytic Mechanism of ATP13A2 (PARK9) from Simulations’, J. Phys. Chem. B, DOI:2021, 125, 11835−11847.

Summary of the Work

Patients diagnosed with Parkinson’s disease (PD), spastic paraplegia (SPG78), Kufor−Rakeb syndrome, neuronal ceroid lipofuscinosis, and other similar neurological disorders often carry a varying range of mutations in the ATP13A2 gene.[1,16–22,111–113] The mechanism through which missense mutations are implicated in Parkinsonism is not always known. Certain protein mutations, which are commonly present in carriers of the condition, such as G504R and F182L, disrupt the vesicular localization of ATP13A2 and promote the mislocalization of the enzyme to the endoplasmic reticulum, thus exposing it to speeded degradation.[15] However, for a large part of the reported missense mutations, the exact mechanism in which they are implicated in pathogenicity, is not clear as they do not alter protein stability or affect subcellular localization.[15] And while the importance of the enzyme in regulating neuronal integrity is established, at the time of the start of this project, there was no three-dimensional structure of this transmembrane enzyme and no consensus on the active site composition and conformation in terms of the number of ions taking part in the catalytic mechanism and the precise mode of ATP binding. This makes it difficult to study how missense mutations

close to the active site might affect the catalytic mechanism of ATP13A2 and whether those mutations disrupt the catalytic mechanism of the protein directly or indirectly. In this chapter, I provide a detailed description of the catalytic reaction leading to the state of the protein where Asp513 is autophosphorylated. The MD and QM/MM simulations provide strong evidence that two Mg^{2+} cations are present at the active site during the catalytic reaction. I also elucidated details of the catalytically competent ATP conformation and the binding mode of the second Mg^{2+} cofactor. The exact role of conserved Arg686 and Lys859 catalytic residues was demonstrated.

Author Contribution

I conceptualized most of this work; wrote the manuscript and performed all the analysis. All bioinformatics research needed for the modeling of this enzyme was done by me, as well as all MD simulations. I performed all QM and QM/MM simulations in this paper and analyzed the results from the QM calculations and potential energy scans. I produced all the figures in the main text and the Supporting Information. My supervisors have approved the manuscript in its final form.

Correction

On p.55, section **Homology modeling**, the correct Uniprot code for the ATP13A2 sequence is Q9NQ11. The sequence used to model the protein in this work is correct, however, the Uniprot code is either incorrectly reported (Q9HD20) in the original paper, or it has changed in the Uniprot database.

THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB Article

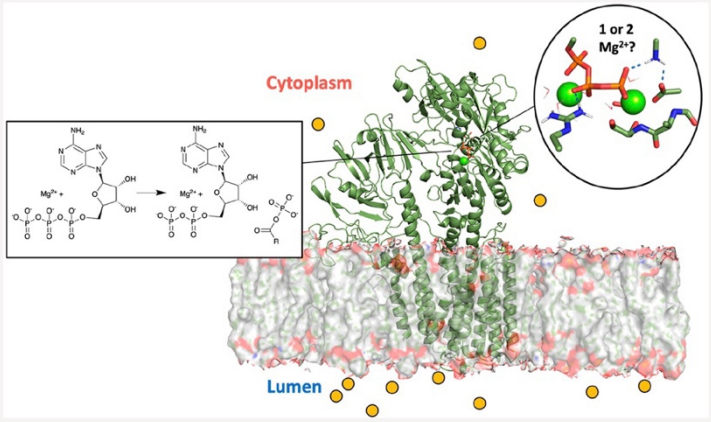
Structural Dynamics and Catalytic Mechanism of ATP13A2 (PARK9) from Simulations

Published as part of The Journal of Physical Chemistry virtual special issue "Dave Thirumalai Festschrift".

Teodora Mateeva, Marco Klähn, and Edina Rosta*

Cite This: *J. Phys. Chem. B* 2021, 125, 11835−11847 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: ATP13A2 is a gene encoding a protein of the P5B subfamily of ATPases and is a PARK gene. Molecular defects of the gene are mainly associated with variations of Parkinson's disease (PD). Despite the established importance of the protein in regulating neuronal integrity, the three-dimensional structure of the protein currently remains unresolved crystallographically. We have modeled the structure and reactivity of the full-length protein in its E1-ATP state. Using molecular dynamics (MD), quantum cluster, and quantum mechanical/molecular mechanical (QM/MM) methods, we aimed at describing the main catalytic reaction, leading to the phosphorylation of Asp513. Our MD simulations suggest that two positively charged Mg^{2+} cations are present at the active site during the catalytic reaction, stabilizing a specific triphosphate binding mode. Using QM/MM calculations, we subsequently calculated the reaction profiles for the phosphoryl transfer step in the presence of one and two Mg^{2+} cations. The calculated barrier heights in both cases are found to be ∼12.5 and 7.5 kcal mol^{-1}, respectively. We elucidated details of the catalytically competent ATP conformation and the binding mode of the second Mg^{2+} cofactor. We also examined the role of the conserved Arg686 and Lys859 catalytic residues. We observed that by significantly lowering the barrier height of the ATP cleavage reaction, Arg686 had major effect on the reaction. The removal of Arg686 increased the barrier height for the ATP cleavage by more than 5.0 kcal mol^{-1} while the removal of key electrostatic interactions created by Lys859 to the γ-phosphate and Asp513 destabilizes the reactant state. When missense mutations occur in close proximity to an active site residue, they can interfere with the barrier height of the reaction, which can halt the normal enzymatic rate of the protein. We also found large binding pockets in the full-length structure, including a transmembrane domain pocket, which is likely where the ATP13A2 cargo binds.

■ INTRODUCTION

The ATP13A2 gene has emerged as one of the genes strongly correlated with Parkinson's disease (PD) and is also known as PARK 9.[1,2] The ATP13A2 gene encodes the P5B ATPase ATP13A2, which has attracted interest as an enzyme implicated in a range of neurodegenerative disorders: spastic paraplegia (SPG78), Kufor−Rakeb syndrome, neuronal ceroid lipofuscinosis, and various other types of neurodegenerative disorders.[2−6] Currently, a multitude of molecular defects associated with the gene have been identified,[1,2,5,7−14] including some loss-of-function missense mutations of the

Received: June 17, 2021
Revised: September 29, 2021
Published: October 22, 2021

https://doi.org/10.1021/acs.jpcb.1c05337
J. Phys. Chem. B 2021, 125, 11835−11847

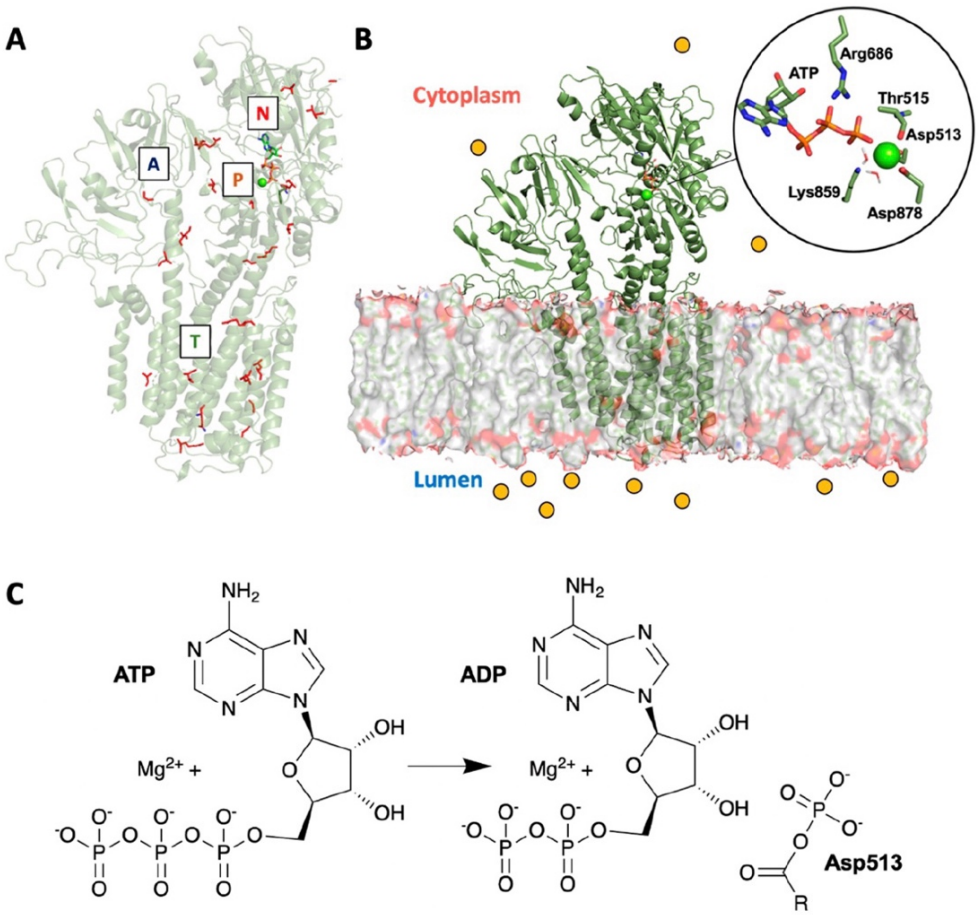


Figure 1. (A) Three-dimensional homology model of the ATP13A2 protein depicting mutations identified clinically[5,9−11,13,18] (red sticks) on the protein (green cartoon) and (B) a homology model of ATP13A2 with the transmembrane domain buried in the lipid-rich membrane while the N, P, and A domains are located in the cytoplasm of the cell. (C) Products of the autophosphorylation reaction in cytoplasmic domains.

protein. The precise role of most missense mutations remains unexplored, as well as the overall structural dynamics of the protein.

ATP13A2 belongs to the haloacid dehydrogenase-like (HAD) superfamily of enzymes that all share a hydrolase fold. The HAD superfamily is very diverse, encompassing phosphoesterases, P-type ATPases, phosphonatases, dehalogenases, and sugar phosphomutases, which act on a wide range of substrates, typically catalyzing carbon or phosphoryl group transfer reactions.[15] Phosphotransferase enzymes typically require Mg^{2+} cofactor for their catalytic activity.[16,17] ATP13A2, in particular, belongs to the big family of P-type ATPases which is split in five distinct subfamilies: P1, P2, P3, P4, and P5.[6] Most of these proteins are well studied and have resolved crystallographic structures, including ones in different functional states. ATP13A2 is part of the least studied subfamily P5B, which remains the only subfamily without any three-dimensional structures resolved.

The cytoplasmic domains of the protein include: Nucleotide-binding domain N, Phosphorylation domain P and an actuator domain A (Figure 1A). Additionally, a transmembrane domain (T) connects the catalytic domains located in the cytoplasm to the extracytoplasmic area (Figure 1B). Interestingly, the various mutations of the ATP13A2 protein currently described in the literature[5,9−11,13,18] are not confined to one spatial region (Figure 1A) but are scattered across the entirety of the protein and encompass all domains. The catalytically active domains, N, P, and A, are involved in ATP binding, ATP cleavage, and auto- and dephosphorylation. ATP13A2 has been classified as a membrane transporter protein[6] with the proposed candidates ranging from heavy metals[2] to Ca^{2+} cations[19] and polyamine spermidine (SPD).[20,21] Recent studies have revealed the role of ATP13A2 in polyamine export.[22] All enzymes belonging to the HAD superfamily contain a specific form of the Rossmannoid fold. This fold has two characteristic features that distinguish it from other superfamilies with Rossmannoid type-folds: a β-hairpin motif (also called a "flap") located immediately downstream of the first β-strand of the core Rossmanoid fold and a single helical turn ("the squiggle").[15] This is important for ATP13A2 and other P-type ATPases, as these motifs provide mobility which allows the protein to alternate between the E1 "open" conformation (before the binding of any cargo) and the E2 "closed" conformation. The E1 state is associated with the

https://doi.org/10.1021/acs.jpcb.1c05337
J. Phys. Chem. B 2021, 125, 11835−11847

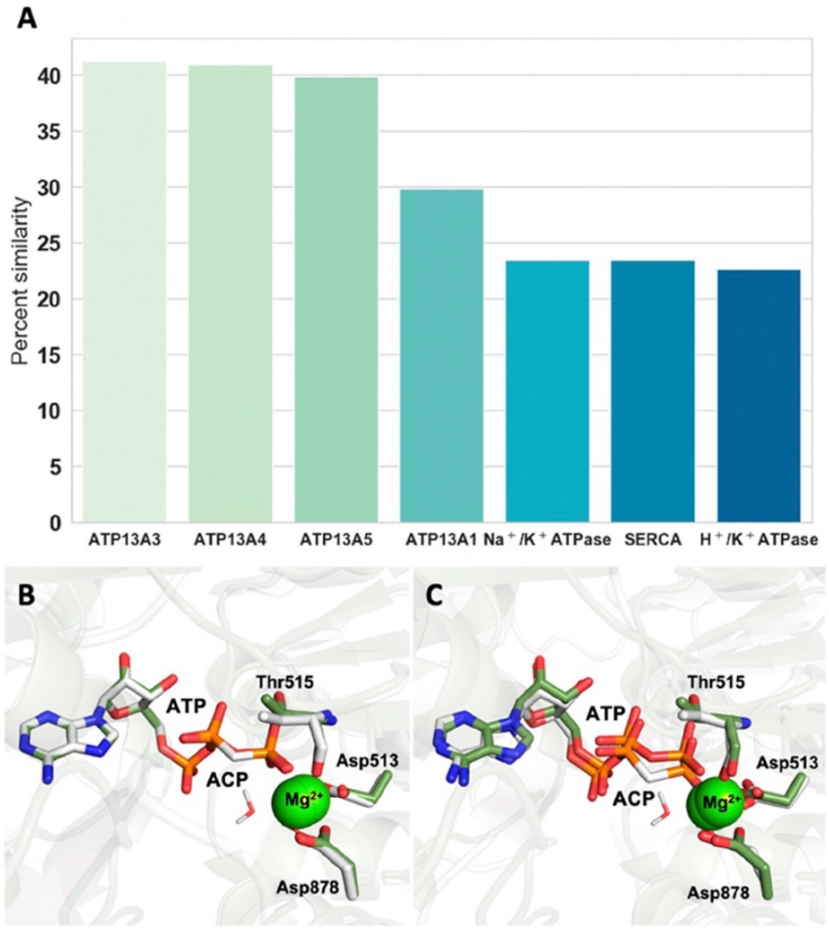


Figure 2. (A) Most similar human proteins to ATP13A2 based on overall fold, percent of sequence similarity, and overall query coverage, ordered from most to least similar. (B) Active site of the endoplasmic reticulum Ca^{2+}−ATPase (SERCA) in the E1 state with bound ACP molecule and one Mg^{2+} cation (PDB code: 3tlm,[26] gray) and our homology model of ATP13A2 (green sticks). (C) Active site of ATP13A1 in the E1 state with bound ACP molecule and 1 Mg^{2+} cation (PDB code: 6xmq,[30] gray sticks) and our homology model of ATP13A2 (green sticks).

binding and subsequent cleavage of ATP (Figure 1C). In this state, the protein has a high affinity for the cargo that is to be transported from the cytoplasm to the other side of the membrane. In ATP13A2, the ATP cleavage reaction results in autophosphorylation of the strictly conserved Asp513. Similarly, the E2 state is associated with the process of dephosphorylation of the aspartate.[6] In this work, we are interested in the change from the E1-ATP to the E1P functional state.

The active site motif DKTGT is strictly conserved among all P-type ATPases.[23] Two other amino acids, which are highly conserved and located immediately in the active site, are Arg686 and Lys859. Arg686 and Lys859 were structurally conserved in all enzymes whose crystal structures were used further in this study,[24−29] Figure S1.

Currently, none of the proteins within the P5B ATPase family have been resolved crystallographically, including ATP13A2, therefore, no three-dimensional structure is resolved in any of the functional states of the protein. Nevertheless, ATPases of the P2A and P2C subfamilies, which are highly homologous, are available with experimentally determined structures. Most recently, a crystal structure of the P5A ATPase ATP13A1 was resolved, which is currently the most homologous protein to ATP13A2 whose three-dimensional structure has been determined.[30] Many of the three-dimensional P-type ATPase structures contain active-site bound ATP-analogues, typically with synthetic nonhydrolyzable derivatives of ATP, such as ACP or AMPPCP.[24−26] Most of these structures feature only one bound active site Mg^{2+} cation.[24−26] However, there are structures obtained with ADP and AlF_3 that feature two Mg^{2+} cations[27,29,31] bound in the active site. This brings up the question whether one or two Mg^{2+} ions are present and/or required for the phosphoryl transfer to proceed? Importantly, due to the different charge distribution of the synthetic derivative analogues, the second ion coordination may be captured incorrectly or not at all. There are no structures which have been crystallized with the catalytically competent ATP, and it is currently unknown what the precise ATP conformation during the phosphoryl transfer reaction is, especially in terms of its proposed second Mg^{2+} coordination.[31] This leaves open the question of what the

https://doi.org/10.1021/acs.jpcb.1c05337
J. Phys. Chem. B 2021, 125, 11835−11847

precise ATP−Mg^{2+} coordination is, as well as the overall Mg^{2+}−Mg^{2+} distance and position.

Currently, the catalytic mechanism is not available using atomistic details for any P5 ATPase. Previous short molecular dynamics (MD) simulations were carried out on the P2A endoplasmic reticulum Ca^{2+}−ATPase;[32] however, those did not provide any detailed insight on the catalytic mechanism or overall conformational dynamics of the protein. Multiscale reactive molecular dynamics (MS-RMD) and free energy sampling have been used to quantify the free energy profile and time scale of the proton transport in SERCA.[33] Quantum mechanical/molecular mechanical (QM/MM) calculations have also previously been performed for the phosphoserine phosphatase, which also belongs to the large HAD-like superfamily of proteins, but it is classified in a distinctly different family.[34] We therefore performed MD, QM cluster, and QM/MM calculations with the aim to describe this important catalytic mechanism and quantify the role of active site residues and active site cations in the phosphoryl transfer reaction in the catalytically competent E1-ATP functional state. Our QM/MM calculations found that Arg686 had a significant effect on the barrier height similarly to arginine fingers, however interacting with the β-phosphate of the ATP backbone.[35] Accordingly, experimental data from mutagenesis studies of the Ca^{2+}−ATPase suggests that this conserved arginine is detrimental to the ATPase activity of the homologous enzyme.[27] We further show the precise effect on the barrier height of Lys859, which is similarly very important through interactions with both the γ-phosphate of the ATP and Asp513 in the reactant state. The results presented in this work can suggest how missense mutations disrupting crucial barrier-lowering interactions can have an abolishing effect on the enzymatic activity of ATP13A2 and other P5B ATPases with a homologous active site, specifically G877R[5].

METHODS

Homology Modeling. The E1-ATP-bound state was modeled based on the Endoplasmic Reticulum Ca^{2+}−ATPase (SERCA), (PDB code: 3tlm).[26] The position of the ATP molecule and the Mg^{2+} cation was based on the position of the ACP moiety and the Mg^{2+} cation, respectively, in this crystal structure.[26] The crystal structure has the position of only one Mg^{2+} resolved; hence, the initial model contained only one Mg^{2+}. The modeling server used was SWISS-MODEL[36] and the sequence of ATP13A2 was obtained from Uniprot[37] for *Homo sapiens* (Uniprot code: Q9HD20). To account for the most recent crystal structure available of ATP13A1, ATP13A2 was also modeled based on the P5A ATPase ATP13A1 (PDB code: 6xmq).[30] Both templates result in the same three-dimensional structure in the active site region of interest (Figure 2).

Molecular Dynamics Simulations. All MD simulations were performed by using the program NAMD.[38] The force field used in the simulations was CHARMM36[39] with periodic boundary conditions and to evaluate the nonbonded long-range interactions the particle mesh Ewald method[40] was utilized with a 12 Å cutoff. The NPT ensemble was maintained with a Langevin thermostat (303 K) and an anisotropic Langevin piston barostat (1 atm). The simulation was repeated at 309.15 K. The system consists of 361707 atoms. The final crystal type of the assembled system is tetragonal with dimensions along the *X*, *Y*, and *Z* axes: 150.7 Å, 150.7 Å, and 170.6 Å, respectively. The angles between all axes are 90 deg.

The water model was TIP3P.[41] To neutralize the system, a 0.15 M KCl solution was added. The ion placing method was by distance. The energy of the system was minimized via steepest descent algorithm, followed by a standard six-step equilibration for membrane-embedded systems with restrained heavy atoms via a standard CHARMM-GUI[42] procedure with a time step of 2 fs. The first step of the equilibration was done with a time step of 1 fs. The SHAKE algorithm[43] was deployed to constraint the covalent bonds involving hydrogen atoms. The equilibration was followed by 100 ns production with all the atoms completely unconstrained and free to move. A second 100 ns MD simulation was performed where the coordinates of the ATP molecule were fixed in their original position (following the crystal structure coordinates of the homologous template) in order to preserve the original coordinates of the crystal structure ACP (PDB code: 3tlm).[29] The protein, solvent and ions were completely unconstrained. The PPM server was used for orientation of the protein in the membrane.[44] The membrane had the following composition: 40% cholesterol, 30% phosphatidylcholine lipids (PC), and 30% phosphatidylethanolamine lipids (PE) to mimic a membrane environment in a lysosome-like cell. The same protocol was repeated for the second homology model of ATP13A2, which was based on the P5A ATPase ATP13A1.

QM Cluster Calculations. All of the QM cluster calculations were performed by using the Gaussian09 program.[45] The QM region was treated with the B3LYP hybrid density functional[46] and the 6-31+G* basis set.[47] The QM region consisted of two Mg^{2+} cations, six water molecules, the side chain of Asp513, the side chain of Asp878, and the full Thr515, as well as the full ATP molecule (Figure S3). The geometry optimization followed standard QM cluster procedure where the C atom where the amino acids are truncated, is frozen. Where a single C−C bond is cut, three H atoms are added to satisfy the C atom valency. The exact atoms which are frozen in the calculation are illustrated in Figure S3. The second Mg^{2+} cation in the starting geometry of G1 (Figure S4A) was placed based on alignment with the crystal structure of the Na^+/K^+-transporting ATPase, which has full sequence conservation within 4.5 Å of the Mg^{2+} ions and is resolved with ADP and two Mg^{2+} cations (PDB code: 3wgu).[29] The starting structure of G3 was taken from a snapshot of the last nanosecond of the unconstrained MD simulation (Figure S4C). In G2 and G4, the second Mg^{2+} ion was placed by a manual initial guess.

Quantum Mechanical/Molecular Mechanical Methods. All QM/MM calculations were performed by Q-Chem,[48] coupled with CHARMM.[49] The QM region contained: the Mg^{2+} cation/s, six water molecules, the phosphate chain of ATP, the side chain of Asp513, the side chain of Asp878, the full Thr515, the side chain of Lys859 and Arg686, the full Gly516, and the main chain of Lys514. The QM region during the RCS was treated with the B3LYP hybrid density functional and 6-31+G* basis set.[46,47] The MM region contained all residues and solvent within 25 Å of the QM region. The residues included in the QM region were separated from the rest of the chain by cleaving homonuclear C−C bonds and introducing link atoms, which were treated as hydrogen atoms in the QM calculations. An initial energy minimization was carried out, which constituted 1000 steps via the SCF DIIS

https://doi.org/10.1021/acs.jpcb.1c05337
J. Phys. Chem. B 2021, 125, 11835−11847

algorithm. Each minimized geometry was supplied for the RCS as a reactant state starting point. The active site containing two Mg^{2+} ions was obtained by aligning the ATP and second Mg^{2+} ion from the QM cluster optimization and translating the optimized geometry to the QM/MM model. The conformation of the ATP molecule and the position of the Mg^{2+} ion in the one Mg^{2+}-model were taken directly from the one observed in the crystal structure of the E1-ATP state of SERCA,[26] however, substituting one carbon atom of the crystal structure ACP molecule to a phosphorus, in order to have the catalytically active ATP. This structure was further minimized for 1000 steps. We defined the reaction coordinate by the distance from the nucleophile to the phosphorus $O_{2D}-P_G$ (R1) and the phosphorus and the leaving group P_G-O_{3B} (R2). Starting from the reactant state and moving along this coordinate, we simultaneously decrease the R1 distance and increase the R2 distance, to reach the product state. The distances were changed linearly. This forward−backward scanning is performed until energy convergence is observed. The solvent molecules in the QM region do not undergo reorganization from reactant to product state so the system has not been constrained additionally. A total of 40 minimization steps were completed each time before a data point was recorded during the RCS. All presented scans are converged and show the forward scan direction, going from the reactant to the product state. Six systems were independently minimized. The minimized structure of each was supplied for a starting structure (reactant) of the RCS. Each system was studied with its corresponding QM-region, and the overall charge and atomic constitution (residues included and number of Mg^{2+} ions) are summarized in Table S1.

Two-Mg^{2+} Active Site Simulations. We also performed unconstrained MD simulations with two Mg^{2+} ions in the active site to probe the overall stability of the system, and in particular the ATP conformation. We performed three replicates of 100 ns duration each. The same protocol was used as the one already described for the one-Mg^{2+} MD simulations. The transmembrane domain was not used in the two-Mg^{2+} MD simulations, only residues 485−930. For the starting structure, the ATP geometry was taken from the optimized QM cluster geometry, which had the most energetically favorable zigzag conformation of the ATP.

Pocket Analysis. Twenty frames were extracted from the 100 ns MD trajectory of the unconstrained simulation. The frames were spaced equidistantly and covered the duration of the MD simulation and were spaced 5 ns apart from each other (Table S2). The 20 biggest pockets were calculated for each frame, using the Pymol[50] plugin PyVOL.[51] To find pockets on the surface of ATP13A2, PyVOL[51] was provided with the protein chain only without the ATP or any of the Mg^{2+} ions. The four biggest pockets in terms of surface area were chosen for further analysis. Upon inspection, it was observed that the biggest pockets for every frame were observed in the N-binding domain where ATP binding normally occurs, and in the transmembrane domain, respectively. For this work, we define a pocket as an "ATP-binding pocket" if the pocket was found in a location within 1.5 Å of the ATP in the respective frame from the simulation. We define a "transmembrane pocket" if a pocket is located within 1.5 Å of where small inorganic ion binding has been observed in the crystal structures[26,29] of homologous proteins. If a pocket was calculated by PyVOL[51] to be occupying this area of interest, it was recorded as found. If it was not calculated within this area, it was recorded as not found (Table S2). In this way, we were able to calculate the frequency of occurrence of those two pockets (Table S2). Other pockets that were found consistently are illustrated in Figure S5. However, the occurrence was not as consistent as for the ATP pocket and the main transmembrane pocket. The results presented are obtained from the MD simulation of the homology model based on SERCA, but the same analysis was performed on the homology model based on ATP13A1 and the main transmembrane pocket was found in this model as well.

RESULTS

Homology Modeling. To identify the closest protein sequences to ATP13A2 and any available structural information on these, we performed BLAST[52] and FASTA[53] searches. The most similar proteins in humans, by sequence similarity, are as follows: ATP13A3, ATP13A4, and ATP13A5 (Figure 2A), which are all part of the P5B ATPase protein subfamily. As was proposed earlier,[22] those proteins have high conservation in the substrate binding domain and are likely transporting and/or interacting with the same cargo within the cell. Unfortunately, none of these proteins have three-dimensional structures deposited in the Protein Data Bank (PDB). The closest available proteins with experimentally resolved structures are ATP13A1 (P5A subfamily), the Na^+/K^+-transporting ATPase (P2C subfamily), the endoplasmic reticulum Ca^{2+}−ATPase (SERCA) (P2A subfamily), and the H^+/K^+−ATPase (P2C subfamily). We selected the endoplasmic reticulum Ca^{2+}−ATPase (SERCA)[26] and the very recently resolved structure of ATP13A1,[30] for modeling the active site (Figure 2, parts B and C). The SERCA ATPase has a complete conservation with ATP13A2 in the active site region of interest (Figure S1) and was already available with several ATP analogues bound, including water molecules crystallographically resolved.[26] Our final active site models match both experimental structures very accurately (Figure 2, parts B and C, for SERCA and ATP13A1, respectively).

We also note that any of the ATPases listed in Figure 2A would be a suitable choice for modeling the active site of ATP13A2 as the Mg^{2+}-binding residues in the catalytic P-domain are highly conserved among P-type ATPases, as well as the other amino acids found in the immediate active site. More structural information on the two homology models is available in Figure S2.

Molecular Dynamics Simulations. To probe the conformational dynamics of the protein, first, MD simulations were performed on the membrane-embedded model based on SERCA. The overall structure and in particular, the active site, was well conserved during the simulations (Figure S6).

We focused on the Mg^{2+} ion coordination, which was initially octahedrally coordinated, coordinating two water molecules, the side chain of the Asp513 residue (monodentate), the side chain of Asp878 (monodentate), the main chain of Thr515 (monodentate), and the γ-phosphate of the ATP. During the simulations, the coordination of the Mg^{2+} ion has remained octahedral, and importantly, the binding mode of the Asp513 residue has remained monodentate. The γ-phosphate of the ATP phosphate chain was also in the correct orientation for the phosphoryl transfer to occur.

Importantly, the ATP molecule no longer preserved its original "zigzag" conformation but immediately adopted a "straight" conformation, from the first nanosecond of the simulation (Figure S7A). This has also been observed before in

https://doi.org/10.1021/acs.jpcb.1c05337
J. Phys. Chem. B 2021, 125, 11835−11847

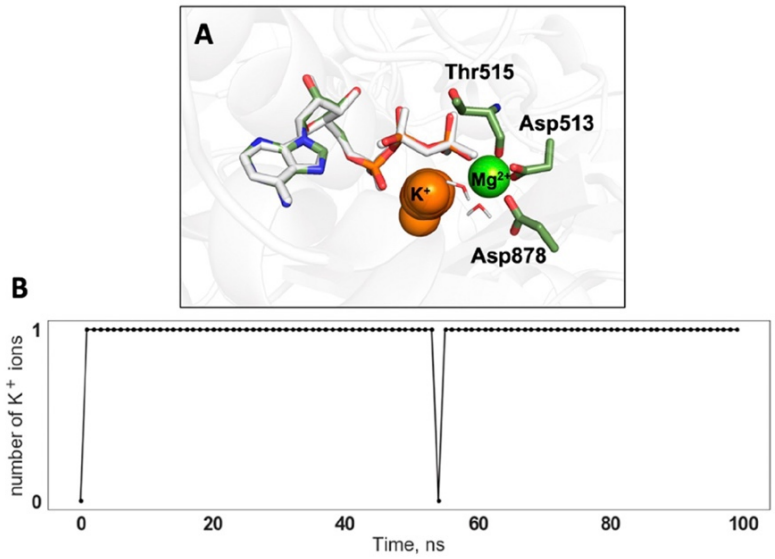


Figure 3. (A) Region of the active site where the K^+ ion clustering is observed during the simulation (orange spheres). (B) Number of K^+ ions in the active site during the duration of the 100 ns simulation.

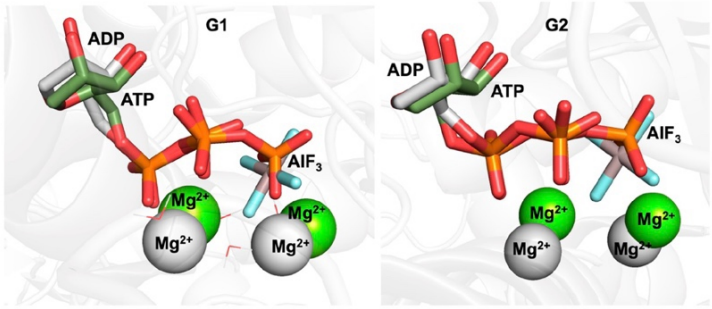


Figure 4. Optimized active site geometries for G1 and G2 (green sticks), aligned to the existing homologous active site crystal structure of the Na^+/K^+-ATPase (gray sticks, pdb: 3wgu[29]), which is resolved with two Mg^{2+} ions (gray spheres) and an ADP molecule (gray sticks). The green spheres represent the two Mg^{2+} ions in the optimized geometries and the gray spheres represent the two Mg^{2+} ions in the crystal structure. Both structures represent the same binding mode to the second Mg^{2+} via two oxygen atoms of the α and β phosphates of the ATP.

the short MD simulations of SERCA.[32] Komuro et al. suggested that the "straight" ATP conformation is incorrect and reparameterization of the ATP molecule is needed, to ensure that the original conformation from the crystal structure ACP is preserved. We also noticed that there was a charge imbalance in the active site that could possibly cause this conformational change. The initial simulation based on the crystal structure involved a single Mg^{2+} cation only in the active site; however, we observed that a second cation constantly occupied a position very close to the α and β-phosphates (Figure S7B). We conducted subsequent MD simulations where the ATP molecule was fixed at its original conformation. The rest of the ions, water, and the protein were free to move without any constraints. In this simulation, we also observed that K^+ ions approached the active site for the full simulation time as before, and were located in the same position where a second Mg^{2+} cation was found in the crystal structures of the homologous SERCA active site TS-like crystal structure[29] (Figure 3A). Throughout the duration of the MD simulations, the K^+ ions remained a constant presence at the active site (Figure 3B).

As crystallographic structures often lack catalytically essential Mg^{2+} ions, we propose that the second Mg^{2+} ion could be needed to stabilize the catalytically active conformation of the ATP. Two Mg^{2+} ions have been resolved in the structures of homologous enzymes, however, only in TS-like states, most likely because the chain of the synthetic derivative analogue has a different charge than the catalytically active ATP.

QM Cluster Calculations. To probe the catalytically active conformation of the ATP molecule at the active site, and, specifically, the second Mg^{2+} ion coordination, we performed QM cluster calculations. We generated four different ATP–Mg^{2+} starting geometries, G1–G4 (Figure S4), by using information from the crystal structures containing two active site cations and ADP[27,29,31] (G1) and from preliminary results of the initial MD simulations with one Mg^{2+} in the active site (G3) or by manual initial guess (G2 and G4). All four starting geometries G1–G4 were optimized and upon convergence

https://doi.org/10.1021/acs.jpcb.1c05337
J. Phys. Chem. B 2021, 125, 11835–11847

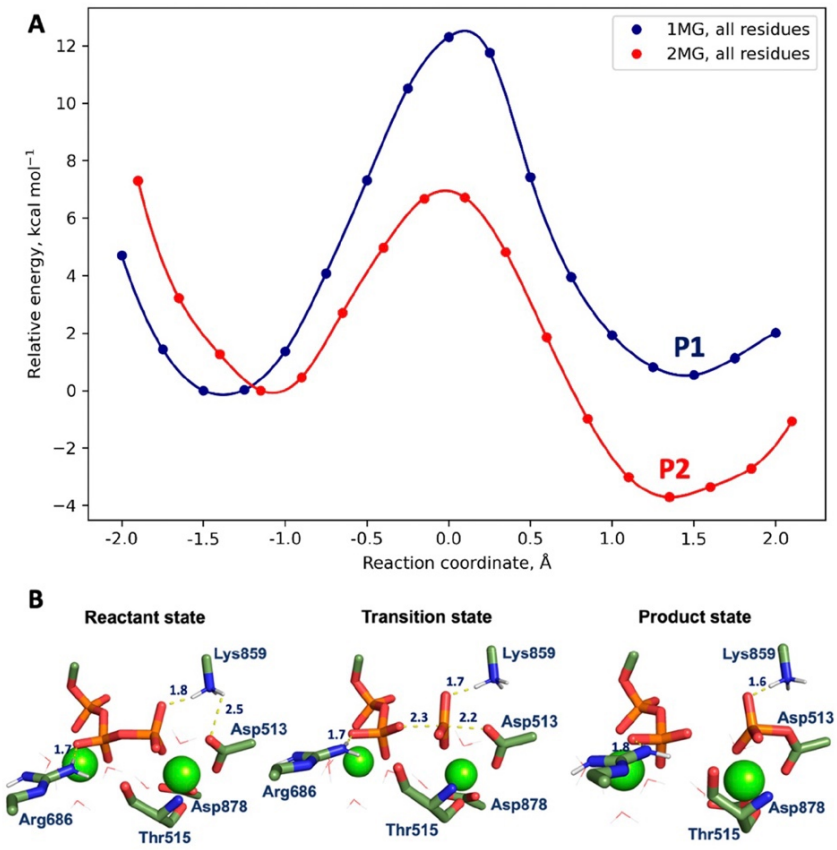


Figure 5. (A) Reaction coordinate scan with 1 (blue) and 2 (red) Mg^{2+} ions in the active site of ATP13A2. The reactant state is set to 0. (B) Reactant, transition state, and product state during the ATP cleavage reaction with two Mg^{2+} ions in the active site (green sticks). Distances are shown between the Lys859 and Arg686 and the ATP molecule. The $O_{2D}-P_G-O_{3B}$ distance is shown in the TS state. Hydrogens on the rest of the amino acids are not shown for clarity. Throughout the RCS, the Mg^{2+} ion coordinating Asp513 is always bound to two water molecules and the Mg^{2+} ion, which is coordinating the α- and β-phosphates of the ATP chain, coordinates four water molecules. Only atoms important for the reaction are shown, for clarity purposes. The full description of the system can be found in the Methods.

yielded three distinct conformations. The optimized structures of G1 and G2 (Figure 4, green sticks) are most energetically favorable and agree particularly well with the conformation and coordination mode observed in the crystal structures containing two Mg^{2+} ions and ADP[29] (Figure 4, gray sticks), where the second Mg^{2+} ion is coordinating only two oxygen atoms coming from the α and β phosphate of the ATP phosphate chain, and four water molecules. Importantly, the phosphate chain in both G1 and G2 is not in a straight conformation such as in geometries G3 and G4. Conformations G3 and G4 in which the phosphate chain is "straight" (Figure S4, parts C and D) have considerably higher energy and were therefore not used in any further QM/MM calculations. Additionally, upon aligning the optimized geometries G3 and G4 with "straight" chain to the crystal structure,[29] larger deviations are clearly visible (Figure S8, parts C and D). Conformations in which the second Mg^{2+} cation is coordinated by three oxygen atoms are not favorable either (Figure S4B) and converge to the coordination mode observed in G1 and G2, Figure 4. This ATP conformation and coordination mode were subsequently used for QM/MM calculations to determine the corresponding reaction profile.

QM/MM Calculations. For the QM/MM calculations, six systems (P1−P6) were created using the same number and species of atoms in the MM region. However, we varied the number of atoms in the QM region to explore various effects related to the role of the number of Mg^{2+} ions, and the Lys859 and Arg686 residues for the reaction. System P1 included one Mg^{2+} ion, six water molecules, the phosphate chain of ATP, the side chain of Asp513, the side chain of Asp878 and Thr515, the side chain of Lys859, Arg686, and Gly516 and the main chain of Lys514 in the QM region. System P2 contained the same number and atom species but also a second Mg^{2+}. P3 and P4 (Table S1) contain one Mg^{2+} ion in the active site and Lys859 or Arg686, respectively, are removed from the QM region and their MM atomic charges are set to 0. P5 and P6 contain two Mg^{2+} ions in the active site (Table S1) with Lys859 or Arg686, respectively, electrostatically removed from the system as detailed above. Reaction scans were performed on all six systems, P1−P6, to determine the potential energy barriers. We performed forward and backward scans until we observed consistent energy profiles (Figure 5A) and the energy minimum was obtained. Both scans show a good convergence. For P1, in the presence of one Mg^{2+} only, the barrier height is ~12.5 kcal mol^{-1}. During the phosphoryl transfer reaction, the

https://doi.org/10.1021/acs.jpcb.1c05337
J. Phys. Chem. B 2021, 125, 11835−11847

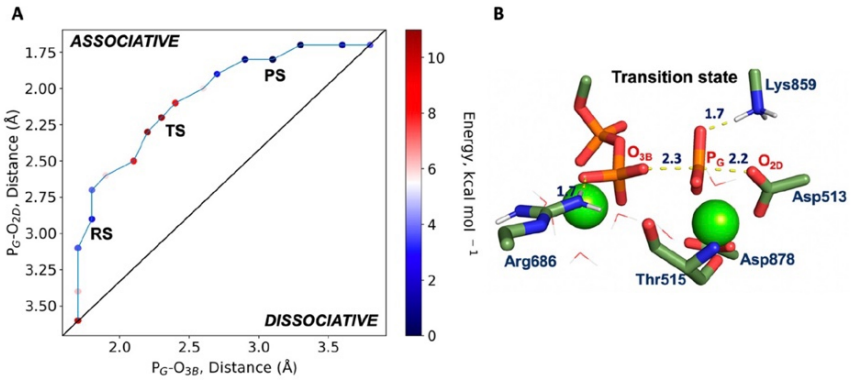


Figure 6. (A) Path adopted for the ATP cleavage. It is associative, based on the distance evolution from the reacting nucleophile O_{2D} to P_G and P_G−O_{3B} (leaving group), respectively. Each data point is colored based on the relative energy. (B) Details of the transition state structure for the reaction. Key residues at the active site (sticks and spheres for Mg^{2+}) and relevant distances (yellow dashed lines) are shown.

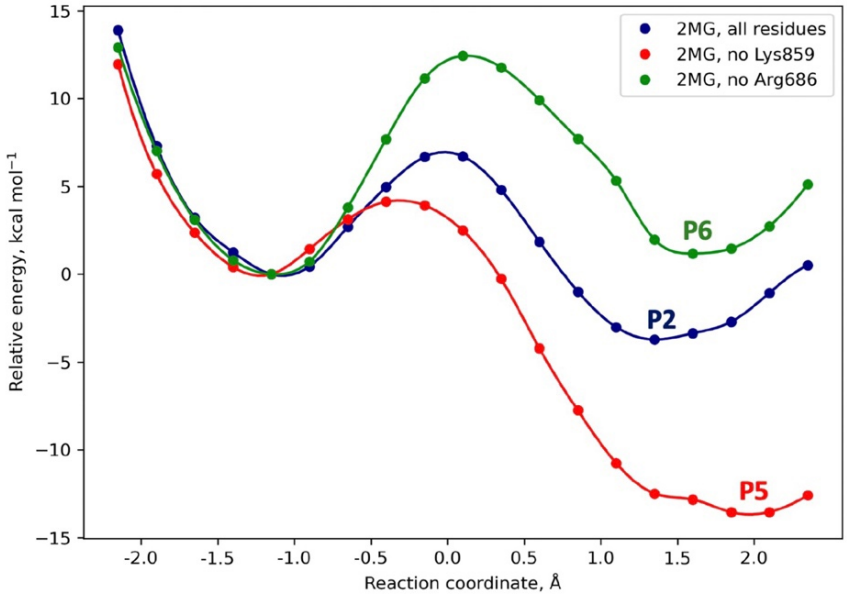


Figure 7. QM/MM reaction coordinate scans of the catalytic reaction of ATP13A2 containing two Mg^{2+} ions. The energy profile depicted in red represents the phosphate transfer reaction without Lys859 (P5). The energy profile in green represents the phosphate transfer reaction without Arg686 (P6). The blue energy profile (P2) shows the reaction with the full QM region present.

ion is octahedrally coordinated via two water molecules, the γ-phosphate of the ATP phosphate chain, the carboxylate side chains of Asp878 and Asp513 (in a monodentate mode), and the main chain carbonyl of Thr515 (Figure 5B). The barrier height for system P2, containing two Mg^{2+} ions, was calculated to be ∼7.5 kcal mol^{-1}. The second Mg^{2+} ion coordinates four water molecules and two oxygen atoms coming from the α- and β-phosphates of the ATP chain. Importantly, it does not coordinate a third oxygen atom from the phosphate chain.

Kinetic studies of the Ca^{2+}-transporting sarcoplasmic reticulum (SR), which has a fully conserved active site with ATP13A2, report a rate constant of 225 s^{-1} for the phosphorylation of the wild type enzyme by ATP in equilibrium conditions.[54] This experimental report agrees with the work of Petithory et al.,[55] which reports a rate constant for formation of the phosphorylated enzyme of 220 s^{-1}. This corresponds to a barrier height of ∼14.3 kcal mol^{-1} using the Eyring equation at 298 K. Inesi et al. reported experiments with a somewhat slower rate (100−150 s^{-1}).[56] Our potential energy barrier height obtained for the one Mg^{2+} case is in excellent agreement with these kinetic experiments. All experimental work agrees that the enzyme phosphorylation reaction is fast, and it is not the rate limiting step, with the expected free energy barrier ranging from ∼14−15 kcal/mol^{-1}. Based on the barriers obtained from our potential energy scans, it is possible that the enzyme operates and is sufficiently active already with only one Mg^{2+} ion. Missing entropic effects and free energy calculations with conformational sampling could also account for the relatively lower potential energy barrier observed using two Mg^{2+} ions in the active site.

