Pázmány Péter Catholic University
Roska Tamás Doctoral School of Sciences
and Technology

Computer modeling of postsynaptic protein complexes

Theses of PhD Dissertation

MISKI Marcell

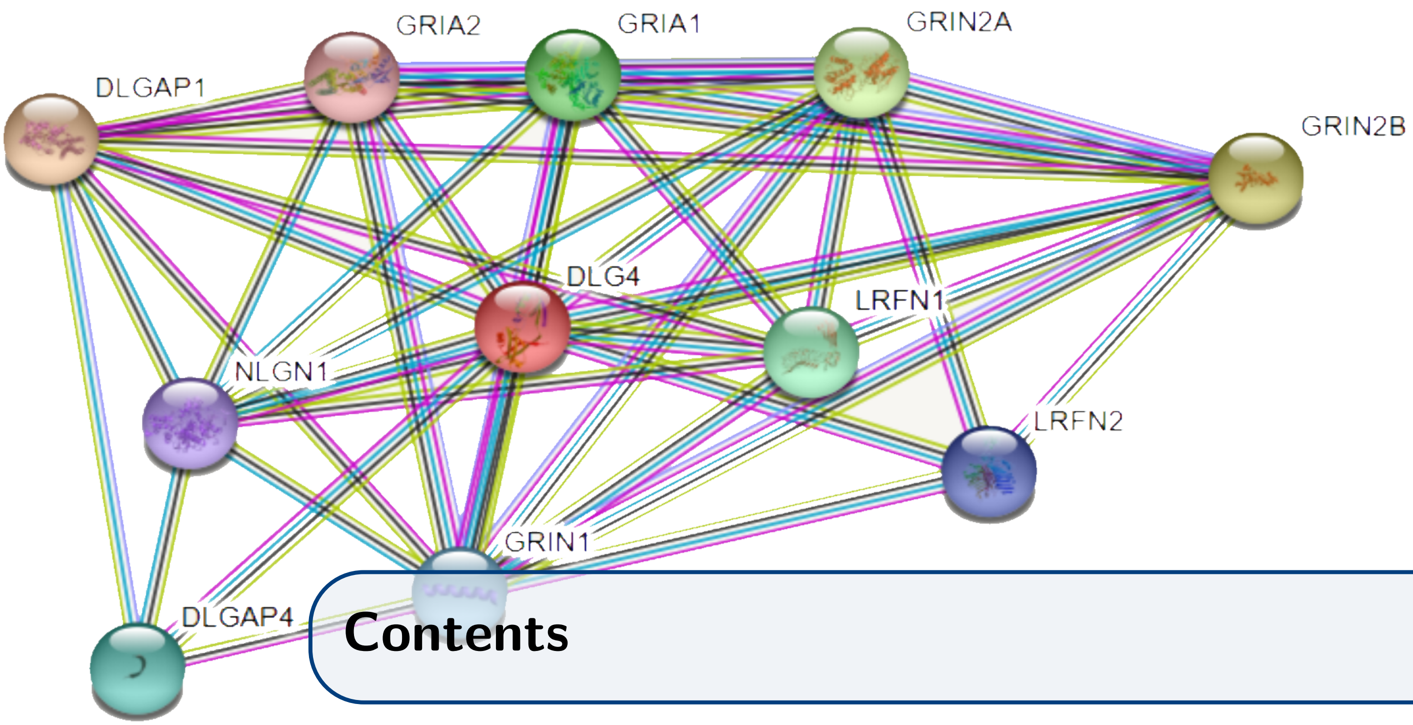
Supervisor: Prof. Dr. Csikász-Nagy Attila DSc