We find that the phosphoryl transfer reaction proceeds without any stable pentavalent phosphate intermediate. The

https://doi.org/10.1021/acs.jpcb.1c05337
J. Phys. Chem. B 2021, 125, 11835−11847

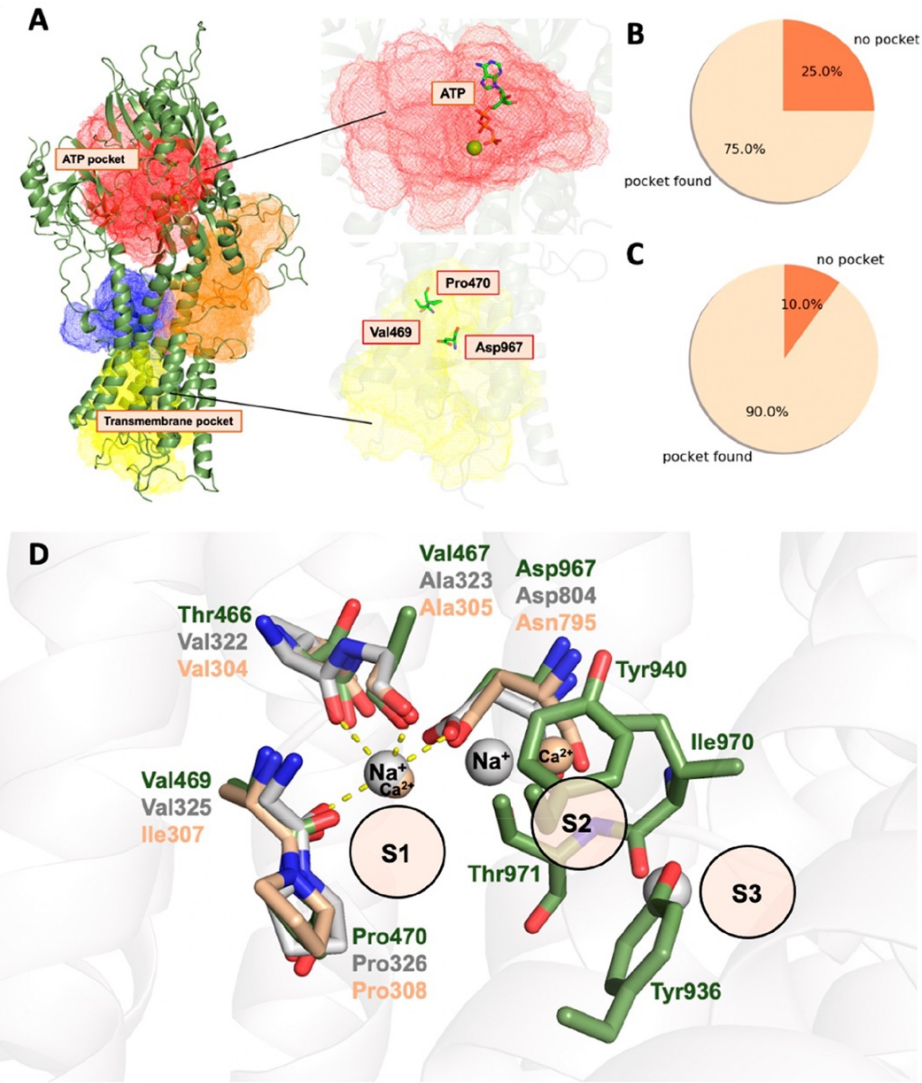


Figure 8. (A) Two main binding pockets (red mesh and yellow mesh) found consistently on the surface of the homology model of ATP13A2 (green cartoon). (B) Frequency of ATP-pocket occurrence calculated from the 20 frames of the unconstrained MD simulations. (C) Frequency of transmembrane pocket occurrence calculated for the same frames. (D) Sequence conservation in the inorganic ion-binding region in the transmembrane between the Na^+/K^+–ATPase (gray sticks), SERCA (wheat sticks), and ATP13A2 (green sticks). S1, S2, and S3 stand for binding sites 1, 2 and 3.

reaction pathway in phosphoryl transfer reactions can be classified as associative, dissociative or concerted.[57,58] In the associative pathway, the attacking nucleophile approaches the phosphorus atom, decreasing the bond length between the attacking nucleophile and the reactive phosphorus while the bond to the leaving group simultaneously increases. In this pathway, the nucleophilic attack occurs before the departure of the leaving group. Alternatively, in the dissociative pathway, the leaving group departure precedes the nucleophilic attack. In the concerted pathway, partial bond formation and bond breaking occur simultaneously in only one step. The reaction pathway is classified based on whether the bond formation or bond cleavage dominates as the transition state is approached.[58] In this ATP cleavage mechanism of ATP13A2, there is an energetically stable leaving group, a diphosphate anion, and a dissociative pathway could be expected to potentially be more favorable.[56] To confirm, we considered the distances between O_{2D}–P_G and P_G–O_{3B} (Figure 6A). Our results show that the path from the reactant state (RS) to the product state (PS) follows an associative pathway in the protein environment, for both one and two Mg^{2+} cases. Furthermore, we note that while the reaction can follow an associative pathway, the TS itself does not have to be associative.[58,59] We calculated the bond order in the transition state and obtained Wiberg bond indices for the P_G–O_{2D} and P_G–O_{3B} bonds of 0.162 and 0.125. We therefore found that

https://doi.org/10.1021/acs.jpcb.1c05337
J. Phys. Chem. B 2021, 125, 11835−11847

while the reaction clearly follows an associative pathway (Figure 6A), the TS itself corresponds to a loose, dissociative structure, where both P−O bonds are already broken (Figure 6B). This is consistent both with computational work demonstrating an associative path[57] and with experimental findings that can capture the loose TS.[60]

To quantify the effects of the interactions between the ATP and Lys859 or Arg686 (P3−P6), we obtained converged reaction coordinate scans by eliminating these interactions (Figure 7). This allows us to observe the reaction coordinate without the stabilizing electrostatic interactions formed between Arg686 and the β-phosphate of the ATP chain; and between Lys859 and both the γ-phosphate and the Asp513. Without Arg686, the barrier height increases by more than 5.0 kcal mol^{-1} (Figure 7), which is very significant in terms of time scales. Without Lys859, which forms key electrostatic interactions to the side chain of Asp513 and the γ-phosphate of ATP (Figure 5B), the reactant state is considerably higher in energy than in the cases when Lys859 is present in the system. While this results in a lower apparent barrier, the reactant state is very destabilized and the ATP binding is likely impaired, which might have additional consequences for the enzyme function and stability.

We performed additional QM/MM calculations to quantify the destabilization of the system caused by the missing Lys859 atomic charges. We calculated the single point energies for the RS, TS, and PS optimized geometries from system P2, but placing the Lys859 in the MM region with (blue) and without (red) its atomic charges present (Figure S11A). This allows us to separate the electrostatic effects of Lys859 from the geometric ones keeping the reaction profile unchanged. Similarly, QM/MM single point calculations were performed for RS, TS, and PS states along the P5 profile with (blue) and without (red) the Lys859 (Figure S11B). These geometries account for some partial geometry relaxation when Lys859 atomic charges were set to zero in the MM region.

Our results demonstrate that the energies of all states are higher without the positively charged Lys859 (Figure S11). The RS is destabilized most significantly compared to TS and PS, and thus, Lys589 is not directly involved in the TS stabilization. Lys859 is likely needed for proper binding of the ATP in the active site, providing overall stabilization to the γ-phosphate of the ATP and the Asp513 residue, through stable hydrogen bonding.

Our observation that the Lys859 residue is very important for the phosphoryl transfer reaction is supported by the kinetic analysis of mutants of the homologous Ca^{2+}−ATPase sarcoplasmic reticulum by Sorensen et al.[54] This work shows that the rate of ATP binding and subsequent phosphoryl transfer in the Lys684Arg mutant (corresponding to Lys859Arg in ATP13A2), was reduced 50-fold, relative to the wild type, thus indicating the importance of this residue. This information combined with our analysis shows that the structural effects are also very important in case of the Lys859Arg substitution, as the electrostatic factors here are unchanged, which we predicted to lower the barrier; therefore, the overall geometrical changes significantly slow down the reaction.

Two-Mg^{2+} Active Site Simulations. We also performed MD simulations with two Mg^{2+} ions present in the active site to probe the overall stability of the system, and in particular the ATP conformation. Starting from our QM cluster optimized ATP conformation, we performed three replicates of unbiased MD simulations, each 100 ns in duration. In all three simulations, the ATP was still found in its most energetically favorable conformation. No K^+ ions approached the active site. The second Mg^{2+} ion was stable at its original position coordinating two oxygen atoms of the α- and β-phosphates of the ATP. However, the catalytic Mg^{2+} did not preserve its coordination fully, and one of the water molecules was replaced by an oxygen atom from Asp513 leading to a bidentate coordination. Figure S12 illustrates the ATP conformation in the simulations after 100 ns of unconstrained MD.

This suggests that no additional parametrization is required to stabilize the ATP conformation in the presence of two active site cations. However, the metal ion coordination is notoriously difficult to maintain in some cases using standard force fields, and either further changes are needed in the active site geometry that stabilize the catalytically competent Mg-coordination, or improvements in the Mg^{2+} force field.

Binding Pockets and Transmembrane Binding Analysis. While the P5A ATPase ATP13A1 has been shown to be a Mn-transporter,[61] it has recently been demonstrated that ATP13A2 is strongly implicated in polyamine export.[22] It has been known that P5B ATPases such as ATP13A2 likely transport different cargo from ATP13A1 due to major differences in the transmembrane domain sequence conservation.[62] To identify binding regions on the surface of ATP13A2, pocket analysis was performed on 20 equidistantly spaced frames from the MD trajectories of the modeled protein, each frame being 5 ns apart from the next one. The four biggest pockets were analyzed for each frame. Not surprisingly, for most of the frames the biggest pocket in terms of surface area is the one where ATP binds in the N-domain (Figure 8A). It is expected that this pocket will be conserved for all homologous P-type ATPases due to the highly conserved ATP binding region and mechanism in the active site.

Interestingly, the second biggest pocket, which appears consistently, was identified in the transmembrane region (Figure 8A), more specifically, where inorganic ion binding has been observed in the crystal structures of homologous enzymes such as the P2C Na^+/K^+−ATPase[28] and the P2A ATPase SERCA.[26] The frequency of occurrence of the two biggest pockets is shown in Figure 8B and C. For the Na^+/K^+−ATPase, the specific amino acid scaffold surrounding the three Na^+ ions which bind in the transmembrane domain consists of: Val322, Ala323, Val325, Pro326, Glu327, Tyr771, Thr774, Ser755, Asn776, Glu 779, Asp804, and Gln924. From this sequence motif, we observe conservation in ATP13A2 for the amino acids that bind one of the inorganic ions in Site 1 (S1), Figure 8D. In ATP13A2, these correspond to Val469, Pro470, and Asp967 (Figure 8D). The remaining amino acids that coordinate the additional two Na^+ ions in the Na^+/K^+−ATPase (Sites 2 and 3, Figure 8D) are not conserved in ATP13A2. Val469 and Pro470 in S1 are also conserved between ATP13A2 and SERCA, which transports Ca^{2+} (Figure 8D), but the amino acids coordinating the second ion in SERCA and Na^+/K^+−ATPase (S2 and S3) are not conserved in ATP13A2. Considering the amino acid conservation in the ion-binding region of the transmembrane (Figure 8D), and overall similar shape of the ion-binding scaffold (Figure 8D) it is possible that ATP13A2 also can bind and transport one inorganic ion, although this is unlikely to be either Na^+ or Ca^{2+}. Additionally, this transmembrane pocket was consistently

https://doi.org/10.1021/acs.jpcb.1c05337
J. Phys. Chem. B 2021, 125, 11835−11847

(90% of the time of the analyzed frames) found within 1.5 Å of ion-binding amino acids (Val469, Pro470, Asp967) in the S1 region, which additionally supports the idea that the protein binds cargo in this area. It is important to note that the surface area of this pocket is considerably larger than expected from an ion binding site alone—the average size of the transmembrane pocket is 2264.61 $Å^3$. This suggests that this region of the protein additionally could interact with a much bigger substrate, likely in the S2 and S3 region, which is not conserved between the Na^+/K^+−ATPase or SERCA, Figure 8D.

Two additional pockets also appeared frequently, however, though not as consistently as the ATP pocket and the main transmembrane pocket. These include a considerably smaller pocket in the upper part of the transmembrane (Figure S5) and an additional pocket around residues 985−995 and 800−807 (Figure S5).

CONCLUSIONS

In this work, we present a structural model of the P5B enzyme ATP13A2 which has been complemented with MD, QM cluster, and QM/MM calculations. This has allowed us to find an accurate conformation for the catalytically competent ATP structure and a reliable position for the second Mg^{2+} ion with respect to the ATP and the other Mg^{2+} active site cation. Using this information, we have subsequently calculated the barrier height for the phosphoryl transfer with one and two Mg^{2+} ions. Additionally, we present the first quantitative analysis of the role of Arg686 and Lys859 on the barrier height of the ATP cleavage. This work can suggest how missense mutations close to important active site interactions in the respective catalytic domains can have a diminishing effect on the catalytic activity of the enzyme.

Additionally, we have analyzed the surface of the protein for binding pockets and found two pockets that occur consistently. The pocket that appeared most consistently was found in the transmembrane domain of the protein. From the sequence analysis performed and the binding pocket calculations, we suggest that ATP13A2 also likely binds a substrate in this part of the transmembrane, other than an inorganic ion, which is consistent with the big size of the calculated pocket in this part of the transmembrane.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.1c05337.

Figures S1−12, showing alignments, data from the MD, pocket analysis, QM structures before and after optimization, transition states for systems P5 and P6, individual potential energy scans, Lys859 destabilization plots, and two-Mg^{2+} MD snapshots, and Tables S1−S3, giving the energy of the optimized geometries, details of QM/MM systems, and additional pocket information from the binding pocket calculations (PDF)

AUTHOR INFORMATION

Corresponding Author

Edina Rosta − *Department of Chemistry, Faculty of Natural & Mathematical Sciences, King's College London, London SE1 1DB, U.K.; Department of Physics and Astronomy, Faculty of Maths & Physical Sciences, University College London, London WC1E 6BT, U.K.;* orcid.org/0000-0002-9823-4766; Email: e.rosta@ucl.ac.uk

Authors

Teodora Mateeva − *Department of Chemistry, Faculty of Natural & Mathematical Sciences, King's College London, London SE1 1DB, U.K.*

Marco Klähn − *Department of Materials Science and Chemistry, Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore 138 632, Singapore*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.1c05337

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We are grateful to Dr Tamás Földes for numerous discussions. We would also like to thank Dr Attila Csikász-Nagy for initial discussions. E.R. acknowledges funding from the ERC (Project 757850 BioNet). T.M. acknowledges funding from the Agency for Science, Technology and Research (A*STAR) Singapore Research Attachment Programme (ARAP) and King's College London's Centre for doctoral studies. We acknowledge the use of the research computing facility at King's College London, Rosalind (https://rosalind.kcl.ac.uk).

REFERENCES

(1) Park, J. S.; Blair, N. F.; Sue, C. M. The role of ATP13A2 in Parkinson's disease: Clinical phenotypes and molecular mechanisms. *Mov. Disord.* **2015**, *30*, 770−9.

(2) van Veen, S.; et al. Cellular function and pathological role of ATP13A2 and related P-type transport ATPases in Parkinson's disease and other neurological disorders. *Front. Mol. Neurosci.* **2014**, *7*, 1−22.

(3) Martin, S.; Holemans, T.; Vangheluwe, P. Unlocking atp13a2/park9 activity. *Cell Cycle* **2015**, *14*, 3341−3342.

(4) Holemans, T.; et al. A lipid switch unlocks Parkinson's disease-associated ATP13A2. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 9040−9045.

(5) Podhajska, A.; et al. Common pathogenic effects of missense mutations in the P-type ATPase ATP13A2 (PARK9) associated with early-onset parkinsonism. *PLoS One* **2012**, *7*, No. e39942.

(6) Bublitz, M.; Morth, J. P.; Nissen, P. P-type ATPases at a glance. *J. Cell Sci.* **2011**, *124*, 3917.

(7) Brüggemann, N.; et al. Recessively inherited parkinsonism: Effect of ATP13A2 mutations on the clinical and neuroimaging phenotype. *Arch. Neurol.* **2010**, *67*, 1357−63.

(8) Schneider, S. A.; et al. ATP13A2 mutations (PARK9) cause neurodegeneration with brain iron accumulation. *Mov. Disord.* **2010**, *25*, 979−84.

(9) Odake, Y.; et al. Identification of a novel mutation in ATP13A2 associated with a complicated form of hereditary spastic paraplegia. *Neurol. Genet.* **2020**, *6*, No. e514.

(10) Di Fonzo, A.; et al. ATP13A2 missense mutations in juvenile parkinsonism and young onset Parkinson disease. *Neurology* **2007**, *68*, 1557−62.

(11) Estrada-Cuzcano, A.; et al. Loss-of-function mutations in the ATP13A2/PARK9 gene cause complicated hereditary spastic paraplegia (SPG78). *Brain* **2017**, *140*, 287−305.

(12) Lin, C. H.; et al. Novel ATP13A2 variant associated with Parkinson disease in Taiwan and Singapore. *Neurology* **2008**, *71*, 1727−32.

(13) Vilariño-Güell, C.; et al. ATP13A2 variability in Parkinson disease. *Hum. Mutat.* **2009**, *30*, 406−10.

(14) Santoro, L.; et al. Novel ATP13A2 (PARK9) homozygous mutation in a family with marked phenotype variability. *Neurogenetics* **2011**, *12*, 33−39.
(15) Burroughs, A. M.; Allen, K. N.; Dunaway-Mariano, D.; Aravind, L. Evolutionary Genomics of the HAD Superfamily: Understanding the Structural Adaptations and Catalytic Diversity in a Superfamily of Phosphoesterases and Allied Enzymes. *J. Mol. Biol.* **2006**, *361*, 1003−34.
(16) Koonin, E. V.; Tatusov, R. L. Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity. Application of an iterative approach to database search. *J. Mol. Biol.* **1994**, *244*, 125−132.
(17) Aravind, L.; Galperin, M. Y.; Koonin, E. V. The catalytic domain of the P-type ATPase has the haloacid dehalogenase fold. *Trends Biochem. Sci.* **1998**, *23*, 127−9.
(18) Abbas, M. M.; Govindappa, S. T.; Sheerin, U. M.; Bhatia, K. P.; Muthane, U. B. Exome Sequencing Identifies a Novel Homozygous Missense ATP13A2Mutation. *Mov. Disord. Clin. Pract.* **2017**, *4*, 132−135.
(19) Narayanaswamy, N.; et al. A pH-correctable, DNA-based fluorescent reporter for organellar calcium. *Nat. Methods* **2019**, *16*, 95−102.
(20) De La Hera, D. P.; Corradi, G. R.; Adamo, H. P.; De Tezanos Pinto, F. Parkinson's disease-associated human P5B-ATPase ATP13A2 increases spermidine uptake. *Biochem. J.* **2013**, *450*, 47−53.
(21) Heinick, A.; et al. Caenorhabditis elegans P 5B -type ATPase CATP-5 operates in polyamine transport and is crucial for norspermidine-mediated suppression of RNA interference. *FASEB J.* **2010**, *24*, 206−217.
(22) van Veen, S.; et al. ATP13A2 deficiency disrupts lysosomal polyamine export. *Nature* **2020**, *578*, 419−424.
(23) Thever, M. D.; Saier, M. H. Bioinformatic characterization of P-Type ATPases encoded within the fully sequenced genomes of 26 eukaryotes. *J. Membr. Biol.* **2009**, *229*, 115−130.
(24) Toyoshima, C.; Nakasako, M.; Nomura, H.; Ogawa, H. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 Å resolution. *Nature* **2000**, *405*, 647−655.
(25) Toyoshima, C.; Mizutani, T. Crystal structure of the calcium pump with a bound ATP analogue. *Nature* **2004**, *430*, 529−35.
(26) Sacchetto, R.; et al. Crystal structure of sarcoplasmic reticulum Ca 2+-ATPase (SERCA) from bovine muscle. *J. Struct. Biol.* **2012**, *178*, 38−44.
(27) Clausen, J. D.; McIntosh, D. B.; Vilsen, B.; Woolley, D. G.; Andersen, J. P. Importance of conserved N-domain residues Thr441, Glu442, Lys515, Arg560, and Leu562 of sarcoplasmic reticulum Ca2+-ATPase for MgATP binding and subsequent catalytic steps. Plasticity of the nucleotide-binding site. *J. Biol. Chem.* **2003**, *278*, 20245−58.
(28) Ogawa, H.; Shinoda, T.; Cornelius, F.; Toyoshima, C. Crystal structure of the sodium-potassium pump (Na+,K + -ATPase) with bound potassium and ouabain. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 13742−13747.
(29) Kanai, R.; Ogawa, H.; Vilsen, B.; Cornelius, F.; Toyoshima, C. Crystal structure of a Na + -bound Na + ,K + -ATPase preceding the E1P state. *Nature* **2013**, *502*, 201−6.
(30) McKenna, M. J.; et al. The endoplasmic reticulum P5A-ATPase is a transmembrane helix dislocase. *Science (Washington, DC, U. S.)* **2020**, *369*, No. eabc5809.
(31) Sørensen, T. L. M.; Møller, J. V.; Nissen, P. Phosphoryl transfer and calcium ion occlusion in the calcium pump. *Science (Washington, DC, U. S.)* **2004**, *304*, 1672−5.
(32) Komuro, Y.; Re, S.; Kobayashi, C.; Muneyuki, E.; Sugita, Y. CHARMM force-fields with modified polyphosphate parameters allow stable simulation of the ATP-bound structure of Ca2+-ATPase. *J. Chem. Theory Comput.* **2014**, *10*, 4133−4142.
(33) Li, C.; Yue, Z.; Espinoza-Fonseca, L. M.; Voth, G. A. Multiscale Simulation Reveals Passive Proton Transport Through SERCA on the Microsecond Timescale. *Biophys. J.* **2020**, *119*, 1033−1040.
(34) Krachtus, D.; Smith, J. C.; Imhof, P. Quantum mechanical/molecular mechanical analysis of the catalytic mechanism of phosphoserine phosphatase. *Molecules* **2018**, *23*, 3342.
(35) Nagy, G. N.; et al. Structural Characterization of Arginine Fingers: Identification of an Arginine Finger for the Pyrophosphatase dUTPases. *J. Am. Chem. Soc.* **2016**, *138*, 15035−15045.
(36) Biasini, M.; et al. SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **2014**, *42*, W252−W258.
(37) Bateman, A.; et al. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480−D489.
(38) Phillips, J. C.; et al. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781−802.
(39) Pastor, R. W.; MacKerell, A. D. Development of the CHARMM force field for lipids. *J. Phys. Chem. Lett.* **2011**, *2*, 1526−1532.
(40) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(41) MacKerell, A. D.; et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(42) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* **2008**, *29*, 1859−65.
(43) Kräutler, V.; Van Gunsteren, W. F.; Hünenberger, P. H. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J. Comput. Chem.* **2001**, *22*, 501−508.
(44) Lomize, M. A.; Lomize, A. L.; Pogozheva, I. D.; Mosberg, H. I. OPM: Orientations of proteins in membranes database. *Bioinformatics* **2006**, *22*, 623−5.
(45) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; et al. *Gaussian 09*, Rev. D.01; Gaussian, Inc.: Wallingford, CT, 2013; DOI: 10.1017/CBO9781107415324.004.
(46) Becke, A. B3LYP. *J. Chem. Phys.* **1993**, *98*, 5648.
(47) Frisch, M. J.; et al. *Gaussian 03*, rev. B.05; Gaussian. Inc.: Pittsburgh, PA, 2003.
(48) Shao, Y.; et al. Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Mol. Phys.* **2015**, *113*, 184−215.
(49) Brooks, B. R.; et al. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545−1614.
(50) Schrödinger LLC. *The PyMOL Molecular Graphics System*, Ver. 2.4; Schrödinger LLC: 2020.
(51) Smith, R. H. B.; Dar, A. C.; Schlessinger, A. PyVOL: A PyMOL plugin for visualization, comparison, and volume calculation of drug-binding sites. *bioRxiv* 2019; DOI: 10.1101/816702. Accessed 24th October 2019.
(52) Altschul, S. F.; et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(53) Pearson, W. R.; Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.* **1988**, *85*, 2444−8.
(54) Sørensen, T. L. M.; Dupont, Y.; Vilsen, B.; Andersen, J. P. Fast kinetic analysis of conformational changes in mutants of the Ca2+-ATPase of sarcoplasmic reticulum. *J. Biol. Chem.* **2000**, *275*, 5400−5408.
(55) Petithory, J. R.; Jencks, W. P. Sequential Dissociation of Ca2+from the Calcium Adenosinetriphosphatase of Sarcoplasmic Reticulum and the Calcium Requirement for Its Phosphorylation by ATP. *Biochemistry* **1988**, *27*, 5553−5564.
(56) Inesi, G.; et al. Equilibrium and Kinetic Studies of Calcium Transport and ATPase Activity in Sarcoplasmic Reticulum. *Z. Naturforsch., C: J. Biosci.* **1982**, *37*, 685−91.
(57) Klähn, M.; Rosta, E.; Warshel, A. On the mechanism of hydrolysis of phosphate monoesters dianions in solutions and proteins. *J. Am. Chem. Soc.* **2006**, *128*, 15310−15323.

https://doi.org/10.1021/acs.jpcb.1c05337
J. Phys. Chem. B 2021, 125, 11835−11847

(58) Kamerlin, S. C. L.; Wilkie, J. The effect of leaving group on mechanistic preference in phosphate monoester hydrolysis. *Org. Biomol. Chem.* **2011**, *9*, 5394−5406.
(59) Lopata, A.; et al. Mutations decouple proton transfer from phosphate cleavage in the dutpase catalytic reaction. *ACS Catal.* **2015**, *5*, 3225−3237.
(60) Zalatan, J. G.; Herschlag, D. Alkaline phosphatase mono- and diesterase reactions: Comparative transition state analysis. *J. Am. Chem. Soc.* **2006**, *128*, 1293−1303.
(61) Cohen, Y.; et al. The yeast P5 type ATPase, Spf1, regulates manganese transport into the endoplasmic reticulum. *PLoS One* **2013**, *8*, No. e85519.
(62) Sørensen, D. M.; Buch-Pedersen, M. J.; Palmgren, M. G. Structural divergence between the two subgroups of P5 ATPases. *Biochim. Biophys. Acta, Bioenerg.* **2010**, *1797*, 846−855.

https://doi.org/10.1021/acs.jpcb.1c05337
J. Phys. Chem. B 2021, 125, 11835−11847

Future Work

Now that there is an established profile for the wild type enzyme, it would be possible to repeat the QM/MM calculations for mutated variants where a mutation implicated in neurodegenerative process is identified in proximity to the active site and can be incorporated into the QM region. The now available crystal structures with the bound substrate and Mg^{2+} ions fully agree with our model (Fig. 3.9), which validates the QM/MM potential energy scans and all of our prior computational calculations. Figure 3.10 shows the mutations with to-be-determined mode of action as red sticks on the 3D crystal structure (now resolved) of ATP13A2 (PDB code: 7N73).[114] It can be seen that some of these mutations are located in the P and N domains, respectively, are near the active site and possibly interfere with the catalytic mechanism.

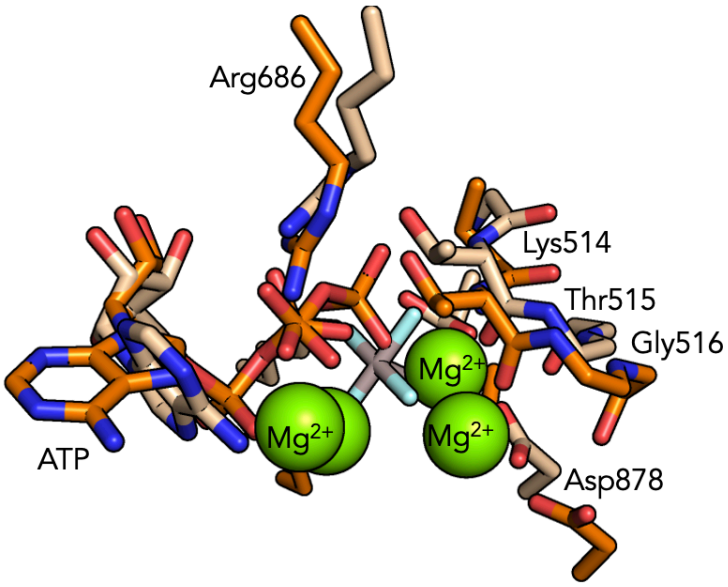


Figure 3.9. Starting structure for the QM/MM potential energy scans (wheat sticks) and the crystal structure of ATP13A2 (orange sticks) which was resolved after our model (PDB code: 7N75).[114] The position of the second Mg^{2+} ion perfectly matches our model, as well as the other residues we modeled. The crystal structure is resolved with AlF_3 which represents the transition from the E1 to E1P state, whereas our model represents the E1 conformational state immediately after ATP binding, so slight misalignment is to be expected.

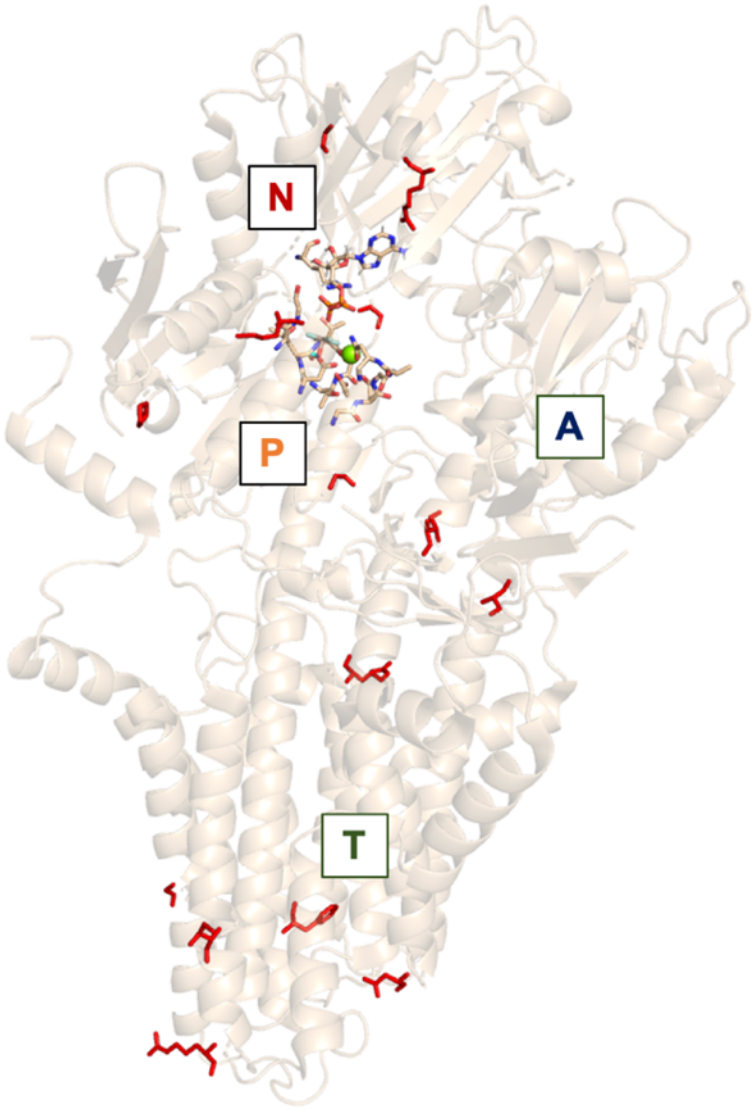


Figure 3.10. Crystal structure of ATP13A2 with red sticks depicting mutations whose effect is to be determined. Some of these mutations are located in the P and N domains, respectively, and can potentially interfere with the ATPase activity of the enzyme.

Chapter 4

Machine Learning Classification Pipeline for Galactose Oxidase Variants based on Transition State Molecular Dynamics

4.1 Introduction

Galactose Oxidase (GO) is a monomeric copper-containing oxidoreductase found in several fungal species.[115–117] It oxidizes the C6-OH hydroxyl of the primary alcohol D-galactose, converting it to its corresponding aldehyde. The wild type (WT) intracellular GO catalyzes the oxidation of only a narrow range of substrates such as D-galactose and galactose-containing oligosaccharides. However, in the last 20 years, the scope of substrates for GO was significantly expanded using engineered variants of the enzyme, with the capability of converting a range of primary alcohols.[118] The substrate scope of GO was also extended to secondary alcohols, [13] which is of key importance for the pharmaceutical industry due to their role in the synthesis of various compounds with global healthcare impact. Most recently, the range of substrates was further expanded to include multiple bulky benzylic alcohols with large side chains.[119]

In the field of GO engineering, variants that retain the optimal catalytic properties of the WT enzyme while accommodating a broader range of benzylic substrates are of key importance. It is often observed, however, that there exists a trade-off between the expanded substrate specificity of a variant and its catalytic capability. Most often, GO variants with widened substrate specificity, do not retain the optimal catalytic rate of the WT enzyme. This work proposes a new way of utilizing Molecular Dynamics (MD) and Machine Learning (ML) to develop a classification pipeline for GO variants based on their predicted catalytic performance. The main aim is to be able to predict the effect on the catalytic rate of a GO variant (positive/neutral or negative), upon the introduction of a combination of missense

mutations. Additional complexity is added to the problem by predicting the effect on the rate from missense mutations and when there is a non-native substrate in the active site. The problem is handled as a binary classification - two classification categories are established. The first category contains all variants with rates falling within the range of the WT enzyme (± 1.0 kcal/mol). The second category contains the variants that slow down the rate of catalysis considerably (≥ 2.9 kcal/mol). The aim is to categorize the variants and predict their catalytic efficiency irrespective of whether the substrate is a non-native primary or secondary alcohol. The Gibbs free energy of activation ($\Delta G^{\ddagger}$) for the rate-limiting step in the WT GO is measured experimentally to be ~13.8 kcal/mol and this is used as a reference value.[120]

A unique feature of the GO enzyme is the presence of a free radical-coupled copper active site,[116,121,122] a property of some copper metalloenzymes, combining the reactivity of a free radical ligand with a redox-active metal centre (Figure 4.1A).[123] The catalytic reaction leading to the aldehyde product involves multiple steps with the copper ion adopting several distinct oxidation states during the process. Crystal structures of GO from different species are available,[124,125] which all display the central copper ion bound in a square pyramidal coordination to Tyr272, Tyr495, His496, His581, and either a water molecule or an azide ion, which is where the alcohol substrate binds (Figure 4.1A). Another important residue in the immediate active site is Cys228 which is linked to Tyr272 through a thioether bond and has a vital role in the catalytic reaction as mutational studies show a 1000-fold decrease in the catalytic rate in the presence of the C228G mutation.[116] Trp290 which π-stacks to the Tyr272-Cys228 moiety is also considered very important for the regulation of entry to the active site (Figure 4.1).[125] The catalytic mechanism was previously explored experimentally with extensive spectroscopic work, isotope substitution experiments[123,126], and theoretically with DFT.[127] There is a general agreement that the alcohol-to-aldehyde conversion is a complex multistep process that proceeds with a proton transfer, followed by a hydrogen atom transfer, subsequent electron transfer, O_2 binding, and reduction. The rate-limiting step (Figure 4.1B) is the hydrogen atom transfer from the substrate to the equatorial modified tyrosyl radical Tyr272. Full details on all catalytic steps can be found in the works of different authors.[123,126,127] In this work, only the rate-limiting step is considered, as we are interested in predicting the effects on the catalytic rate. The 3D structure of one of the mutated variants (M3-5) is shown in Figure 4.1C with red sticks showing missense mutations on this variant, to

illustrate the general location distribution of mutations in variants that convert a wider range of alcohols. The TS with one of the non-native substrates is shown in Figure 4.1D. The non-native alcohols are expected to bind in the active site in the same way as D-galactose, as well as other alcohols with similar chemical composition. The active site in GO is not buried deep in the protein, as is the case in some other metalloenzymes.

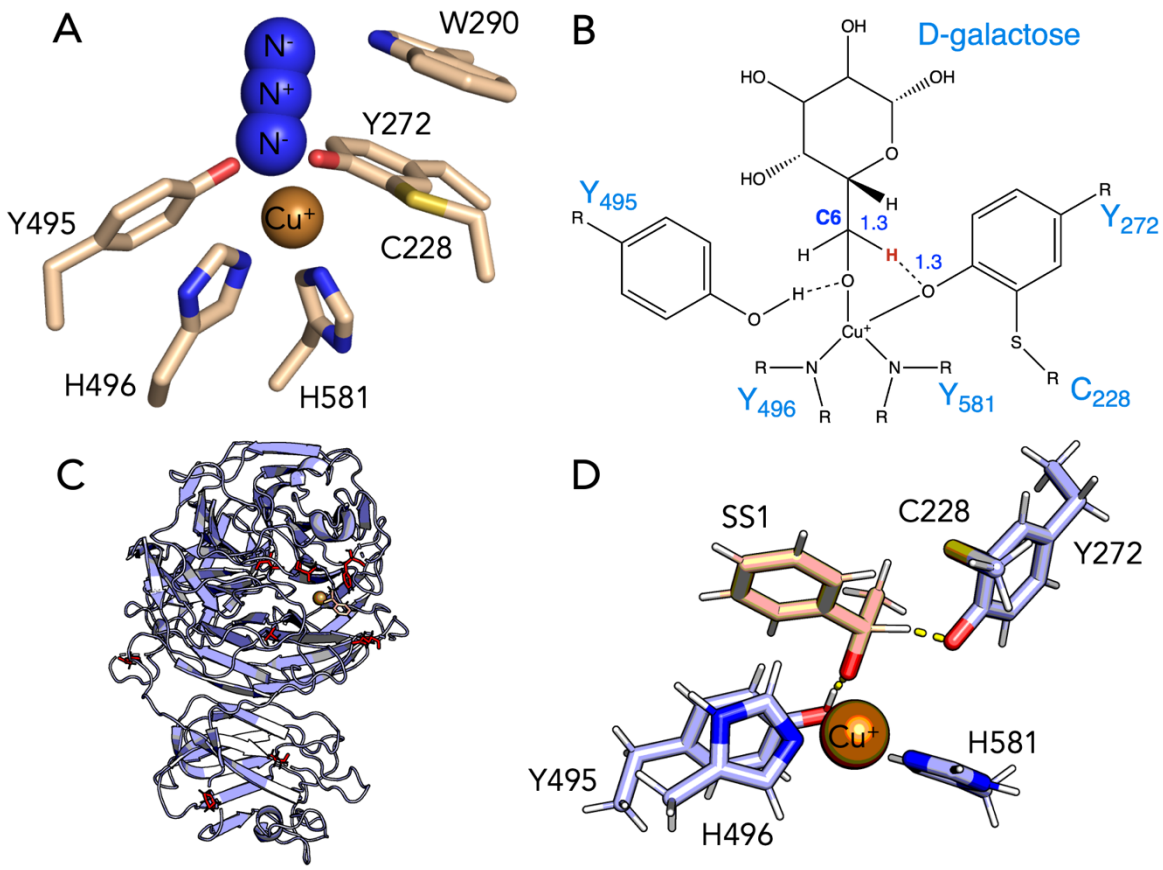


Figure 4.1. (A) Active site of the wild type crystal structure of GO (PDB code: 2EIE)[128] with copper ion bound in a square pyramidal geometry and an azide ion. (B) 2D representation of the TS for the rate-limiting step in the catalytic mechanism of the WT enzyme. The hydrogen atom that is transferred from the substrate alcohol to the tyrosyl radical is highlighted in red. Some of the hydrogen atoms are not shown for clarity. (C) 3D cartoon representation of the M35 variant of GO, mutated residues are represented with red sticks. (D) 3D representation of the TS with 1-phenylethanol (SS1) as the substrate.

The proposed pipeline involves conducting short MD simulations for 31 variants of GO at/around the rate-limiting step of the catalytic mechanism with the following substrates bound in the active site: D-galactose (primary alcohol, Figure 4.2A), 1-phenylethanol (SS1) and α-Tetrol (S128), (secondary alcohols, Figure 4.2B and C). The reason for conducting the MD simulations at/around the TS is that we are interested in predicting the effect on the catalytic rate, which is directly related to the rate-limiting step of the enzymatic reaction. The TS for the wild type enzyme with D-galactose in the active site was previously found and verified using the DFT cluster TS search approach. It should be noted that this TS was verified against the accepted catalytic mechanism and the 3D structure of the active site agreed very well.[127] The Cartesian coordinates of the TS structure were used to restrain the active site to ensure that subsequent MD simulations sampled in proximity of this rate-limiting TS. The last frames from the equilibration run were extracted and for all variants, it was verified that the protein adopted a TS-like structure before starting the production run. Throughout the MD simulations, the three-dimensional coordinates of the enzyme's main active site residues are kept in a configuration closely resembling the rate-limiting step of the catalytic mechanism through the application of restraints (added harmonic potentials), which allows to sample the dynamics near the TS, but at the same time keeps the protein at a TS-like structure. I used DFT calculations to define the charge distribution in the active site at the TS of the rate-limiting step and used the obtained partial charges to re-parameterize the FF accordingly.

This approach centers on the hypothesis that since some of the protein mutations in the simulated variants occur near the active site, these will affect the three-dimensional conformation of the TS structure at/around the rate-limiting step. Affecting the active site geometry together with changes in the active site charge distribution is expected to affect the energy barrier for the respective step in the catalytic mechanism. By starting the MD simulations at the TS conformation of the enzyme, I aim to capture potential displacements of the active site relative to the WT GO enzyme. Mutations in variants that slow down the reaction considerably are expected to distort the active site more significantly. The underlying assumption is that variants with similar catalytic rates to the WT enzyme will experience little to no displacement during the rate-limiting step, or similar "behavior" during the MD simulations. Consequently, key distances within the active site should remain similar to the ones observed for the WT GO. We ran simulations with three replicates for each variant,

extracted features from the simulation trajectories, and averaged the values for the features over all replicas. Subsequently, we utilized Random Forest and other decision tree-based algorithms to classify each variant based on these key features obtained from the TS MD simulations.

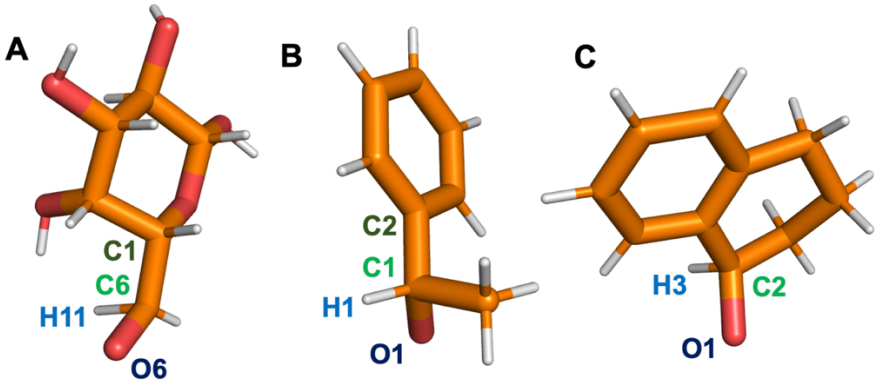


Figure 4.2. 3D structures of the alcohol substrates used in this work: (A) D-galactose, (B) 1-phenylethanol (SS1), (C) α-Tetrol (S128). Key atoms in the hydrogen transfer step are labeled. The original atom names from the PDB databank are used.

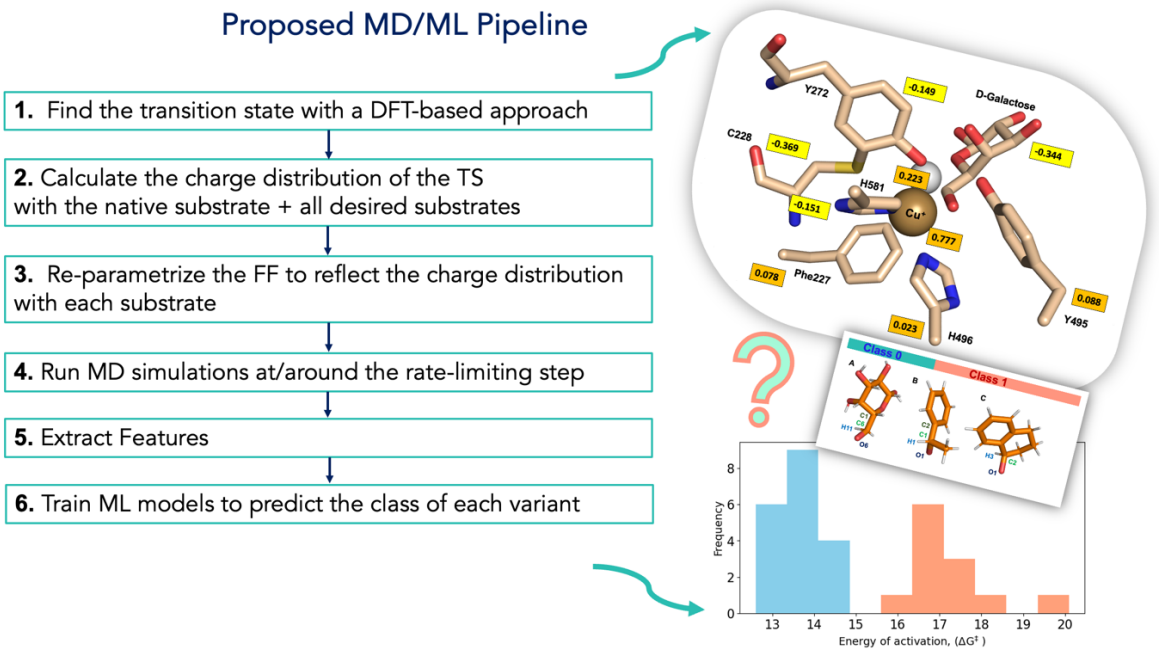


Figure 4.3. Proposed pipeline for the classification of GO variants based on their predicted catalytic rate.

4.2 Methods

4.2.1 Modeling the GO variants

The structure of the wild type Galactose Oxidase (GO) enzyme was obtained from the Protein Data Bank (PDB),[69] PDB code: 2EIE.[128] All GO mutant structures were generated with Pymol's mutagenesis tool.[129] Variants were then compared to crystal structures, where such were available, and showed excellent agreement for the side chain conformation of the mutated residue.[120,125] This confirms that the Pymol mutagenesis tool can reliably predict the most likely rotamer of the mutated residue.

To place the substrate in the active site, the 3D structure of D-galactose was aligned to the position of the azide ion in the wild type GO crystal structure,[125] and it was also aligned to the TS geometry previously obtained by a DFT-TS search. The TS was not obtained by me; therefore, the approach is not discussed in detail here. The TS structure was only used as a reference for the MD simulations and as an input for the charge distribution calculations. All subsequent substrates modeled in the active site were aligned to the C6-OH hydroxyl group and the H11 atom of D-galactose to make sure the crucial reactive hydroxyl group occupies the same space for all substrates and the starting point for the MD simulations is conserved. Any potential clashes of the surrounding residues with the non-native substrates were resolved during the minimization steps of the MD.

All variants in this work are modeled based on the crystal structure of the WT GO enzyme in Fusarium graminearum. First, the M1 variant was created which differs from the WT by five missense mutations and one silent mutation, Figure 4.4. M1 was then solvated, minimized, and equilibrated and a production run was performed according to the procedure described in the MD setup section. The last frame from the production run of the M1 variant was extracted and this structure was used for building all M1_383 variants shown in Figure 4.4 with all possible mutations at position 383. The same procedure was followed for variants

W290F, W290G, and W290H. All other variants had a different substrate in the active site (S128 or SS1, Figure 4.4), therefore, a structure of the protein already simulated with D-galactose in the active site could not be used. Instead, the WT crystal structure was used to introduce the mutations and then all variants were simulated for 20 ns, to have the same overall production run time for all variants. All features extracted from the MD simulations were used after each variant had been subjected to the same overall simulation time. The exact mutations present in each variant and the modeled substrate are summarized in Figure 4.4, as well as in Appendix B (Table SI1). Residues shown in bold orange represent new mutations that were not present in the parent variant. In-house Python and bash scripts were developed to automate the process and make the inclusion of new mutations straightforward. The same applies to the feature extraction process. All variants created from Goh1001b modeled in this work (Figure 4.4, grey background) were suggested by and come from the work of Yeo W. et al.[119]

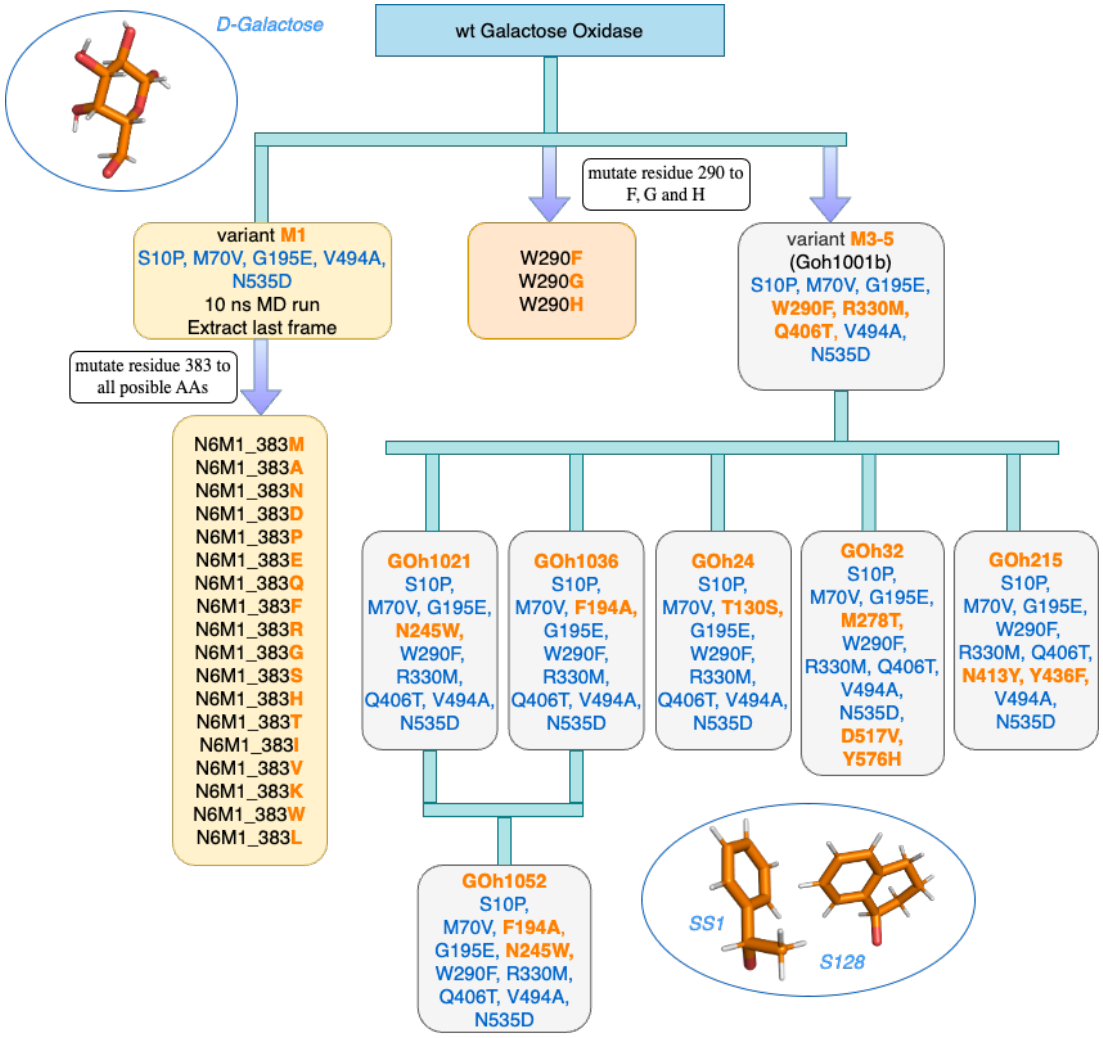


Figure 4.4. Diagram illustrating how the GO variants are modeled starting from the WT crystal structure.

4.2.2 Parametrization of the active site based on QM calculations

The charge distribution of the active site was evaluated with the electrostatic potential (ESP) calculation as implemented in Gaussian 09 ES64L-G09RevE.01.[130] All DFT ESP calculations were performed at B3LYP/Def2TZVP level of theory with GD3 empirical dispersion.[131,132] To obtain accurate charge distribution for the TS of the WT protein, the ESP calculation used the TS coordinates of the active site as input with the following residues: Y272, C228, F227, Y495, H496, H581, the D-galactose substrate, and the copper ion.

The charges obtained from DFT were then used to re-parametrize the original force field, to reflect the more accurate active site charge distribution at the rate-limiting step. D-galactose was replaced with S128 and SS1, respectively, and the charge distribution was re-calculated for each substrate. The respective charges were used in the MD simulations. The charge of each substrate and active site residue, according to the DFT calculations, are shown in Appendix B. Corrections had to be made for all the linker atoms, where applicable, which added additional charge to the system. Backbone C and O atoms were generally kept at the charge provided in the original FF library. After those corrections, it was made sure that mutual charge transfer as predicted by DFT was preserved as well as intramolecular charge polarization, while at the same time charges on protein backbone atoms were retained to keep the protein FF overall consistent.

4.2.3 Molecular Dynamics setup

All MD simulations were performed with GROMACS version 2020.6.[133] The force field used to model the systems was CHARMM36.[84] The protein was solvated in a cubic water box and the water model was TIP3P.[134] For the equilibration the Berendsen pressure coupling[90,91] was combined with the V-rescale thermostat.[94] For the production run the Parrinello-Rahman pressure coupling[94] was combined with the Nose-Hoover temperature coupling[89,92] (T = 298K). The Verlet cut-off scheme was employed to generate pair-lists and the electrostatic interactions were evaluated with the Particle Mesh Ewald.[135,136] Minimization, equilibration, and production steps were completed. The time step of the equilibration was 1 fs for a total of 50,000 steps. The time step of the production run was 2 fs for a total of 5,000,000 steps. It was further extended for another 5,000,000 steps, for a total of 20 ns simulation time for the variants directly modeled from the crystal structure of the WT GO. Considering that the simulations were conducted by enforcing the protein's active site to adopt a TS geometry, our objective was to use the shortest simulation times feasible while still capturing any active site displacement resulting from the presence of missense mutations. This approach anticipates a trade-off between applying restraints to maintain the active site at a TS closely matching the one obtained by DFT and simultaneously observing displacements of atomic positions relative to the WT enzyme simulation. The force constants were decreased iteratively until the minimum force could be used which would keep the simulations at an active site

conformation similar to the one from the DFT (see Appendix B for force constants). The active site structure was evaluated after every different production run (the force constants starting from a conservative force and decreasing, depending on the bond type). Overall, 186 MD simulations were completed. First, 93 simulations (31 variants, 3 replicas each) were run with one set of restraints (Appendix B, Table SI3 and SI4). Then, a further 93 simulations were run with a different set of restraints (Appendix B, Table SI3 and SI4). The reason for the second set of simulations was because the originally used set of restraints included substrate atoms which might not be explicitly present in future substrates. This would limit the possibility of expanding the dataset to other bulky secondary alcohols. For this reason, the simulations were re-run with the second set of restraints to make sure that the future dataset can be inclusive of substrates with a more diverse chemical composition.

4.2.4 Machine Learning

Both machine learning models utilized for classification of the variants in this work are decision tree-based algorithms. One of the models is the Random Forest (RF) algorithm. RF is based on bootstrap aggregation which means that it divides the dataset into subsets and builds trees for the different data subsets. The final classification label comes from the majority label predicted by most individual decision trees.

The other model that was tested is the Gradient Boosted Decision Trees (GBDT) algorithm which works similarly. However, it does not obtain the result from the majority vote of individual trees but rather builds one strong model from the weaker decision trees by correcting the error of each previous tree.[108,137]

All code was written in Python 3.9.6. The Random Forest Regressor, Random Forest Classifier, and Gradient Boosted Decision Trees models in this work were used with the implementation from scikit-learn 1.2.2.[110] The parameters used in the final models are described below for each case.

The loss function used for the RF Regressor is the Mean Squared Error (MSE). The MSE is calculated in the following way:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2 \qquad (4.1)$$

Where n is the number of data points, Y_i are the observed values, and $\widehat{Y}_i$ are the predicted values.

The loss function used for the RF Classifier and GBDT Classifier is the log loss. It is also known as binomial deviance or binary cross-entropy. The log loss is calculated as:

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n} y_i \cdot \log\left(p\left(y_i\right)\right) + (1-y)\log\left(1-p\left(y_i\right)\right) \qquad (4.2)$$

Where n is the number of samples, y is the binary label (0 or 1), $p(y)$ is the probability of the data point being 1 for all n samples.

For the RF regressor, RF and GBDT Classifiers the parameters were set to default in scikit-learn.

The RepeatedKFold cross validator, as implemented in scikit-learn 1.2.2,[110] was used to split the data into 100 folds. It was used for both of the described models. The parameters specified were n_splits and n_repeats. n_splits was set to 2 and n_repeats was set to 50, which results in 100 unique training and testing sets. To calculate the accuracy with different splits, n_splits was also set to 3 and n_repeats set to 50 which gives rise to 150 unique folds with 67% training and 33% testing data. The accuracy did not change considerably from the use of different ratios. The reported accuracy in the Results section is the one from 100 folds.

Metrics

The metrics used to evaluate the performance of the classification models are discussed below.

The true positive rate (TPR) represents the proportion of actual positive instances that are correctly identified by the model. It is calculated as:

$$TPR = \frac{TP}{TP + FN}$$

where TP is the number of true positives, and FN is the number of false negatives.

The false positive rate (FPR) is the proportion of actual negative instances that are incorrectly identified as positive by the model. It is calculated as:

$$FPR = \frac{FP}{FP + TN}$$

where FP is the number of false positives, and TN is the number of true negatives.

A Receiver Operating Characteristic (ROC) curve shows the trade-off between sensitivity and specificity. Classifier models that have curves that reach toward the top-left corner demonstrate more accurate classification. The Area Under the Curve (AUC) is a quantitative measure derived from the ROC curve which represents the area under the ROC curve and provides a single scalar value to assess the overall performance of a classification model. A higher AUC indicates better model performance, with a value of 1 indicating perfect classification performance and a value of 0.5 indicating random guessing.

Classification accuracy, precision, and recall are also reported to evaluate the performance of each predictive model. They are represented as:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2\left(\frac{precision\ x\ recall}{precision\ +\ recall}\right)$$

4.2.4.1 Target variable

The experimentally measured k_{cat} value was converted to $\Delta G^{\ddagger}$. Since the rate is only dependent on the temperature at which the reaction is taking place, and the $\Delta G^{\ddagger}$, to get the barrier for the hydrolysis reaction in kcal/mol, k_{cat} was converted to $\Delta G^{\ddagger}$ by rearranging the Eyring equation for $\Delta G^{\ddagger}$ where: k_{cat} is the catalytic constant, k_B is the Boltzmann constant, h is Planck's constant, T is the temperature, and $\Delta G^{\ddagger}$ is the Gibbs free energy of activation:

$$k_{cat} = \frac{k_B T}{h} e^{-\frac{\Delta G^{\ddagger}}{RT}} \qquad (4.3)$$

$$\Delta G^{\ddagger} = \left(-\log\left(k_{cat} \cdot \frac{h}{k_B T}\right) \cdot k_B T\right) \qquad (4.4)$$

$\Delta G^{\ddagger}$ values were used as a target variable in the ML models. To perform binary classification, those were converted to two label classes. 0 or 1, which is further discussed in the 4.3 Results section.

4.3 Results and Discussion

The distribution of the $\Delta G^{\ddagger}$ values for all GO variants is shown in Figure 4.5. Based on this distribution, the cutoff point to sort the variants into a binary classification system was chosen to be the following: all variants with $\Delta G^{\ddagger}$ below or at 14.4 kcal/mol were labeled as Class 0, whereas all mutant variants with $\Delta G^{\ddagger}$ above 16.7 kcal/mol were labeled as Class 1. First, MD simulations were performed with the WT GO enzyme with its native substrate D-galactose.

All variants labeled Class 0 had D-galactose in the active site (22 variants) while the rest of the variants (9) had SS1 or S128 and were labeled Class 1.

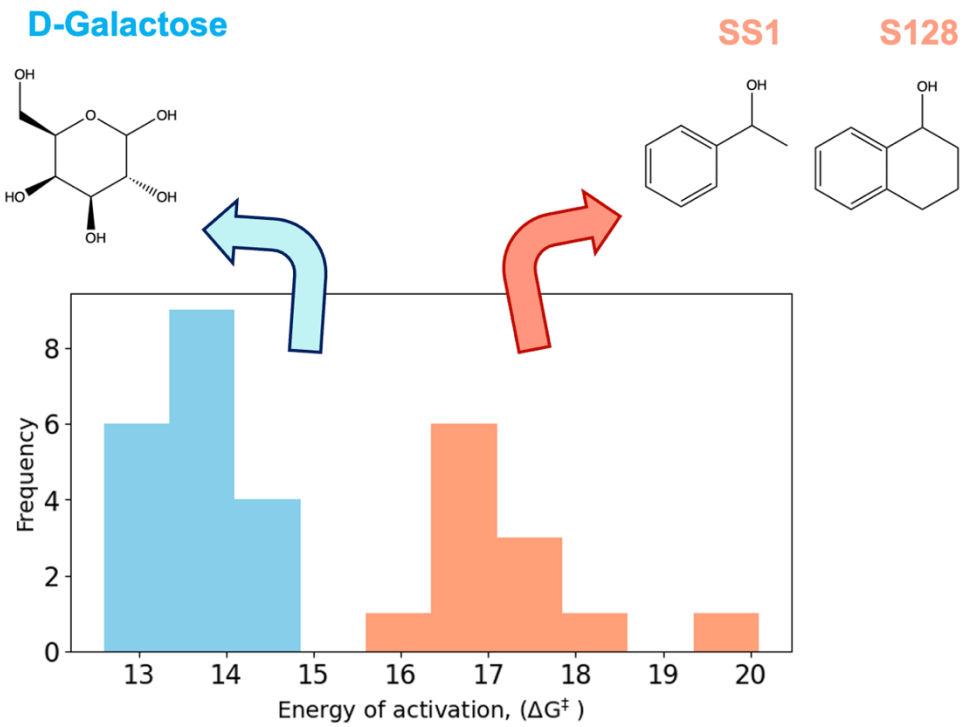


Figure 4.5. Histogram of the GO variants based on $\Delta G^{\ddagger}$ values for the catalytic reaction. All variants with $\Delta G^{\ddagger}$ below 14.4 kcal/mol had D-galactose in the active site. All variants with $\Delta G^{\ddagger}$ > 16.7 kcal/mol had SS1 or S128 as the substrate.

Since $\Delta G^{\ddagger}$ for the WT enzyme is estimated to be ~13.8 kcal/mol and there is usually an experimental error of 0.5 kcal/mol in the measurement, for a variant to be classified as different from the WT, one would consider at least 1.0 kcal/mol difference in the barriers.

4.3.1 Feature selection

The aim was to develop an ML model that can predict the effect of mutations on the catalytic rate of variants with different substrates in the active site. For this reason, the features utilized by the model need to be present in all substrates and all tested variants. Therefore, features had to come only from residues which will remain unmutated in all variants, as well as from atoms which will be present in all of the substrates. The original approach was to

extract inter-residue distance combinations within 6 Å of the copper ion, as well as angles, using a random approach. Then, followed by the extraction of the features, to use an ML model to find the most important descriptors, based on supervised learning with labeled training data. The idea is that the model will find the pattern which will fit x and create a function f(x) that can predict y for a new x.

The cutoff distance of 6 Å was chosen as it contained all the copper-coordinating residues and other residues that are generally not used in the directed evolution for GO, except for W290, which was not included in any feature selection as it does get mutated in some variants. New synthetic features were also generated by using the displacement from the WT, using the respective distance in the WT GO as a reference distance, and then subtracting the same distance in each respective variant. Other features which could be easily calculated from the MD trajectories and were tested as features include the Root Mean Square Deviation (RMSD) which was obtained by calculating the deviation of the α-carbons of the protein backbone in the starting structure versus the last frame of the production run; and the Root Mean Square Fluctuation (RMSF). The overall RMSD and RMSF for the full protein did not appear as useful features and were not included in the final feature set. The final set of features that used contained only the interatomic distances coming from the active site residues within 6 Å of the copper cation. For a table of all features refer to Appendix B, (Tables SI5 and 6).

The final dataset had the following form:

$$\begin{bmatrix} X1_1 & X2_1 & \cdots & XN_1 & Y_1 \\ X1_2 & X2_2 & \cdots & XN_2 & Y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X1_{31} & X2_{31} & \cdots & XN_{31} & Y_{31} \end{bmatrix}$$

With 31 representing the number of rows (GO variants) and *N* representing the number of columns (features).

The initial strategy for dimensionality reduction was to conduct a PCA on the feature dataset, which was followed by k-means clustering. However, this approach yielded 14 variants inaccurately clustered, indicating that employing linear dimensionality reduction through PCA is not particularly beneficial in this scenario (see Appendix B).

Instead, to find out the most important features to fit the model, a Random Forest Regressor (RFR) was employed. The RFR was used as implemented in scikit-learn[110] and the scikit-learn attribute, feature importance was used, which selects the top features the model learns the most from. Since the dataset is very small, features need to be filtered out before a model can be fitted in order to prevent overfitting. For this reason, only the top 5 most important features were pre-selected. The following 5 features (interatomic distances) emerged as the most important: Cu-CZ(Y272), O(Y495)-O6(Sub), O(Y272)-NE2(H581), SG(C228)-C6(Sub), SG(C228)-O(Y405), Figure 4.6. The importance is defined as the frequency with which a feature is selected by the RFR averaged over all of the folds. For example, a feature that is selected by the RFR 100 times over 100 folds of training and testing data, will be the most important one.

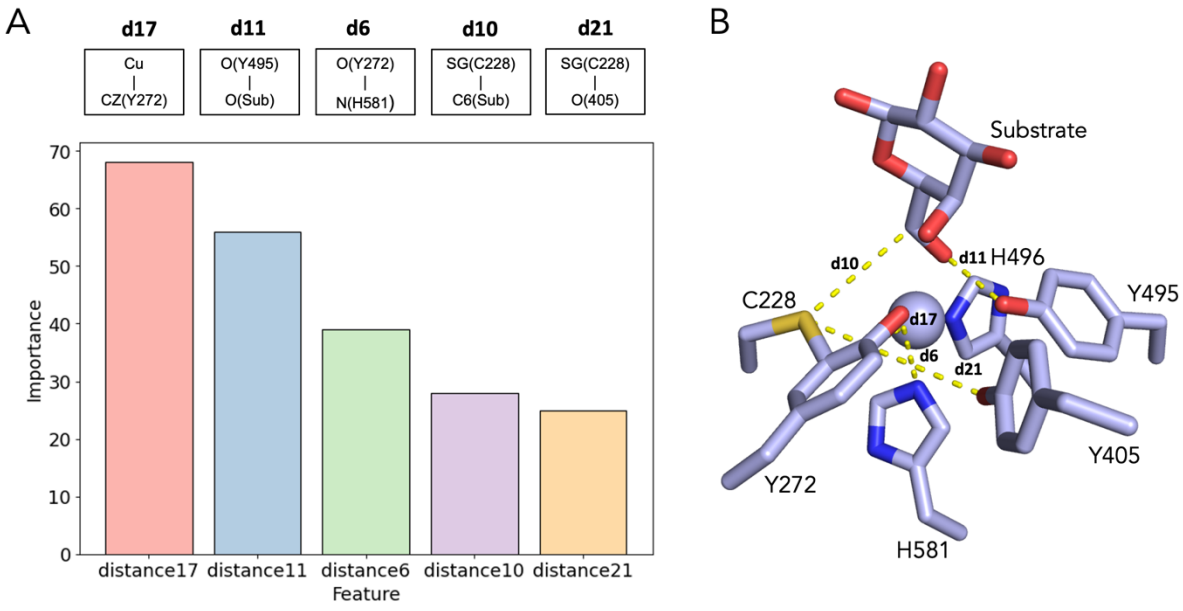


Figure 4.6. (A) The most important features selected by the RFR, ranked by importance. (B) The active site residues are shown as purple sticks, and the most important distances are shown with yellow dashed lines.

As expected, most of the key distances come from the tyrosyl radical which accepts the H atom. It is not surprising as this is the key residue in the catalytic reaction and its displacement during the rate-limiting step is expected to be detrimental for the catalytic rate. In the TS of the WT GO, the distance from the O(Y272) to the HX atom is 1.3 Å and 3.8 Å from the O(Y272) to the copper cation. It was observed that with most variants that were predicted to slow down the catalytic rate, O(Y272) had a shorter distance to the copper cation and a slightly tilted overall orientation. It had also moved from the reference position in the WT and was on average further away from the Cu-coordinating nitrogen of H581, relative to the variants which do not slow down the rate considerably. The orientation of the C228 residue seems to be quite important as well; it is directly linked to Y272 through a thioether bond. This residue seems to be critical for the proper functioning of the enzyme, which is confirmed by mutational studies,[116] and distances from this residue to other residues in the active site need to be maintained similar to as in the WT GO enzyme.

The same analysis was also performed for the original 93 MD simulations which had one extra restraint between atoms CZ(Y272) and C1(Sub), (details on the restrains are available in Appendix B, Tables SI3 and 4). The following five distances emerged as the most important: O(Y495)-O(substrate), O(Y272)-NE2(H581), O(Y272)-NE2(H496), O(Y405)-Cu, and SG(C228)-Cu. Notably, two of the most important distances are the same - O(Y272)-N(H581) and O(Y495)-O(Sub) as in the other set of simulations, with the Y272 residue orientation generally appearing to be the most important across the two feature datasets.

4.3.2 Performance evaluation

Different models for binary classification were tested to find the one achieving the highest accuracy. The most successful model was based on an RF Classifier.

The results from the last 93 simulations are presented (as they contain the reduced restraints set that can be used for new substrates also). The average accuracy of the classification models was 78 for the RF and 73 for the GBDT model, respectively. On average, for the best model, 24-25 of the 31 variants are classified correctly and 6-7 are classified incorrectly. The reported values are the average of 100 testing sets. The mean ROC is displayed in Fig. 4.7.

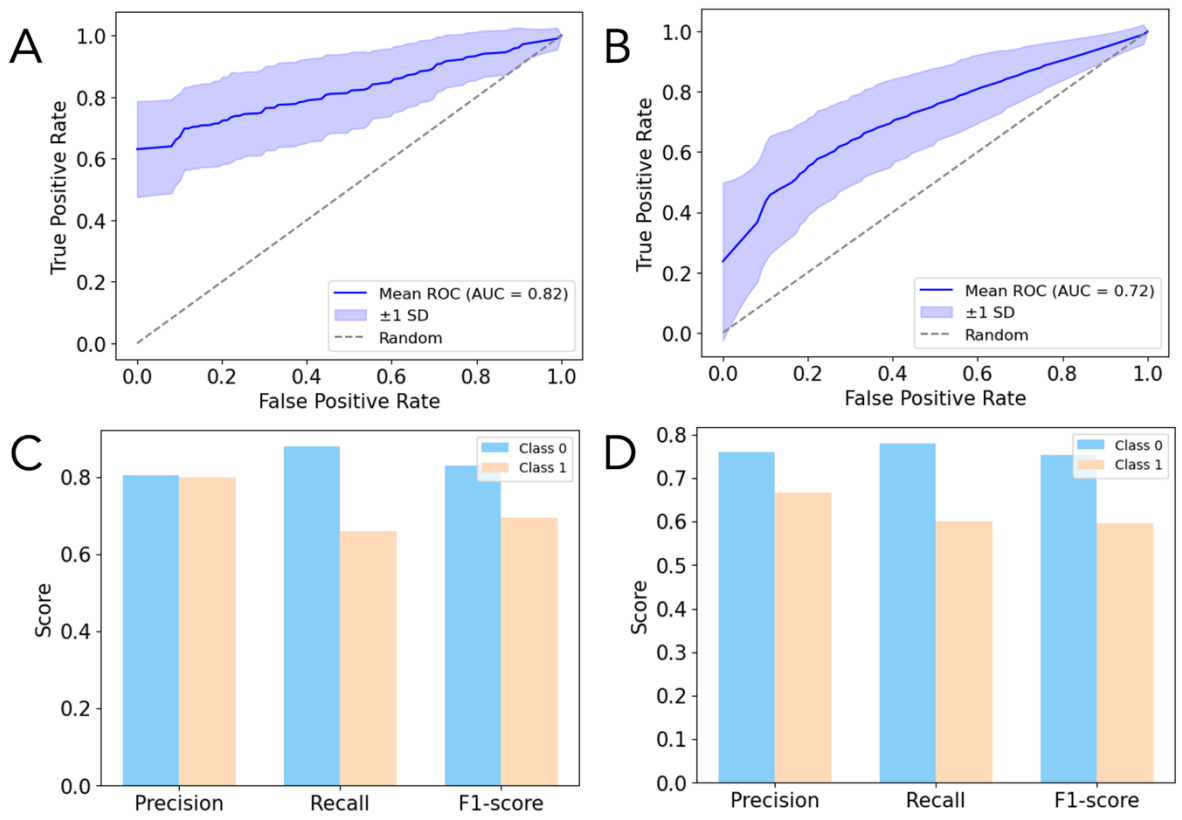


Figure 4.7. (A) TPR and TFR and ROC curve based on the Random Forest Classifier (RFC) model. (B) TPR and TFR and ROC curve based on the Gradient Boosted Decision Trees (GBDT) model. (C) Precision, recall, and F1 metrics for the RFC. (D) Precision, recall, and F1 metrics for the GBDT model.

If we look at the individual classification results for each variant, we can gain more insight about the effect of single protein mutations. According to the results shown in Figure 4.8, the variants with D-galactose in the active site, which have significantly lower experimental k_{cat} values (raising the value of $\Delta G^{\ddagger}$ considerably), in particular, N6M1_W, are hardly recognized by the model as slowing down the rate. N6M1_R is also classified correctly only 14% of the time. This arginine brings a charge close to the active site. Misclassification could occur because Arg383-induced polarization is neglected. Potentially, Arg383 changes the protonation states of other nearby residues, which is not accounted for in our simulations. These two variants being misclassified suggests that this type of TS MD approach is likely not sensitive enough to identify active site displacements from variants that differ from the rest of the dataset by only 1 mutation (here meaning the other N6M1 variants). The likely reason

is also the strong restraints used to keep the coordinates resembling the 3D structure of the transition state. However, interestingly W290G and W290F which have only 1 mutation as a whole, relative to the WT enzyme, are correctly labeled by the ML model. W290F, which has similar k_{cat} to the WT enzyme arguably does not change a lot the overall catalytic rate as phenylalanine has a very similar structure to tryptophan, which probably contributes to it having similar π-stacking activity, and it behaves similarly in the MD simulations, upon inspection of individual active site distances. It should be noted that in the force field used, there is no specific term that accounts for π-stacking. The fact that W290G is generally correctly classified could mean that the mutation at position 290 is very important, which is also previously discussed in mutational studies of GO,[125] and due to its significance, it is recognized by the ML models. It should also be noted that it is also the amino acid, which is closest to the active site, out of the ones which are subject to mutation. It is therefore expected that variants that have mutation in position 290 will displace the active site more considerably, or at least show more distinct dynamics relative to the ones that do not. Unfortunately, W290H is not picked up by the model as Class 1 as the extracted distances show it behaved similarly during the MD simulations as W290F. The reason for this could be that the protonation state for W290H could be altered and differ from the one we used in our simulations. The protonation states for active site histidines can be very difficult to predict accurately. The mutation at position 330 is in the immediate location to the active site, and it also forms contact with the D-galactose substrate. Since M3-5 and its derivative variants contain mutations at both position 290 and 330, it is expected for them to show different dynamics during the MD simulations, or at least to present a more significant displacement at the active site, relative to the variants that do not have those two mutations, especially considering this as a synergistic effect. All of the variants which have a non-native substrate in the active site are correctly classified as Class 1. It could be argued in this case that the model does not identify the displacement coming from the protein mutations but rather from the presence of the non-native substrate. This could be established more concretely once variants with non-native substrates with similar rates to WT GO are included in the dataset.

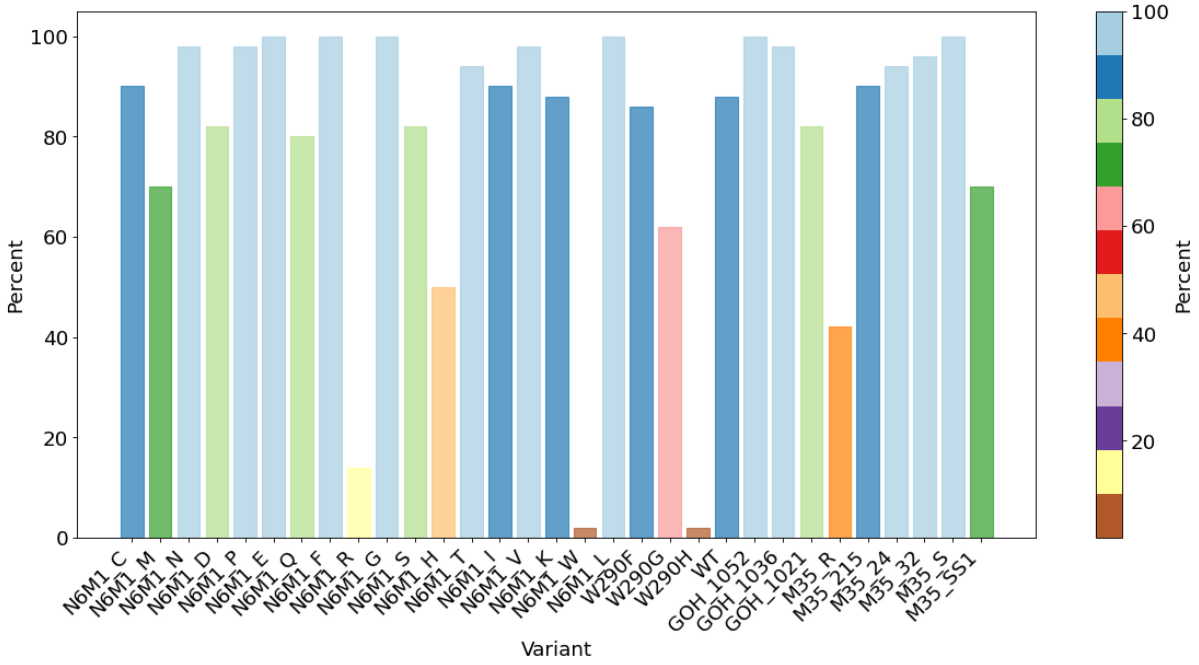


Figure 4.8. Percent of times each variant is classified correctly by the RF classification model. Light blue and blue bars show that the variant is generally always classified correctly (~80% of the time), and green bars show that the variant is predicted correctly most of the time (~70% of the time). Variants that are never or rarely predicted correctly are displayed as brown or yellow bars.

Another point of consideration is why this problem was handled as a classification and not a regression task. This type of data is possibly more suitable for regression or predicting the exact k_{cat} value for each variant. However, at the time when the data was curated, there were and still are some conflicting k_{cat} values reported in the literature for some of the variants, in some cases having three different k_{cat} values reported for the same variant. This can be due to various experimental conditions and factors. This results in $\Delta G^{\ddagger}$ in one of the cases at 13.8, 14.4 or 14.7 kcal/mol, respectively. However, despite the three different reported values, all of the three variants would still fall within Class 0. For this reason, upon the evaluation of the data, it was decided to handle this problem as a classification task. There are possible shortcomings which need to be mentioned. Most notably, all of the variants should be modeled directly from the WT GO crystal structure and simulated for 20 ns. In the case of the N6M1_X variants, those were modeled after the M1 variant which was already subject to 10 ns MD simulations (Figure 4.4). For this reason, the starting structure for all N6M1_X variants is slightly different from the WT, M3-5, and M3-5-derived variants. The slightly different

starting point for the ML simulations may create a bias in some of the interatomic distances which is then picked by the ML models. Ideally, all variants should be modeled from the same crystal structure, simulated with the same restraints, for the same overall simulation time, so that the extracted distances are comparable. However, it should be pointed out that the distances equilibrate in the course of the 20 ns MD and this introduced bias is not likely a significant factor. This pipeline offers an excellent alternative for highly accurate variants classification, to the more time-consuming QM/MM-based methods, and is also easily transferable to other enzymes.

4.4 Conclusion

31 variants of GO were modeled and 3 MD simulations at/around the rate-limiting step were performed with each one. Interatomic distances were extracted from the MD trajectories and used as features to create an ML model that can predict, based on the dynamical behavior during the MD simulations, whether unseen variants will have a similar catalytic rate to the WT GO enzyme or the active site will be affected in a way which will slow down the rate of catalysis. The best model achieving the highest accuracy of classification was based on a Random Forest. This classification approach has advantages over QM-based methods as it offers the opportunity to sample conformations, and it can be significantly faster, allowing for an easily automated and high throughput approach.

Future work

Currently, the dataset for this ML pipeline is very small, which limits the type of ML that can be utilized. I would ideally like to considerably expand the modeled variants and substrates to include more secondary alcohols with more diverse structures. Another disadvantage of the current dataset, apart from the small sample size, is that all variants for which the rate of catalysis is similar to the native enzyme, have the same substrate class in the active site. It would be interesting to model variants with non-native substrates where the rate is comparable to WT GO or faster. There is currently more experimental data available, compared to when this project was started, and with the current automated pipeline, it would

be straightforward to simulate more variants and substrates, even from older studies. It might be interesting to also include features from MD simulations of the reactant and product states and investigate whether that allows the models to achieve higher classification accuracy.

Chapter 5

Combining Data Integration and Molecular Dynamics for Target Identification in α-Synuclein-Aggregating Neurodegenerative Diseases: Structural Insights into Synapotojanin-1 (Synj1)

This Chapter was published in The Journal of Computational and Structural Biotechnology in 2020 and is reproduced here with permission from the authors, 'Combining data integration and molecular dynamics for target identification in α-Synuclein-aggregating neurodegenerative diseases: Structural insights on Synaptojanin-1 (Synj1)', Computational and Structural Biotechnology Journal, DOI: 10.1016/j.csbj.2020.04.010. Copyright Computational and Structural Biotechnology Journal.

Summary of the Work

The aim of this work is to integrate genomic and proteomic data from toxicity studies of α-synuclein and to identify protein targets for neurodegenerative diseases. One of the proteins identified, which is independently shown to be strongly implicated in Parkinson's disease (PD), is Synapotojanin-1 (synj-1). A wide range of mutations in the gene coding for the protein are long known. In this study, we report the full atomistic model of the 5-phosphatase domain of synaptojanin-1, embedded in a membrane and show its binding to the substrate (PIP_2). Details on the binding of PIP_2 to the 5-phosphatase domain are needed for targeting of the protein in diseases where synj-1 is overexpressed.

Author Contribution

This work was originally conceptualized by Edina Rosta, Attila Csikász-Nagy, Paola Piccotti and Kirsten Jenkins. Modeling and MD simulations of synj-1 were originally done by Kirsten Jenkins/István Szabó but further modeled and refined by me. The bioinformatics analysis and final homology models used in the manuscript were done by me, as well as the 2-Mg^{2+} MD simulations used in the final manuscript. Figures were done by me with the exception of Figures 5.5 and 5.6 which were produced by István Szabó. The trajectory analysis used in this work was done by me, Kirsten Jenkins and István Szabó. Most of the methodology, data curation, visualization, writing of the original draft, review & editing was done by me and Edina Rosta. All experimental proteomics data collection was done in the lab of Paola Picotti. If experimental data was obtained from any other authors/sources, it is mentioned in the text.

The Supporting Information for the article is available in Appendix C.

Computational and Structural Biotechnology Journal 18 (2020) 1032–1042

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Combining data integration and molecular dynamics for target identification in α-Synuclein-aggregating neurodegenerative diseases: Structural insights on Synaptojanin-1 (Synj1)

Kirsten Jenkins [a], Teodora Mateeva [b], István Szabó [b], Andre Melnik [c], Paola Picotti [c,1], Attila Csikász-Nagy [a,d], Edina Rosta [b,*]

[a] Randall Division of Cell and Molecular Biophysics, Institute for Mathematical and Molecular Biomedicine, King's College London, London SE1 1UL, UK
[b] Department of Chemistry, King's College London, London SE1 1DB, UK
[c] Institute of Biochemistry, Department of Biology, ETH Zurich, CH-8093 Zurich, Switzerland
[d] Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, 1083 Budapest, Hungary

ARTICLE INFO

Article history:
Received 31 January 2020
Received in revised form 15 April 2020
Accepted 18 April 2020
Available online 22 April 2020

Keywords:
Data integration
Molecular dynamics (MD)
Neurodegenerative diseases
Parkinson's disease (PD)
Synaptojanin-1
α-Synuclein

ABSTRACT

Parkinson's disease (PD), Alzheimer's disease (AD) and Amyotrophic lateral sclerosis (ALS) are neurodegenerative diseases hallmarked by the formation of toxic protein aggregates. However, targeting these aggregates therapeutically have thus far shown no success. The treatment of AD has remained particularly problematic since no new drugs have been approved in the last 15 years. Therefore, novel therapeutic targets need to be identified and explored. Here, through the integration of genomic and proteomic data, a set of proteins with strong links to α-synuclein-aggregating neurodegenerative diseases was identified. We propose 17 protein targets that are likely implicated in neurodegeneration and could serve as potential targets. The human phosphatidylinositol 5-phosphatase synaptojanin-1, which has already been independently confirmed to be implicated in Parkinson's and Alzheimer's disease, was among those identified. Despite its involvement in PD and AD, structural aspects are currently missing at the molecular level. We present the first atomistic model of the 5-phosphatase domain of synaptojanin-1 and its binding to its substrate phosphatidylinositol 4,5-bisphosphate (PIP_2). We determine structural information on the active site including membrane-embedded molecular dynamics simulations. Deficiency of charge within the active site of the protein is observed, which suggests that a second divalent cation is required to complete dephosphorylation of the substrate. The findings in this work shed light on the protein's binding to phosphatidylinositol 4,5-bisphosphate (PIP_2) and give additional insight for future targeting of the protein active site, which might be of interest in neurodegenerative diseases where synaptojanin-1 is overexpressed.

1. Introduction

Age-related diseases are rapidly increasing in their frequency due to longer life expectancy and can have devastating effects upon the quality of life of sufferers [1,2]. At a cellular level, Parkinson's disease (PD) and other neurodegenerative diseases, including Alzheimer's disease (AD) and Amyotrophic lateral sclerosis (ALS) are linked to toxic protein aggregation [3,4]. However, targeting these protein aggregates has not led to successful drug therapies. Small drug molecules are ineffective towards them and no new therapies for Alzheimer's disease have been approved in the last 15 years. It is, therefore, becoming increasingly important to identify novel targets for protein-aggregating neurodegenerative diseases [5,6].

In PD, α-synuclein is of particular importance as it is the primary aggregating protein [7–9], its gene amplifications and mutations may lead to PD [10–12]. Human neurons are complex cells with long lifespans, therefore, α-synuclein toxicity has been explored in the model eukaryotic organism, *Saccharomyces cerevisiae* (budding yeast) [13–15]. Budding yeast cells do not have a homologue to α-synuclein, therefore protein expression has been

* Corresponding author.
E-mail address: edina.rosta@kcl.ac.uk (E. Rosta).
[1] Current affiliation: Institute of Molecular Systems Biology, Department of Biology, ETH Zurich, Switzerland.

https://doi.org/10.1016/j.csbj.2020.04.010

induced using a galactose inducible promoter, showing toxic aggregation in yeast which leads to cell death [16].

The abundance of biological data from various experimental sources (genomics, proteomics, metabolomics) offers unprecedented opportunities for data integration approaches for novel target identification [17]. Importantly, data integration is particularly useful in analysing networks of protein interactions and is widely used in developing understanding of how various cellular processes are altered [18]. Importantly, it can inform novel targets for atomistic studies, which is yet underutilized [19]. In this study, we employed data integration upon two complementary studies of α-synuclein toxicity: (i) genomics study by Khurana et al. and (ii) proteomics study by Melnik et al. [15,20]. The first quantified the effect of deletion and overexpression of various proteins on the toxicity of α-synuclein in buddying yeast cells [14]. Khurana et al. compared the lifespan of yeast cells that were modified to express α-synuclein, with cells that expressed α-synuclein but had one protein deleted or overexpressed. When the deletion or overexpression of a protein significantly affected the lifetime of the cells, this protein was labelled a disease modifier: cell death enhancer or suppressor. The second dataset was collected from a proteomic analysis of the perturbation in protein-wide concentrations of α-synuclein induced aggregation in buddying yeast, compared to yeast that did not express the aggregating protein [15,20]. By integrating data from both studies, we identified 17 potential human protein targets.

Among the proteins identified to be of interest, we chose the protein polyphosphatidylinositol phosphatase INP53 for further investigation and in particular, its human homologue, synaptojanin-1 (synj1). Apart from being a cell death enhancer when deleted in α-synuclein expressing cells and simultaneously showing to be downregulated when α-synuclein was overexpressed, it has also already been independently identified that the gene coding for the protein SYNJ1, is a particular PARK locus, PARK 20 [21]. Additionally, synj1 is implicated not only in PD but also in AD [22–25]. The primary substrates of synj1 are phosphoinositides (PIPs) with phosphatidylinositol 4,5-bisphosphate PIP_2 and phosphatidylinositol 3,4,5-trisphosphate PIP_3 being among the most important signalling lipids in membrane trafficking. An imbalance in PIPs has previously been identified to be crucial in many protein aggregating diseases, namely AD and PD [26,27]. The imbalance of phosphoinositides is heavily correlated to malfunctions in synj1 activity, and mutations of synj1 itself are implicated in various neurodegenerative diseases [22–26].

Synj1 has three domains. The main catalytic inositol 5-phosphatase domain, the N-terminal Sac1 inositol phosphatase domain, and a C-terminal proline-rich domain that plays a role in protein–protein interactions related to vesicle endocytosis [9,28]. Mutations in the Sac1 domain have already been linked to the downregulation of PIPs and malfunctions in autophagy [29].

Currently, experimentally resolved structures of the first two domains of human synj1 are unavailable. We present here the first atomistic model of the 5-phosphatase catalytic domain of the protein both in membrane-free and membrane-embedded molecular dynamics (MD) simulations. Additionally, we propose that the protein active site involves two divalent cations. It is well accepted that 5-phosphoinositide phosphatases are Mg-dependent enzymes [30,31], with catalytic activity supported by Mg^{2+} or Mn^{2+}, however inhibited by Ca^{2+} and other divalent cations [32]. This behaviour is often observed in phosphate catalytic enzymes, demonstrating apoptotic regulatory role of Ca^{2+} [33]. We suggest that one of the Mg^{2+} ions has a role in activating the water nucleophile, whereas the second Mg^{2+} stabilizes the leaving group, similar to other enzymes using a two-metal ion catalytic mechanism [34].

2. Methods

2.1. Data integration

Two data sets were used for the data integration. The first dataset was obtained by Khurana et al. [14] and was generated by comparison of the survival rate of yeast cells (*S. cerevisiae*) that were modified to express α-synuclein to cells that expressed α-synuclein but had one protein deleted or overexpressed. The proteins were labelled as either toxicity 'Suppressor' (S) or 'Enhancer' (E), based on their toxicity modulating effect on the α-synuclein expressing cells [14].

The second dataset was obtained by Melnik et al. [20]. The dataset was generated using mass spectrometry-based label-free shotgun proteomics. α-Synuclein was expressed in yeast cells (*S. cerevisiae*) by a galactose-induced promoter and the overall changes of the protein abundancies in the proteome were compared to control cells proliferating at the same time length but transformed with an empty vector (EV). Protein abundance changes were monitored at 6 h, 12 h, 18 h and 24 h after the expression of α-synuclein. Proteins which had significantly perturbed abundance at 12 h and 18 h were selected in this work (Table1 ESI and Fig. 3 ESI). Proteins perturbed at 6 h were not included as very few proteins were observed to be altered at this time suggesting that it is too early to observe the toxic effect on the cell. The results at 24 h were also omitted as the cells were dying and therefore many pathways were malfunctioning. The median ratio for the protein concentration (α-synuclein expressing cells vs. control) was the parameter used to classify the proteins as up or down regulated. The value of 1.00 was chosen as a cut-off point. If the average of the mean ratio value for the concentration of the proteins between 12 h and 18 h was above 1.00, the protein was classified 'upregulated', and if below 1.00, 'downregulated', Fig. 3 ESI.

Following this, a combined protein dataset was generated. All of the proteins that did not appear in the two initial datasets were removed. The list of proteins that had significant results in both studies were then further reduced by selecting only proteins with human homologues, using the Yeast Mine Database [35]. Finally, it was confirmed whether the protein had been previously linked to aggregation diseases using Malacards database [36].

2.2. Molecular dynamics simulations of Synaptojanin-1

The main catalytic 5-phosphatase domain of the human protein synj1 does not currently have an experimentally resolved structure in the protein data bank (PDB). Therefore, a homology modelling server was used (SWISS-MODEL) [37] to create the three-dimensional structure of the protein, using the amino acid sequence from the Uniprot database [38] (code: O43426). The 3D structure obtained from SWISS-MODEL was used for the MD simulations of synj1.