2024

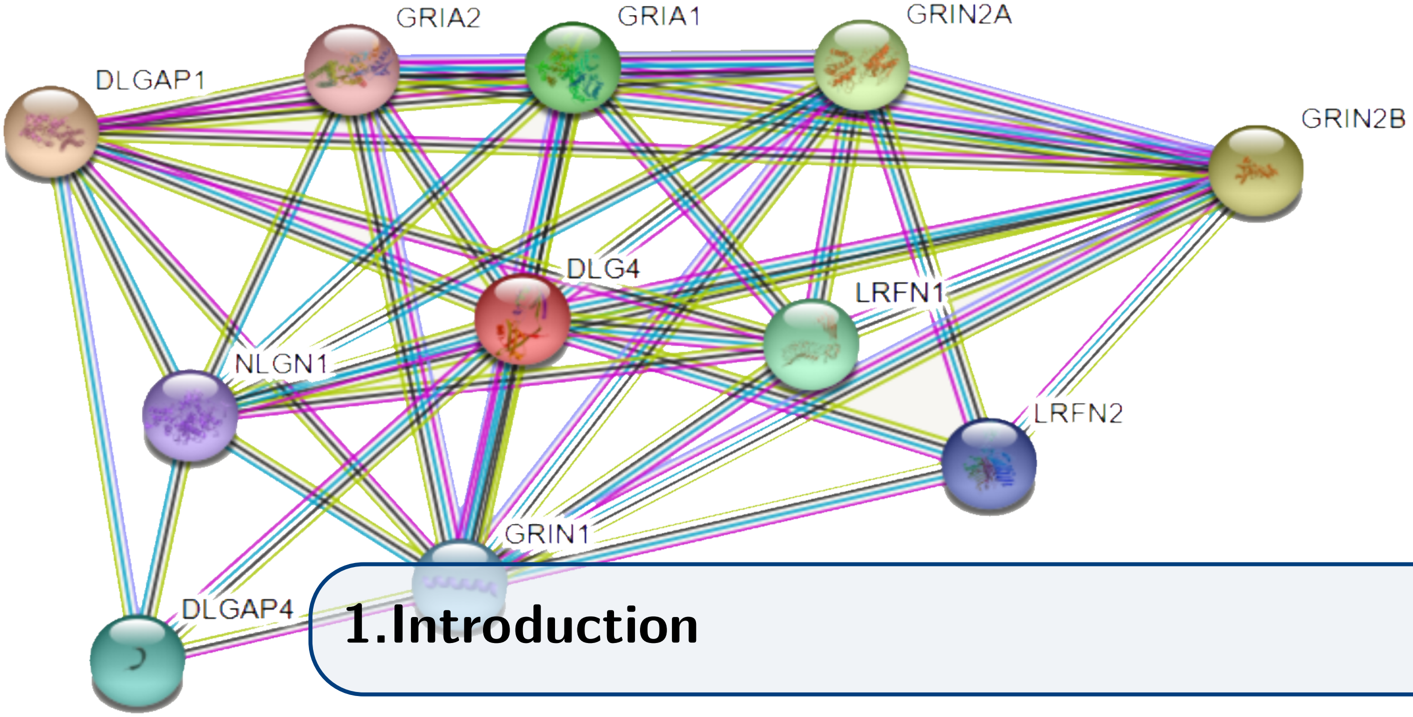
Pázmány Péter Katolikus Egyetem,
Roska Tamás Műszaki és Természettudományi Doktori Iskola

A *PhD* fokozat megszerzésére benyújtott dolgozat

2024. július 25.


GRIA2
GRIA1
GRIN2A
DLGAP1
GRIN2B
DLG4
LRFN1
NLGN1
LRFN2
GRIN1
DLGAP4

Contents



1.Introduction

1.1 The postsynaptic density

The postsynaptic density (PSD) is the protein-rich part of the dendrites of the postsynaptic neuron close to its membrane [1]. It appears as a dark ("electron dense") band on electron microscopic images, hence its name. Among the proteins of the PSD are many receptors and other scaffolding proteins that are crucial in the process of electrochemical signaling [2]. The importance of this is also indicated by the correlation that in the case of long-term activity (long term potentiation (LTP)) the size of the postsynaptic density increases and becomes stronger [3]. The reactive reorganization of the protein network is considered an important process of learning and memory [4]. Mutations in PSD proteins can be associated with various neurological diseases, one of the most important of which is ASD (Autism Spectrum Disorder). Understanding the organization of the PSD is currently a serious challenge, as we currently cannot obtain a high-resolution picture of it with experimental methods precisely because of its diversity and dynamic nature. Therefore, the use of modeling procedures is extremely important in this area.