The 5-phosphatase domain of synj1 (residues 500–899) was modelled using the template OCRL-1 in complex with a phosphate ion (PDB code 4CMN) [30]. Residues 517–894 had a sequence identity of 36.47% to OCRL and a global model quality estimate of 0.64. A ligand with an identical phosphate head group but shortened tails was positioned manually along with the coordinating residues and water molecules, for an initial comparison of the ligand to PIP_2. A single magnesium ion was added by visual inspection of known crystal structures based on the alignment of the conserved catalytic sites from within the 5-phosphatase family. Magnesium ion was chosen as the catalytic ion in the active site, as human 5-phosphoinositide phosphatases are Mg^{2+}-dependent [30], and Zn^{2+}, Ca^{2+} and other divalent ions (except for Mn^{2+}) typically inhi-

bit catalytic activity. The reference structure for positioning the PIP_2 ligand and the magnesium ion was chosen to be the inositol polyphosphate 5-phosphatase domain (IPP5C) of SPsynaptojanin available in complex with inositol (1,4)-bisphosphate and a calcium ion (PDB code 1I9Z) [39].

All Molecular Dynamics simulations were performed by using the program NAMD [40]. The force field used in the simulations was CHARMM36 with periodic boundary conditions and to evaluate the non-bonded long-range interactions the particle mesh Ewald method was utilised with a 12 Å cutoff [41,42]. The NPT ensemble was maintained with a Langevin thermostat (310 K) and an anisotropic Langevin piston barostat (1 atm). CHARMM-GUI was used to set up the simulation box of side length 107.4 Å; neutralise and solvate the system; and determine the charged state of all ionisable residues using a standard protocol [43,44]. Ions randomly replaced water molecules using a Monte Carlo method to neutralise the system using 3 K^+ ions, then additional 111 K^+ and 111Cl^- ions were added to create a salt concentration of 0.15 M. Equilibration was completed using the standard CHARMM-GUI protocol [43,44], with the addition of constraints upon the distance between the Mg^{2+} ion and: (i) the phosphate group on the fifth carbon of the inositol ring (5-P); (ii) Asp-359; (iii) Glu-92; to be approximately 3 Å [40]. 8 ns of constrained molecular dynamics simulations were completed using the constraints above, and 92 ns of non-constrained MD was run to test the stability of the membrane-free structure.

The tails of the PIP_2 were then reinserted to the structure of the completed membrane-free simulation, and the whole structure (including the bound PIP_2) was uploaded to the Orientation of Proteins in the Membrane (OPM) server which gave a membrane alignment for the system [45]. This alignment was then input into CHARMM-GUI to add the membrane [43,44]. The membrane was comprised of 90% phosphatidylcholine (PC), 5% phosphatidylserine (PS) and 5% PIP_2. To solvate the system the protein was inserted into cubic pre-equilibrated TIP3P water box of with a dimensions 127.029 Å × 127.029 Å × 129.809 Å. Ions randomly replaced water molecules using a Monte Carlo method to neutralise the system, then an additional ions (231 K^+ and 113Cl^- in total) were added to create a salt concentration of 0.15 M. Six equilibration steps were conducted based on standard CHARMM-GUI protocol [43,44]. 10 ns of constrained molecular dynamics was run, where constraints were added upon the distance of 2 Å between the Mg^{2+} ion and the following: (i) 5-P; (ii) Asp-359; (iii) Glu-92; (iv) Asn-44; the phosphate group on the fourth carbon of the inositol ring. These additional constraint for the Asn-44 residue were added to establish if additional residues were required to stop the potassium ions approaching the catalytic site. Two independent simulations of unconstrained molecular dynamics each lasting 300 ns were completed from the constrained molecular dynamics in order to obtain final structures for the Synaptojanin-PIP_2 system.

3. Results and discussion

3.1. Data integration

We combined datasets from genome-wide and proteomic studies where: (i) the effects of protein deletion and overexpression on cell death was studied, and where (ii) overall protein perturbation levels were measured, in α-synuclein-expressing yeast cells. We selected those proteins that were: (i) suppressors or enhancers of cell death when deleted or overexpressed and (ii) had their concentration perturbed at 12 h and 18 h post α-synuclein expression. This data integration highlighted 62 proteins of potential interest in Parkinson's, Alzheimer's or other neurodegenerative diseases, based on the proteins' toxicity modulating effect and concentration, as quantified in α-synuclein-expressing yeast cells. We then considered whether these proteins had human homologues, whether they have already been implicated in any protein aggregation diseases, and the approximate function of the protein, if known, in yeast. Upon removal of yeast specific proteins, which are not of interest to human neurodegenerative disease, the pool of proteins of interest was reduced to 47. The combined integrated data is visually represented in Fig. 1.

We further narrowed down the list of proteins to have most significance by proposing that the candidates of most interest for us would be those that either: enhance toxicity when deleted and are significantly downregulated in α-synuclein induced cells; suppress toxicity when overexpressed and their concentration is downregulated in α-synuclein induced cells; enhance toxicity when overexpressed and are significantly upregulated. These proteins are represented in Fig. 1 within the circled area.

This method of data integration highlights 17 proteins (Table1 ESI). Four of the 17 proteins (INP53, RAD27, YPK9 and POR1) have already been independently confirmed to be implicated in Parkinson's, Alzheimer's or other neurodegenerative diseases caused by protein aggregation [36]. Therefore, our analysis demonstrates that data integration is indeed useful in locating existing and novel protein targets that directly impact the toxicity of α-synuclein in humans, as well as in yeast, and therefore might play an important role in neurodegenerative diseases such as PD or AD.

Note that some of the proteins appear in multiple sections: YPK9, RTS1, RPS14A. For all three, cell toxicity is enhanced when the proteins are deleted, and suppressed when they are overexpressed. This is consistent with their roles as being overall needed by the cells to survive in the α-synuclein-rich environment. Interestingly, however, while RTS1, RPS14A are accordingly upregulated by the cells, YPK9 appears downregulated. YPK9 therefore has a key function, which appears to be impaired by α-synuclein overexpression, as the cells are unable to produce enough YPK9 to help cell survival.

YPK9's human homologue, ATP13A2, is also identified by various independent measures as a key protein in PD. It is one of the PARK genes identified in human disease, PARK9, its mutations are associated with Spastic Paraplegia (SPG78), Kufor-Rakeb syndrome and neuronal ceroid lipofuscinosis [47].

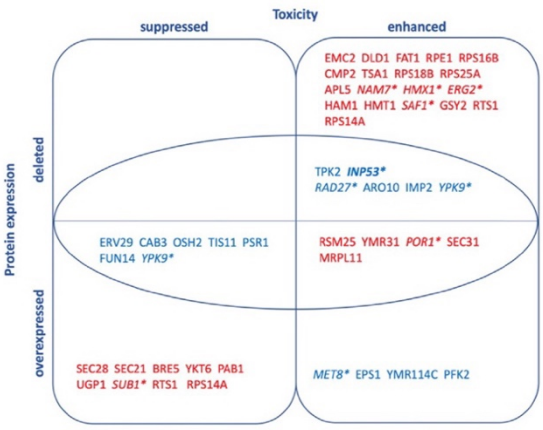


Fig. 1. Effects of protein expression on α-synuclein cell toxicity. All proteins in the diagram have human homologues. Protein downregulation (blue) or upregulation (red) is also indicated in α-synuclein expressing cells. The circled area contains proteins identified to be proteins of interest. The human homologues of the proteins in italics with an asterisk are known to be involved in Parkinson's or other neurodegenerative diseases [46]. INP53 (bold) is the protein chosen for further molecular dynamics modelling in this work. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The RPS14A gene's human homologue encodes 40S ribosomal protein S14. It is a member of the ribosome, a central protein of the ribosomal protein subunit S40. It has many diverse roles and it is required for ribosome assembly and 20S pre-rRNA processing, therefore this might lead to its consistent role needed for cell survival [48–51].

RTS1 is a homologue of the mammalian B′ subunit of PP2A and encodes a serine/threonine-protein phosphatase [52]. It is a central protein with several diverse roles: it is required for maintenance of septin ring organization during cytokinesis, for ring disassembly in G1 and for dephosphorylation of septin [53]. Similarly to RPS14A, a diverse regulatory function might be the reason for it being consistently beneficial for cell survival.

3.2. Synaptojanin-1 as potential drug target

Next, we selected synj1 to investigate further using atomistic molecular simulations. In our data integration, synj1 showed strong correlation with α-synuclein toxicity in the following ways: (i) when deleted the protein increased the rate of cell death; (ii) when α-synuclein was expressed in buddying yeast cells, the concentration of synj1 was downregulated compared to empty vector (EV) control cells that did not express α-synuclein. Based on these results the protein shows to be directly or indirectly involved in the toxicity of the aggregating protein α-synuclein. Furthermore, it is also independently confirmed by genetic analysis of PD patients' genome that mutations of the synj1 gene have strong correlation to Parkinson's disease [23]. Synj1 is also a PARK gene (PARK20) [21]. In addition, mutations of synj1 are also correlated with Alzheimer's disease suggesting that it is a crucial protein in neurodegenerative diseases [25].

Synj1 does not currently have a crystallographically resolved structure except for its proline-rich domain, therefore structural studies will offer valuable insights for future drug discovery projects. We were also able to identify suitably accurate homology model template for the main catalytic 5-phosphatase domain, with an active site that is almost identical within the same phosphatase phosphoinosidtide subfamily. Other proteins of interest without available structure had lower sequence identity to template structures in the Protein Data Bank, therefore, they were less suitable candidates for molecular dynamics simulations at the time of the project start.

We first determined the protein–protein interaction network of synj1 using the STRING database [54] (Fig. 2, full list of proteins and their corresponding function in Table 2 ESI). This network suggests that synj1's primary functional partners are proteins involved indirectly or directly in synaptic vesicle endocytosis/vesicle trafficking, either through their role in PIPs regulation (signalling kinases) or in the cascade towards synaptic vesicle endocytosis. Synj1 is therefore likely mainly implicated in neurodegenerative diseases via its role in phosphatidylinositol signalling dynamics, and does not have direct effect on protein aggregation [27,29,55]. As part of the synaptic vesicle trafficking pathway, one of its primary catalytic functions is the dephosphorylation of the 5′P of the PIP moiety in Phosphatidylinositol 4,5-bisphosphate (PIP_2) and Phosphatidylinositol 3,4,5-trisphosphate (PIP_3). In this work, we have focused on the 5-phosphatase domain of the protein, which mainly dephosphorylates PIP_2, a crucial phosphoinositide for healthy nerve function, with known effects on neurodegeneration [56,57].

3.3. Homology modelling of Synaptojanin-1

We created a homology model of synaptojanin-1 based on the human inositol polyphosphate 5-phosphatase OCRL-1 (4CMN) [58], which has very high active site sequence identity to synaptojanin-1. There are six residues within 5 Å of the Mg^{2+} of our model that are all conserved between the model and the template protein (Fig. 1 ESI). This evidence demonstrates that the active site of synj1 can be reliably modelled based on OCRL. Additionally, charged residues are also highly conserved, which is an expected outcome for proteins with identical functionality, and additionally supports the likelihood of a reliable homology model. The sequence of synj1 (5-phosphatase domain) was also aligned

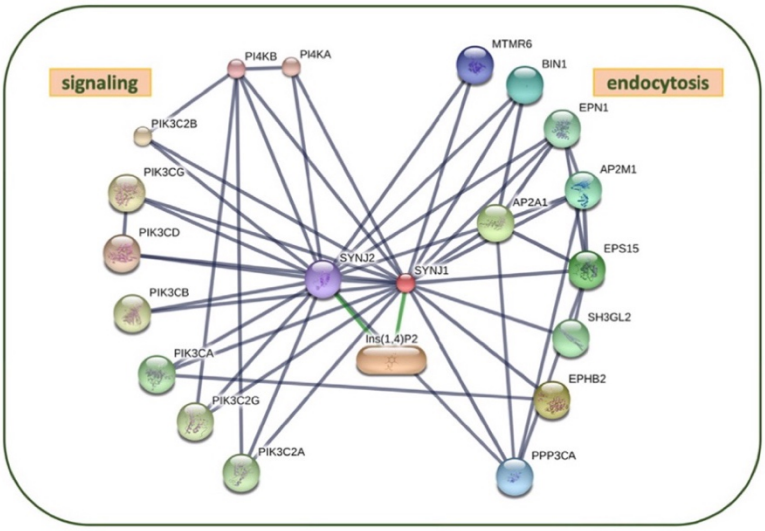


Fig. 2. Predicted close functional partners of synj1. All proteins shown in larger nodes with cartoon have determined 3D structure, the small nodes represent proteins of unknown 3D structure. Grey lines: protein–protein interaction; green: protein-chemical. Active interaction sources: experiments, gene fusion, databases, co-occurrence, co-expression. Generated in high confidence (0.700) [54]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

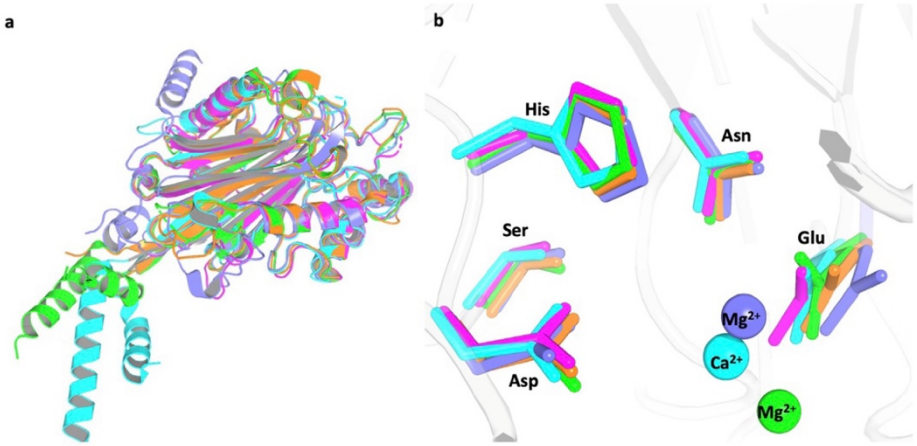


Fig. 3. The overall conserved fold (a) and the evolutionary conserved residues (b) of 5-phosphoinositide phosphatase proteins with their code from the Protein Data Bank (PDB): yeast fission synaptojanin (1I9Z, cyan) [39], human OCRL-1 (4CMN, green) [58], human SHIP2 (4A9C, magenta) [62], human I5P2 (3MTC, orange) [58] and human Synj1 model (purple) [40,58,62]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and compared to other 5-phosphoinositide phosphatase proteins (Sequence in Fig. 2 ESI) [59–61]. Comparison of 5-phosphatases within the same subfamily with defined crystal structures: yeast fission synaptojanin (1I9Z) [39], human inositol polyphosphate 5-phosphatase OCRL-1 (4CMN) [58], human phosphatidylinositol 3,4,5-trisphosphate 5-phosphatase 2 SHIP2 (4A9C) [62] and human Type II inositol 1,4,5-trisphosphate 5-phosphatase I5P2 (3MTC) [58] shows the conserved overall three-dimensional fold (Fig. 3a) and evolutionary conserved residues in close proximity to the active site (Fig. 3b).

The completed homology model of the 5-phosphatase domain was compared to fission yeast synaptojanin (PDB:1I9Z) [39] with an overall very similar fold (Fig. 4a). It was observed that the Asp, His and Glu, the primary conserved active site residues, are located in the binding pocket for both the fission yeast synaptojanin crystal structure and the homology model of human synj1 (Fig. 4b). The conserved residues correspond to Asp-359, His-360 and Glu-92 in the homology model.

4. Molecular dynamics simulations

4.1. Membrane-free molecular dynamics simulation

The flexibility of the whole protein was determined by RMSD calculations, confirming that the most flexible parts lie outside of the catalytic domain, (RMSD in Fig. 4 and Fig. 5 ESI). Fig. 5 shows the protein coloured according to the RMSD value with red signifying flexible regions and blue the more rigid parts. The flexible regions most likely belong to areas that are involved in protein–protein interactions within the synj1 or with external binding partners, as the simulations only use one domain of the protein and the interacting partners are missing from the simulations. This can be seen by the more flexible behaviour occurring at the surface of the system, mainly involving loops. This does not affect the active site or the PIP_2 interaction as the highly flexible regions are not within significant proximity of the active site. The conformation of the synj1 active site was first probed in a membrane-free simulation

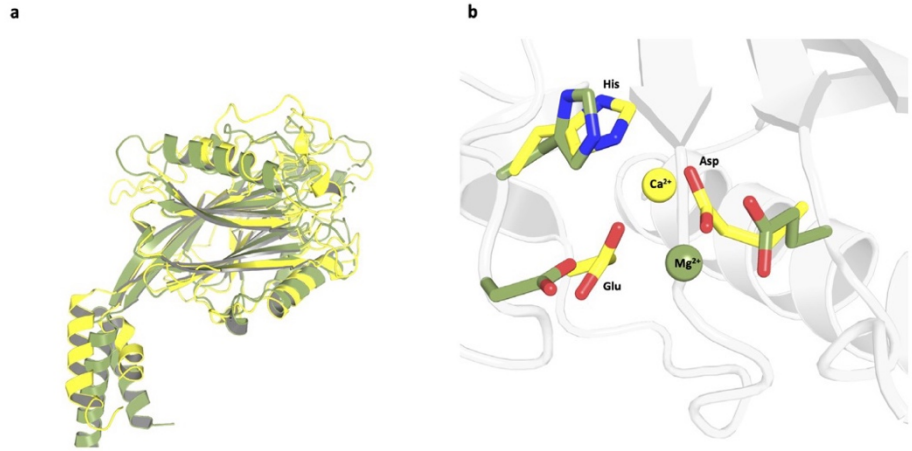


Fig. 4. Comparison of the homology model of the 5-phosphatase domain of human synaptojanin-1 (green) with the crystal structure of fission yeast synaptojanin (yellow), (PDB: 1I9Z) [39]. Overall fold of the 5-phosphatase domain in both proteins (a), conserved residues within active site (b). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

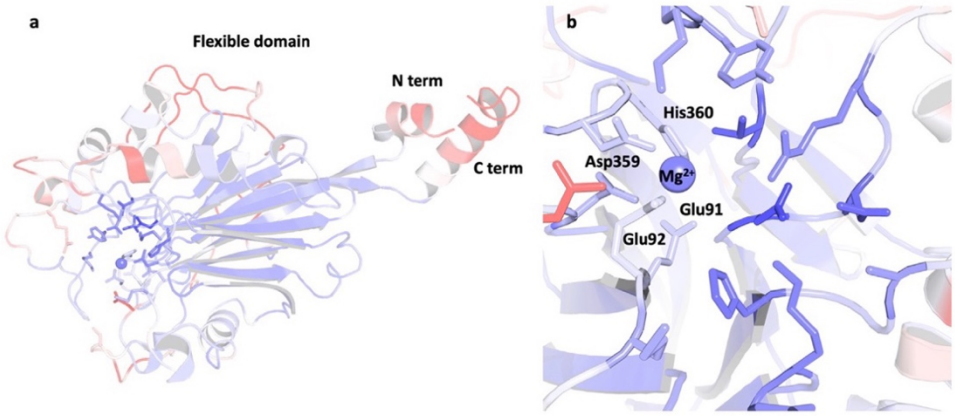


Fig. 5. Protein coloured according to RMSD calculations of the full system (a) and catalytic domain only (b).

and remained stable throughout the MD simulations, with an RMSD of 5 ± 1 Å. It was observed that potassium ions often appeared very close to the active site and remained there for extended periods of time. This was a surprising result, as currently the crystal structures of 5-phosphatases within the 5-phosphoinositide phosphatase family have not observed two metal ions at the active site [31,39,58].

4.2. Membrane-embedded simulations

The unexpected presence of the potassium ions within the active site was further investigated in membrane-embedded simulations. The phospholipid bilayer consisted of 90% phosphatidylcholine (PC), 5% phosphatidylserine (PS) and 5% PIP_2. The simulation setup of the system is shown in Fig. 6, it included the lipid bilayer, PIP_2 ligand, protein, and the single Mg^{2+} cation. The PIP_2 tail was indeed embedded in the membrane, and the protein located on top of the lipid bilayer allowed the phosphosugar headgroup of the PIP_2 to bind to the synj1 active site.

Two independent simulations were conducted over 300 ns each. The stability of the system was assessed via the radius of gyration and solvent accessible surface area of the protein (Fig. 7). The independent simulations present variation in values within a narrow range, suggesting the simulations are stable. Further analysis of the system found that the distances between both the bilayer centre of mass and the protein centre of mass, and the PIP_2 and the magnesium ion respectively, remained stable throughout the simulation (Fig. 7). Therefore, the protein did not penetrate the bilayer, neither did the PIP_2 ligand penetrate further into the protein.

It was observed during both simulations that potassium ions appeared for prolonged periods of time within the active site, as quantified in Fig. 8. The active site is defined as distance of 4.5 Å or less to the 5′P atom of the PIP_2 ligand. Consequently, if a potassium ion appears within 4.5 Å of the 5′P, it is considered to reside within the active site. There were three cases in the course of the simulation – either 0, 1 or 2 potassium ions appeared within the active site. We found that during the first and second simulations,

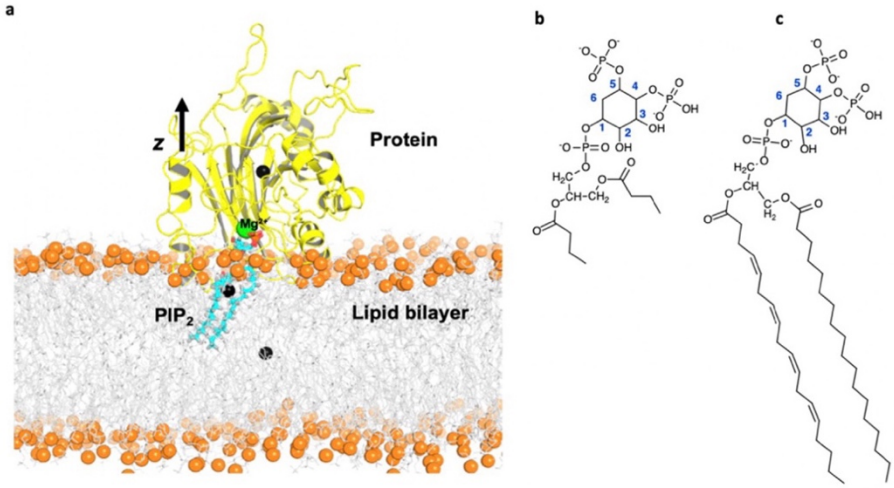


Fig. 6. Simulation setup for the membrane-embedded simulations (a). The centres of mass (black spheres) for each component of the simulation: lipid bilayer (orange spheres and grey sticks), bound PIP_2 ligand (blue sticks), and protein (yellow cartoon) are shown. In membrane-free simulations the ligand was modified (b) from the structure of PIP_2 (c). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

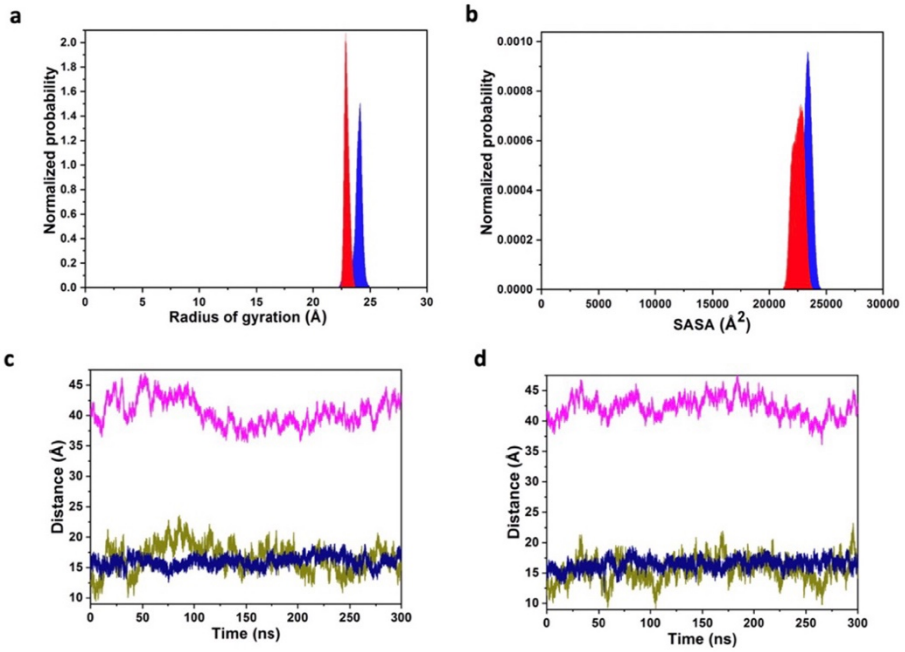


Fig. 7. Stability measures of the synaptojanin-1 complex. (a) Radius of gyration and (b) solvent accessible surface area (SASA) distributions of the protein with a resolution of 0.03 Å and 20.0 Å, respectively. The area under each probability distribution curve is normalized to unity. (c and d) Distances between centre of mass positions projected onto the z axis for simulations 1 and 2, respectively (pink – lipid bilayer and protein; olive – lipid bilayer and PIP_2; blue – Mg^{2+} and PIP_2). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

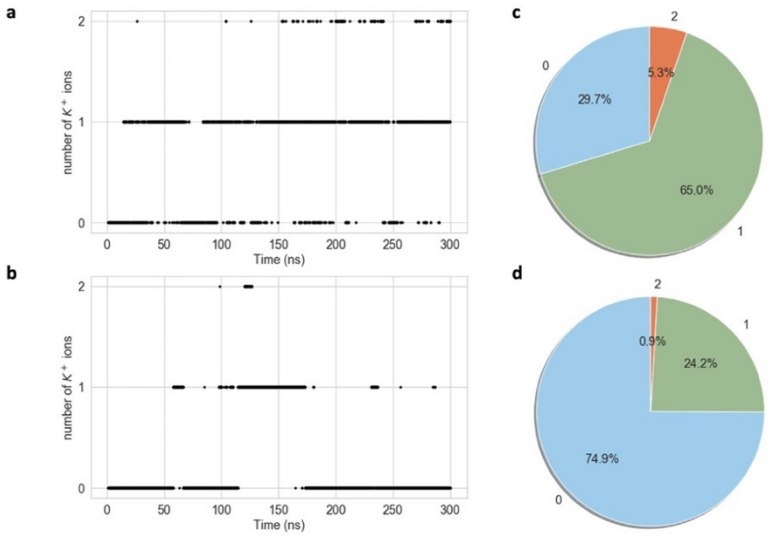


Fig. 8. K^+ ions within 4.5 Å of the active site during the course of simulation one (a) and simulation two (b). Percentage of 0, 1, and 2 K^+ ions, respectively, during simulation one (c) and simulation two (d).

over 70% and 25% of the time respectively, there was at least one potassium ion within 4.5 Å of 5′P. This suggests that a second positive ion is required to balance the negative charge within the binding pocket. As can be seen in Fig. 8a and b, the potassium ions approach the active site and then go away, with the number of potassium ions fluctuating constantly between 0, 1 or 2. It was observed that the localisation of the potassium cations within the active site was dependent on the orientation of the 4′phosphate group of the PIP_2 ligand. During the second simulation the 4′phosphate group undergoes rotation, which alters the site where

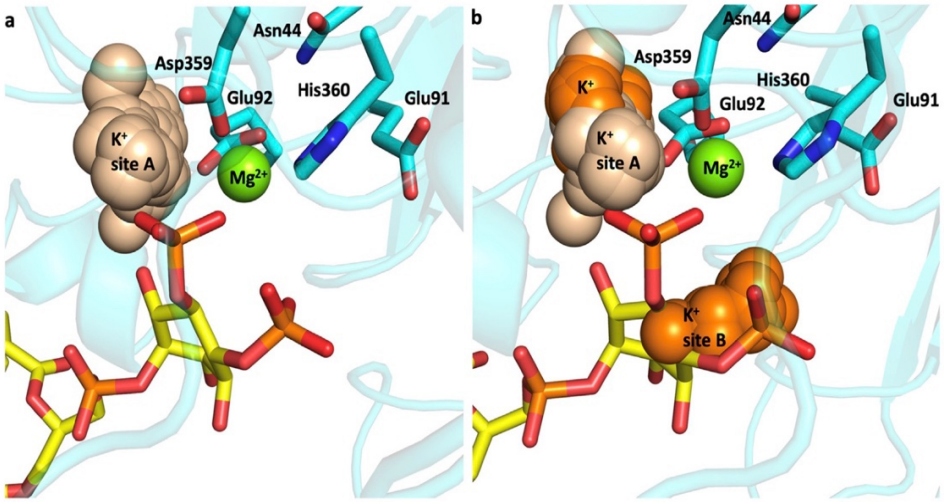


Fig. 9. Main K^+ binding sites A and B during simulation 1 (a) and simulation 1 and 2 (b). K^+ population from simulation one is coloured in wheat, K^+ population from simulation two is coloured in orange. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.) The positions of the K^+ ions are shown relative to the active site Mg^{2+}-binding residues and the Mg^{2+} cation of the first simulation frame.

the K^+ cations localise within the active site. Fig. 9 illustrates the two preferential binding sites for the potassium ions, defined here as binding site A and B. If the 4′phosphate group of the PIP_2 remains in its original orientation for the entire 300 ns of the simulation, then the cations preferentially cluster within the same spatial region in the active site and form only one binding site (Fig. 9a). In a rotated orientation where the phosphate groups of the PIP_2 4′ and 5′ point in opposite directions, the K^+ cations cluster in two locations, as depicted in Fig. 9b.

4.3. PIP_2 dephosphorylation requires 2-metal-ion active site

Our MD simulations showed that the catalytic site of synj1 has a positive charge deficiency, attracting potassium ions to approach and remain within 4.5 Å of the catalytic site. Given the consistent location of these potassium ions, this suggests that synj1 functions as a two-metal ion catalytic system. Currently, all known crystal structures of phosphoinositide 5-phosphatases have been resolved with only a single Mg^{2+} ion bound in the active site [30,39,58]. However, our simulations used the catalytically active ligand, PIP_2, as opposed to the inactive protein–ligand complexes with synthetic derivatives resolved structurally. Furthermore, we also included the membrane environment not present in these crystallographic structures. We found that PIP_2 does not significantly change its conformation with respect to the protein or the lipids but rather the potassium ions do approach the highly conserved catalytic residues, further supporting the suggestion that one magnesium ion may be insufficient. The occupancy of the potassium binding pocket within the active site continuously changes between 0, 1 and 2 ions during the course of the simulations, highlighting the openness of the binding site, which is likely only stable once the catalytic complex is correctly assembled. This could provide an explanation for the lack of crystallographic observation of the second metal ion.

Independently from our current work, various studies suggest that 5-phosphatases operate via the same catalytic mechanism as Mg^{2+}-dependent DNA restriction endonucleases, including the members DNase I and the apurinic/apyrimidinic base excision repair endonuclease Ape1 [63,64]. The conserved catalytic mechanism of the same 5-phosphate-type cleavage reaction is supported by evolutionarily strongly conserved amino acid sequence motifs within the active site (Fig. 6, ESI) [64]. Various endonuclease structures have been resolved with 2 cations within the active site, further suggesting that synj1 might also operate as a two-ion catalytic system [34,65]. Previous MD simulations of Ape1 also suggested the possible transient transfer between the two metal ion locations, termed "moving metal mechanism" [66], however, we do not observe evidence for such a mechanism in our simulations.

Additionally, biochemical experiments using various Mg^{2+} and Ca^{2+} concentrations also support the two-metal ion catalytic mechanism. Two metal-binding sites, each with a distinct binding affinity, are expected to yield biphasic inhibition curves when titrated with a non-productive metal. These bimodal effects were observed for APE1 further supporting that two metal ions are required for the catalytic reaction [34].

4.4. Synaptojanin-1 binding to PIP_2 and the potential for drug therapy

The importance of understanding the binding mechanism between synaptojanin-1 and the phosphatidylinositol phosphates has been already established. This understanding creates the potential for a new drug target. The MD simulations discussed in this work have achieved new insight into this binding process. It was shown that the PIP_2 bound to the synaptojanin-1 complex is stable and relatively open as the protein needs to also interact with the membrane surface to bind to PIP_2. This opens the possibility of drug molecules potentially interfering with the binding process, which could be used to decrease the activity of the protein in neurodegenerative diseases where upregulation increases the pathogenicity of the disease, for example in Alzheimer's disease. The decreased expression of synaptojanin-1 in AD has been shown to be protective and aids in amyloid-beta clearance [9,22]. Any drug created would need to be carefully administered as uncontrolled downregulation of the protein can also be harmful, as seen in our data integration results. The drug target would also need to interact preferentially with synj1 over the other phosphatidylinositol 5-phosphatases, all of which have very similar catalytic sites. Due to this, it may be worthwhile investigating whether targeting other regions of synj1 may be preferential. Alternatively, a drug target may bind to a synj1-specific surface that interferes with the membrane interactions, preventing PIP_2 binding.

4.5. Using yeast to predict key proteins in neurodegenerative diseases

The utilisation of yeast to predict the most important proteins in neurodegenerative diseases in humans has been found to have many benefits. As yeast is a much simpler cell than a neuron and is a single-celled organism, it significantly reduces the complexity of the problem. It also has a much shorter lifespan making it easier to study and collect sufficient data upon [67]. In humans, we generally use post-mortem samples or positron emission tomography (PET), which are potentially not very effective methods for identifying early markers and causative processes of a disease, as they are not single cell methods [68]. Ideally, preferred therapies intervene before significant cell death, cognitive decline and bradykinesia occur, enabling a higher quality of life for patients. Many proteins that have been discovered to have an effect on human disease progression are identified by mutations that cause harmful effects in the protein, and subsequently increase the likelihood or speed of disease progression [69]. Using mutations to identify proteins related to disease while useful, does not necessarily aid in understanding the sporadic disease, or general disease pathway. It is possible to identify proteins that suppress disease progression in wild type cells, but when mutated are unable to perform their function and lead to increased disease progression, as well as those that are already actively exacerbating the disease in wild type cells. Using yeast where high throughput genomic and proteomic studies are regularly conducted, it is possible to combine multiple datasets in the hope to provide more insight into the effect of the non-mutated proteins on neurodegenerative diseases' progression [13,14,16]. However, arguably, the most significant drawback of this method is that neurodegenerative diseases are often developing at the synapse, which is not present in yeast. For this reason, any yeast-based method, including the data integration found in this work, cannot identify any neuron-specific proteins or pathways but rather generic cell pathways that are conserved in both humans and yeast, and so invariably they will be proteins that are highly conserved across all eukaryotes. This is the underlying reason why the 17 candidate proteins found are primarily involved in processes or organelles that are ubiquitous across eukaryotes; with many linked to the mitochondria and its associated processes. Data integration is still a very powerful tool as it has been possible not only to investigate the effect of α-synuclein aggregation upon protein concentrations in the cell, but also how these perturbations in protein concentration may be altering the toxicity of aggregation [14,20].

5. Conclusions

The wealth of biological information currently being produced requires new approaches to interpret and utilise the data so that we maximally filter useful information. Data integration is one possibility that could enable us to reuse data that is currently under-utilised. This is particularly beneficial as it does not require conducting more experiments to gain more information. Using this principle of data integration, two large scale studies of α-synuclein induction in budding yeast were analysed and used to identify 17 proteins that could be of interest in human PD and AD. Most of these 17 proteins were found to be related to human diseases, either directly or indirectly.

Among those, we chose to investigate further the 5-phosphatase domain of the regulatory lipid phosphatase synaptojanin-1. By dysregulation of various PIPs, the malfunction of synj1 is linked to the decrease of cell health and increase of proteomic stress. Synaptojanin-1 dephosphorylates the phosphatidylinositol PIP_2 at the synapse membrane. The catalytic function is carried out through an interaction with an essential coordinating magnesium cofactor. Through all-atom MD simulations including the membrane and the PIP_2-bound protein, we observed that the proposed catalytic site was stable, but potassium ions persistently approached the binding pocket. This suggests that another positive charge is required for a catalytically active complex. Therefore, we propose that synj1 is likely using a two-metal ion catalytic mechanism for its phosphatase function. Current human phosphoinositide 5-phosphatases are all resolved crystallographically with only a single metal ion at the active site. Future work on synj1 could confirm our results via high-resolution crystal structures, or by biochemical measurements on the effects of mutations at the catalytic site or using Mg^{2+} concentration-dependent catalytic rate measurements. This would be particularly beneficial in future targeting of the active site.

Our work identifies potential novel targets for α-synuclein aggregating diseases. Furthermore, it provides the first atomistic investigation of the human synj1 main 5-phosphatase catalytic domain. Our novel structural information could potentially enable the design of a small molecule inhibitor that could prevent or destabilise PIP_2 binding, leading to a novel avenue for disease therapy where decreasing synj1 activity can be beneficial.

CRediT authorship contribution statement

Kirsten Jenkins: Conceptualization, Methodology, Data curation, Visualization, Writing - original draft. **Teodora Mateeva:** Methodology, Data curation, Visualization, Writing - original draft, Writing - review & editing. **István Szabó:** Methodology, Visualization. **Andre Melnik:** Methodology, Data curation. **Paola Picotti:** Conceptualization, Methodology, Writing - review & editing, Supervision. **Attila Csikász-Nagy:** Conceptualization, Methodology, Writing - original draft, Writing - review & editing, Supervision. **Edina Rosta:** Conceptualization, Methodology, Writing - original draft, Supervision, Writing - review & editing, Visualization.

Acknowledgement

We acknowledge the EPSRC Centre for Doctoral Training in Cross-Disciplinary Approaches to Non-Equilibrium Systems (CANES, EP/L015854/1), EPSRC Grant No. EP/R013012/1 and ERC Project Nos. 757850 BioNet and 866004 Proteomes-in-3D. PP also acknowledges the Swiss National Science Foundation, Sinergia grant (SNSF grant, CRSII5_177195) and the Personalized Health and Related Technologies grant (PHRT-506). TM acknowledges funding from the Agency for Science, Technology and Research (A*STAR) Singapore Research Attachment Programme (ARAP). We acknowledge the use of the research computing facility at King's College London, Rosalind (https://rosalind.kcl.ac.uk).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.04.010.

References

[1] Lewis F, Schaffer SK, Sussex J, O'Neil P, Cockcroft L. The trajectory of dementia in the UK - Making a difference. Technical Report 2014.
[2] National Life Tables, United Kingdom - Office for National Statistics n.d. http://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/bulletins/nationallifetablesunitedkingdom/2014-09-25 (accessed September 25, 2014).
[3] Emamzadeh FN, Surguchov A. Parkinson's disease: Biomarkers, treatment, and risk factors. Front Neurosci 2018. https://doi.org/10.3389/fnins.2018.00612.
[4] Ross CA, Poirier MA. Protein aggregation and neurodegenerative disease. Nat Med 2004. https://doi.org/10.1038/nm1066.

[5] Briggs R, Kennelly SP, O'Neill D. Drug treatments in Alzheimer's disease. Clin Med J R Coll Physicians London 2016. https://doi.org/10.7861/clinmedicine.16-3-247.

[6] Casey DA, Antimisiaris D, O'Brien J. Drugs for Alzheimer's disease: Are they effective?. P T 2010.

[7] Surguchov A. Intracellular dynamics of synucleins: "Here, There and Everywhere". Int Rev Cell Mol Biol 2015. https://doi.org/10.1016/bs.ircmb.2015.07.007.

[8] Hipp MS, Park SH, Hartl UU. Proteostasis impairment in protein-misfolding and -aggregation diseases. Trends Cell Biol 2014. https://doi.org/10.1016/j.tcb.2014.05.003.

[9] Drouet V, Lesage S. Synaptojanin 1 Mutation in Parkinson's disease brings further insight into the neuropathological mechanisms. Biomed Res Int 2014. https://doi.org/10.1155/2014/289728.

[10] Chartier-Harlin MC, Kachergus J, Roumier C, Mouroux V, Douay X, Lincoln S, et al. α-synuclein locus duplication as a cause of familial Parkinson's disease. Lancet 2004. https://doi.org/10.1016/S0140-6736(04)17103-1.

[11] Singleton AB, Farrer M, Johnson J, Singleton A, Hague S, Kachergus J, et al. α-Synuclein Locus Triplication Causes Parkinson's disease. Science (80-) 2003. https://doi.org/10.1126/science.1090278.

[12] Krüger R, Kuhn W, Müller T, Woitalla D, Graeber M, Kösel S, et al. Ala30Pro mutation in the gene encoding α-synuclein in Parkinson's disease. Nat Genet 1998. https://doi.org/10.1038/ng0298-106.

[13] Khurana V, Lindquist S. Modelling neurodegeneration in Saccharomyces cerevisiae: why cook with baker's yeast?. Nat Rev Neurosci 2010. https://doi.org/10.1038/nrn2809.

[14] Khurana V, Peng J, Chung CY, Auluck PK, Fanning S, Tardiff DF, et al. Genome-scale networks link neurodegenerative disease genes to α-synuclein through specific molecular pathways. Cell Syst 2017. https://doi.org/10.1016/j.cels.2016.12.011.

[15] Duda JE, Lee VM-Y, Trojanowski JQ. Neuropathology of synuclein aggregates. J Neurosci Res 2000. https://doi.org/10.1002/1097-4547(20000715)61:2<121::aid-jnr1>3.0.co;2-4.

[16] Willingham S, Outeiro TF, DeVit MJ, Lindquist SL, Muchowski PJ. Yeast genes that enhance the toxicity of a mutant Huntingtin fragment or α-synuclein. Science (80-) 2003. https://doi.org/10.1126/science.1090389.

[17] Wanichthanarak K, Fahrmann JF, Grapov D. Genomic, proteomic, and metabolomic data integration strategies. Biomark Insights 2015. https://doi.org/10.4137/BMI.S29511.

[18] Gligorijević V, Pržulj N. Methods for biological data integration: perspectives and challenges. J R Soc Interface 2015. https://doi.org/10.1098/rsif.2015.0571.

[19] Katsila T, Spyroulias GA, Patrinos GP, Matsoukas MT. Computational approaches in target identification and drug discovery. Comput Struct Biotechnol J 2016. https://doi.org/10.1016/j.csbj.2016.04.004.

[20] MelnikA, Cappellutti V, Vaggi F, Piazza I, Tognetti M, Soste M, de Souza N, Csikasz-Nagy A, Piccotti P. In Preparation 2019.

[21] Chen KH, Wu RM, Lin HI, Tai CH, Lin CH. Mutational analysis of SYNJ1 gene (PARK20) in Parkinson's disease in a Taiwanese population. Neurobiol Aging 2015. https://doi.org/10.1016/j.neurobiolaging.2015.06.009.

[22] McIntire LBJ, Berman DE, Myaeng J, Staniszewski A, Arancio O, Di Paolo G, et al. Reduction of synaptojanin 1 ameliorates synaptic and behavioral impairments in a mouse model of Alzheimer's disease. J Neurosci 2012. https://doi.org/10.1523/JNEUROSCI.2034-12.2012.

[23] Olgiati S, De Rosa A, Quadri M, Criscuolo C, Breedveld GJ, Picillo M, et al. PARK20 caused by SYNJ1 homozygous Arg258Gln mutation in a new Italian family. Neurogenetics 2014. https://doi.org/10.1007/s10048-014-0406-0.

[24] Quadri M, Fang M, Picillo M, Olgiati S, Breedveld GJ, Graafland J, et al. Mutation in the SYNJ1 gene associated with autosomal recessive, early-onset parkinsonism. Hum Mutat 2013. https://doi.org/10.1002/humu.22373.

[25] Miranda AM, Herman M, Cheng R, Nahmani E, Barrett G, Micevska E, et al. Excess synaptojanin 1 contributes to place cell dysfunction and memory deficits in the aging hippocampus in three types of Alzheimer's disease. Cell Rep 2018. https://doi.org/10.1016/j.celrep.2018.05.011.

[26] Waugh MG. PIPs in neurological diseases. Biochim Biophys Acta - Mol Cell Biol Lipids 2015. https://doi.org/10.1016/j.bbalip.2015.02.002.

[27] Vanhauwaert R, Kuenen S, Masius R, Bademosi A, Manetsberger J, Schoovaerts N, et al. The SAC 1 domain in synaptojanin is required for autophagosome maturation at presynaptic terminals. EMBO J 2017. https://doi.org/10.15252/embj.201695773.

[28] Montesinos ML, Castellano-Muñoz M, García-Junco-Clemente P, Fernández-Chacón R. Recycling and EH domain proteins at the synapse. Brain Res Rev 2005. https://doi.org/10.1016/j.brainresrev.2005.06.002.

[29] Krebs CE, Karkheiran S, Powell JC, Cao M, Makarov V, Darvish H, et al. The sac1 domain of SYNJ1 identified mutated in a family with early-onset progressive parkinsonism with generalized seizures. Hum Mutat 2013. https://doi.org/10.1002/humu.22372.

[30] Hsu FS, Mao Y. The structure of phosphoinositide phosphatases: Insights into substrate specificity and catalysis. Biochim Biophys Acta - Mol Cell Biol Lipids 2015;1851:698–710. https://doi.org/10.1016/j.bbalip.2014.09.015.

[31] Berta D, Buigues PJ, Badaoui M, Rosta E. Cations in motion: QM/MM studies of the dynamic and electrostatic roles of H+ and Mg2+ ions in enzyme reactions. Curr Opin Struct Biol 2020. https://doi.org/10.1016/j.sbi.2020.01.002.

[32] Chi Y, Zhou B, Wang WQ, Chung SK, Kwon YU, Ahn YH, et al. Comparative mechanistic and substrate specificity study of inositol polyphosphate 5-phosphatase Schizosaccharomyces pombe synaptojanin and SHIP2. J Biol Chem 2004. https://doi.org/10.1074/jbc.M406416200.

[33] Rosta E, Yang W, Hummer G. Calcium inhibition of ribonuclease H1 two-metal ion catalysis. J Am Chem Soc 2014. https://doi.org/10.1021/ja411408x.

[34] Beernink PT, Segelke BW, Hadi MZ, Erzberger JP, Wilson DM, Rupp B. Two divalent metal ions in the active site of a new crystal form of human apurinic/apyrimidinic endonuclease, Ape 1: implications for the catalytic mechanism. J Mol Biol 2001. https://doi.org/10.1006/jmbi.2001.4529.

[35] Balakrishnan R, Park J, Karra K, Hitz BC, Binkley G, Hong EL, et al. YeastMine-An integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. Database 2012. https://doi.org/10.1093/database/bar062.

[36] Rappaport N, Nativ N, Stelzer G, Twik M, Guan-Golan Y, Stein TI, et al. MalaCards: an integrated compendium for diseases and their annotation. Database 2013. https://doi.org/10.1093/database/bat018.

[37] Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 2014. https://doi.org/10.1093/nar/gku340.

[38] Hancock JM, Zvelebil MJ, Zvelebil MJ. UniProt. Dict. Bioinforma. Comput Biol 2004. https://doi.org/10.1002/9780471650126.dob0721.pub2.

[39] Tsujishita Y, Guo S, Stolz LE, York JD, Hurley JH. Specificity determinants in phosphoinositide dephosphorylation: crystal structure of an archetypal inositol polyphosphate 5-phosphatase. Cell 2001. https://doi.org/10.1016/S0092-8674(01)00326-9.

[40] Phillips JC, Braun R, Wang W, Gumbart J, Tajkhorshid E, Villa E, et al. Scalable molecular dynamics with NAMD. J Comput Chem 2005. https://doi.org/10.1002/jcc.20289.

[41] Huang J, Mackerell AD. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. J Comput Chem 2013. https://doi.org/10.1002/jcc.23354.

[42] Darden T, York D, Pedersen L. Particle mesh Ewald: an N·log(N) method for Ewald sums in large systems. J Chem Phys 1993. https://doi.org/10.1063/1.464397.

[43] Jo S, Kim T, Iyer VG, Im W. CHARMM-GUI: a web-based graphical user interface for CHARMM. J Comput Chem 2008. https://doi.org/10.1002/jcc.20945.

[44] Lee J, Cheng X, Swails JM, Yeom MS, Eastman PK, Lemkul JA, et al. CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field. J Chem Theory Comput 2016. https://doi.org/10.1021/acs.jctc.5b00935.

[45] Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI. OPM: orientations of proteins in membranes database. Bioinformatics 2006. https://doi.org/10.1093/bioinformatics/btk023.

[46] Rappaport N, Twik M, Plaschkes I, Nudel R, Stein TI, Levitt J, et al. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. Nucleic Acids Res 2017. https://doi.org/10.1093/nar/gkw1012.

[47] Estrada-Cuzcano A, Martin S, Chamova T, Synofzik M, Timmann D, Holemans T, et al. Loss-of-function mutations in the ATP13A2/PARK9 gene cause complicated hereditary spastic paraplegia (SPG78). Brain 2017. https://doi.org/10.1093/brain/aww307.

[48] Larkin JC, Thompson JR, Woolford JL. Structure and expression of the Saccharomyces cerevisiae CRY1 gene: a highly conserved ribosomal protein gene. Mol Cell Biol 1987. https://doi.org/10.1128/mcb.7.5.1764.

[49] Moritz M, Paulovich AG, Tsay YF, Woolford JL. Depletion of yeast ribosomal proteins L16 or rp59 disrupts ribosome assembly. J Cell Biol 1990. https://doi.org/10.1083/jcb.111.6.2261.

[50] Lecompte O, Ripp R, Thierry JC, Moras D, Poch O. Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. Nucleic Acids Res 2002. https://doi.org/10.1093/nar/gkf693.

[51] Jakovljevic J, De Mayolo PA, Miles TD, Nguyen TML, Léger-Silvestre I, Gas N, et al. The carboxy-terminal extension of yeast ribosomal protein S14 is necessary for maturation of 43S preribosomes. Mol Cell 2004. https://doi.org/10.1016/S1097-2765(04)00215-1.

[52] Shu Y, Yang H, Hallberg E, Hallberg R. Molecular genetic analysis of Rts1p, a B' regulatory subunit of Saccharomyces cerevisiae protein phosphatase 2A. Mol Cell Biol 1997. https://doi.org/10.1128/mcb.17.6.3242.

[53] Dobbelaere J, Gentry MS, Hallberg RL, Barral Y. Phosphorylation-dependent regulation of septin dynamics during the cell cycle. Dev Cell 2003. https://doi.org/10.1016/S1534-5807(03)00061-3.

[54] Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P. STITCH: interaction networks of chemicals and proteins. Nucleic Acids Res 2008. https://doi.org/10.1093/nar/gkm795.

[55] Vijayan V, Verstreken P. Autophagy in the presynaptic compartment in health and disease. J Cell Biol 2017. https://doi.org/10.1083/jcb.201611113.

[56] Pierzynowska K, Gaffke L, Cyske Z, Puchalski M, Rintz E, Bartkowski M, et al. Autophagy stimulation as a promising approach in treatment of neurodegenerative diseases. Metab Brain Dis 2018. https://doi.org/10.1007/s11011-018-0214-6.

[57] Landman N, Jeong SY, Shin SY, Voronov SV, Serban G, Kang MS, et al. Presenilin mutations linked to familial Alzheimer's disease cause an imbalance in phosphatidylinositol 4,5-bisphosphate metabolism. Proc Natl Acad Sci U S A 2006. https://doi.org/10.1073/pnas.0604954103.

[58] Trésaugues L, Silvander C, Flodin S, Welin M, Nyman T, Gräslund S, et al. Structural basis for phosphoinositide substrate recognition, catalysis, and membrane interactions in human inositol polyphosphate 5-phosphatases. Structure 2014. https://doi.org/10.1016/j.str.2014.01.013.

[59] Clustalw U, To C, Multiple DO. ClustalW and ClustalX. Options 2003. https://doi.org/10.1002/0471250953.bi0203s00.
[60] Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004. https://doi.org/10.1093/nar/gkh340.
[61] Gouet P, Courcelle E, Stuart DI, Métoz F. ESPript: Analysis of multiple sequence alignments in PostScript. Bioinformatics 1999. https://doi.org/10.1093/bioinformatics/15.4.305.
[62] Mills SJ, Persson C, Cozier G, Thomas MP, Trésaugues L, Erneux C, et al. A synthetic polyphosphoinositide headgroup surrogate in complex with SHIP2 provides a rationale for drug discovery. ACS Chem Biol 2012. https://doi.org/10.1021/cb200494d.
[63] Dlakić M. Functionally unrelated signalling proteins contain a fold similar to Mg2+-dependent endonucleases. Trends Biochem Sci 2000. https://doi.org/10.1016/S0968-0004(00)01582-6.
[64] Whisstock JC, Romero S, Gurung R, Nandurkar H, Ooms LM, Bottomley SP, et al. The inositol polyphosphate 5-phosphatases and the apurinic/apyrimidinic base excision repair endonucleases share a common mechanism for catalysis. J Biol Chem 2000. https://doi.org/10.1074/jbc.M006244200.
[65] Pingoud A. Structure and function of type II restriction endonucleases. Nucleic Acids Res 2001. https://doi.org/10.1093/nar/29.18.3705.
[66] Oezguen N, Schein CH, Peddi SR, Power TD, Izumi T, Braun W. A "moving metal mechanism" for substrate cleavage by the DNA repair endonuclease APE-1. Proteins Struct Funct Genet 2007. https://doi.org/10.1002/prot.21397.
[67] Cell biology by the numbers. Choice Rev Online 2016. https://doi.org/10.5860/choice.196525.
[68] Brettschneider J, Del Tredici K, Lee VMY, Trojanowski JQ. Spreading of pathology in neurodegenerative diseases: a focus on human studies. Nat Rev Neurosci 2015. https://doi.org/10.1038/nrn3887.
[69] Kalinderi K, Bostantjopoulou S, Fidani L. The genetic background of Parkinson's disease: current progress and future prospects. Acta Neurol Scand 2016. https://doi.org/10.1111/ane.12563.

Chapter 6

Direct Calculation of Electron Transfer Rates with the Binless Dynamic Weighted Histogram Analysis Method

This Chapter was published in The Journal of Physical Chemistry Letters in 2023 and is reproduced here with permission from: Zsuzsanna Koczor-Benda, Teodora Mateeva, and Edina Rosta, 'Direct Calculation of Electron Transfer Rates with the Binless Dynamic Histogram Analysis Method', J. Phys. Chem. B, DOI:10.1021/acs.jpclett.3c02624. Copyright Journal of Physical Chemistry Letters 2023.

Summary of the Work

Umbrella sampling simulations are commonly employed for situations in which one is interested in events that are difficult to visit by unbiased sampling. In these cases, artificial bias is applied along a Reaction Coordinate (RC) to visit events that would otherwise remain unvisited. The bias can be removed, to obtain global free energy profiles for the respective event. An existing method that is commonly applied and achieves this is the Weighted Histogram Analysis Method (WHAM), however, WHAM disregards time sequence and kinetic information. An alternative method that provides kinetic information is DHAM. Here we present Binless DHAM, which extends the applicability of DHAM to high-dimensional and Hamiltonian-based biasing, enabling the study of electron transfer (ET) processes. By using classical Hamiltonian-based umbrella sampling simulations and electronic coupling values from quantum chemistry calculations, Binless DHAM successfully provides ET rates for both adiabatic and non-adiabatic ET reactions, with excellent agreement with experimental results.

Author Contribution

I set up and performed all MD simulations for the ferrous-ferric model system in water and the $(Q\text{-}TTF\text{-}Q)^-$ anion in different solvents, which were used to test the binless DHAM method. All QM-level calculations were run by Dr. Zsuzsanna Koczor-Benda. All H_{ab} coupling parameters were obtained by Dr. Zsuzsanna Koczor-Benda. The theoretical foundations of the method were developed by Prof. Edina Rosta. The original Matlab code used to implement this method was developed by Prof. Edina Rosta and later edited, optimized, and tested by Dr. Zsuzsanna Koczor-Benda and me. I have later contributed to a Python version of this code. All authors contributed to the writing and editing of this text and approved the manuscript in its final form. The figures in this manuscript and the corresponding SI information were generated with equal contributions from me and Dr. Zsuzsanna Koczor-Benda.

The Supporting Information for this manuscript is available in Appendix D.

THE JOURNAL OF PHYSICAL CHEMISTRY LETTERS
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

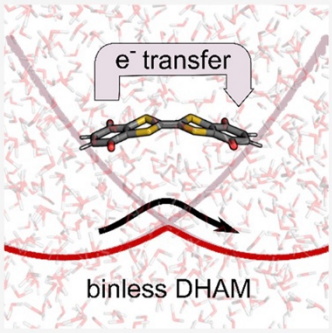
pubs.acs.org/JPCL
Letter

Direct Calculation of Electron Transfer Rates with the Binless Dynamic Histogram Analysis Method

Zsuzsanna Koczor-Benda,[▽] Teodora Mateeva,[▽] and Edina Rosta*

Cite This: J. Phys. Chem. Lett. 2023, 14, 9935−9942
Read Online
ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Umbrella sampling molecular dynamics simulations are widely used to enhance sampling along the reaction coordinates of chemical reactions. The effect of the artificial bias can be removed using methods such as the dynamic weighted histogram analysis method (DHAM), which in addition to the global free energy profile also provides kinetic information about barrier-crossing rates directly from the Markov matrix. Here we present a binless formulation of DHAM that extends DHAM to high-dimensional and Hamiltonian-based biasing to allow the study of electron transfer (ET) processes, for which enhanced sampling is usually not possible based on simple geometric grounds. We show the capabilities of binless DHAM on examples such as aqueous ferrous-ferric ET and intramolecular ET in the radical anion of benzoquinone−tetrathiafulvalene−benzoquinone (Q-TTF-Q)$^{-}$. From classical Hamiltonian-based umbrella sampling simulations and electronic coupling values from quantum chemistry calculations, binless DHAM provides ET rates for adiabatic and nonadiabatic ET reactions alike in excellent agreement with experimental results.



The calculation of free energy profiles is central for modeling chemical reactions. In molecular dynamics (MD) simulations, it is common to employ biasing functions to facilitate the exploration of otherwise rarely visited regions of the free energy surface. To overcome barriers in free energy surfaces, the umbrella sampling (US) method and analogous biased simulations are often used, where the free energy profile is estimated along one or more collective variables (CVs).[1,2] In chemical reactions, there is usually one or a small number of nuclear coordinate changes that describe the transition from reactants to products. In electron transfer (ET) reactions that are not coupled to other chemical changes (e.g., proton transfer), this is not the case. In outer-sphere ET reactions, for example, the reactants and products are different only in the rearrangement of the electron density and the corresponding complex changes in the environment. For ET reactions, instead of nuclear coordinates, the reaction coordinate is better defined as the energy gap (i.e., difference) between diabatic charge localized states.[3,4]

To unbias US-type enhanced sampling simulations and to construct the free energy profile along one (or a few) reaction coordinate(s), the weighted histogram analysis method (WHAM)[5] is commonly used. However, WHAM disregards the time sequence information within simulation trajectories and therefore kinetic information is lost. To obtain molecular kinetics, the dynamic histogram analysis method (DHAM) was developed,[6] as well as its more robust implementation using rate matrices instead of transition matrices via DHAMed.[7] However, when simulations are biased along many coordinates or via biasing functions that may not be related to the reaction coordinate, DHAM needs to be reformulated. We introduce here a modification of DHAM, called binless DHAM, that approximates the unbiased Markov matrix and thus allows for unbiasing in such cases. The term "binless" is used to reflect the similarity to the multistate Bennett acceptance ratio estimator (MBAR),[8] which is an analogous binless implementation of WHAM.[9] The key advantage of binless DHAM over MBAR is that it also directly provides reaction rates. This provides an alternative to Eyring's transition state theory (TST) or Marcus theory for nonadiabatic ET, which calculates rates from activation free energies or Marcus parameters (driving force, reorganization energy, and electronic coupling), respectively. Another method that reports being able to obtain rates directly is dTRAM.[10] Similarly to DHAM, dTRAM does not require data to be sampled from global equilibrium and provides maximum-likelihood estimates of stationary quantities. However, no rates have been reported to be calculated for model systems. While dTRAM in principle can also provide kinetics from multiensemble simulations, this requires that unbiased simulation data are also included, which is typically not available for ET simulations and in many other cases.[11]

We demonstrate the binless DHAM method on various systems, focusing on condensed-phase ET reactions, where a dynamical description of the solvent is essential. To sample

Received: September 18, 2023
Revised: October 18, 2023
Accepted: October 19, 2023

https://doi.org/10.1021/acs.jpclett.3c02624
J. Phys. Chem. Lett. 2023, 14, 9935−9942

different ensembles of configurations and build diabatic free-energy surfaces, we perform Hamiltonian-based US MD, where we vary the charges of donor and acceptor subunits incrementally. This US technique for ET processes has been previously applied in semiclassical and ab initio MD studies[12−14] and is also similar to λ-dynamics[15] used in the context of protein−ligand binding.

Our first example is the ferrous-ferric self-exchange ET process, an often-used test case for new methodologies and an example for nonadiabatic ET. Previous molecular dynamics simulations of this system used classical force fields,[13,16] ab initio MD (Car−Parrinello, CPMD),[14] or quantum mechanics/molecular mechanics (QM/MM) MD[17] to determine free energy profiles and Marcus parameters. The electronic coupling has been investigated with various quantum chemistry methods such as fractional occupation number density functional theory (FON-DFT),[18] restricted open-shell Hartree−Fock ROHF,[19] and a model considering a quantal electron and classical Fe^{3+} ions.[13] Here we use frozen density embedding (FDE)[20,21] to calculate the electronic coupling on frames from MD simulations.

The second example is the intramolecular electron transfer (IET) within the radical anion of the benzoquinone−tetrathiafulvalene−benzoquinone triad $(Q\text{-}TTF\text{-}Q)^-$ in four different solvents: *tert*-butyl alcohol (tBOH), dichloromethane (DCM), ethyl acetate (ETA), and water. The $(Q\text{-}TTF\text{-}Q)^-$ anion is a type II compound according to the Robin-Day classification scheme,[22] in polar solvents, meaning that its ground state is charge-localized and ET between the two parts of the molecule is well approximated by the adiabatic mechanism. Organic compounds capable of IET, such as tetrathiafulvalene (TTF) derivatives, are gathering interest for their potential as organic conductors.[23] The understanding of the IET in the $(Q\text{-}TTF\text{-}Q)^-$ anion and other TTF derivatives could enable the engineering of the ET process which has potential applications in the design of molecular wires and other applications in nanotechnology.[24,25] However, the estimation of the IET currently represents a challenging task for computational methods.[26] The correct description of the system poses a challenge for quantum chemistry methods.[27−34] The electronic coupling was previously calculated with CDFT in the gas phase,[27] with CDFT on ab initio MD simulation frames for the unconstrained charge delocalized state including explicit solvent,[28] directly with CDFT MD,[35] and with time-dependent (TD) DFT.[29] We add to this variety of techniques by determining the coupling with an equation-of-motion coupled cluster (EOM-CC) approach as well.

For the ferrous-ferric ET, we achieve excellent agreement with the experimental rate ($5.2 \times 10^2\ s^{-1}$ calculated vs $7.9 \times 10^2\ s^{-1}$ experimental[36]). For the IET in $(Q\text{-}TTF\text{-}Q)^-$, the calculated rates are within one order of magnitude of the experimentally reported ones.

The relation between biased and unbiased Markov transition probability matrices M can be expressed by solving the Smoluchowski diffusion equation[37] for transition probabilities $p(i \to j,\tau)$ from state i to j within a lag time τ as follows:

$$\frac{M_{ji}^k}{M_{ji}^0} = \frac{p(i \to j,\ \tau)^k}{p(i \to j,\ \tau)^0} = \frac{\exp\left(-\left((x_j - x_i) + \gamma\tau\frac{U_j^k - U_i^k + U_j^0 - U_i^0}{x_j - x_i}\right)^2/4D\tau\right)}{\exp\left(-\left((x_j - x_i) + \gamma\tau\frac{U_j^0 - U_i^0}{x_j - x_i}\right)^2/4D\tau\right)} \tag{1}$$

with superscipt k denoting biased simulation k and 0 denoting the unbiased case. DHAM assumes a shared diffusion coefficient D between biased and unbiased simulations, which can nevertheless be position-dependent. U is the potential energy along the x reaction coordinate, and $\gamma = D/k_BT$ is the mobility of the system. Expanding the squared terms in eq 1 and omitting all τ^2 terms lead to the square root approximation at short lag times,[38]

$$\frac{M_{ji}^k}{M_{ji}^0} \approx \exp(-(U_j^k - U_i^k)/2k_BT) \tag{2}$$

In regular DHAM,[39] the unnormalized Markov matrix is defined as

$$M_{ji} = \sum_{k=1}^{N} \frac{T_{ji}^k}{\sum_{l=1}^{N} n_i^l \exp(-(u_j^l(c_j) - u_i^l(c_i))/2k_BT} \tag{3}$$

where data is binned along x, and

$$T_{ji}^k = \sum_{t}^{L^k-\tau} \delta(x^k(t) \in i)\delta(x^k(t+\tau) \in j)$$

gives the transition count from bin i to bin j in simulation window k, with data saved and analyzed at the frequency of τ (lag time) from the overall L^k length of simulation k. $n_i^k = \sum_j T_{ji}^k$ is the number of transitions initiating from bin i. The bias $u_i^l = U_i^l - U_i^0$ is evaluated at each bin center c_i, assuming that the biasing is also done along x.

The formally exact expression in eq 3 can be approximated by calculating the bias at the actual value of the reaction coordinate for each frame (x_t^k) instead of c_i, similarly to the binless formulation of WHAM.[9] This approximation becomes exact in the limit of very small bin sizes. With this binless formulation, it is then straightforward to obtain M_{ji} for any bias along arbitrary coordinates q_t^k

$$M_{ji} = \sum_{k=1}^{N} \sum_{t}^{L^k-\tau} \frac{\delta(x^k(t) \in i)\delta(x^k(t+\tau) \in j)}{\sum_{l=1}^{N} n_i^l \exp(-(u^l(q_{t+\tau}^k) - u^l(q_t^k))/2k_BT} \tag{4}$$

Equation 3 can also be approximated by evaluating $u^l(q_t^k \in i)$ for all q_t^k data points that fall into bin i and using the average or median values in the denominator (see section S1 of the Supporting Information). This was also used to re-weight free energies in a binless form of the conformational states for the Ala5 peptide with DHAMed.[40]

After normalizing the columns of M_{ji}, its right eigenvector corresponding to eigenvalue 1 gives the normalized equilibrium probabilities p_i, from which the free energy profile is calculated as $G_i = -k_BT\ln p_i$. In the ET examples below, we

https://doi.org/10.1021/acs.jpclett.3c02624
J. Phys. Chem. Lett. 2023, 14, 9935−9942

calculate the biasing energy with respect to the adiabatic ground state energy.

$$E_g = \frac{1}{2}(E_1 - E_2) - \sqrt{4H_{ab}^2 + (E_1 - E_2)^2} \tag{5}$$

Here $E_{1,2}$ are the charge localized diabatic states and H_{ab} is the electronic coupling between them.

Within semiclassical TST the ET rate can be written[41] as

$$k^{sc} = \kappa\Gamma\nu\exp\left(-\frac{\Delta G^{\ddagger}}{k_B T}\right) \tag{6}$$

where $\Delta G^{\ddagger}$ is the activation free energy, ν is the nuclear frequency factor that gives the frequency of reaching the transition state (TS), and κ is the electronic transmission coefficient that describes the probability of electron transfer at the transition state. Γ is the quantum correction factor accounting for nuclear quantum effects such as nuclear tunneling that can enhance the reaction rate, and it is usually considered to be 1; thus, we leave it out from the following formulas to be consistent with previous studies. The magnitude of κ depends on the electronic interaction between the redox pairs; when their interaction is sufficiently strong, then $\kappa \approx 1$ and the reaction is labeled as adiabatic, and when their coupling is small then $\kappa < 1$ and the reaction is nonadiabatic.

In contrast, in binless DHAM the reaction rate (k^M) is calculated directly from the second largest eigenvalue (m) of the normalized M_{ji}:

$$k^M = -\frac{\ln(m)}{\tau} \tag{7}$$

The rate calculated this way is equivalent to the adiabatic rate from TST (eq 6, $\kappa = 1$ case), providing a new way to determine the pre-exponential factor ν as

$$\nu = k^M \exp\left(\frac{\Delta G^{\ddagger}}{k_B T}\right) \tag{8}$$

This can be compared to the common approximation of ν as $k_B T/h$ or as the frequency of the vibrational mode transforming reactants to products (when such mode can be identified). To access nonadiabatic rates as well, only a correction by κ is needed, which can be calculated from Landau−Zener theory.[42−44]

$$\kappa = \begin{cases} \dfrac{2P_{LZ}}{1 + P_{LZ}} & \text{if } \Delta G^{\ddagger} \geq -\lambda \\ 2P_{LZ}(1 - P_{LZ}) & \text{if } \Delta G^{\ddagger} < -\lambda \end{cases} \tag{9.1}$$

$$P_{LZ} = 1 - \exp(-2\pi\gamma) \tag{9.2}$$

$$2\pi\gamma = \frac{\pi^{3/2} H_{ab}^2}{h\nu\sqrt{\lambda k_B T}} \tag{9.3}$$

Since the reorganization energy λ can be determined from the diabatic free energy profiles, the only external parameter needed to determine the nonadiabatic rate κk^M through eqs 7 and 9.1—9.3 is H_{ab}, which is already required to build the adiabatic ground state (eq 5).

In the Condon approximation,[45,46] H_{ab} is constant along the reaction coordinate, and its value is half the energy gap of the two adiabatic potential energy surfaces at the transition state. Calculating H_{ab} directly from the excitation energy is usually not reliable with single reference methods, which break near degeneracies of the ground and excited states. However, to ensure a balanced description of the interacting states,[47] one can take a well-behaved state such as the ground state of the neutral Q-TTF-Q as a starting point and use the electron attachment variant of equation-of-motion coupled cluster theory (EOM-EA-CC)[48] to get the ground and first excited states of the radical anion $(Q\text{-}TTF\text{-}Q)^-$.

A disadvantage of EOM-CC methods is that the solvent can be considered only implicitly due to the high computational cost. Explicit consideration of solvent is possible with DFT methods; however, DFT functionals are prone to electron delocalization error,[49] giving an overstabilized adiabatic state and thus overestimating the coupling.[27,29] Instead, the electronic coupling is better calculated with constrained density functional theory (CDFT)[27] or frozen density embedding (FDE),[20,21] which have a smaller delocalization error due to using only localized diabatic states. However, these methods are not always reliable either, or in some cases H_{ab} can be particularly sensitive to the fraction of exact exchange in the functional, e.g., CDFT-CI is known to give erroneous couplings for the ferrous-ferric ET reaction due to fractional charge transfer.[50]

METHODS

Details of the Monte Carlo simulations for the 1D two-state analytical potential are given in section S2. For ferrous-ferric ET, charges and van der Waals radii were interpolated between the reactant (Fe^{2+}−Fe^{3+}) and product (Fe^{3+}−Fe^{2+}) for 11 simulation windows. For $(Q\text{-}TTF\text{-}Q)^-$, reactant and TS structures were optimized at the B3LYP/TZVP level using the CPCM implicit water model with Gaussian 09.[51] CHELPG atomic charges for the two structures (Table S1) were linearly interpolated to set up a total of four simulation windows. Classical MD simulations with Amber force field[52] and TIP3P water model[53,54] were run for 2.5 (ferrous-ferric ET) and 2 ns (IET in $(Q\text{-}TTF\text{-}Q)^-$), respectively, with 2 fs step size. Longer simulations were run in the organic solvents to ensure the proper equilibration of the systems. For further details see the sections S2 and S3 of the SI. For ferrous-ferric ET, the electronic coupling was calculated with FDE for 10 MD frames near the TS including only the first solvation shell. Calculations were run with PBE functional, TZP basis set, and PW91k for the nonadditive kinetic energy using the ADF software.[55] For $(Q\text{-}TTF\text{-}Q)^-$, the electronic coupling was calculated at the B3LYP/TZVP TS structure using the back-transformed PNO-based EOM-EA-CCSD method[56] available in ORCA[57] with the CPCM implicit water model, aug-cc-pVTZ basis set, and corresponding auxiliary bases.

First, we apply binless DHAM to simple umbrella sampling simulations for two examples, namely (i) a model potential and (ii) Na^+ passage through an ion channel, to test how it compares to regular DHAM and WHAM methods. We then present applications that are beyond reach for these methods: ferrous-ferric ET and IET in $(Q\text{-}TTF\text{-}Q)^-$. We compare free energy profiles to MBAR results in these cases, and present rates calculated directly from the Markov matrix. The results are then compared to experimental ET rates and Marcus parameters determined in previous works.

Binless DHAM reconstructs the exact free energy profile successfully for the 1-D model potential, giving a profile closely matching the regular DHAM (Figure S1). For the passage of Na^+ ions through the transmembrane pore of the GLIC

https://doi.org/10.1021/acs.jpclett.3c02624
J. Phys. Chem. Lett. 2023, 14, 9935−9942

channel (Figure 1A, simulations by Zhu and Hummer[58]), binless DHAM results are in very good agreement with both

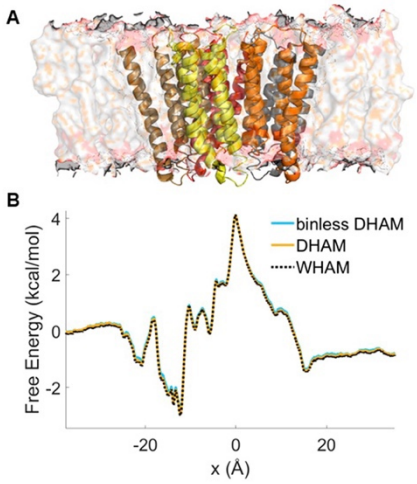


Figure 1. Reconstruction of the free energy profile from umbrella sampling simulations for Na^+ ion passage through the GLIC ion channel. (A) Unit cell of the simulation system. The five transmembrane units of GLIC are shown in different colors, as per the original depiction in ref 58. (B) Binless DHAM (blue) and DHAM (orange) profiles are plotted against the WHAM profile (black dashed lines) obtained by Zhu and Hummer.[58]

DHAM and WHAM results (Figure 1B) using a lag time of 100 fs and bin number of 1000. Our tests show that the convergence of the profile with respect to bin size and lag time needs to be verified in each case[59] (Figures S2–S3). For DHAM, and Markov state model-based methods in general, smaller bin sizes provide more accurate results, as the diffusion process is closer approximated with better discretization.[60,61]

For the ferrous-ferric ET reaction (Figure 2A), the diabatic and adiabatic free energy profiles unbiased with binless DHAM are shown in Figure 2B. For unbiasing high-energy states such as the diabatic states, high numerical precision is needed.[6] We also tested the alternative approach using the mean bias (eq S1), but we only see a difference in performance for a significantly reduced number of data points, where it performs slightly worse than eq 4 (see Figure S4). Binless DHAM gives a very similar free energy profile to MBAR (Figure S5), and the diabatic states are well approximated by a quadratic function (Figure S6), in line with Marcus theory. The reorganization energy λ is calculated from the fitted curves to be 53.1 kcal/mol (see section S8), which is only slightly higher than the experimental 48.4 kcal/mol.[36,62] In contrast, other classical MD simulations significantly overestimate λ, giving about 83 kcal/mol.[13,16] Our improved estimate of λ is probably due to varying the van der Waals radii between Fe^{2+} and Fe^{3+}. Quantum chemical description of the system was shown to provide even more accurate λ; DFT with the four-point approach[63] gives 48.7 kcal/mol,[36] while CPMD with a penalty function spin-polarized DFT approach gives 46 kcal/mol.[14]

FDE calculations on 10 snapshots from the simulation give H_{ab} = 0.24 ± 0.03 kcal/mol, which is in good agreement with values reported in the literature: 0.25 ± 0.06 kcal/mol with FON-DFT+U,[18] 0.28 kcal/mol with ROHF,[19] and 0.35 kcal/mol with a model considering an electron in the pseudopotential field of two classical Fe^{3+} ions.[13] The ET rate as a function of H_{ab} is shown in Figure 2C. Our binless DHAM methodology with the mean FDE coupling yields a rate of 5.2 × 10^2 s^{-1}, in excellent agreement with the experimental ET rate 7.9 × 10^2 s^{-1}.[36,64]

The activation free energy $\Delta G^{\ddagger}$ is 12.8 kcal/mol, somewhat higher than 11.3 kcal/mol with penalty function DFT.[14] The nuclear frequency factor ν = 8.87 × 10^{13} s^{-1} is also higher than 1.16 × 10^{13} s^{-1} calculated in ref 36 from the symmetric Fe–O stretching frequency. In comparison, k_BT/h is 6.32 × 10^{12} s^{-1} at a temperature of 303.15 K. Since κ is dependent on ν, it is not surprising that our calculated κ = 0.013 is also different from previously reported values 0.06–0.0679[36] and 0.15;[19] nevertheless, it is in line with the nonadiabatic nature of this reaction. The agreement with ref 36 is much improved if we look at the prefactors ($\kappa\nu$) directly. We note that although the quantum correction factor Γ is often assumed to be 1, previous studies indicate that for this reaction it can be as high as 10–70,[65−67] increasing the calculated rate, which would worsen the agreement with experimental rates.

Both the RS and TS structures of $(Q\text{-}TTF\text{-}Q)^-$ are nonplanar. The adiabatic ground state charge distribution is shown via the molecular orbitals occupied by the excess electron (Figure 3A and B for RS and TS, respectively), as calculated with EOM-EA-CCSD. From the energy difference of the adiabatic states at the TS, we obtain 1.0 kcal/mol coupling. In comparison, different CDFT-based approaches yielded an H_{ab} of 3.0 kcal/mol in gas phase,[27] while with explicit water solvent H_{ab} is calculated as 4.2 kcal/mol[28] (CDFT on frames from unconstrained MD) or 2.0 kcal/mol[35] (CDFT MD with PBE0 functional). The excitation energy approach with TDDFT and D-COSMO-RS solvent model for

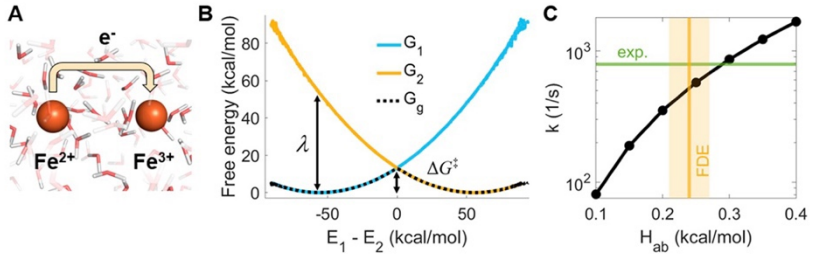


Figure 2. (A) Depiction of the ferrous-ferric electron transfer reaction in water. (B) Binless DHAM free energy profiles of diabatic states ($G_{1,2}$) and the adiabatic ground state (G_g). The reorganization energy λ and the activation free energy $\Delta G^{\ddagger}$ are also shown. (C) ET rates calculated using binless DHAM as a function of H_{ab}. The experimental rate[36] is shown in green, while H_{ab} values calculated with FDE (mean and standard error) are shown in yellow.

https://doi.org/10.1021/acs.jpclett.3c02624
J. Phys. Chem. Lett. 2023, 14, 9935–9942

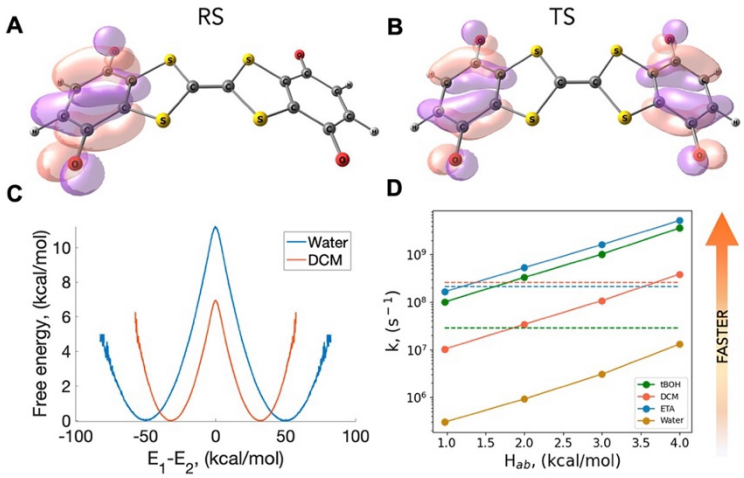


Figure 3. IET in $(Q\text{-}TTF\text{-}Q)^{-}$. Dominant molecular orbitals describe electron attachment to neutral Q-TTF-Q to form (A) the reactant state (RS) and (B) the transition state (TS) of the $(Q\text{-}TTF\text{-}Q)^{-}$ anion. Pink and purple colors represent the different phases of the wave function. (C) Binless DHAM free energy profiles plotted using H_{ab} coupling values from our EOM-CC calculations for water (blue line) and DCM (red line) as an example of the very different rates of ET. (D) Calculated ET rates as a function of H_{ab}. Experimental rates for tBOH (green), DCM (red), and ETA (blue) are also shown as dashed lines.[68]

Table 1. Calculated Energy Barriers from the First Eigenvector of the Markov Matrix, Calculated Rates from the Second Eigenvalue of the Markov Matrix, Derived Pre-Exponential Factors and Reorganization Energies[a] vs Experimentally Measured Rates for the Respective Solvents, and Measured Dielectric Constants (ε)

solvent	energy barrier (kcal/mol)	calculated rate (s^{-1})	pre-exponential factor (s^{-1})	reorganization energy (λ, kcal/mol)	experimental rate (s^{-1})[b]	dielectric constant ε[c]
tBOH	6.61	9.97×10^{07}	5.77×10^{12}	29.69	2.89×10^{07}	10.9
ETA	5.91	1.69×10^{08}	3.04×10^{12}	26.97	2.10×10^{08}	6.02
DCM	6.94	1.03×10^{07}	1.04×10^{12}	30.79	2.58×10^{08}	8.93
Water	11.23	3.00×10^{05}	3.73×10^{13}	48.24	n/a	80.1

[a]Derived using a coupling of H_{ab} = 0.97 kcal/mol for the IET in four solvent environments. [b]See ref 68. [c]See ref 71.

10:1 ethyl acetate/*tert*-butyl alcohol resulted in 2.0 kcal/mol coupling.[29] As H_{ab} values are not unique, there is no standard method of determining these. Here, we compared calculated and experimental rates[68] obtained with various choices of H_{ab} using different solvents.

We calculated the free energy profiles for the intramolecular electron transfer in four solvent environments: *tert*-butyl alcohol (tBOH), ethyl acetate (ETA), dichloromethane (DCM), and water. The binless DHAM free energy profiles for water and dichloromethane (DCM) are shown in Figure 3C. Binless DHAM has excellent agreement with MBAR in all cases (Figure S5). The energy barriers, the calculated rates (using a coupling of 0.97 kcal/mol), and the reorganization energies are summarized in Table 1, together with the experimental dielectric constants and the measured IET rates for all solvents except for water.[68] Our calculations suggest that the process follows similar rates in tBOH, ETA and DCM, but it is considerably slower in water. Our predicted rates are within an order of magnitude of the experimental rates, using the H_{ab} values from around 1−3 (Figure 3D), which demonstrates a good agreement in general and shows that our method could be used to determine H_{ab} values if the rates are known or vice versa, that experimental rates can be determined if H_{ab} values are calculated. The dielectric constant, ε, is much higher for water than the rest of the organic solvents we modeled (Table 1), which also corresponds to the much slower rate we observe for the IET in water. The dielectric constants are broadly similar for the three organic solvents, as are the ET rates, within about an order of magnitude for both the calculated and experimental values (Table 1). We note that while the precise ordering correlates perfectly between the calculated rates and the reorganization energy λ, it does not perfectly correlate across the experimental rates and measured dielectric constants (Table 1). Experimentally, λ is estimated from a broad intervalence charge transfer band to be around 22 kcal/mol in 10:1 ethyl acetate/*tert*-butyl alcohol,[64] which is also matched well by TDDFT predictions of 16.1 kcal/mol for the same solvent mixture.[24] In line with the adiabatic classification of the reaction, the calculated κ is 1.00 for all solvent environments.

Using the calculated barrier heights and the relaxation times from the eigenvalues of the unbiased Markov matrices, we also calculated the pre-exponential factor ν for the adiabatic rates in the form of eq 8. We have an excellent agreement with the standard kT/h values (6.32×10^{12} s^{-1} at 303.15 K, Table 1), demonstrating that our reaction coordinate correctly captures the rate limiting factors for this process. Using low dimensional reaction coordinates that miss key relevant degrees of freedom could result in too low free energy barriers, even if the sampling is perfect.[69] This could result in an apparent pre-exponential factor that is significantly different from the kT/h value, as observed in, e.g., umbrella sampling MD simulations of small molecules membrane permeation ($\nu \sim 10^{8}$).[70] Analogously, using reaction coordinates that better capture

https://doi.org/10.1021/acs.jpclett.3c02624
J. Phys. Chem. Lett. 2023, 14, 9935−9942

the rate limiting process for the IET could increase the barrier heights in IET simulations and thus could result in better agreement with experimental rates without invoking changes in the nuclear tunneling effects.

We derived a binless formulation of the dynamic histogram analysis method that can be used to build the free energy profile from molecular dynamics simulations biased along many arbitrary coordinates, such as Hamiltonian-based biasing. It is especially suited for the investigation of electron transfer (ET) reactions, which we demonstrated on two examples, ferrous-ferric ET and IET in $(Q\text{-}TTF\text{-}Q)^{-}$. With binless DHAM, reaction rates can be directly calculated from the Markov transition probability matrix, also providing an alternative route to determine the nuclear frequency factor of the transition state theory. The only external parameter needed to access adiabatic or nonadiabatic ET rates is the electronic coupling between redox pairs, readily calculated with frozen density embedding, constrained density functional theory, or excited state methods.

Our method gives nearly identical results to DHAM and WHAM on simulations biased along a low-dimensional reaction coordinate and to MBAR when biasing is along arbitrary coordinates, provided that the profile is converged with respect to bin size and lag time. Importantly, using the binless DHAM, the pre-exponential factor can be calculated from the unbiased Markov matrix estimate; hence, not only the free energy but also the kinetic rates are directly obtained from biased simulations.

Here, we demonstrate that using a binless DHAM for unbiasing ET simulations, the rates can be directly determined from MD simulations using different model Hamiltonians. Using appropriate coupling values, we obtained excellent agreement with experimental rates for both adiabatic and nonadiabatic ET reactions. We obtain IET rates within an order of magnitude of the experimental rates for $(Q\text{-}TTF\text{-}Q)^{-}$ in three different organic solvents using our H_{ab} coupling value determined using EOM-CC. Additionally, our calculated reorganization energies are also in good agreement with experimental estimates. Apart from ET reactions, binless DHAM can also be potentially used to calculate kinetic rates in cases where different Hamiltonians are used for sampling and energy calculations, e.g., higher level QM calculations on classical MD frames, or different force fields.[7]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpclett.3c02624.

Computational details of simulations, validation of binless DHAM results, calculation of reorganization energies (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Edina Rosta − *Department of Physics and Astronomy, University College London, London WC1E 6BT, United Kingdom;* orcid.org/0000-0002-9823-4766; Email: e.rosta@ucl.ac.uk

Authors

Zsuzsanna Koczor-Benda − *Department of Physics and Astronomy, University College London, London WC1E 6BT, United Kingdom; Department of Chemistry, University of Warwick, Coventry CV4 7AL, United Kingdom*

Teodora Mateeva − *Department of Physics, King's College London, London WC2R 2LS, United Kingdom*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.3c02624

Author Contributions

[V]Equal contributions.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Prof. Jochen Blumberger for valuable comments and suggestions. We acknowledge funding from the EPSRC (EP/R013012/1, EP/N020669/1) and ERC project 757850 BioNet. We are grateful to the UK Materials and Molecular Modelling Hub, which is partially funded by EPSRC (EP/P020194/1), for computational resources. T.M. acknowledges funding from the Agency for Science, Technology and Research (A*STAR) Singapore Research Attachment Programme (ARAP) and funding from King's College London (KCL)'s Centre for Doctoral Studies. T.M. acknowledges the use of the High Performance Computing System pluto at A*STAR.

■ REFERENCES

(1) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23* (2), 187.

(2) Liao, Q. Enhanced Sampling and Free Energy Calculations for Protein Simulations. *Progress in Molecular Biology and Translational Science* **2020**, *170*, 177.

(3) Warshel, A. Dynamics of Reactions in Polar Solvents. Semiclassical Trajectory Studies of Electron-Transfer and Proton-Transfer Reactions. *J. Phys. Chem.* **1982**, *86*, 2218.

(4) Zusman, L. D. Outer-Sphere Electron Transfer in Polar Solvents. *Chem. Phys.* **1980**, *49* (2), 295.

(5) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. THE Weighted Histogram Analysis Method for Free-energy Calculations on Biomolecules. I. The Method. *J. Comput. Chem.* **1992**, *13* (8), 1011.

(6) Rosta, E.; Hummer, G. Free Energies from Dynamic Weighted Histogram Analysis Using Unbiased Markov State Model. *J. Chem. Theory Comput* **2015**, *11* (1), 276−285.

(7) Stelzl, L. S.; Kells, A.; Rosta, E.; Hummer, G. Dynamic Histogram Analysis To Determine Free Energies and Rates from Biased Simulations. *J. Chem. Theory Comput* **2017**, *13* (12), 6328−6342.

(8) Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples from Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129* (12), 124105 DOI: 10.1063/1.2978177.

(9) Tan, Z.; Gallicchio, E.; Lapelosa, M.; Levy, R. M. Theory of Binless Multi-State Free Energy Estimation with Applications to Protein-Ligand Binding. *J. Chem. Phys.* **2012**, *136* (14), 144102 DOI: 10.1063/1.3701175.

(10) Wua, H.; Paul, F.; Wehmeyer, C.; Noé, F. Multiensemble Markov Models of Molecular Thermodynamics and Kinetics. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113* (23), E3221−E3230.

(11) Badaoui, M.; Kells, A.; Molteni, C.; Dickson, C. J.; Hornak, V.; Rosta, E. Calculating Kinetic Rates and Membrane Permeability from Biased Simulations. *J. Phys. Chem. B* **2018**, *122* (49), 11571.

(12) Hwang, J. K.; Warshel, A. Microscopic Examination of Free-Energy Relationships for Electron Transfer in Polar Solvents. *J. Am. Chem. Soc.* **1987**, *109* (3), 715.