1.2 The synaptome theory

The synaptome theory was developed by Seth Grant and his research group [5]. This theory modifies and expands the idea that, for the time being, many biology textbooks hold that different brain functions are caused by different wiring of neurons. Nevertheless, wiring, neuron-neuron connections do not provide a complete answer to genetic neurodegenerative disorders, such as Alzheimer's and the Autism spectrum. According to the central dogma, genes directly encode proteins [6], not the "wires" themselves, neuron extensions, so according to the theory, proteins must be the missing link between wires, coded functions and genes [5]. So, according to the synaptome theory, the specific neuron type is determined by the expressed proteins. Thus, proteins indirectly influence brain functions and can change behavior. I set the goal of my PhD Dissertation to get to know and examine this theory in more detail with the help of computer simulations and models because the interesting thing about the theory is that it also covers the learning process itself [7]: that learning occurs by modifying the synapse proteome [8] does not require long-term potentiation (LTP) of synaptic weight or the growth of new synapses, and theory predicts that LTP modulates recall of information [7] . From the point of view of later three-dimensional modeling, what is interesting, as Seth Grant also notes [5]: the spatial architecture of the synaptome, protein networks, originates from the underlying molecular hierarchy that links it into complexes and supercomplexes through the supramolecular assembly of the proteins encoded by the genome. This molecular hierarchy explains how evolution at the genome level results in changes in an organism's behavioral repertoire. It's this disruption mutations alter the architecture of synapse protein networks, potentially accounting for behavioral phenotypes associated with neurological and psychiatric disorders [9], [10].

1.3 Agent-based simulation of protein complex formation

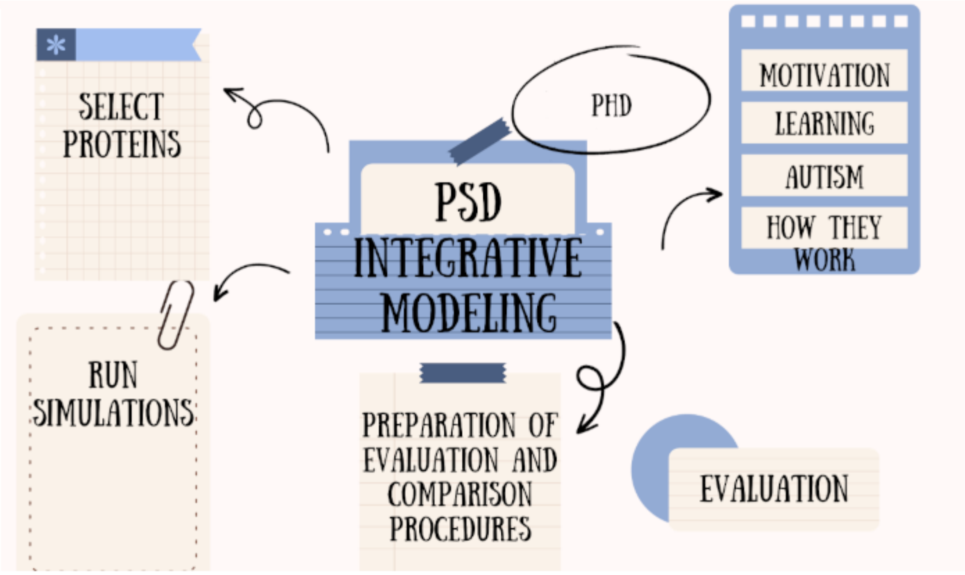
Many modeling and simulation procedures can be used to understand the processes in the neuron, from the logical description of the network of signal-transmitting processes to the modeling of the neuron's activation patterns [11]. My research examines changes at the level of protein complexes.

The formation of protein complexes can be described in the same way as a chemical reaction: two molecules move randomly and then meet. During this, if the conditions are right, they react with each other and a new molecule is created. The essence of the simulation method is that the molecules are moved on an agent basis, and if they get close enough, it is considered an encounter event. During the encounter, we create a connection between them with a given probability (the two separate proteins become a complex - two agents transform into one agent). This given probability is biologically described by binding strengths and affinities. At the same time, the complex may disintegrate in a subsequent step with a probability determined by the dissociation constant.

1.4 Modeling the 3D structure of protein complexes

The biggest disadvantage of agent-based modeling is that it may require a lot of memory, so it is not realistic that the actual three-dimensional structure of proteins or we can represent the resulting constraints in such a simulation. A protein, a complex agent appears as a given point - even if its binding sites existed virtually. At the same time, there are also software that can be used to model the structure and shape of protein complexes. In the course of my work, I consider it an important aspect to be able to view the results of quantitative agent-based simulations and the results of qualitative 3D modeling together and enable their integration in a joint process later on.