(13) Kuharski, R. A.; Bader, J. S.; Chandler, D.; Sprik, M.; Klein, M. L.; Impey, R. W. Molecular Model for Aqueous Ferrous-Ferric Electron Transfer. *J. Chem. Phys.* **1988**, *89* (5), 3248−3257.
(14) Sit, P. H. L.; Cococcioni, M.; Marzari, N. Realistic Quantitative Descriptions of Electron Transfer Reactions: Diabatic Free-Energy Surfaces from First-Principles Molecular Dynamics. *Phys. Rev. Lett.* **2006**, *97* (2), 028303 DOI: 10.1103/PhysRevLett.97.028303.
(15) Knight, J. L.; Brooks, C. L. λ-Dynamics Free Energy Simulation Methods. *J. Comput. Chem.* **2009**, *30* (11), 1692−1700.
(16) Ando, K. Solvent Nuclear Quantum Effects in Electron Transfer Reactions. III. Metal Ions in Water. Solute Size and Ligand Effects. *J. Chem. Phys.* **2001**, *114* (21), 9470.
(17) Zeng, X.; Hu, H.; Hu, X.; Cohen, A. J.; Yang, W. Ab Initio Quantum Mechanical/Molecular Mechanical Simulation of Electron Transfer Process: Fractional Electron Approach. *J. Chem. Phys.* **2008**, *128* (12), 124510 DOI: 10.1063/1.2832946.
(18) Migliore, A.; Sit, P. H. L.; Klein, M. L. Evaluation of Electronic Coupling in Transition-Metal Systems Using DFT: Application to the Hexa-Aquo Ferric-Ferrous Redox Couple. *J. Chem. Theory Comput* **2009**, *5* (2), 307−323.
(19) Logan, J.; Newton, M. D. Ab Initio Study of Electronic Coupling in the Aqueous Fe 2+-Fe3+ Electron Exchange Process. *J. Chem. Phys.* **1983**, *78* (6), 4086−4091.
(20) Jacob, C. R.; Neugebauer, J.; Visscher, L. Software News and Update a Flexible Implementation of Frozen-Density Embedding for Use in Multilevel Simulations. *J. Comput. Chem.* **2008**, *29* (6), 1011.
(21) Wesolowski, T. A.; Warshel, A. Frozen Density Functional Approach for Ab Initio Calculations of Solvated Molecules. *J. Phys. Chem.* **1993**, *97* (30), 8050.
(22) Robin, M. B.; Day, P. Mixed Valence Chemistry-A Survey and Classification. *Adv. Inorg. Chem. Radiochem.* **1968**, *10* (C), 247.
(23) Kalinowski, J.; Berski, S.; Gordon, A. J. Electron Localization Function Study on Intramolecular Electron Transfer in the QTTFQ and DBTTFI Radical Anions. *J. Phys. Chem. A* **2011**, *115* (46), 13513.
(24) Joachim, C.; Gimzewski, J. K.; Aviram, A. Electronics Using Hybrid-Molecular and Mono-Molecular Devices. *Nature.* **2000**, *408*, 541.
(25) Chiorboli, C.; Indelli, M. T.; Scandola, F. Photoinduced Electron/Energy Transfer across Molecular Bridges in Binuclear Metal Complexes. *Top Curr. Chem.* **2005**, *257*, 63.
(26) Šrut, A.; Lear, B. J.; Krewald, V. The Marcus Dimension: Identifying the Nuclear Coordinate for Electron Transfer from Ab Initio Calculations. *Chem. Sci.* **2023**, *14*, 9213.
(27) Wu, Q.; Van Voorhis, T. Extracting Electron Transfer Coupling Elements from Constrained Density Functional Theory. *J. Chem. Phys.* **2006**, *125* (16), 164105 DOI: 10.1063/1.2360263.
(28) Oberhofer, H.; Blumberger, J. Electronic Coupling Matrix Elements from Charge Constrained Density Functional Theory Calculations Using a Plane Wave Basis Set. *J. Chem. Phys.* **2010**, *133* (24), 244105 DOI: 10.1063/1.3507878.
(29) Renz, M.; Kaupp, M. Predicting the Localized/Delocalized Character of Mixed-Valence Diquinone Radical Anions. Toward the Right Answer for the Right Reason. *J. Phys. Chem. A* **2012**, *116* (43), 10629−10637.
(30) Wu, Q.; Van Voorhis, T. Direct Calculation of Electron Transfer Parameters through Constrained Density Functional Theory. *J. Phys. Chem. A* **2006**, *110* (29), 9212−9218.
(31) Řezáč, J.; Lévy, B.; Demachy, I.; De La Lande, A. Robust and Efficient Constrained DFT Molecular Dynamics Approach for Biochemical Modeling. *J. Chem. Theory Comput* **2012**, *8* (2), 418−427.
(32) Calbo, J.; Aragó, J.; Ortí, E. Theoretical Study of the Benzoquinone-Tetrathiafulvalene- Benzoquinone Triad in Neutral and Oxidized/Reduced States. *Theor. Chem. Acc.* **2013**, *132* (3), 1−10.
(33) Vydrov, O. A.; Scuseria, G. E. Assessment of a Long-Range Corrected Hybrid Functional. *J. Chem. Phys.* **2006**, *125* (23), 234109 DOI: 10.1063/1.2409292.
(34) Kalinowski, J.; Berski, S.; Gordon, A. J. Electron Localization Function Study on Intramolecular Electron Transfer in the QTTFQ and DBTTFI Radical Anions. *J. Phys. Chem. A* **2011**, *115* (46), 13513.
(35) Holmberg, N.; Laasonen, K. Efficient Constrained Density Functional Theory Implementation for Simulation of Condensed Phase Electron Transfer Reactions. *J. Chem. Theory Comput* **2017**, *13* (2), 587−601.
(36) Rosso, K. M.; Rustad, J. R. Ab Initio Calculation of Homogeneous Outer Sphere Electron Transfer Rates: Application to M(OH2)63+/2+ Redox Couples. *J. Phys. Chem. A* **2000**, *104* (29), 6718−6725.
(37) Smoluchowski, M. V. Über Brownsche Molekularbewegung Unter Einwirkung Äußerer Kräfte Und Deren Zusammenhang Mit Der Verallgemeinerten Diffusionsgleichung. *Ann. Phys.* **1916**, *353* (24), 1103.
(38) Wigner, E. P. Derivations of Onsager's Reciprocal Relations. *J. Chem. Phys.* **1954**, *22* (11), 1912.
(39) Wu, H.; Mey, A. S. J. S.; Rosta, E.; Noé, F. Statistically Optimal Analysis of State-Discretized Trajectory Data from Multiple Thermodynamic States. *J. Chem. Phys.* **2014**, *141* (21), 214106 DOI: 10.1063/1.4902240.
(40) Stelzl, L. S.; Kells, A.; Rosta, E.; Hummer, G. Dynamic Histogram Analysis To Determine Free Energies and Rates from Biased Simulations. *J. Chem. Theory Comput* **2017**, *13* (12), 6328.
(41) Brunschwig, B. S.; Logan, J.; Newton, M. D.; Sutin, N. A Semiclassical Treatment of Electron-Exchange Reactions. Application to the Hexaaquoiron(II)-Hexaaquoiron(III) System. *J. Am. Chem. Soc.* **1980**, *102* (18), 5798.
(42) Landau, L. D. Zur Theorie Der Energieübertragung. II. *Phys. Z. Sowjetunion* **1932**, *2*, 46−51.
(43) Zener, C. Non-Adiabatic Crossing of Energy Levels. *Proc. R. Soc. Long. A* **1932**, *137* (833), 696−702.
(44) Zener, C. Dissociation of Excited Diatomic Molecules by External Perturbations. *Proc. R. Soc. Lond. A* **1933**, *140* (842), 660−668.
(45) Condon, E. A Theory of Intensity Distribution in Band Systems. *Phys. Rev.* **1926**, *28* (6), 1182.
(46) Condon, E. U. Nuclear Motions Associated with Electron Transitions in Diatomic Molecules. *Phys. Rev.* **1928**, *32* (6), 858.
(47) Krylov, A. I. Equation-of-Motion Coupled-Cluster Methods for Open-Shell and Electronically Excited Species: The Hitchhiker's Guide to Fock Space. *Ann. Rev. Phys. Chem.* **2008**, *59*, 433.
(48) Stanton, J. F.; Bartlett, R. J. The Equation of Motion Coupled-Cluster Method. A Systematic Biorthogonal Approach to Molecular Excitation Energies, Transition Probabilities, and Excited State Properties. *J. Chem. Phys.* **1993**, *98* (9), 7029.
(49) Zhang, Y.; Yang, W. A Challenge for Density Functionals: Self-Interaction Error Increases for Systems with a Noninteger Number of Electrons. *J. Chem. Phys.* **1998**, *109* (7), 2604.
(50) Mavros, M. G.; Van Voorhis, T. Communication: CDFT-CI Couplings Can Be Unreliable When There Is Fractional Charge Transfer. *J. Chem. Phys.* **2015**, *143* (23), 231102 DOI: 10.1063/1.4938103.
(51) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. v; Cioslowski, J.; Fox, D. J. *Gaussian 09*, rev. D.01; Gaussian, Inc.: Wallingford, CT, 2009.

https://doi.org/10.1021/acs.jpclett.3c02624
J. Phys. Chem. Lett. 2023, 14, 9935−9942

(52) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25* (9), 1157.
(53) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79* (2), 926.
(54) Price, D. J.; Brooks, C. L. A Modified TIP3P Water Potential for Simulation with Ewald Summation. *J. Chem. Phys.* **2004**, *121* (20), 10096.
(55) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Fonseca Guerra, C.; van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. Chemistry with ADF. *J. Comput. Chem.* **2001**, *22* (9), 931.
(56) Dutta, A. K.; Neese, F.; Izsák, R. Towards a Pair Natural Orbital Coupled Cluster Method for Excited States. *J. Chem. Phys.* **2016**, *145* (3), 034102 DOI: 10.1063/1.4958734.
(57) Neese, F. The ORCA Program System. *WIREs Mol. Sci.* **2012**, *2* (1), 73.
(58) Zhu, F.; Hummer, G. Convergence and Error Estimation in Free Energy Calculations Using the Weighted Histogram Analysis Method. *J. Comput. Chem.* **2012**, *33* (4), 453−465.
(59) Martini, L.; Kells, A.; Covino, R.; Hummer, G.; Buchete, N. V.; Rosta, E. Variational Identification of Markovian Transition States. *Phys. Rev. X* **2017**, *7* (3), 031060 DOI: 10.1103/PhysRevX.7.031060.
(60) Prinz, J. H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov Models of Molecular Kinetics: Generation and Validation. *J. Chem. Phys.* **2011**, *134* (17), 174105.
(61) Schütte, C.; Sarich, M. A Critical Appraisal of Markov State Models. *European Physical Journal: Special Topics.* **2015**, *224*, 2445.
(62) Delahay, P.; Von Burg, K.; Dziedzic, A. Photoelectron Emission Spectroscopy of Inorganic Cations in Aqueous Solution. *Chem. Phys. Lett.* **1981**, *79* (1), 157.
(63) Nelsen, S. F.; Blackstock, S. C.; Kim, Y. Estimation of Inner Shell Marcus Terms for Amino Nitrogen Compounds by Molecular Orbital Calculations. *J. Am. Chem. Soc.* **1987**, *109* (3), 677.
(64) The experimentally observed rate for ferrous-ferric ET includes the diffusion of reactants and products as well as the intrinsic ET rate. The former can be separated based on the equilibrium constant for formation of the precursor complex as done in ref 15, resulting in an intrinsic ET rate at optimal separation of the ions.
(65) Bader, J. S.; Kuharski, R. A.; Chandler, D. Role of Nuclear Tunneling in Aqueous Ferrous-Ferric Electron Transfer. *J. Chem. Phys.* **1990**, *93* (1), 230.
(66) Song, X.; Marcus, R. A. Quantum Correction for Electron Transfer Rates. Comparison of Polarizable versus Nonpolarizable Descriptions of Solvent. *J. Chem. Phys.* **1993**, *99* (10), 7768.
(67) Fang, W.; Zarotiadis, R. A.; Richardson, J. O. Revisiting Nuclear Tunnelling in the Aqueous Ferrous-Ferric Electron Transfer. *Phys. Chem. Chem. Phys.* **2020**, *22* (19), 10687−10698.
(68) Gautier, N.; Dumur, F.; Lloveras, V.; Vidal-Gancedo, J.; Veciana, J.; Rovira, C.; Hudhomme, P. Intramolecular Electron Transfer Mediated by a Tetrathiafulvalene Bridge in a Purely Organic Mixed-Valence System. *Angew. Chem. Int. Ed.* **2003**, *42* (24), 2765.
(69) Rosta, E.; Woodcock, H. L.; Brooks, B. R.; Hummer, G. Artificial Reaction Coordinate "Tunneling" in Free-Energy Calculations: The Catalytic Reaction of RNase H. *J. Comput. Chem.* **2009**, *30* (11), 1634−1641.
(70) Badaoui, M.; Kells, A.; Molteni, C.; Dickson, C. J.; Hornak, V.; Rosta, E. Calculating Kinetic Rates and Membrane Permeability from Biased Simulations. *J. Phys. Chem. B* **2018**, *122* (49), 11571.
(71) Ghazali, A. R.; Inayat-Hussain, S. H. N,N-Dimethylacetamide. *Encyclopedia of Toxicology*, 3rd ed.; Wiley, 2014; pp 594−597.

https://doi.org/10.1021/acs.jpclett.3c02624
J. Phys. Chem. Lett. 2023, 14, 9935−9942

Chapter 7

CONCLUSION

This thesis aimed to add a new understanding to the role missense mutations play in several specific cases, with examples from disease and in enzymes that are of significant importance to industrial applications. The thesis achieves this through the use of several computational methods such as Molecular Dynamics simulations, DFT calculations, QM/MM simulations, and the application of several Machine Learning algorithms. Additionally, a novel method for the calculation of the rate of electron transfer, which can be applied in a biological context, was developed, and tested on several different model systems.

In the 3rd Chapter, I showed that ATP13A2 conducts the autophosphorylation reaction with the assistance of two Mg^{2+} cations in the active site, as well as the exact mode of ATP binding in the E1 conformational state. The QM/MM potential energy scans supported the evidence from the MD simulations that the catalytic reaction likely proceeds with two cations in the active site, as evidenced by the lower barrier height (7.5 vs. 12.5 kcal/mol). This is now supported by crystal structures of the enzyme which capture the two Mg^{2+} cations,[114] demonstrating the validity of my model. Additionally, the QM/MM potential energy scans describe the crucial role of Arg686 and Lys581 in stabilizing the transition and reactant states, respectively. The active site Arg and Lys have been shown to be of similar importance in other ATPases.[53] By providing a full picture of the active site composition and conformation, this modeling and simulation work allows the study of the effect of mutations near the active site, by expanding the QM region to incorporate more residues. I also found several binding pockets in different domains of the protein (P, N, T), after analyzing the dynamics of the protein, from MD trajectories. This analysis suggests where the potential substrates of ATP13A2 can bind. Some of the pockets, in particular in the transmembrane domain, agree with the now experimentally verified binding locations of ATP13A2 cargo.[138]

In the 4th Chapter, I proposed a new method for the classification of enzyme variants, based on the predicted effect the protein mutations have on the catalytic rate. I utilized Random Forest and Gradient Boosted Decision Trees algorithms, with features extracted from Molecular Dynamics simulations of the Galactose Oxidase enzyme at/around the rate-limiting step of the alcohol conversion reaction, achieving ~78% in accuracy of classification, as well as excellent precision and recall. Predicting the effect on the catalytic rate is particularly suited to this type of ML approach, when one is limited by the experimental data available. In contrast, Deep Learning models based on, for example, ensembles of CNNs need thousands of data points, to be able to achieve accurate classification for this type of problem and are therefore not as suitable as the proposed methodology/classification pipeline since in most similar situations the experimental data is of limited size. Additionally, MD simulations at/around the TS also offer a less time-consuming alternative to QM/MM simulations.

In the 5th Chapter, I modeled and conducted Molecular Dynamics simulations of the 5-phosphatase domain of synaptojanin-1 (Synj-1), and more specifically its binding to its substrate phosphatidylinositol-4,5-bisphosphate (PIP_2). This enzyme, similarly, to ATP13A2, was independently identified through the integration of genomic and proteomic data, to be implicated in various neurodegenerative diseases. Similarly to ATP13A2, mutations in the protein give rise to various pathological processes. In this work, we provided the first three-dimensional structure of the 5-phosphatase domain, embedded in a membrane, and bound to its substrate, before the wide availability of tools such as AlphaFold.[71] Currently, what Deep Learning models still fail to predict accurately is the exact conformation of the active site substrates during the different stages of catalytic reactions, as well as the number and role of the active site cations. The bioinformatics analysis I performed, homology modeling, and MD simulations, identified that the active site residues were highly conserved between Synj-1 and some of the Mg^{2+}-dependent DNA restriction endonucleases,[139,140] and Synj-1 likely also exhibits a two-metal ion catalytic mechanism. Both my MD simulations, and the conserved active site, support the hypothesis that this enzyme completes a two-ion dephosphorylation of PIP_2.

In the 6th Chapter, I modeled several systems to test and apply a novel method for the calculation of the rate of electron transfer called Binless Dynamic Weighted Histogram

Analysis Method (DHAM). The ferrous-ferric (Fe^{2+}-Fe^{3+}) intermolecular electron transfer reaction and the intramolecular electron transfer reaction in the $(Q\text{-}TTF\text{-}Q)^-$ anion represent examples of diabatic and adiabatic coupling mechanisms, respectively. It was demonstrated that Binless DHAM achieves excellent agreement with experimental measurements in both types of electron transfer, achieving a predicted rate of electron transfer of $5.2 \times 10^2\ s^{-1}$ for the ferrous-ferric system in water and $9.97 \times 10^7\ s^{-1}$ for $(Q\text{-}TTF\text{-}Q)^-$ anion in tBOH, respectively. This method for the estimation of the rate of electron transfer rate has not been applied to the study of biological systems yet, but it could be tested on enzyme active sites where the catalytic reaction involves the transfer of electrons.

Bibliography

1. Estrada-Cuzcano, A. *et al.* Loss-of-function mutations in the ATP13A2/PARK9 gene cause complicated hereditary spastic paraplegia (SPG78). *Brain* **140**, 287–305 (2017).
2. Esteves-Ferreira, A. A. *et al.* Cyanobacterial nitrogenases: Phylogenetic diversity, regulation and functional predictions. *Genetics and Molecular Biology* vol. 40.
3. Berta, D., Gehrke, S., Nyíri, K., Vértessy, B. G. & Rosta, E. Mechanism-Based Redesign of GAP to Activate Oncogenic Ras. *J Am Chem Soc* (2023). doi:10.1021/jacs.3c04330.
4. Zhang, L., Wang, L., Liang, Y. & Yu, J. FgPEX4 is involved in development, pathogenicity, and cell wall integrity in Fusarium graminearum. *Curr Genet* **65**, (2019).
5. de Majo, M. *et al.* ALS-associated missense and nonsense TBK1 mutations can both cause loss of kinase function. *Neurobiol Aging* **71**, (2018).
6. Roland, B. P. *et al.* Missense variant in TPI1 (Arg189Gln) causes neurologic deficits through structural changes in the triosephosphate isomerase catalytic site and reduced enzyme levels in vivo. *Biochim Biophys Acta Mol Basis Dis* **1865**, (2019).
7. Starke, E. L., Zius, K. & Barbee, S. A. FXS causing missense mutations disrupt FMRP granule formation, dynamics, and function. *PLoS Genet* **18**, (2022).
8. Worth, A. J. J. *et al.* Disease-associated missense mutations in the EVH1 domain disrupt intrinsic WASp function causing dysregulated actin dynamics and impaired dendritic cell migration. *Blood* **121**, (2013).
9. Miller, D. C., Athavale, S. V. & Arnold, F. H. Combining chemistry and protein engineering for new-to-nature biocatalysis. *Nature Synthesis* **1**, (2022).
10. Birmingham, W. R. & Turner, N. J. A Single Enzyme Oxidative 'cascade' via a Dual-Functional Galactose Oxidase. *ACS Catal* **8**, (2018).
11. Birmingham, W. R. *et al.* Toward scalable biocatalytic conversion of 5-hydroxymethylfurfural by galactose oxidase using coordinated reaction and enzyme engineering. *Nat Commun* **12**, (2021).
12. Turner, N. J. Directed evolution drives the next generation of biocatalysts. *Nature Chemical Biology* vol. 5, (2009).
13. Escalettes, F. & Turner, N. J. Directed evolution of galactose oxidase: Generation of enantioselective secondary alcohol oxidases. *ChemBioChem* **9**, (2008).
14. Mazurenko, S., Prokop, Z. & Damborsky, J. Machine Learning in Enzyme Engineering. *ACS Catalysis* vol. 10, (2020).
15. Podhajska, A. *et al.* Common pathogenic effects of missense mutations in the P-type ATPase ATP13A2 (PARK9) associated with early-onset parkinsonism. *PLoS One* **7**, e39942 (2012).
16. Santoro, L. *et al.* Novel ATP13A2 (PARK9) homozygous mutation in a family with marked phenotype variability. *Neurogenetics* **12**, 33–39 (2011).
17. Schneider, S. A. *et al.* ATP13A2 mutations (PARK9) cause neurodegeneration with brain iron accumulation. *Movement Disorders* **25**, 979–84 (2010).
18. Di Fonzo, A. *et al.* ATP13A2 missense mutations in juvenile parkinsonism and young onset Parkinson disease. *Neurology* vol. 68 1557–62 (2007).
19. Brüggemann, N. *et al.* Recessively inherited parkinsonism: Effect of ATP13A2 mutations on the clinical and neuroimaging phenotype. *Arch Neurol* **67**, 1357–63 (2010).

20. Abbas, M. M., Govindappa, S. T., Sheerin, U. M., Bhatia, K. P. & Muthane, U. B. Exome Sequencing Identifies a Novel Homozygous Missense ATP13A2 Mutation. *Mov Disord Clin Pract* **4**, 132–135 (2017).
21. Behrens, M. I. *et al.* Clinical spectrum of Kufor-Rakeb syndrome in the Chilean kindred with ATP13A2 mutations. *Movement Disorders* **25**, 1929–37 (2010).
22. Odake, Y. *et al.* Identification of a novel mutation in ATP13A2 associated with a complicated form of hereditary spastic paraplegia. *Neurol Genet* **6**, e514 (2020).
23. Wu, S., Akhtari, M. & Alachkar, H. Characterization of Mutations in the Mitochondrial Encoded Electron Transport Chain Complexes in Acute Myeloid Leukemia. *Sci Rep* **8**, (2018).
24. Li, M., Goncearenco, A. & Panchenko, A. R. Annotating mutational effects on proteins and protein interactions: Designing novel and revisiting existing protocols. in *Methods in Molecular Biology* vol. 1550 (2017).
25. Maistro, S. *et al.* Germline mutations in BRCA1 and BRCA2 in epithelial ovarian cancer patients in Brazil. *BMC Cancer* **16**, (2016).
26. Loveday, C. *et al.* Analysis of rare disruptive germline mutations in 2135 enriched BRCA-negative breast cancers excludes additional high-impact susceptibility genes. *Annals of Oncology* **33**, (2022).
27. Jeong, A. R., Forbes, K., Orosco, R. K. & Cohen, E. E. W. Hereditary oral squamous cell carcinoma associated with CDKN2A germline mutation: a case report. *Journal of Otolaryngology - Head and Neck Surgery* **51**, (2022).
28. Kampmeyer, C. *et al.* Blocking protein quality control to counter hereditary cancers. *Genes Chromosomes and Cancer* vol. 56, (2017).
29. Casadio, R., Vassura, M., Tiwari, S., Fariselli, P. & Luigi Martelli, P. Correlating disease-related mutations to their effect on protein stability: A large-scale analysis of the human proteome. *Hum Mutat* **32**, (2011).
30. Petrosino, M. *et al.* Analysis and interpretation of the impact of missense variants in cancer. *International Journal of Molecular Sciences* vol. 22, (2021).
31. Singh, S. M., Kongari, N., Cabello-Villegas, J. & Mallela, K. M. G. Missense mutations in dystrophin that trigger muscular dystrophy decrease protein stability and lead to cross-β aggregates. *Proc Natl Acad Sci U S A* **107**, (2010).
32. Nielsen, S. V. *et al.* Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. *PLoS Genet* **13**, (2017).
33. Shan, X., Dunbrack, R. L., Christopher, S. A. & Kruger, W. D. Mutations in the regulatory domain of cystathionine β-synthase can functionally suppress patient-derived mutations in cis. *Hum Mol Genet* **10**, (2001).
34. Lewis, D. *et al.* Distinctive features of catalytic and transport mechanisms in mammalian sarco-endoplasmic reticulum Ca2+ ATPase (SERCA) and Cu+ (ATP7A/B) ATPases. *Journal of Biological Chemistry* **287**, (2012).
35. Schymkowitz, J. *et al.* The FoldX web server: An online force field. *Nucleic Acids Res* **33**, (2005).
36. Leman, J. K. *et al.* Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nature Methods* vol. 17, (2020).
37. Blaabjerg, L. M. *et al.* Rapid protein stability prediction using deep learning representations. *Elife* **12**, (2023).
38. Pancotti, C. *et al.* A deep-learning sequence-based method to predict protein stability changes upon genetic variations. *Genes (Basel)* **12**, (2021).

39. Li, B., Yang, Y. T., Capra, J. A. & Gerstein, M. B. Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. *PLoS Comput Biol* **16**, (2020).
40. Tachikawa, M., Kanagawa, M., Yu, C. C., Kobayashi, K. & Toda, T. Mislocalization of fukutin protein by disease-causing missense mutations can be rescued with treatments directed at folding amelioration. *Journal of Biological Chemistry* **287**, (2012).
41. Nishibori, Y. *et al.* Disease-causing missense mutations in NPHS2 gene alter normal nephrin trafficking to the plasma membrane. *Kidney Int* **66**, (2004).
42. Ugolino, J., Fang, S., Kubisch, C. & Monteiro, M. J. Mutant Atp13a2 proteins involved in parkinsonism are degraded by ER-associated degradation and sensitize cells to ER-stress induced cell death. *Hum Mol Genet* **20**, (2011).
43. Thireou, T. & Reczko, M. Bidirectional long short-term memory networks for predicting the subcellular localization of eukaryotic proteins. *IEEE/ACM Trans Comput Biol Bioinform* **4**, (2007).
44. Liao, Z., Pan, G., Sun, C. & Tang, J. Predicting subcellular location of protein with evolution information and sequence-based deep learning. *BMC Bioinformatics* **22**, (2021).
45. Jiang, Y., Wang, D., Wang, W. & Xu, D. Computational methods for protein localization prediction. *Computational and Structural Biotechnology Journal* vol. 19, (2021).
46. Schurmann, K. *et al.* Molecular basis for the reduced catalytic activity of the naturally occurring T560m mutant of human 12/15-lipoxygenase that has been implicated in coronary artery disease. *Journal of Biological Chemistry* **286**, (2011).
47. Korasick, D. A. & Tanner, J. J. Impact of missense mutations in the ALDH7A1 gene on enzyme structure and catalytic function. *Biochimie* vol. 183, 016 (2021).
48. Morgan, C. T., Tsivkovskii, R., Kosinsky, Y. A., Efremov, R. G. & Lutsenko, S. The Distinct Functional Properties of the Nucleotide-binding Domain of ATP7B, the Human Copper-transporting ATPase. *Journal of Biological Chemistry* **279**, (2004).
49. Molina, P., Knegtel, R. M. A. & Macher, B. A. Site-directed mutagenesis of glutamate 317 of bovine α-1,3Galactosyltransferase and its effect on enzyme activity: Implications for reaction mechanism. *Biochim Biophys Acta Gen Subj* **1770**, (2007).
50. Lopata, A. *et al.* Mutations decouple proton transfer from phosphate cleavage in the dutpase catalytic reaction. *ACS Catal* **5**, 3225–3237 (2015).
51. Spataro, R. *et al.* Mutations in ATP13A2 (PARK9) are associated with an amyotrophic lateral sclerosis-like phenotype, implicating this locus in further phenotypic expansion. *Hum Genomics* (2019). doi:10.1186/s40246-019-0203-9.
52. Santoro, L. *et al.* Novel ATP13A2 (PARK9) homozygous mutation in a family with marked phenotype variability. *Neurogenetics* **12**, (2011).
53. Junop, M. S., Obmolova, G., Rausch, K., Hsieh, P. & Yang, W. Composite active site of an ABC ATPase: MutS uses ATP to verify mismatch recognition and authorize DNA repair. *Mol Cell* **7**, (2001).
54. Guarnera, E. & Berezovsky, I. N. On the perturbation nature of allostery: sites, mutations, and signal modulation. *Current Opinion in Structural Biology* vol. 56, (2019).
55. Guarnera, E. & Berezovsky, I. N. Toward Comprehensive Allosteric Control over Protein Activity. *Structure* **27**, (2019).

56. Fuchs, J. E., Muñoz, I. G., Timson, D. J. & Pey, A. L. Experimental and computational evidence on conformational fluctuations as a source of catalytic defects in genetic diseases. *RSC Adv* **6**, (2016).
57. Yang, G. M. *et al.* Structures of the human Wilson disease copper transporter ATP7B. *Cell Rep* **42**, (2023).
58. Scheffzek, K. *et al.* The Ras-RasGAP complex: Structural basis for GTPase activation and its loss in oncogenic ras mutants. *Science (1979)* **277**, (1997).
59. Thoden, J. B., Wohlers, T. M., Fridovich-Keil, J. L. & Holden, H. M. Crystallographic evidence for Tyr 157 functioning as the active site base in human UDP-galactose 4-epimerase. *Biochemistry* **39**, (2000).
60. Ogrizek, M., Janežič, M., Valjavec, K. & Perdih, A. Catalytic Mechanism of ATP Hydrolysis in the ATPase Domain of Human DNA Topoisomerase IIα. *J Chem Inf Model* **62**, (2022).
61. Luna, E., Rodríguez-Huete, A., Rincón, V., Mateo, R. & Mateu, M. G. Systematic Study of the Genetic Response of a Variable Virus to the Introduction of Deleterious Mutations in a Functional Capsid Region. *J Virol* **83**, (2009).
62. Acevedo-Rocha, C. G. *et al.* P450-Catalyzed regio- and diastereoselective steroid hydroxylation: Efficient directed evolution enabled by mutability landscaping. *ACS Catal* **8**, (2018).
63. McCullum, E. O., Williams, B. A. R., Zhang, J. & Chaput, J. C. Random mutagenesis by error-prone PCR. *Methods in Molecular Biology* **634**, (2010).
64. Reetz, M. T. & Carballeira, J. D. Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nat Protoc* **2**, (2007).
65. Cadet, X. F., Gelly, J. C., van Noord, A., Cadet, F. & Acevedo-Rocha, C. G. Learning Strategies in Protein Directed Evolution. in *Methods in Molecular Biology* vol. 2461 (2022).
66. She, W. *et al.* Rapid and Error-Free Site-Directed Mutagenesis by a PCR-Free in Vitro CRISPR/Cas9-Mediated Mutagenic System. *ACS Synth Biol* **7**, (2018).
67. Hancock, J. M., Zvelebil, M. J. & Zvelebil, M. J. UniProt. in *Dictionary of Bioinformatics and Computational Biology* (2004). doi:10.1002/9780471650126.dob0721.pub2.
68. Bateman, A. *et al.* UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res* **49**, D480–D489 (2021).
69. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* vol. 28, (2000).
70. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, (2020).
71. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, (2021).
72. Meynard-Piganeau, B., Feinauer, C., Weigt, M., Walczak, A. M. & Mora, T. TULIP — a Transformer based Unsupervised Language model for Interacting Peptides and T-cell receptors that generalizes to unseen epitopes. doi:10.1101/2023.07.19.549669.
73. Hudson, D., Fernandes, R. A., Basham, M., Ogg, G. & Koohy, H. Can we predict T cell specificity with digital biology and machine learning? *Nat Rev Immunol* **23**, (2023).
74. Banerjee, A. *et al.* BATMAN: Improved T cell receptor cross-reactivity prediction benchmarked on a comprehensive mutational scan database. doi:10.1101/2024.01.22.576714.
75. Daniya, T., Geetha, M. & Kumar, K. S. Classification and regression trees with gini index. *Advances in Mathematics: Scientific Journal* **9**, (2020).

76. Tangirala, S. Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications* (2020). doi:10.14569/ijacsa.2020.0110277.
77. O'Donnell, T. J. *et al.* MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. *Cell Syst* **7**, (2018).
78. Xu, Y. *et al.* Deep Dive into Machine Learning Models for Protein Engineering. *J Chem Inf Model* **60**, (2020).
79. Almagro Armenteros, J. J. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol* **37**, (2019).
80. Jung, F., Frey, K., Zimmer, D. & Mühlhaus, T. DeepSTABp: A Deep Learning Approach for the Prediction of Thermal Protein Stability. *Int J Mol Sci* **24**, (2023).
81. Chen, J., Zheng, S., Zhao, H. & Yang, Y. Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map. *J Cheminform* **13**, (2021).
82. Krapp, L. F., Abriata, L. A., Cortés Rodriguez, F. & Dal Peraro, M. PeSTo: parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nat Commun* **14**, (2023).
83. Wang, P., Zhang, G., Yu, Z. G. & Huang, G. A Deep Learning and XGBoost-Based Method for Predicting Protein-Protein Interaction Sites. *Front Genet* **12**, (2021).
84. Huang, J. & Mackerell, A. D. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J Comput Chem* (2013) doi:10.1002/jcc.23354.
85. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general Amber force field. *J Comput Chem* **25**, (2004).
86. The classical equation of state of gaseous helium, neon and argon. *Proc R Soc Lond A Math Phys Sci* **168**, (1938).
87. Verlet, L. Computer 'experiments' on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Physical Review* (1967). doi:10.1103/PhysRev.159.98.
88. Van Gunsteren, W. F. & Berendsen, H. J. C. A Leap-Frog Algorithm for Stochastic Dynamics. *Mol Simul* **1**, (1988).
89. Victor, R. Berendsen and Nose-Hoover thermostats. *Americal Journal Physics* (2007).
90. Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., Dinola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J Chem Phys* **81**, (1984).
91. Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *J Chem Phys* **72**, (1980).
92. Nosé, S. A molecular dynamics method for simulations in the canonical ensemble. *Mol Phys* **52**, (1984).
93. Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys Rev A (Coll Park)* **31**, (1985).
94. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *Journal of Chemical Physics* **126**, (2007).
95. Hamelberg, D., Mongan, J. & McCammon, J. A. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. *Journal of Chemical Physics* **120**, (2004).
96. Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* **314**, (1999).

97. Badaoui, M. *et al.* Combined Free-Energy Calculation and Machine Learning Methods for Understanding Ligand Unbinding Kinetics. *J Chem Theory Comput* **18**, (2022).
98. Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H. & Kollman, P. A. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J Comput Chem* **13**, (1992).
99. Hub, J. S., De Groot, B. L. & Van Der Spoel, D. G-whams-a free Weighted Histogram Analysis implementation including robust error and autocorrelation estimates. *J Chem Theory Comput* **6**, (2010).
100. Stelzl, L. S., Kells, A., Rosta, E. & Hummer, G. Dynamic Histogram Analysis To Determine Free Energies and Rates from Biased Simulations. *J Chem Theory Comput* **13**, (2017).
101. Rosta, E. & Hummer, G. Free energies from dynamic weighted histogram analysis using unbiased Markov state model. *J Chem Theory Comput* **11**, (2015).
102. Smoluchowski, M. V. Über Brownsche Molekularbewegung unter Einwirkung äußerer Kräfte und deren Zusammenhang mit der verallgemeinerten Diffusionsgleichung. *Ann Phys* **353**, (1916).
103. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Physical Review* (1964). doi:10.1103/PhysRev.136.B864.
104. Jones, L. O., Mosquera, M. A., Schatz, G. C. & Ratner, M. A. Embedding Methods for Quantum Chemistry: Applications from Materials to Life Sciences. *J Am Chem Soc* **142**, (2020).
105. Senn, H. M. & Thiel, W. QM/MM methods for biomolecular systems. *Angewandte Chemie - International Edition*, (2009).
106. Naeem, S., Ali, A., Anam, S. & Ahmed, M. M. An Unsupervised Machine Learning Algorithms: Comprehensive Review. *International Journal of Computing and Digital Systems* **13**, (2023).
107. Sinaga, K. P. & Yang, M. S. Unsupervised K-means clustering algorithm. *IEEE Access* **8**, (2020).
108. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann Stat* **29**, (2001).
109. Hastie, T., Tibshirani, R. & Friedman, J. Boosting and Additive Trees. in (2009). doi:10.1007/b94608_10.
110. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, (2011).
111. van Veen, S. *et al.* Cellular function and pathological role of ATP13A2 and related P-type transport ATPases in Parkinson's disease and other neurological disorders. *Front Mol Neurosci* **7**, 1–22 (2014).
112. Lin, C. H. *et al.* Novel ATP13A2 variant associated with Parkinson disease in Taiwan and Singapore. *Neurology* **71**, 1727–32 (2008).
113. Park, J. S., Blair, N. F. & Sue, C. M. The role of ATP13A2 in Parkinson's disease: Clinical phenotypes and molecular mechanisms. *Movement Disorders* **30**, 770–9 (2015).
114. Sim, S. I., von Bülow, S., Hummer, G. & Park, E. Structural basis of polyamine transport by human ATP13A2 (PARK9). *Mol Cell* **81**, (2021).
115. AMARAL, D., BERNSTEIN, L., MORSE, D. & HORECKER, B. L. Galactose oxidase of Polyporus circinatus: a copper enzyme. *J Biol Chem* **238**, (1963).
116. Baron, A. J. *et al.* Structure and mechanism of galactose oxidase. The free radical site. *Journal of Biological Chemistry* **269**, (1994).

117. Fong, J. K. & Brumer, H. Copper radical oxidases: galactose oxidase, glyoxal oxidase, and beyond! *Essays Biochem* (2022). doi:10.1042/ebc20220124.
118. Sun, L., Bulter, T., Alcalde, M., Petrounia, I. P. & Arnold, F. H. Modification of galactose oxidase to introduce glucose 6-oxidase activity. *ChemBioChem* **3**, (2002).
119. Yeo, W. L. *et al.* Directed Evolution and Computational Modeling of Galactose Oxidase toward Bulky Benzylic and Alkyl Secondary Alcohols. *ACS Catal* **13**, 16088–16096 (2023).
120. Wilkinson, D. *et al.* Structural kinetic studies of a series of mutants of galactose oxidase identified by directed evolution. *Protein Engineering, Design and Selection* **17**, (2004).
121. Barry, B. A., El-Deeb, M. K., Sandusky, P. O. & Babcock, G. T. Tyrosine radicals in photosystem II and related model compounds. Characterization by isotopic labeling and EPR spectroscopy. *Journal of Biological Chemistry* **265**, (1990).
122. Parikka, K., Master, E. & Tenkanen, M. Oxidation with galactose oxidase: Multifunctional enzymatic catalysis. *Journal of Molecular Catalysis B: Enzymatic* vol. 120, (2015).
123. Whittaker, J. W. Free radical catalysis by galactose oxidase. *Chem Rev* **103**, (2003).
124. Ito, N. *et al.* Novel thioether bond revealed by a 1.7 Å crystal structure of galactose oxidase. *Nature* **350**, (1991).
125. Rogers, M. S. *et al.* The stacking tryptophan of galactose oxidase: A second-coordination sphere residue that has profound effects on tyrosyl radical behavior and enzyme catalysis. *Biochemistry* **46**, (2007).
126. Whittaker, M. M. & Whittaker, J. W. Catalytic reaction profile for alcohol oxidation by galactose oxidase. *Biochemistry* **40**, (2001).
127. Himo, F., Eriksson, L. A., Maseras, F. & Siegbahn, P. E. M. Catalytic mechanism of galactose oxidase: A theoretical study. *J Am Chem Soc* **122**, (2000).
128. Rogers, M. S. *et al.* The stacking tryptophan of galactose oxidase: A second-coordination sphere residue that has profound effects on tyrosyl radical behavior and enzyme catalysis. *Biochemistry* **46**, 4606–4618 (2007).
129. Schrödinger LLC. The PyMOL Molecular Graphics System, Version 2.4. *Schrödinger LLC*, (2020).
130. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Had, and D. J. F. *et al.* Gaussian 09, Revision D.01. *Gaussian, Inc., Wallingford*, (2013).
131. Becke, A. B3LYP. *J. Chem. Phys.* **98**, 5648 (1993).
132. Weigend, F. & Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics* **7**, (2005).
133. van der Spoel, D. *et al.* GROMACS: Fast, flexible, and free. *Journal of Computational Chemistry* vol. 26, (2005).
134. Price, D. J. & Brooks, C. L. A modified TIP3P water potential for simulation with Ewald summation. *Journal of Chemical Physics* **121**, (2004).
135. Andricioaei, I. & Karplus, M. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems The. *Statistical Mechanics of Fluid Mixtures The Journal of Chemical Physics* **115**, (2001).

136. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J Chem Phys* **98**, (1993).
137. Friedman, J. H. Stochastic gradient boosting. *Comput Stat Data Anal* **38**, (2002).
138. van Veen, S. *et al.* ATP13A2 deficiency disrupts lysosomal polyamine export. *Nature* **578**, 419–424 (2020).
139. Whisstock, J. C. *et al.* The inositol polyphosphate 5-phosphatases and the apurinic/apyrimidinic base excision repair endonucleases share a common mechanism for catalysis. *Journal of Biological Chemistry* (2000). doi:10.1074/jbc.M006244200.
140. Dlakić, M. Functionally unrelated signalling proteins contain a fold similar to Mg2+-dependent endonucleases. *Trends Biochem Sci* (2000). doi:10.1016/S0968-0004(00)01582-6.

Appendix A

Supporting Information for

Structural Dynamics and Catalytic Mechanism of ATP13A2 (PARK9) from Simulations

Teodora Mateeva[1], Marco Klähn[2], Edina Rosta[1,3*]

[1]Department of Chemistry, Faculty of Natural & Mathematical Sciences, King's College London, London SE1 1DB, United Kingdom
[2]Department of Materials Science and Chemistry, Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore 138 632, Singapore
[3]Department of Physics and Astronomy, Faculty of Maths & Physical Sciences, University College London, London WC1E 6BT, United Kingdom

*Corresponding author: E-mail: e.rosta@ucl.ac.uk

Homology Modelling

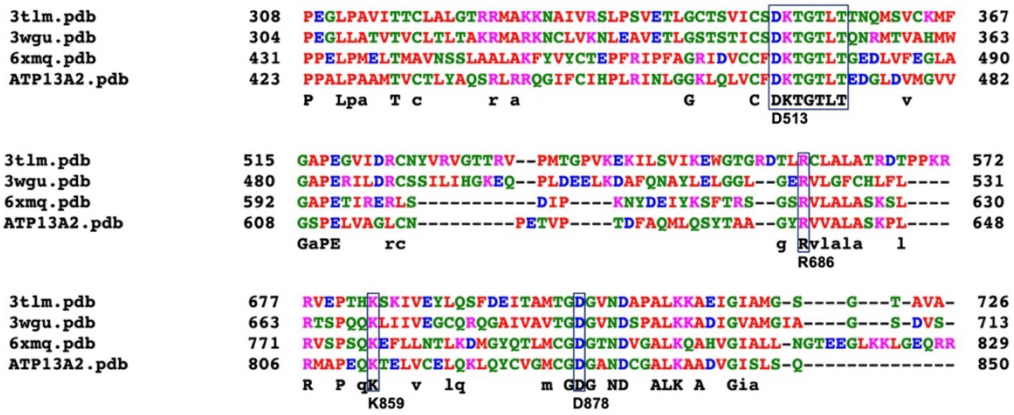


Figure S1. Alignment of the sequences of the most homologous proteins to ATP13A2 which have crystallographic structures deposited in the Protein Data Bank (PDB). Only regions of interest in this work are shown. The **DKTGTLT** motif is conserved among all protein structures used in this work. The Mg^{2+}-coordinating active site residues D513 and T515, correspond to **DKT**GTLT in this motif. The active site residues R686 and K859 are also strictly conserved among all P-type ATPases used in this work. D878 is the second aspartate amino acid which coordinates the catalytic Mg^{2+} ion during the cleavage of ATP. The proteins aligned here are: SERCA (PDB code: 3tlm)[1], Na^+/K^+- ATPase (PDB code: 3wgu)[2], ATP13A1 (PDB code: 6xmq)[3] and the homology model of ATP13A2. The numbering below the conserved residues corresponds to the number of the residue in the sequence of ATP13A2 for Homo Sapiens (Uniprot[4] code: Q9NQ11).

The active site is conserved between the two homology models (Fig. S2D and S2E). Residues 1-189 could not be modelled with SERCA as the template. These correspond to: 1-44 cytoplasmic, 45-65 intramembrane and 66-188 cytoplasmic regions of the protein, respectively. This segment of the enzyme has some significant differences in sequence conservation between ATP13A2, ATP13A1, as well as SERCA, therefore, neither of the templates available to us could model this region of the protein ideally. The crystal structure of ATP13A1 in the E1-ATP state did not have any solvent molecules resolved, which is very important for modelling the active site. Additionally, Lys797 (corresponding to Lys859 in ATP13A2) was resolved further away in the crystal structures of ATP13A1. We are mainly interested in the active site, therefore we used SERCA for the modelling template as it had crystallographic waters and the active site residues were positioned correctly for the ATP cleavage reaction. The RMSD of the CA atoms between the two homology models of the full protein is 4.54 Å. The RMSD of the CA atoms for residues in the active site within 5 Å of the ATP is 0.66 Å.

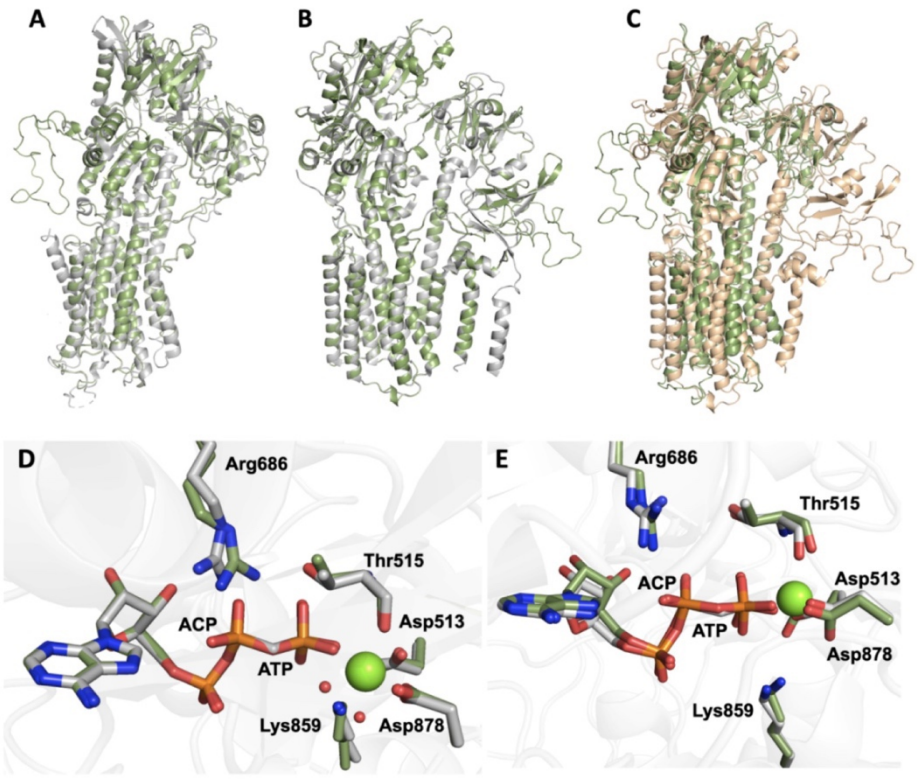


Figure S2. (A) Three-dimensional model of ATP13A2 (green cartoon) based on SERCA (grey cartoon, pdb code: 3tlm). **(B)** Three-dimensional model of ATP13A2 (green cartoon) based on ATP13A1 (grey cartoon, pdb code: 6xmq). **(C)** Aligned initial structures of ATP13A2 based on SERCA and on ATP13A1. **(D)** and **(E)** show the active sites of the two models, respectively for the ATP13A2 model (green sticks) and the template crystal structures (grey sticks). **(D)** shows the water molecules resolved in the crystal structure, whereas the crystal structure in **(E)** did not have any water resolved.