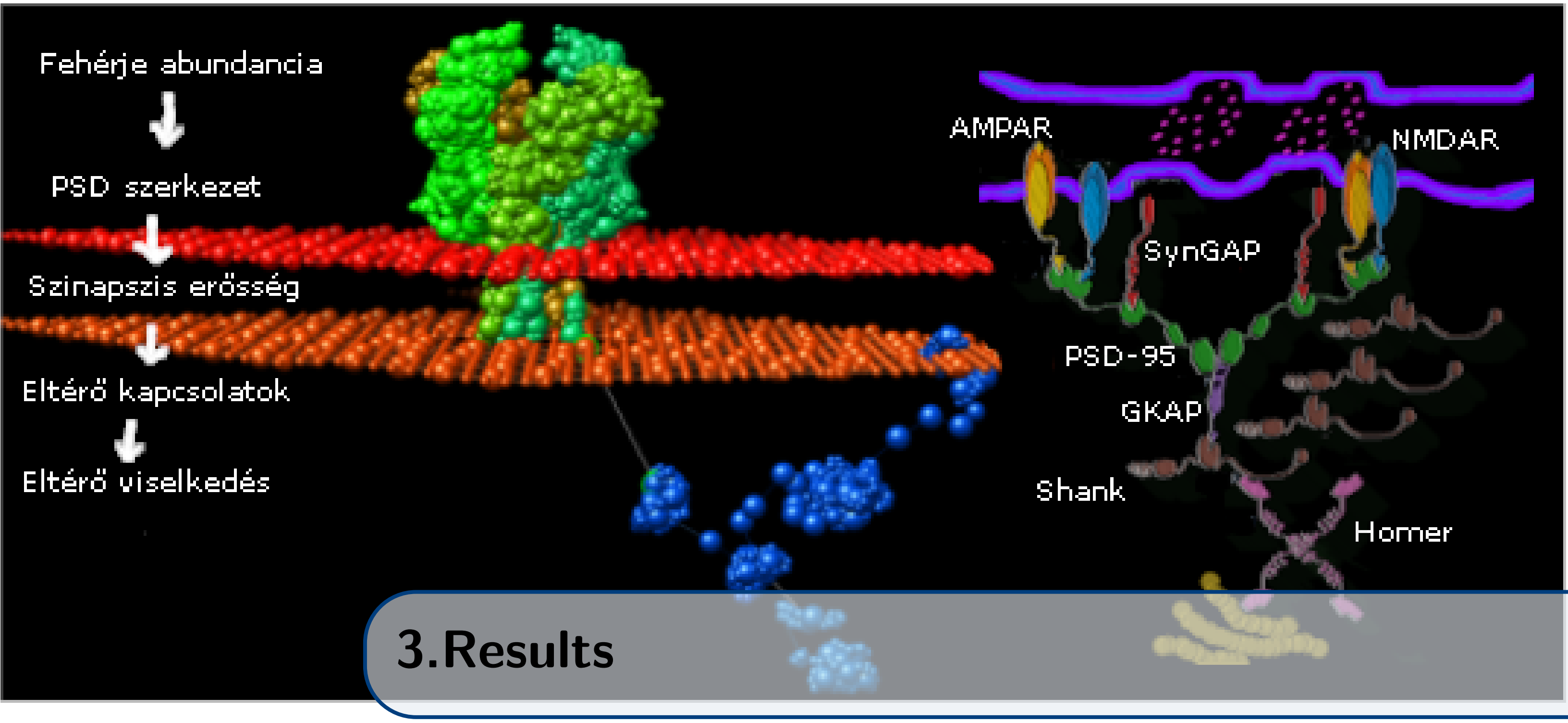
As the name of the IMP (Integrative Modeling Platform) program implies, it was created for the purpose of integrative modeling, which means that it can take into account input data from multiple sources and different types of representations of certain parts of a given system in an integrated way during the modeling process. The platform enables the creation of three-dimensional structural models of large-scale complexes using a large number of data together.



2. Goals

The main goal of my research was to analyze and apply the basic procedures necessary for the integrated modeling of postsynaptic density. According to my plans, with the help of integrated modeling, in addition to the probabilities of the formation of individual protein complexes, we can also create structural models, taking into account the information obtained from simplified simulations. To achieve this, I aimed to implement the following steps:

- Development of a simplified, manageable size model of the postsynaptic density built up by the complex interaction of many proteins, with the help of which agent-based simulations can be performed. This requires the identification and representation of the most common proteins, their number of copies and the interactions that can be formed with each other in such a way that it can serve as input for the planned simulations.
- Development of the necessary procedures for the evaluation of agent-based simulations performed with the help of the model, including the clear identification of the resulting protein complexes and the comparison of simulations started with different input parameters.
- Completing a large number of simulations using many different input parameters using literature protein frequency data, then evaluating and interpreting the simulations.
- Modeling the effect of an important disease causing interaction-weakening mutation using the framework.
- In parallel with all of this, the development and testing of a workflow for the three-dimensional modeling of large, multi-component postsynaptic protein complexes.