QM Cluster Calculations

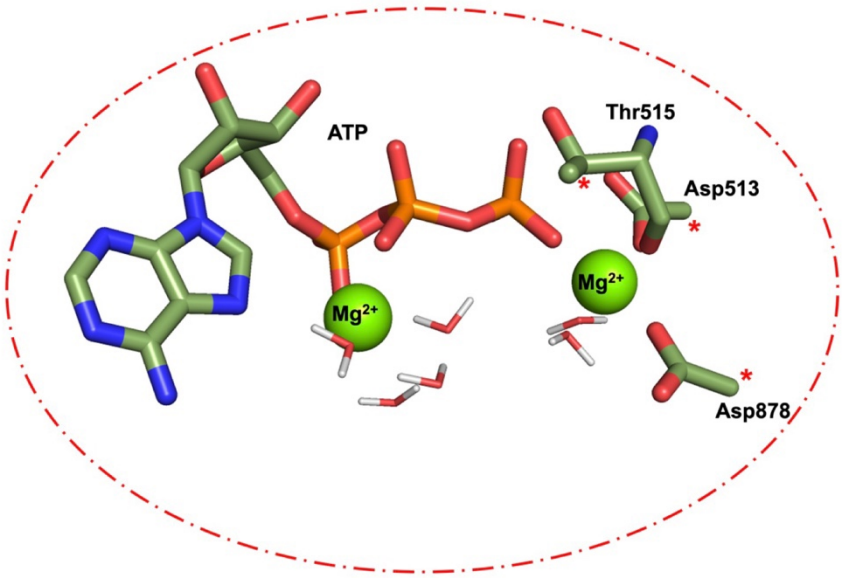


Figure S3. Heavy atoms included in the QM cluster geometry optimizations. Red asterisk marks where an amino acid residue has been truncated, as well as which atom was frozen during the geometry optimization. Each calculation contained six water molecules, two Mg^{2+} cations, the full ATP and the three amino acids depicted.

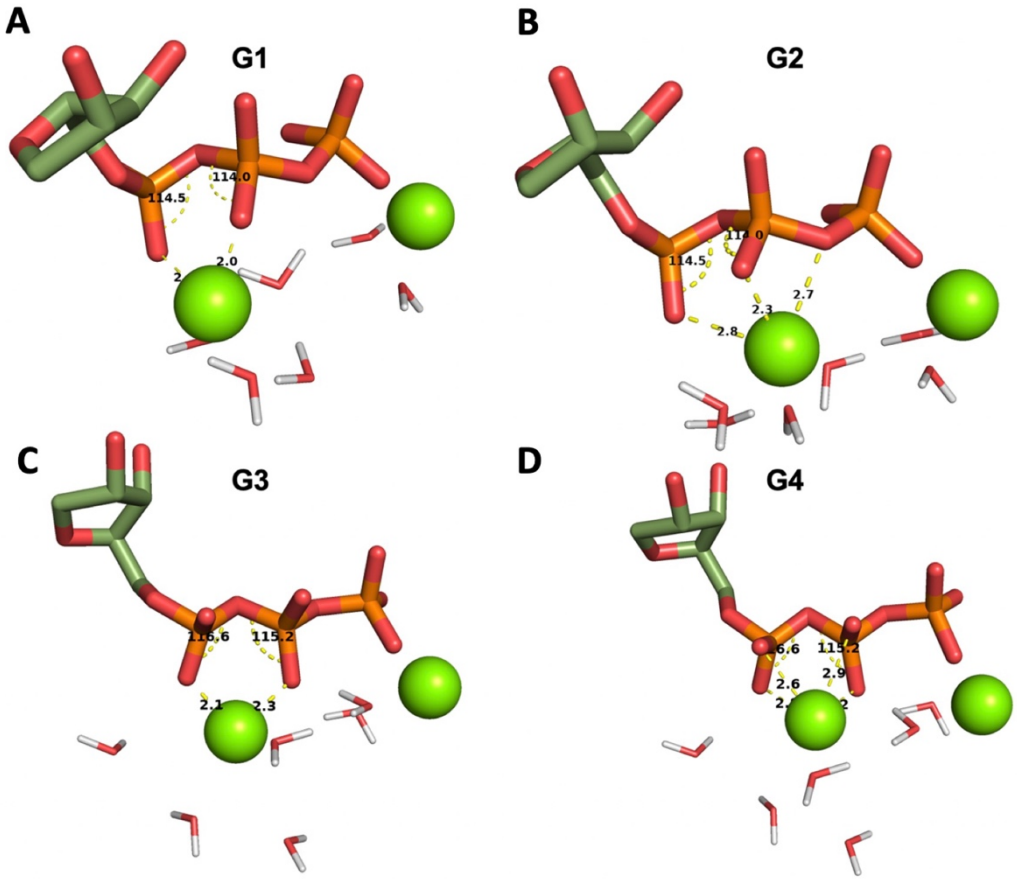


Figure S4. Starting geometries G1-G4 for the QM cluster calculations (before optimization). Starting structure G1 **(A)** was created based on alignment with the crystal structure resolved with ADP and 2 Mg^{2+} ions (PDB code: 3wgu[2]). Structure G2 **(B)** exhibits an alternative binding but has the same conformation as **(A)**, here the second Mg^{2+} ion is coordinating three oxygen atoms instead of two. G3 **(C)** was taken from the last frame of the original MD simulation (after 100 ns) and G4 **(D)** represents a more unorthodox geometry that has not been observed by us in any P-type ATPase crystal structures but was generated to explore different possibilities. The "straight" conformation of the phosphate chain G3 and G4 in **(C)** and **(D)** is unfavorable energetically.

QM/MM Reaction Coordinate Scans (RCSs)

Six systems were created (P1-P6) which were initially minimized for 1000 steps. To describe the QM region, "all amino acids" refers to Asp513, Asp878, Thr515, Lys859, Arg686, Gly516 and Lys514. The minimized structure of each system was supplied as a starting point for the reaction coordinate scan (RCS). P1 refers to the full system containing all amino acids described to be contained in the QM region, including Arg686 and Lys859, with one Mg^{2+} in the active site. P2 refers to the full system containing all amino acids described to be contained in the QM region, including Arg686 and Lys859, with two Mg^{2+} ions in the active site. P3 and P4 have one Mg^{2+} ion in the active site and Arg686 or Lys859 missing, respectively. P5 and P6 have two Mg^{2+} ions in the active site with Arg686 or Lys859 missing, respectively. The aim of obtaining a converged RCS for each system P1-P6 was to show the effect of Lys859 and Arg686, as well as the profile of the wild type system where all amino acids are present and not mutated.

Table S1. Six QM/MM systems (P1-P6) were initially minimized, and subsequently QM/MM reaction coordinate scans were run. The overall charge and atoms of each QM region, as well as the number of Mg^{2+} ions in each active site are shown below. To test the effects of residues missing from the active site, we neutralized the atomic charges of selected residues as indicated.

System	Number of Mg^{2+}	Overall Charge	Residues Missing
P1	1	-2	none
P2	2	0	none
P3	1	-3	Lys859
P4	1	-3	Arg686
P5	2	-1	Lys859
P6	2	-1	Arg686

Pocket Analysis

Twenty equidistant frames from the 100 ns unconstrained MD simulation were extracted. The twenty biggest pockets were calculated in each frame and the four biggest pockets in terms of surface area were further analyzed. An ATP pocket was found in 75% of all frames and a transmembrane pocket, within 1.5 Å of residues Val469, Pro470 and Asp967, was found 90% of the time. The size of each pocket in the corresponding frame is listed in Table 2. Two other pockets were found consistently, and their location is shown in Figure S5. The blue pocket is relatively small in surface area and is found exclusively in the first 25 ns of the simulation, whereas the orange pocket is found in the later part of the simulation (after 25 ns).

Table S2. Frames extracted from the 100 ns unconstrained MD simulation and the size of two biggest pockets found.

Frame	ATP pocket size (Å^3)	Transmembrane pocket size (Å^3)
50	3750.0	1278.0
100	**Not found**	975.0
150	745.0	**Not found**
200	3519.0	**Not found**
250	2907.0	1541.0
300	**Not found**	2197.0
350	1093.0	1045.0
400	**Not found**	2998.0
450	1035.0	1818.0
500	**Not found**	1928.0
550	1725.0	2110.0
600	5576.0	3185.0
650	6687.0	1731.0
700	6743.0	4021.0
750	5621.0	3838.0
800	5067.0	3508.0
850	5867.0	2901.0
900	12411.0	1886.0
950	14078.0	1835.0
1000	**Not found**	1968.0

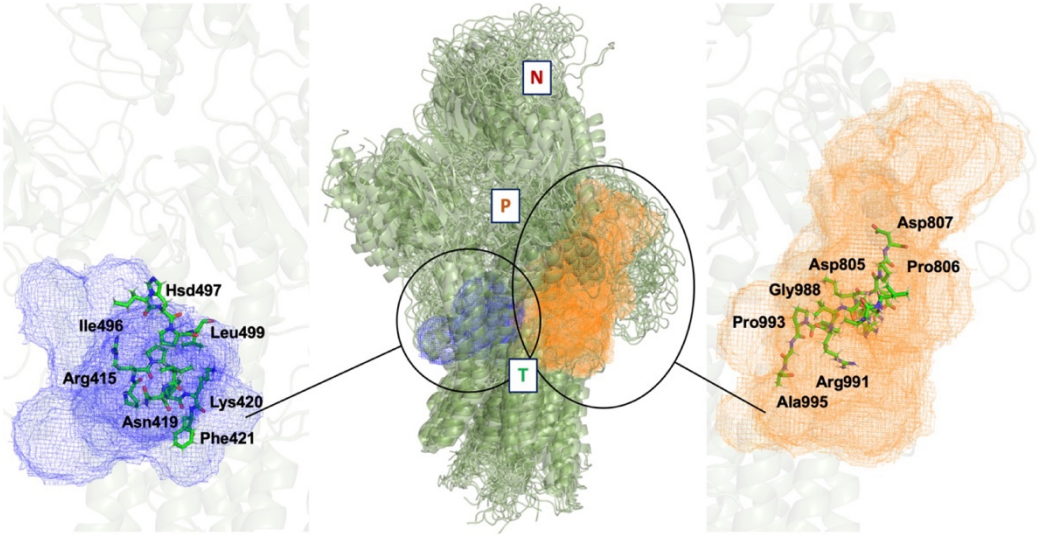


Figure S5. Two additional pockets (blue mesh and orange mesh) which were found on the surface of the ATP13A2 homology model (green cartoon).

One Mg^{2+} Active Site Molecular Dynamics (MD) Simulations

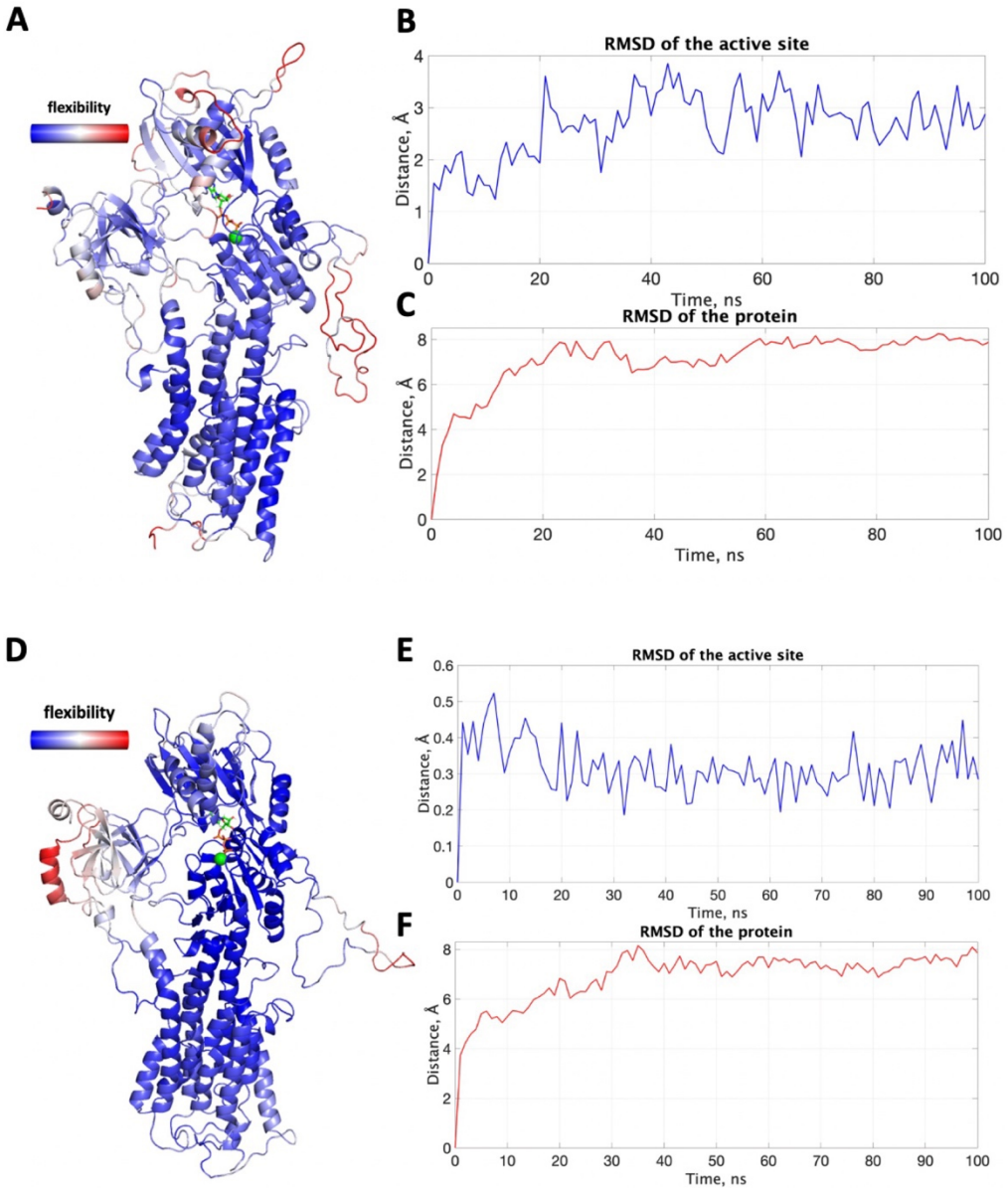


Figure S6. (A) The protein colored by flexibility in the first unconstrained MD simulation. Red regions signify more flexible parts while blue the more rigid regions. The membrane and water molecules/ions are not included in this visualization for clarity. The RMSD of the active site features Asp513, Thr515, Asp878, the Mg^{2+} cation and the ATP molecule. **(B)** shows the RMSD of the active site alone and **(C)** the RMSD of protein. **(D), (E)** and **(F)** show the same parameters for the second simulation where the ATP molecule was fixed to its original coordinates.

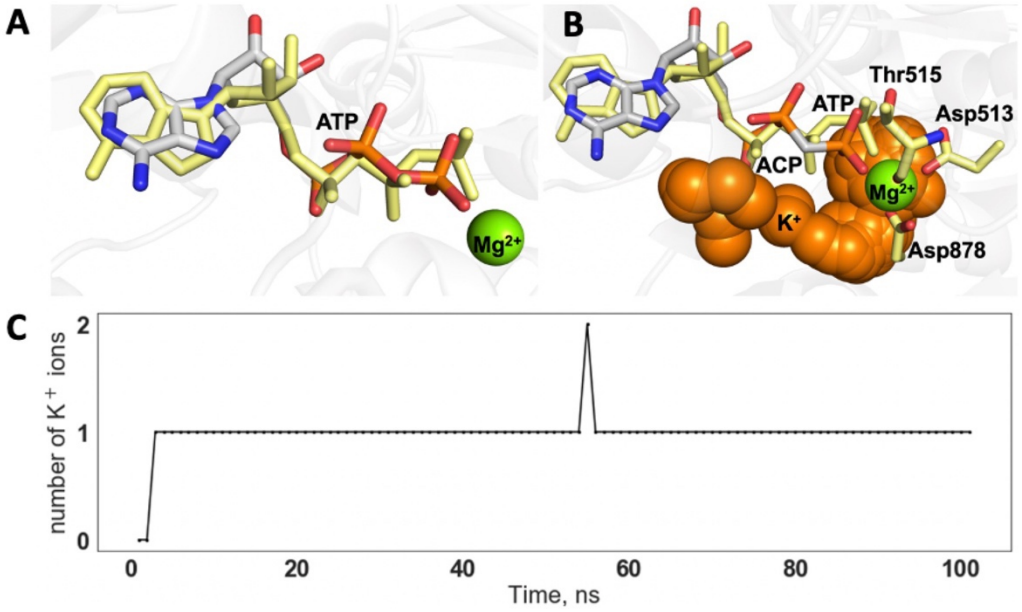


Figure S7. (A) Conformation of the ATP molecule (grey sticks) at the start of the MD simulation and after 1 ns (yellow sticks). **(B)** Region of the active site where the K^+ ion clustering is observed during the unconstrained simulation (orange spheres). The constant presence of the K^+ ions **(C)** suggests insufficiency of positive charge in the active site. Since the ATP has not preserved its original "zig-zag" conformation, the K^+ ions cluster ununiformly. **(C)** Number of K^+ ions in the active site throughout the 100 ns unconstrained simulation.

QM Cluster Calculations

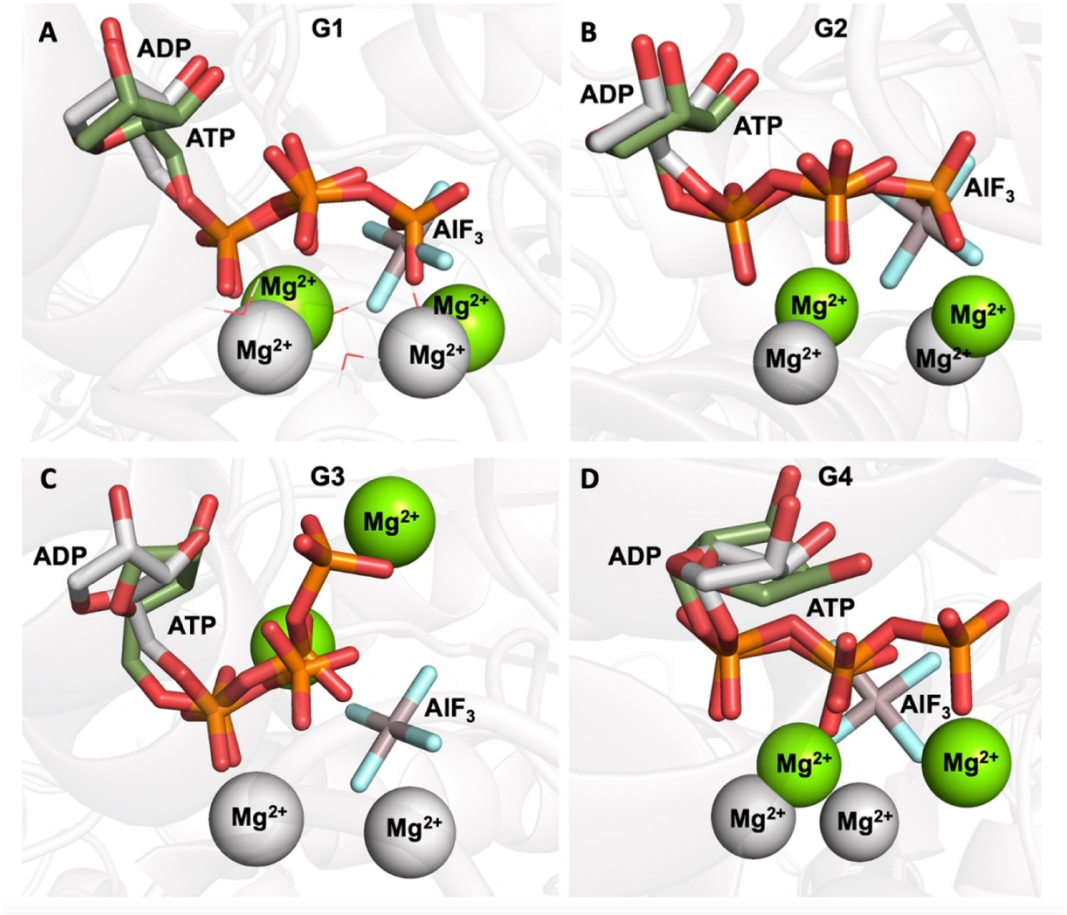


Figure S8. The four optimized geometries (green sticks) aligned to the crystal structure (grey sticks) resolved with ADP and two Mg^{2+} ions[2]. The conformation of G1 in **(A)** and G2 in **(B)** agree very well with the crystal structure of the transition state, unlike the "straight" phosphate chain conformations of G3 and G4 **(C)** and **(D)**, respectively.

Starting Geometries G1-G4 (Fig. S4) were optimized and upon convergence yielded three distinct conformations. G1 and G2 optimize to the same type of conformation and coordination mode. All optimized geometries with straight phosphate chain (G3 and G4) are energetically unfavourable (Table S3) and were therefore not used in any further QM/MM reaction coordinate scans (RCSs).

Table S3. Energy of the QM cluster optimized structures.

Optimized conformation	**Energy (kcal/mol)**
G1	-2725927.76
G2	-2725951.10
G3	-2725162.55
G4	-2725180.81

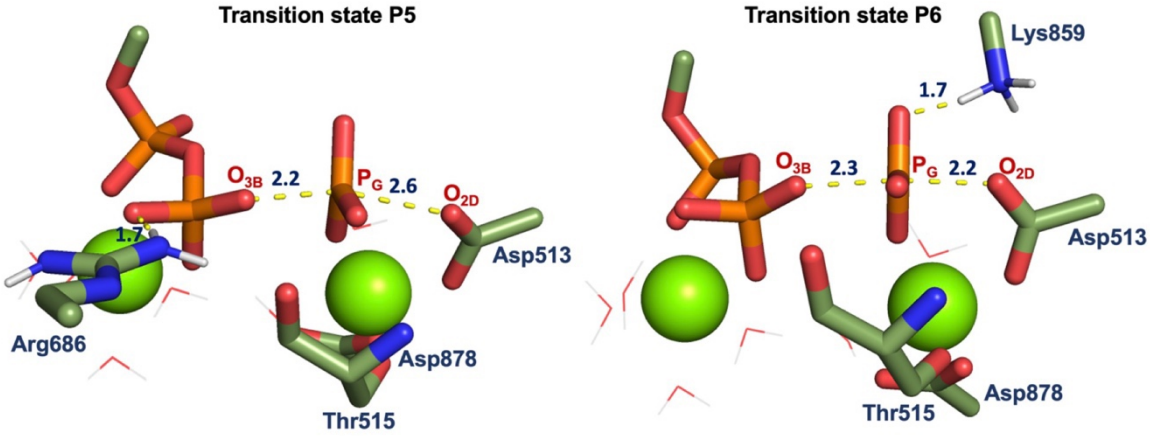


Figure S9. Transition state for systems P5 and P6. The transition state where Lys859 is missing is considerably affected (P5) while in system P6 the missing Arg686 does not influence the distances between the reacting nucleophile O_{2D} and P_G and P_G and O_{3B}.

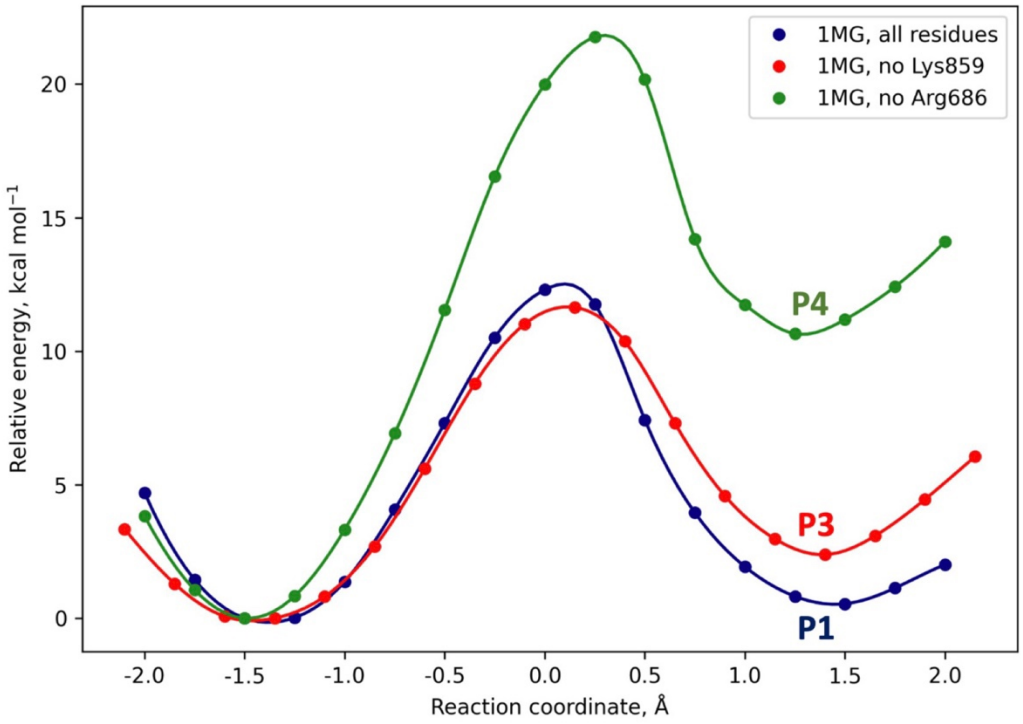


Figure S10. Reaction coordinate scans (RCS) for the active site of ATP13A2 containing one Mg^{2+} ion. The minimum energy profile along the RCS depicted in red represents the phosphate transfer reaction without Lys859 (P5). The profile in green represents the phosphate transfer reaction without Arg686 (P6). The blue energy profile shows the reaction in the active site when all amino acids required for the normal enzymatic activity are present.

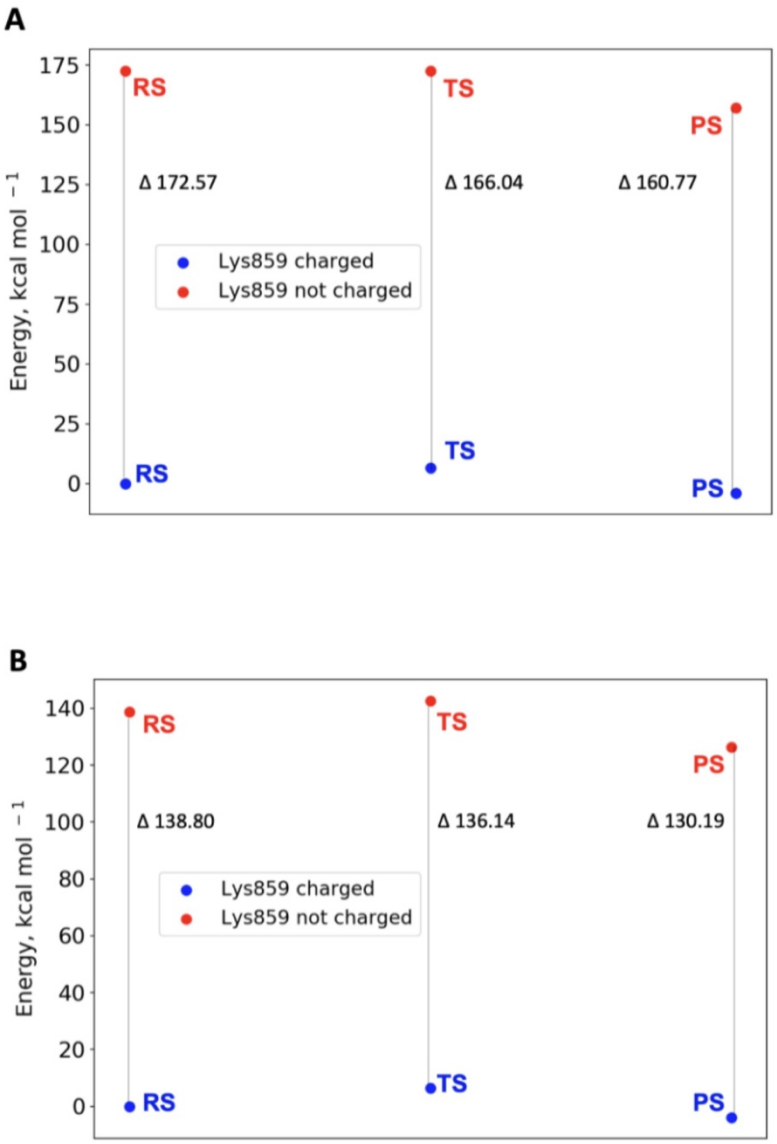


Figure S11. (A) Electrostatic effects of the missing Lys859 atomic charges on the destabilization of the RS, TS and PS using the geometries obtained with the full system (P2). **(B)** Geometric effects of the missing Lys859 atomic charges (geometries taken from scan P5) cause additional destabilization of the system in comparison to when all Lys859 is charged, in particular for the RS.

Two Mg^{2+} Active Site MD Simulations

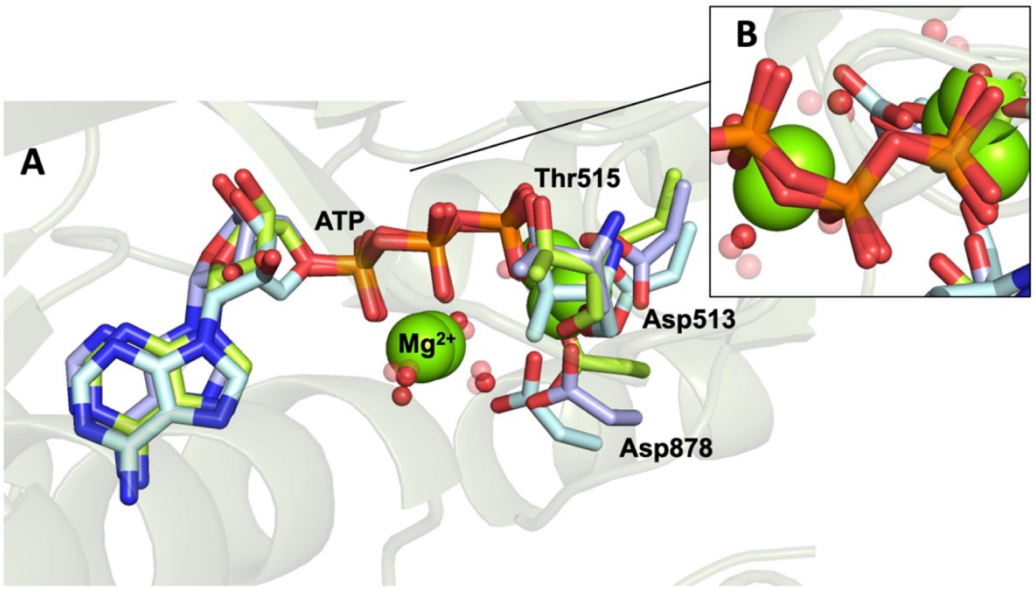


Figure S12. (A) Final active site structures of the 3 replicas after 100 ns of unconstrained MD simulations. Water molecules coordinating the Mg^{2+} ions are shown as red spheres. The second Mg^{2+} ion was coordinated by four water molecules and the catalytic Mg^{2+} ion was coordinated by one after 100 ns. The Mg-coordinating residues are shown as sticks. **(B)** Top view of the ATP phosphate chain, which was still zigzag after 100 ns of unconstrained MD.

References

1. Sacchetto, R. *et al.* Crystal structure of sarcoplasmic reticulum Ca 2+-ATPase (SERCA) from bovine muscle. *J. Struct. Biol.* **178**, 38–44 (2012).
2. Kanai, R., Ogawa, H., Vilsen, B., Cornelius, F. & Toyoshima, C. Crystal structure of a Na + -bound Na +,K + -ATPase preceding the E1P state. *Nature* **502**, 201–6 (2013).
3. McKenna, M. J. *et al.* The endoplasmic reticulum P5A-ATPase is a transmembrane helix dislocase. *Science (80-.).* **369**, eabc5809 (2020).
4. Bateman, A. *et al.* UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).

Appendix B

Table SI1. All variants modeled and subject of MD simulations in this work. The second column shows the mutations present; the third column shows the substrate in the active site, the fourth column shows the experimentally obtained k_{cat} value and the last column shows the target label used in the ML binary classification. All reported k_{cat} values come from the works of the cited authors,[13,101,105] as well as from collaborators at A*STAR, Singapore.[119] If more than one k_{cat} value is reported, it means that there was more than one experimental reporting of it. Values for k_{cat} listed as * were reported in $\Delta G^{\ddagger}$ from collaborators and started from 16.7 kcal/mol.

Mutant	Missense mutations	Substrate	k_{cat} [s^{-1}]	Class
WT	none	D-galactose	1094 503±16.2	0
W290F	W290F	D-galactose	371 ± 43.0	0
W290G	W290G	D-galactose	1.66 ± 0.28	1
W290H	W290H	D-galactose	0.24 ± 0.004	1
N6M1	S10P, M70V, P136(silent), G195E, V494A, N535D	D-galactose	1100 ± 41	0
N6M1_C383M	S10P, M70V, P136(silent), G195E, C383M, V494A, N535D	D-galactose	510 ± 15	0
N6M1_C383A	S10P, M70V, P136(silent), G195E, C383A, V494A, N535D	D-galactose	1200 ± 50	0
N6M1_C383N	S10P, M70V, P136(silent), G195E, C383N, V494A, N535D	D-galactose	410 ± 17	0
N6M1_C383D	S10P, M70V, P136(silent), G195E, C383D, V494A, N535D	D-galactose	440 ± 3.8	0
N6M1_C383P	S10P, M70V, P136(silent), G195E, C383P, V494A, N535D	D-galactose	490 ± 17	0
N6M1_C383E	S10P, M70V, P136(silent),	D-galactose	550 ± 9.0	0

	G195E, C383E, V494A, N535D			
N6M1_C383Q	S10P, M70V, P136(silent), G195E, C383Q, V494A, N535D	D-galactose	170 ± 12	0
N6M1_C383F	S10P, M70V, P136(silent), G195E, C383F, V494A, N535D	D-galactose	190 ± 13	0
N6M1_C383R	S10P, M70V, P136(silent), G195E, C383M, V494A, N535D	D-galactose	8.8 ± 0.3	1
N6M1_C383G	S10P, M70V, P136(silent), G195E, C383G, V494A, N535D	D-galactose	1100 ± 12	0
N6M1_C383S	S10P, M70V, P136(silent), G195E, C383S, V494A, N535D	D-galactose	1100 ± 30	0
N6M1_C383H	S10P, M70V, P136(silent), G195E, C383H, V494A, N535D	D-galactose	210 ± 5.7	1
N6M1_C383T	S10P, M70V, P136(silent), G195E, C383T, V494A, N535D	D-galactose	3400 ± 300	0
N6M1_C383I	S10P, M70V, P136(silent), G195E, C383I, V494A, N535D	D-galactose	260 ± 8.5	0
N6M1_C383V	S10P, M70V, P136(silent), G195E, C383V, V494A, N535D	D-galactose	360 ± 7.8	0
N6M1_C383K	S10P, M70V, P136(silent), G195E, C383K, V494A, N535D	D-galactose	1100 ± 30	0
N6M1_C383W	S10P, M70V, P136(silent),	D-galactose	0.011 ± 0.0004	1

	G195E, C383W, V494A, N535D			
N6M1_C383L	S10P, M70V, P136(silent), G195E, C383L, V494A, N535D	D-galactose	450 ± 17	0
M35	S10P, M70V, G195E, W290F, R330M, Q406T, V494A, N535D	S128	3.5 ± 0.04	1
M35	S10P, M70V, G195E, W290F, R330M, Q406T, V494A, N535D	SS1	*	1
GOH_1052	S10P, M70V, G195E, W290F, R330M, Q406T, V494A, N535D, F194A, N245W	S128	*	1
GOH_1036	S10P, M70V, G195E, W290F, R330M, Q406T, V494A, F194A	S128	*	1
GOH_1021	S10P, M70V, G195E, W290F, R330M, Q406T, V494A, F194A	S128	*	1
M35_24	S10P, M70V, G195E, W290F, R330M, E406T, V494A, N535D, T130S	SS1	3.6 ± 0.07	1
M35_32	S10P, M70V, G195E, W290F, R330M, E406T, V494A, N535D, M278T, D517V, Y576H	SS1	3.1 ± 0.05	1
M35_215	S10P, M70V, G195E, W290F, R330M, E406T, V494A, N535D, D413Y, Y436F	SS1	2.9 ± 0.05	1

Table SI2. Total charges of the active site residues and substrates calculated for the TS structure of GO with D-Galactose, SS1, and S128. The values are reported in units of electron charge (e).

Residue	D-Galactose	SS1	S128
Cu	0.777	0.779	0.411
Substrate	-0.344	-0.612	-0.081
HX	0.223	0.590	0.548
C228	-0.369	-0.234	-0.225
Y272	-0.149	-0.283	-0.411
H496	0.023	-0.122	0.014
H581	-0.151	-0.026	0.097
Y495	-0.088	-0.070	-0.157
P290	0.078	-0.020	-0.197

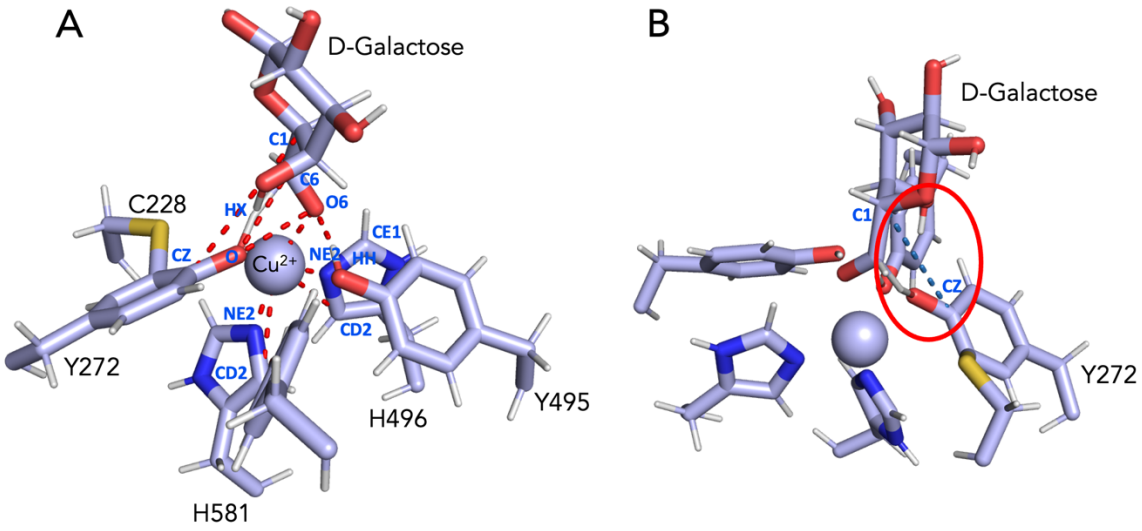


Figure SI1. (A) Atoms with additional harmonic potentials. (B) The second set of simulations did not contain the harmonic restraint between atoms CZ(Y272) and C1 (substrate); the distance is shown with blue dashed lines.

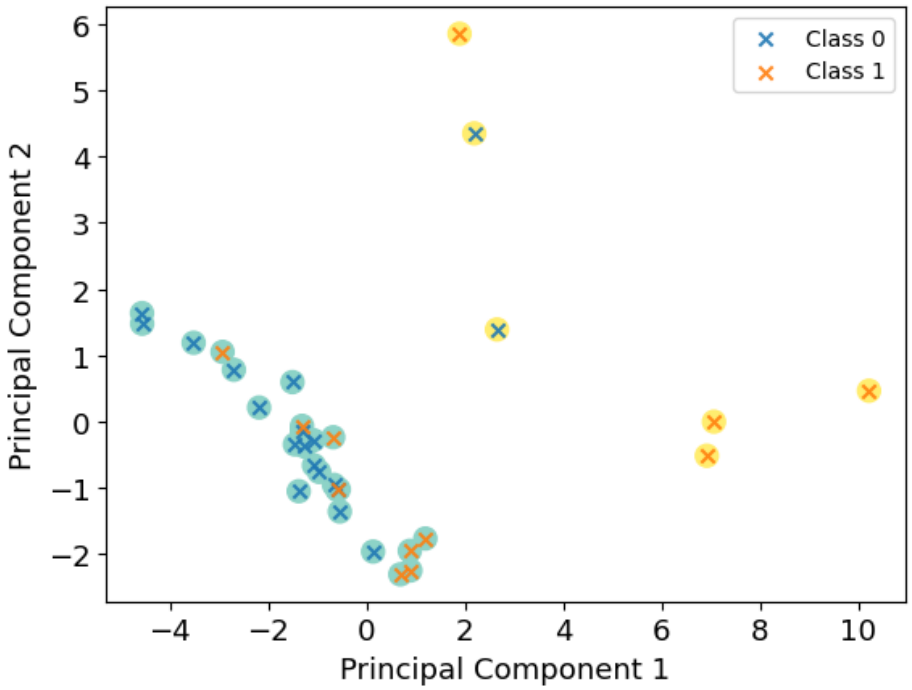


Figure SI2. Data points clustered based on the first two principal components using k-means clustering.

Table SI3. Restraints on interatomic distances used during the MD simulations. The distance marked with an asterisk * is not restrained during the second set of 93 simulations. The units are in kJ mol^{-1} nm^{-1}.

Atoms	Force constant	Reference distance (Å)
HX(Sub)-O(TYX)	10000	1.3
HX(Sub)-C6(Sub)	10000	1.3
O6(Sub)-Cu	100000	2.1
O(TYX)-C6(Sub)	100000	2.6
O(TYX)-O6(Sub)	100000	3.3
CZ(TYX)-C1(Sub)*	100000	3.6
HH(TRR495)-O6(Sub)	5000	1.7
NE2(HIS581)-CU	10000	1.9
NE2(HIS496)-CU	10000	1.9
CD2(HIS496)-CU	10000	3.0

CD2(HIS581)-CU	10000	2.9

Table SI4. Restraints on angles during the MD simulations.

Angle	Force constant	Reference angle
NE2-CU-NE2(H496, H581)	359.0	153.0
NE2-O6-NE2(H496, Sub, H581)	359.0	73.2
CU-O6-C6(Sub)	569.0	129.6
CE1-NE2-CU(His496)	359.0	124.4
OH(Y272)-O6-C6(Sub)	359.0	49.7

Table SI5. Reduced set of interatomic distances and distances created from displacement (WT reference).

Cu-SG(C228) d1	WT d2 – Variant d2
Cu-HX(Substrate) d2	WT d3 – Variant d3
OH(Y272)-Cu d3	WT d4 – Variant d4
OH(Y495)-Cu d4	WT d5 – Variant d5
OH(Y272)-NE2(H496) d5	WT d6 – Variant d6
OH(Y272)-NE2(H581) d6	WT d7 – Variant d7
OH(Y405)-Cu d7	WT d8 – Variant d9
CZ(Y272)-C6(Substrate) d8	WT d10 – Variant d10
CG(Y272)-C6(Substrate) d9	WT d11 – Variant d112
SG(C228)-C6(Substrate) d10	WT d12 – Variant d12
OH(Y495)-O6(Substrate) d11	WT d13 – Variant d13
OH(Y495)-C6(Substrate) d12	WT d14 – Variant d14
OH(Y405)-O6(Substrate) d13	WT d15 – Variant d15
CZ(Y272)-SG(C228) d14	WT d16 – Variant d16
O6(Substrate)-CE1(H496) d15	WT d17 – Variant d17
Cu-CZ(Y495) d16	WT d18 – Variant d18
Cu-CZ(C228) d17	WT d19 – Variant d19
CB(C228)-C6(Substrate) d18	WT d20 – Variant d20
CZ(Y272)-OH(Y495) d19	WT d21 – Variant d21

CZ(Y495)-SG(YC228) d20	WT d22 – Variant d22
SG(C228)-OH(Y495) d21	WT d23 – Variant d23
SG(C228)-CZ(Y495) d22	WT d24 – Variant d24
SG(C228)-OH(Y495) d23	RMSD (First – Last Frame)
OH(Y405)-OH(Y272) d24	RMSF (First-Last Frame)
WT d1 – Variant d1	

Table SI6. Set of interatomic distances and distances created from displacement (WT reference).

Cu-SG(C228) d1	WT d1 – Variant d1
Cu-HX(Substrate) d2	WT d2 – Variant d2
OH(Y272)-Cu d3	WT d3 – Variant d3
OH(Y495)-Cu d4	WT d4 – Variant d4
OH(Y272)-NE2(H496) d5	WT d5 – Variant d5
OH(Y272)-NE2(H581) d6	WT d6 – Variant d6
C1(Substrate)-Cu d7	WT d7 – Variant d7
OH(Y405)-Cu d8	WT d8 – Variant d8
C1(Substrate)-CZ(Y272) d9	WT d9 – Variant d9
CZ(Y272)-C6(Substrate) d10	WT d10 – Variant d10
OH(Y272)-C1(Substrate) d11	WT d11 – Variant d11
CG(Y272)-C6(Substrate) d12	WT d12 – Variant d12
CG(Y272)-C1(Substrate) d13	WT d13 – Variant d13
SG(C228)-C6(Substrate) d14	WT d14 – Variant d14
SG(C228)-C1(Substrate) d15	WT d15 – Variant d15
OH(Y495)-O6(Substrate) d16	WT d16 – Variant d16
OH(Y495)-C6(Substrate) d17	WT d17 – Variant d17
OH(Y495)-C1(Substrate) d18	WT d18 – Variant d18
OH(Y405)-C1(Substrate) d19	WT d19 – Variant d19
OH(Y405)-O6(Substrate) d20	WT d20 – Variant d20
CZ(Y272)-SG(C228) d21	WT d21 – Variant d21
O6(Substrate)-CE1(H496) d22	WT d22 – Variant d22
Cu-CZ(Y495) d23	WT d23 – Variant d23
Cu-CZ(C228) d24	WT d24 – Variant d24

CB(C228)-C6(Substrate) d25	WT d25 – Variant d25
CZ(Y272)-OH(Y495) d26	WT d26 – Variant d26
CZ(Y495)-SG(YC228) d27	WT d27 – Variant d27
SG(C228)-OH(Y495) d28	WT d28 – Variant d28
SG(C228)-CZ(Y495) d29	WT d29 – Variant d29
SG(C228)-OH(Y495) d30	WT d30 – Variant d30
OH(Y405)-OH(Y272) d31	RMSD (First – Last Frame)
WT d1 – Variant d1	RMSF (First – Last Frame)

Appendix C

Electronic Supplementary Information (ESI)

Combining data integration and molecular dynamics for target identification in ∝-synuclein-aggregating neurodegenerative diseases: Structural insights on Synj1

Kirsten Jenkins,[a] Teodora Mateeva,[b] István Szabó,[b] Andre Melnik,[c] Paola Picotti,[c] Attila Csikász-Nagy,[a,d] Edina Rosta[b*]

[a]Randall Division of Cell and Molecular Biophysics, Institute for Mathematical and Molecular Biomedicine, King's College London, London SE1 1UL, UK

[b]Department of Chemistry, King's College London, London SE1 1DB, UK

[c]Institute of Biochemistry, Department of Biology, ETH Zurich, CH-8093 Zurich, Switzerland

[d]Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, 1083 Budapest, Hungary

*Corresponding author:

E-mail: edina.rosta@kcl.ac.uk

Electronic Supplementary Tables

Table 1 ESI. Proteins chosen to be of interest in this work. The corresponding human homologue of the yeast protein is shown in the second column [1]. Third and fourth column show the disease modulating effect of the protein on ∝-synuclein toxicity, as found in the Khurana *et. al* study [2]. Fifth and sixth column show the median ratio for the protein concentration between ∝-synuclein expressing cells and control empty vector (EV) cells. The final column shows the average of the median ratio value (∝-synuclein expressing vs. control at 12h and 18h) from the Melnik *et. al* study [3]. Values coloured in red signify upregulated proteins, blue - downregulated.

Yeast Protein	Human Homologue	Deletion Modulator	Overexpression Modulator	Median ratio value for the protein c. at 12h	Median ratio value for the protein c. at 8h	Average of median ratio value between 12h and 18h
ERV29	SURF4		Suppressor	0.9487	0.7617	0.8552
CAB3	PPCDC		Suppressor	0.9913	0.2275	0.6094
OSH2	OSBP		Suppressor	0.7693	0.7696	0.7695
TIS11	ZFP36		Suppressor	0.5446	1.0161	0.7803
PSR1	CTDSP2		Suppressor	0.7297	0.3574	0.5436
FUN14	FUNDC1		Suppressor	0.5588	0.9406	0.7497
YPK9	ATP13A2	Enhancer	Suppressor	0.5094	0.6475	0.5785
TPK2	PRKACB	Enhancer		0.6505	0.6821	0.6663
INP53	SYNJ1	Enhancer		0.7353	1.0839	0.9096
RAD27	FEN1	Enhancer		0.8404	0.5271	0.6837
ARO10	ILVBL	Enhancer		0.7201	0.9930	0.8566
IMP2	IMPP2L	Enhancer		0.3738	0.3737	0.3738
RSM25	MRPS23		Enhancer	1.6853	1.5705	1.6279
YMR31	MRPS36		Enhancer	1.6649	1.7560	1.7105
POR1	VDAC1		Enhancer	1.1816	1.5695	1.3756
SEC31	SEC31B		Enhancer	1.3086	1.1884	1.2485
MRPL11	MRPL10		Enhancer	1.6600	1.6863	1.6732

Table 2 ESI. Close functional partners to synaptojanin-1. Generated using STITCH [4].

Protein name	Function
PIK3CA	Phosphatidylinositol-4,5-bisphosphate 3-kinase, **catalytic subunit alpha**; Phosphoinositide-3-kinase (PI3K) that phosphorylates PtdIns (Phosphatidylinositol), PtdIns4P (Phosphatidylinositol 4- phosphate) and PtdIns(4,5)P2 (Phosphatidylinositol 4,5- bisphosphate) to generate phosphatidylinositol 3,4,5-trisphosphate (PIP_3). PIP_3 plays a key role by recruiting PH domain-containing proteins to the membrane, including AKT1 and PDPK1, activating signalling cascades involved in cell growth, survival, proliferation, motility and morphology.
PIK3CB	Phosphatidylinositol-4,5-bisphosphate 3-kinase, **catalytic subunit beta**; Phosphoinositide-3-kinase (PI3K) that phosphorylates PtdIns (Phosphatidylinositol), PtdIns4P (Phosphatidylinositol 4- phosphate) and PtdIns(4,5)P2 (Phosphatidylinositol 4,5- bisphosphate) to generate phosphatidylinositol 3,4,5-trisphosphate (PIP3). PIP3 plays a key role by recruiting PH domain-containing proteins to the membrane, including AKT1 and PDPK1, activating signalling cascades involved in cell growth, survival, proliferation, motility and morphology.
PIK3CD	Phosphatidylinositol-4,5-bisphosphate 3-kinase, **catalytic subunit delta**; Phosphoinositide-3-kinase (PI3K) that phosphorylates PftdIns(4,5)P2 (Phosphatidylinositol 4,5-bisphosphate) to generate phosphatidylinositol 3,4,5-trisphosphate (PIP_3). PIP_3 plays a key role by recruiting PH domain-containing proteins to the membrane, including AKT1 and PDPK1, activating signalling cascades involved in cell growth, survival, proliferation, motility and morphology. Mediates immune responses. Plays a role in B-cell development, proliferation, migration, and function.
PIK3CG	Phosphatidylinositol-4,5-bisphosphate 3-kinase, **catalytic subunit gamma**; Phosphoinositide-3-kinase (PI3K) that phosphorylates PtdIns(4,5)P2 (Phosphatidylinositol 4,5-bisphosphate) to generate phosphatidylinositol 3,4,5-trisphosphate (PIP_3). PIP_3 plays a key role by recruiting PH domain-containing proteins to the membrane, including AKT1 and PDPK1, activating signalling cascades involved in cell growth, survival, proliferation, motility and morphology. Links G-protein coupled receptor activation to PIP_3 production.
EPHB2	EPH receptor B2; Receptor tyrosine kinase which binds promiscuously transmembrane ephrin-B family ligands residing on adjacent cells, leading to contact-dependent bidirectional signalling into neighbouring cells. The signalling pathway downstream of the receptor is referred to as forward signalling while the signalling pathway downstream of the ephrin ligand is referred to as reverse signalling. Functions in axon guidance during development. Involved in the guidance of commissural axons, that form a major interhemispheric connection between the 2 temporal lobes of the cerebral cortex.
SH3GL2	Endophilin-A3; Implicated in endocytosis. May recruit other proteins to membranes with high curvature.
EPS15	Epidermal growth factor receptor substrate 15; Involved in cell growth regulation. May be involved in the regulation of mitogenic signals and control of cell proliferation. Involved in the internalization of ligand-inducible receptors of the receptor tyrosine kinase (RTK) type, in particular EGFR. Plays a role in the assembly of clathrin-coated pits (CCPs). Acts as a clathrin adapter required for post-Golgi trafficking. Seems to be involved in CCPs maturation including invagination or

	budding. Involved in endocytosis of integrin beta-1 (ITGB1) and transferrin receptor (TFR).
EPN1	Epsin 1; Binds to membranes enriched in phosphatidylinositol 4,5- bisphosphate (PtdIns(4,5)P2). Modifies membrane curvature and facilitates the formation of clathrin-coated invaginations.
BIN1	Myc box-dependent-interacting protein 1; May be involved in regulation of synaptic vesicle endocytosis. May act as a tumor suppressor and inhibits malignant cell transformation.
AP2A1	AP-2 complex subunit alpha-1; Component of the adaptor protein complex 2 (AP-2). Adaptor protein complexes function in protein transport via transport vesicles in different membrane traffic pathways. Adaptor protein complexes are vesicle coat components and appear to be involved in cargo selection and vesicle formation. AP-2 is involved in clathrin-dependent endocytosis in which cargo proteins are incorporated into vesicles surrounded by clathrin (clathrin-coated vesicles, CCVs) which are destined for fusion with the early endosome.
AP2M1	Adaptor-related protein complex 2, mu 1 subunit; Component of the adaptor protein complex 2 (AP-2). Adaptor protein complexes function in protein transport via transport vesicles in different membrane traffic pathways. Adaptor protein complexes are vesicle coat components and appear to be involved in cargo selection and vesicle formation. AP-2 is involved in clathrin-dependent endocytosis in which cargo proteins are incorporated into vesicles surrounded by clathrin (clathrin- coated vesicles, CCVs) which are destined for fusion with the early endosome.
MTMR6	Myotubularin related protein 6; Phosphatase that acts on lipids with a phosphoinositol headgroup. Acts as a negative regulator of KCNN4/KCa3.1 channel activity in CD4+ T-cells possibly by decreasing intracellular levels of phosphatidylinositol 3 phosphatase. Negatively regulates proliferation of reactivated CD4+ T-cells.
SYNJ2	Synaptojanin 2; Inositol 5-phosphatase which may be involved in distinct membrane trafficking and signal transduction pathways. May mediate the inhibitory effect of Rac1 on endocytosis.
PI4KB	Phosphatidylinositol 4-kinase, catalytic, **beta**; Phosphorylates phosphatidylinositol (PI) in the first committed step in the production of the second messenger inositol- 1,4,5,-trisphosphate (PIP). May regulate Golgi disintegration/reorganization during mitosis, possibly via its phosphorylation.
PI4KA	Phosphatidylinositol 4-kinase, catalytic, **alpha**; Acts on phosphatidylinositol (PtdIns) in the first committed step in the production of the second messenger inositol- 1,4,5,-trisphosphate.
PIK3C2B	Phosphatidylinositol-4-phosphate 3-kinase, catalytic subunit type 2 beta; Phosphorylates PtdIns and PtdIns4P with a preference for PtdIns. Does not phosphorylate PtdIns(4,5)P2. May be involved in EGF and PDGF signaling cascades.
PIK3C2G	Phosphatidylinositol-4-phosphate 3-kinase, catalytic subunit type 2 gamma; Generates phosphatidylinositol 3-phosphate (PtdIns3P) and phosphatidylinositol 3,4-bisphosphate (PtdIns(3,4)P2) that act as second messengers.
PIK3C2A	Phosphatidylinositol-4-phosphate 3-kinase, catalytic subunit type 2 alpha; Generates phosphatidylinositol 3-phosphate (PtdIns3P) and phosphatidylinositol 3,4-bisphosphate (PtdIns(3,4)P2) that act as second messengers.
PPP3CA	Protein phosphatase 3, catalytic subunit, alpha isozyme; Calcium-dependent, calmodulin-stimulated protein phosphatase. This subunit may have a role in the calmodulin activation of calcineurin. Dephosphorylates DNM1L, HSPB1 and SSH1.

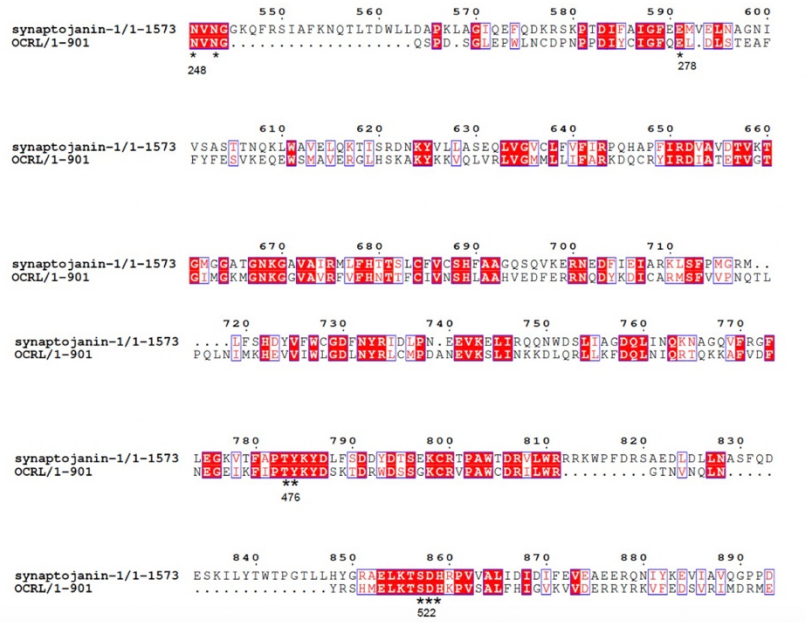


Fig. 1 ESI. The 5-phosphatase domain of human Synj1 (Synaptojanin-1, uniport code: O43426) and OCRL (Inositol polyphosphate 5-phosphatase, uniport code: Q01968) are aligned. The conserved residues in the active site are marked with an asterisk. The numbering is based on synj1. The corresponding numbering of the active site residues in the OCRL protein are shown below.

Sequence analysis was performed on proteins within the 5-phosphoinositide phosphatase family (OCRL, I5P2, SYNJ1, SHIP2) [5–8] using the algorithm ClustalW and Muscle [9,10] (as implemented in Jalview) [11] and visualised with Espript 3.0 [12]. The important conserved residues within the active site are marked with a red asterisk. The Mg cation-coordinating residues, corresponding to His 360, Asp 359 and Glu 92 in the homology model, are highly conserved in this family and are marked with asterisk. This sequence alignment explores only the 5-phosphatase domain of the proteins.

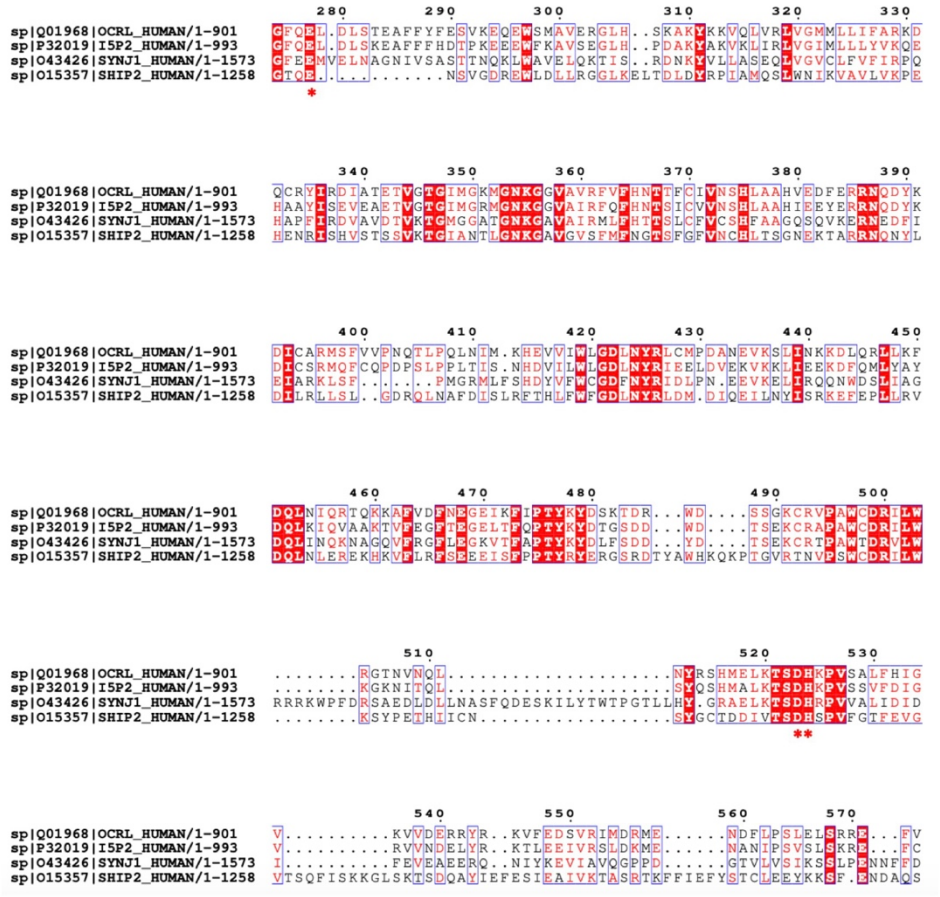


Fig. 2 ESI. Amino acid sequence comparisons of the 5-phosphatase domain of human 5-phosphoinositide phosphatases. Alignments were performed using the Clustal W and Muscle algorithm as implemented in Jalview and visualisation was done with Espript 3.0 [9–12]. Residue numbering is based on OCRL. The conserved residues in the active site Asp (**D**), His (**H**) and Glu (**E**) found for all 5-phosphatases (SHIP2, Synj1, OCRL, INPPB5) [5,6,8] are marked with an asterisk. These are Mg-coordinating and correspond to His-360, Asp-359 and Glu-92 in the homology model. All sequences are obtained from Uniprot, using the Homo sapiens sequence [13].

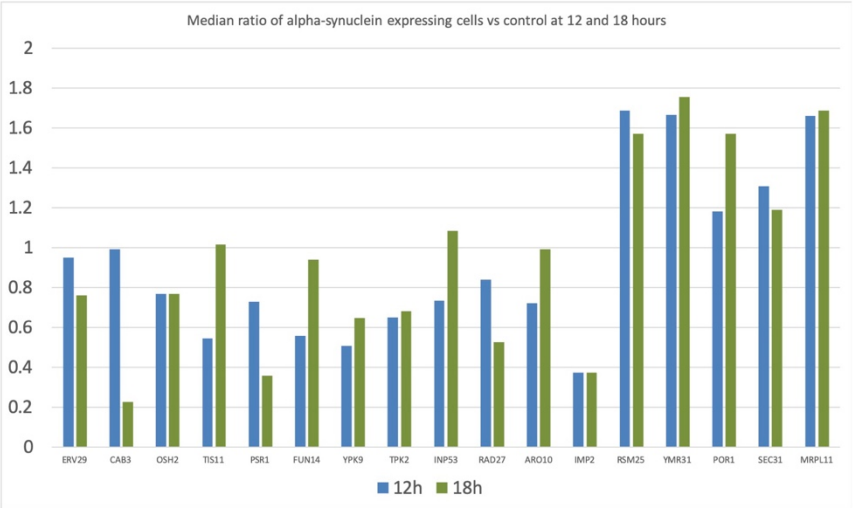


Fig. 3 ESI. Median ratio of protein concentration in ∝-synuclein expressing cells vs. control at 12h and 18h. Proteins which have a median ratio value of above 1 (averaged between 12h and 18 h), have been defined as 'upregulated', proteins with a value below 1 have been defined as downregulated in ∝-synuclein expressing cells.

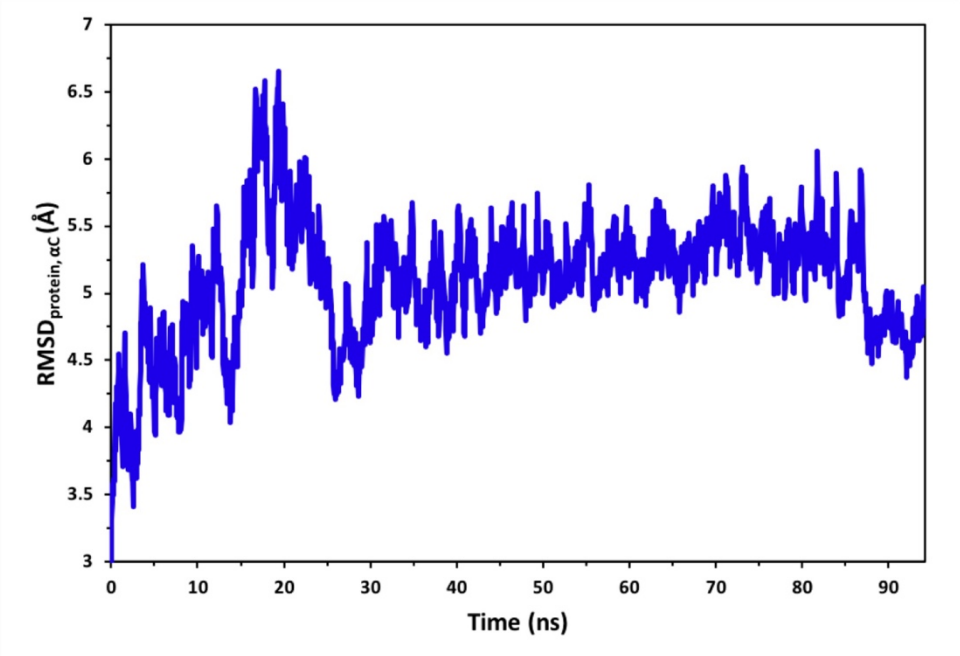


Fig. 4 ESI. RMSD of the membrane-free simulation.

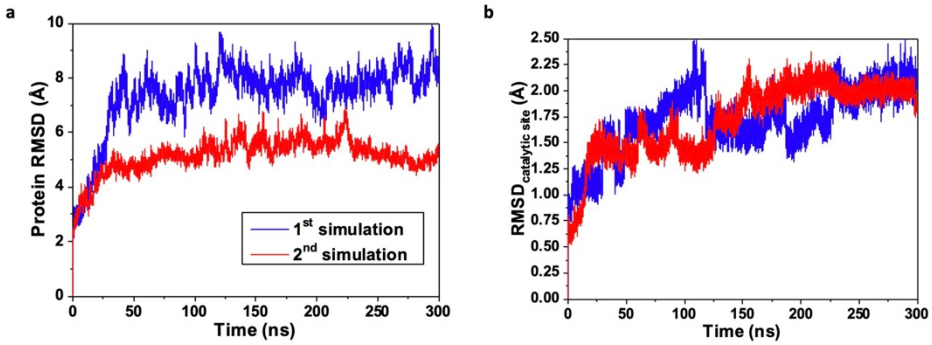


Fig. 5 ESI. RMSD of the membrane-embedded protein (**a**) in simulation 1 and simulation 2 and of the active site only (**b**).

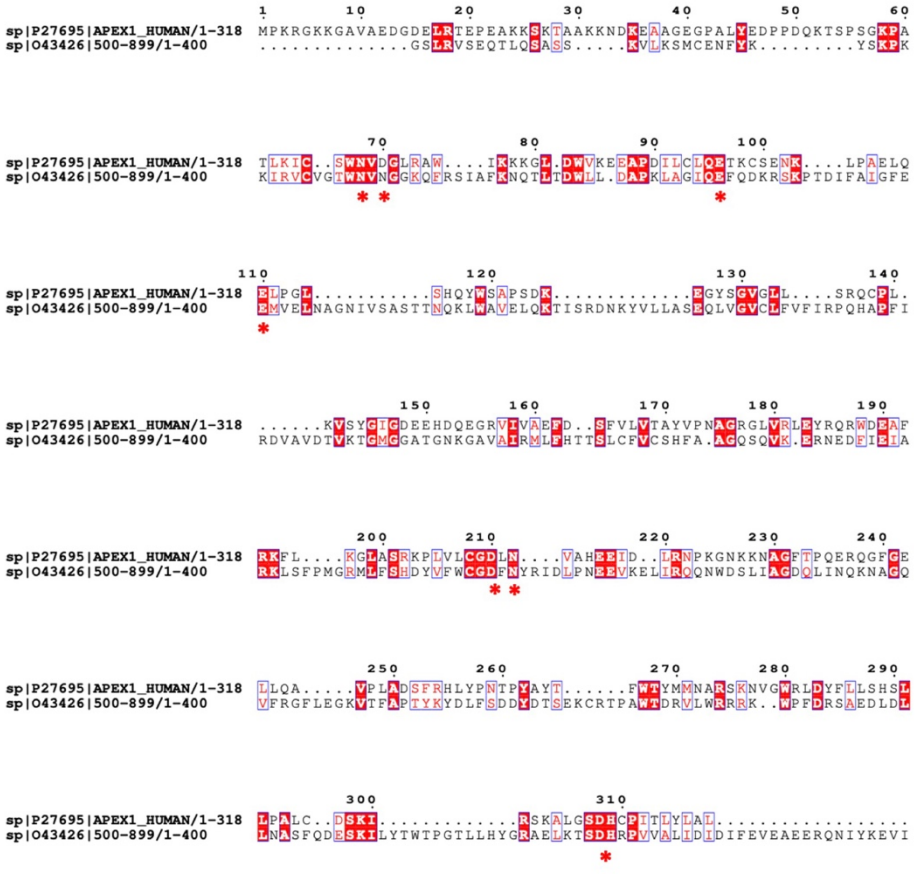


Fig. 6 ESI. Amino acid sequence of the 5-phosphatase domain of synj1 aligned to the apurinic/apyrimidinic base excision repair endonuclease Ape1. All sequences are obtained from Uniprot [13], and correspond to Homo sapiens. Visualised with Espript 3.0 [12].

References

[1] Balakrishnan R, Park J, Karra K, Hitz BC, Binkley G, Hong EL, et al. YeastMine-An integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. Database 2012. https://doi.org/10.1093/database/bar062.

[2] Khurana V, Peng J, Chung CY, Auluck PK, Fanning S, Tardiff DF, et al. Genome-Scale Networks Link Neurodegenerative Disease Genes to ∝-Synuclein through Specific Molecular Pathways. Cell Syst 2017. https://doi.org/10.1016/j.cels.2016.12.011.

[3] Melnik, A., Cappellutti, V., Vaggi, F., Piazza, I., Tognetti, M., Soste, M., de Souza, N., Csikasz-Nagy, A., Piccotti P. In Preparation 2019.

[4] Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P. STITCH: Interaction networks of chemicals and proteins. Nucleic Acids Res 2008. https://doi.org/10.1093/nar/gkm795.

[5] Tsujishita Y, Guo S, Stolz LE, York JD, Hurley JH. Specificity determinants in phosphoinositide dephosphorylation: Crystal structure of an archetypal inositol polyphosphate 5-phosphatase. Cell 2001. https://doi.org/10.1016/S0092-8674(01)00326-9.

[6] Trésaugues L, Silvander C, Flodin S, Welin M, Nyman T, Gräslund S, et al. Structural basis for phosphoinositide substrate recognition, catalysis, and membrane interactions in human inositol polyphosphate 5-phosphatases. Structure 2014. https://doi.org/10.1016/j.str.2014.01.013.

[7] Hsu FS, Mao Y. The structure of phosphoinositide phosphatases: Insights into substrate specificity and catalysis. Biochim Biophys Acta - Mol Cell Biol Lipids 2015;1851:698–710. https://doi.org/10.1016/j.bbalip.2014.09.015.

[8] Mills SJ, Silvander C, Cozier G, Trésaugues L, Nordlund P, Potter BVL. Crystal Structures of Type-II Inositol Polyphosphate 5-Phosphatase INPP5B with Synthetic Inositol Polyphosphate Surrogates Reveal New Mechanistic Insights for the Inositol 5-Phosphatase Family. Biochemistry 2016. https://doi.org/10.1021/acs.biochem.5b00838.

[9] Clustalw U, To C, Multiple DO. ClustalW and ClustalX. Options 2003. https://doi.org/10.1002/0471250953.bi0203s00.

[10] Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004. https://doi.org/10.1093/nar/gkh340.

[11] Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics 2009. https://doi.org/10.1093/bioinformatics/btp033.

[12] Gouet P, Courcelle E, Stuart DI, Métoz F. ESPript: Analysis of multiple sequence alignments in PostScript. Bioinformatics 1999. https://doi.org/10.1093/bioinformatics/15.4.305.

[13] Hancock JM, Zvelebil MJ, Zvelebil MJ. UniProt. Dict. Bioinforma. Comput. Biol., 2004. https://doi.org/10.1002/9780471650126.dob0721.pub2.

Appendix D

Correction

On p. S3, MD simulations of ferrous-ferric ET, this part of the sentence is redundant ".. and the LINCS constraint algorithm was used to constrain bonded hydrogens".

Direct Calculation of Electron Transfer Rates with the Binless Dynamic Histogram Analysis Method

Zsuzsanna Koczor-Benda,[1,2†] Teodora Mateeva,[3†] Edina Rosta[1*]

[1] Department of Physics and Astronomy, University College London, London, WC1E 6BT, United Kingdom

[2] The Department of Chemistry, University of Warwick, Coventry, CV4 7AL, United Kingdom

[3] Department of Physics, King's College London, London, WC2R 2LS, United Kingdom

[†] Equal contributions

[*] e-mail: e.rosta@ucl.ac.uk

S1. Alternative binless formulation

Instead of using the bias calculated for each data point as in Equation 4 of the manuscript, a binless formulation of DHAM can also be achieved by using the average (or median) bias $\bar{u}_i^l$ of all data points falling in bin *i* in simulation window *l* according to

$$M_{ji} = \sum_{k=1}^{N} \frac{T_{ji}^k}{\sum_{l=1}^{N} n_i^l \exp(-(\bar{u}_j^l - \bar{u}_i^l)/2k_BT)}. \qquad \text{(S1)}$$

We test this approach, which is analogous to that presented in Ref. [1] for the application of Ala5, on the ferrous-ferric ET example in Section S8.

S2. 1-D model potential

Monte Carlo (MC) simulations were carried out on an analytical model potential (details can be found in Ref.[2]) using 50 uniformly distributed umbrella windows in the range [0.05, 1.55], with K = 200 kcal/mol biasing spring constant for 5000 steps. The average of 20 repeated simulations was used to construct the free energy profile using 500 bins.

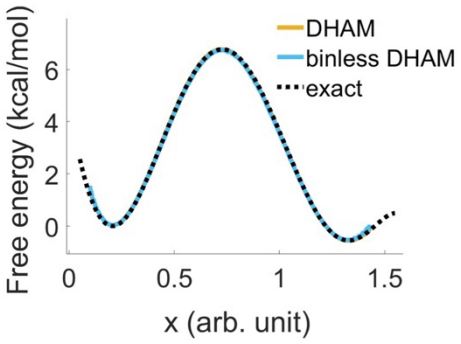


Figure S1. Free energy profiles for the 1-D model potential reconstructed with binless DHAM (blue) and DHAM (orange) compared to the exact profile (black dashed line).

S3. MD simulations for ferrous-ferric ET

All MD simulations in this work were performed with GROMACS version 2019.4 [3] and the Amber force field was used to model the systems [4]. The distance between the cations was fixed at 5.5 Å, the optimal separation of redox centers as determined in previous studies [5–7]. A cubic box with 567 water molecules and approximate size of 25 x 25 x 25 Å was used to solvate the system. The solvent was described with the TIP3P water model [8,9] and the LINCS constraint algorithm [10] was used for constraining bonded hydrogens. Minimization, equilibration and production steps were completed. The equilibration consisted of 500,000 steps using a step size of 1 fs. The production run consisted of 1,250,000 steps using a step size of 2 fs. Each frame of the production run was recorded and used in the analysis. The production step was completed in the constant-temperature, constant-volume ensemble (NVT). The temperature of 298 K was maintained with the Nose-Hoover thermostat. The Verlet cut-off scheme was employed to generate pair lists and the electrostatic interactions were evaluated with the Particle Mesh Ewald [11].

For the ET umbrella sampling calculations, charges were changed linearly in increments of 0.1 between reactants (Fe^{2+} + Fe^{3+}) and product (Fe^{3+} + Fe^{2+}), resulting in 11 independent simulations. The Van der Waals radius of the cations was also interpolated linearly. The potential energy was then re-evaluated for every window with every possible charge combination, resulting in 11 energy values for every MD frame, in total 1,250,000 frames for each umbrella window. The potential energy of each frame was re-evaluated using the rerun feature of mdrun.

S4. MD simulations for IET in $(Q\text{-}TTF\text{-}Q)^{-}$

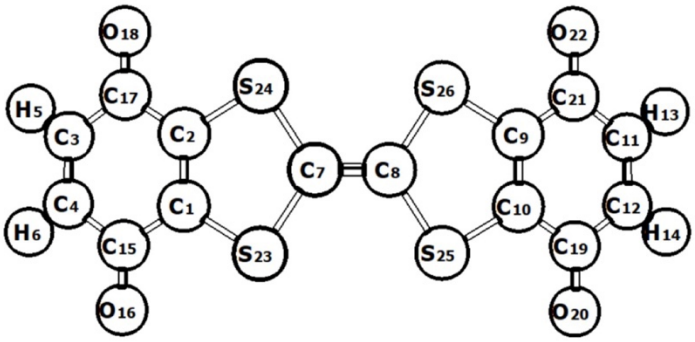
A cubic box with 1112 water molecules and approximate size of 30 x 30 x 30 Å was used to solvate the system. The solvent was described with the TIP3P water model [8,9] and the LINCS constraint algorithm [10] was used for constraining bonded hydrogens. The equilibration step size was 1 fs for a total of 2,000,000 steps. The production run was completed with a step size of 2 fs for a total of 1,000,000 steps. The Nose-Hoover temperature coupling was used (303.15 K) with the Parrinello-Rahman pressure coupling for the production step. The Verlet cut-off scheme was employed to generate pair lists and the electrostatic interactions were evaluated with the Particle Mesh Ewald [11].

At the TS, the atomic charges of the two sides of the $(Q\text{-}TTF\text{-}Q)^{-}$ anion are symmetric (see Table 1), therefore, the same atom types can be used for each side. However, that is not the case for the reactant state, or the intermediate windows. For the first and for the intermediate windows, additional atom types were created. The charge of each atom at each window can be found in Table S1.

For the organic solvents tBOH, ETA and DCM the same protocol was used but this time the simulations were run longer to ensure the bulkier polar solvents were fully equilibrated. The production run was completed with a step size of 1 fs for a total of 40,000,000 steps. Only the last 10 ns were used to re-evaluate the potential energy, ensuring the respective system was fully equilibrated at this point. 30 x 30 x 30 Å cubic box was used to solvate the systems with the respective number of particles corresponding to the experimental density of each solvent (781 kg/m^3 for tBOH, 1322 kg/m^3 for DCM and 902 kg/m^3 for ethyl acetate).

Table S1. CHELPG atomic charges of the $(Q\text{-}TTF\text{-}Q)^{-}$ anion in the 4 simulation windows. The atom numbering is shown in the picture insert.

ATOM	WINDOW 1 (MINIMUM)	WINDOW 2	WINDOW 3	WINDOW 4 (TS)
C_1	-0.091513	-0.092180667	-0.092848333	-0.093516
C_2	-0.091513	-0.092180667	-0.092848333	-0.093516
C_3	-0.212622	-0.222454333	-0.232286667	-0.242119
C_4	-0.212622	-0.222454333	-0.232286667	-0.242119
H_5	0.157216	0.152515333	0.147814667	0.143114
H_6	0.157216	0.152515333	0.147814667	0.143114
C_7	0.001152	0.008055333	0.014958667	0.021862
C_8	0.042544	0.03565	0.028756	0.021862
C_9	-0.076267	-0.082016667	-0.087766333	-0.093516
C_{10}	-0.076267	-0.082016667	-0.087766333	-0.093516
C_{11}	-0.277497	-0.265704333	-0.253911667	-0.242119
C_{12}	-0.277497	-0.265704333	-0.253911667	-0.242119
H_{13}	0.132185	0.135828	0.139471	0.143114
H_{14}	0.132185	0.135828	0.139471	0.143114
C_{15}	0.660902	0.639588	0.618274	0.59696
O_{16}	-0.538238	-0.563756	-0.589274	-0.614792
C_{17}	0.660902	0.639588	0.618274	0.59696
O_{18}	-0.538238	-0.563756	-0.589274	-0.614792
C_{19}	0.518638	0.544745333	0.570852667	0.59696
O_{20}	-0.687509	-0.66327	-0.639031	-0.614792
C_{21}	0.518638	0.544745333	0.570852667	0.59696
O_{22}	-0.687509	-0.66327	-0.639031	-0.614792
S_{23}	-0.023899	-0.032792	-0.041685	-0.050578
S_{24}	-0.023899	-0.032792	-0.041685	-0.050578
S_{25}	-0.083244	-0.072355333	-0.061466667	-0.050578
S_{26}	-0.083244	-0.072355333	-0.061466667	-0.050578



S5. Effects of changing the number of bins on free energy profiles

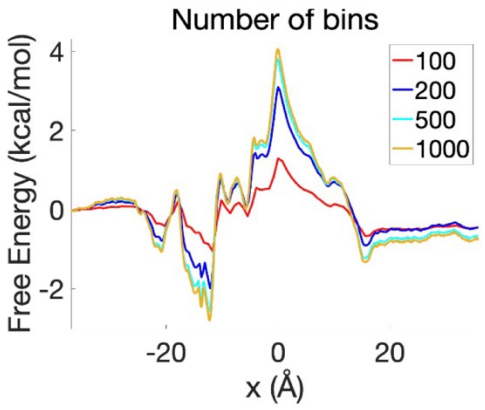


Figure S2. Binless DHAM free energy profiles with different number of bins for Na^+ passage through the GLIC ion channel.

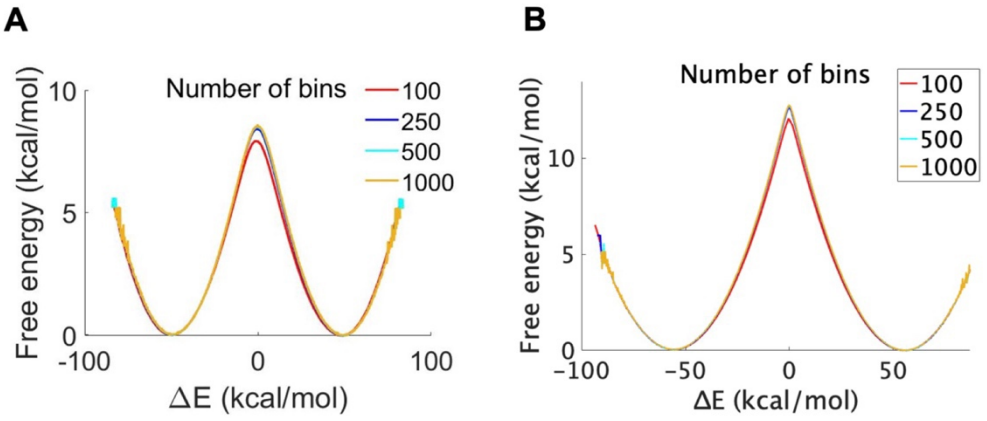


Figure S3. (A) Binless DHAM free energy profiles with different number of bins for ET in $(Q\text{-}TTF\text{-}Q)^{-}$, using 2 fs lag time and H_{ab} = 4.2 kcal/mol and (B) free energy profiles with different number of bins for ET in the ferrous-ferric system, using lag time 2 fs and H_{ab} = 0.2 kcal/mol.

S6. Average vs instantaneous bias values

Calculating the mean bias for all data points in bin *i* for each simulation window (according to Equation (S1)) instead of each data point (Equation (4) of manuscript) has a negligible effect on the results with lag time 2 fs. However, if the lag time is increased to 20 fs, the small number of observations of large energy gaps causes a noisy free energy profile, with the original formulation (Figure S4/A) giving numerically more stable results than the mean (Figure S4/B).

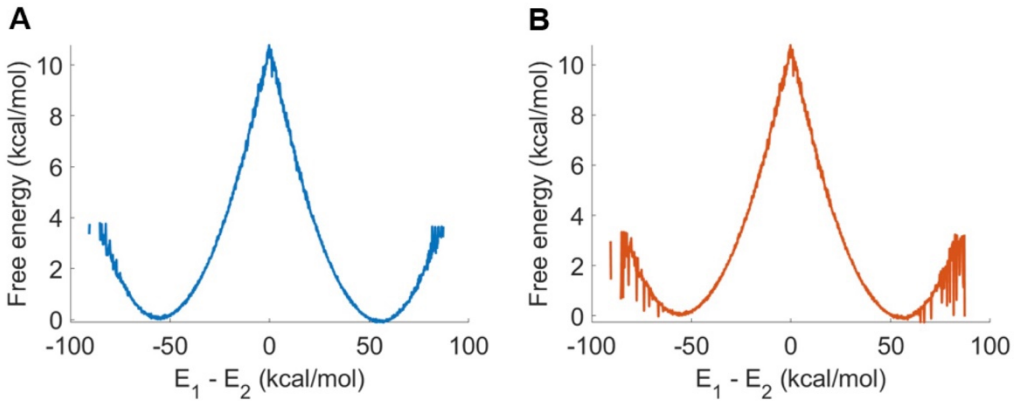


Figure S4. Binless DHAM profiles unbiased at the actual datapoints (blue) vs. the mean of all datapoints in the corresponding bin (red) for the ferrous-ferric ET. The lag time was increased to 20 fs to investigate the numerical performance of the approaches, while the number of bins was kept at 1000.

S7. Comparison of binless DHAM and MBAR profiles

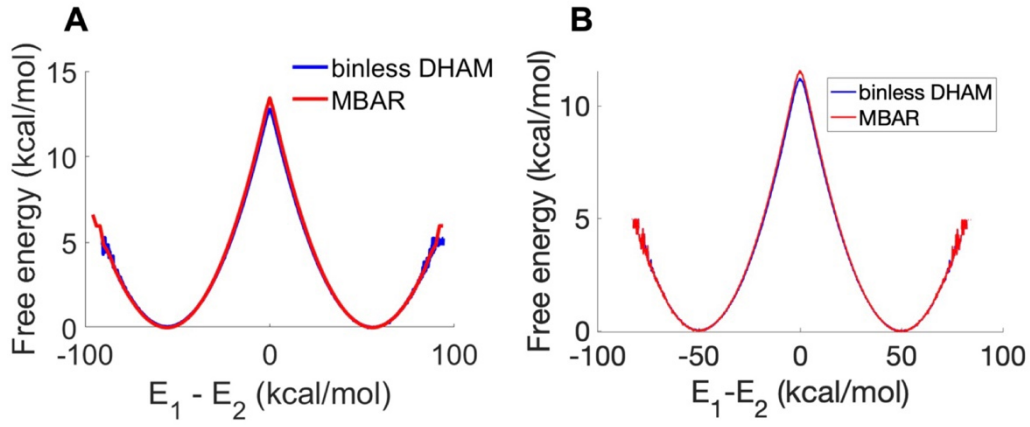


Figure S5. Binless DHAM (blue) and MBAR (red) free energy profiles for (A) ferrous-ferric ET and (B) ET in $(Q\text{-}TTF\text{-}Q)^{-}$ in water. For both reactions, 1000 bins and 2 fs lag time were used with binless DHAM, and 100 bins with MBAR. H_{ab} values of (A) 0.2 kcal/mol and (B) 0.97 kcal/mol have been used.

S8. Determining the reorganization energy from diabatic free energy profiles

To determine reorganization energy λ, quadratic functions are fitted to $G_{1,2}$ (Figure S6), and the free energy difference is taken between the reactant and product minimum structures for each curve. For ferrous-ferric ET (Figure S6/A), we get λ values of 53.0 and 53.2 kcal/mol, from state 1 and 2 respectively. For further calculations we use their average value, 53.1 kcal/mol. For IET in $(Q\text{-}TTF\text{-}Q)^{-}$ (Figure S6/B) we get λ = 48.4 kcal/mol from both curves.

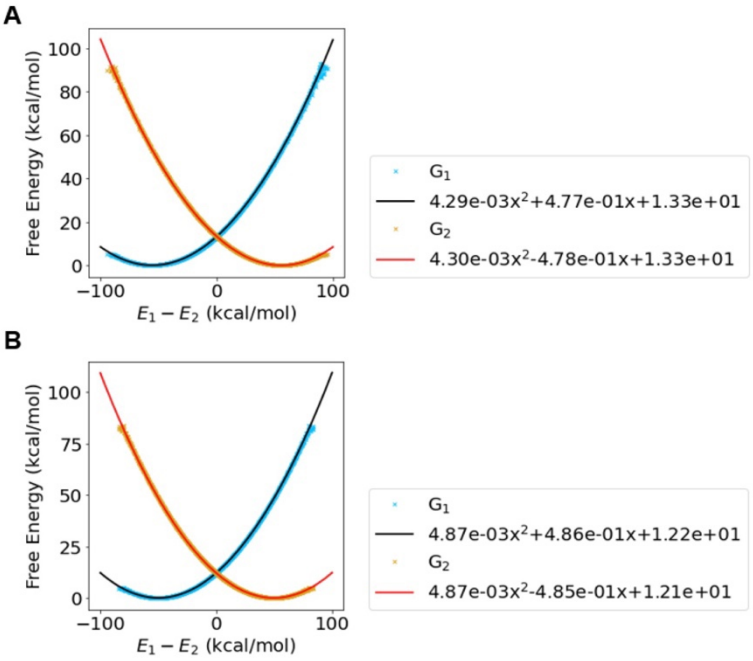


Figure S6. Binless DHAM free energy profiles for (A) ferrous-ferric ET and (B) IET in (Q-TTF-Q)$^{-}$ depicting diabatic states 1 (blue) and 2 (orange) as well as quadratic fits (black and red) on data between bins 40 and 960.

References

(1) Stelzl, L. S.; Kells, A.; Rosta, E.; Hummer, G. Dynamic Histogram Analysis To Determine Free Energies and Rates from Biased Simulations. *J Chem Theory Comput* **2017**, *13* (12). https://doi.org/10.1021/acs.jctc.7b00373.

(2) Rosta, E.; Hummer, G. Free Energies from Dynamic Weighted Histogram Analysis Using Unbiased Markov State Model. *J Chem Theory Comput* **2015**, *11* (1), 276–285. https://doi.org/10.1021/ct500719p.

(3) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, Flexible, and Free. *Journal of Computational Chemistry*. **2005**. https://doi.org/10.1002/jcc.20291.

(4) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J Comput Chem* **2004**, *25* (9). https://doi.org/10.1002/jcc.20035.

(5) Sit, P. H. L.; Cococcioni, M.; Marzari, N. Realistic Quantitative Descriptions of Electron Transfer Reactions: Diabatic Free-Energy Surfaces from First-Principles Molecular Dynamics. *Phys Rev Lett* **2006**, *97* (2). https://doi.org/10.1103/PhysRevLett.97.028303.

(6) Logan, J.; Newton, M. D. Ab Initio Study of Electronic Coupling in the Aqueous Fe 2+-Fe3+ Electron Exchange Process. *J Chem Phys* **1983**, *78* (6), 4086–4091. https://doi.org/10.1063/1.445136.

(7) Kuharski, R. A.; Bader, J. S.; Chandler, D.; Sprik, M.; Klein, M. L.; Impey, R. W. Molecular Model for Aqueous Ferrous-Ferric Electron Transfer. *J Chem Phys* **1988**, *89* (5), 3248–3257. https://doi.org/10.1063/1.454929.