3.Results

3.0.1 Preparation of the model system

During literature research, I collected experimental data for protein frequencies broken down into specific brain regions of given subjects [12]. These seven proteins are NMDAR, AMPAR, SynGAP, PSD-95, GKAP, Shank3, Homer1 already presented in the introduction. However, I did not have an adequate amount of direct data on the frequency of these proteins, so I had to convert the found mRNA expression levels into numbers.

In the literature, RPKM (reads per kilobase million.) from RNA-Seq experiments I found data containing values in the publication [12].

From these values, I generated number data by comparing them to the known number of PSD-95. I have prepared the data on the number of proteins as follows:

In the beginning, I worked with Shank3, then I examined Shank1 for the mutation analysis. With regard to the simulations performed with the Cytocast method, the relative position of the domains does not mean a change in the domain structure of the protein - therefore, the two proteins can be easily replaced. The estimated frequency of each protein varies greatly in each data set. Average amounts were:

The model is suitable for creating a primary estimate. When approximating the amount

protein	min	max	average	distribution
NMDAR	0	95	16.84	17.87
AMPAR	1	688	126.81	78.39
PSD-95	36	1067	328.09	159.21
SynGAP	11	359	1102.21	60.13
GKAP	2	367	82.62	65.50
Shank1	1	388	69.20	59.40
Homer1	1	124	21.98	17.20

Table 3.1: Mean and standard deviation of protein frequency data used as input

of protein with mRNA expression, it must be taken into account that differences between regions can already be observed at the level of mRNA synthesis, however, other differences and important groupings can also develop during the translation of proteins.

However, in addition to the protein frequency, the probability that two proteins encounter each other also depends on the simulation grid size and the simulation time. In the case of most complexes, I found that 40 repetitions proved to be enough to smooth out the stochastic fluctuations to a usable extent and narrow the radius of the confidence interval to 1 complex.

Calculating the effects of mutations with binding strength prediction

In the course of my work, I also wanted to investigate the effect of a given mutation, but experimental binding data for a protein (or protein domain) containing a given mutation is very rarely available, moreover, the Gillespie algorithm handles the association and dissociation rates separately, for which there is hardly any data for wild-type proteins as well our data. The direct effect of a mutation on partner binding is not easy to estimate, but several methods are available that estimate changes in the stability of a given protein, so I tried to modify the binding constants in the simulation system based on this.

A point mutation can change the stability of the protein and therefore modify the accessibility of the protein. I calculated the stability change ($\Delta\Delta G$) of biologically relevant point mutations using NeEMO [13].

The bond dissociation rate ($K = \frac{[denaturated]}{[naturated]}$) can be derived from $\Delta\Delta G$. The rate of change shows how the dissociation rate is changed by the mutation according to the predicted $\Delta\Delta G$ values.

$$\Delta\Delta G = -RT\ln(K_{\mathrm{WT}}) + RT\ln(K_{\mathrm{MT}}) \tag{3.1}$$

$$\text{rate of change} = \frac{K_{\mathrm{MT}}}{K_{\mathrm{WT}}} = e^{\frac{\Delta\Delta G}{RT}} \tag{3.2}$$

In my model, the probability of complex formation is inversely proportional to the change in the dissociation rate (K). To achieve this, I modified the dissociation constant in the parameter set.

Selection and modeling of the Shank1 R743H mutation

For our specific modeling purpose, I chose the R743H mutation in Shank1. Although I selected from the COSMIC[14] database, which contains cancer-related mutations, my intention was to use a well-described mutation in a globular domain that could affect a structurally well-characterized interaction in our model. In our case, this is the connection between the PDZ domain of Shank1 and the C-terminal region of GKAP. The effect of the mutation on complex formation was estimated from the destabilization of the PDZ domain as described above with the NeEMO [13] procedure According to this, the mutation causes a 5.5-fold increase in the dissociation rate of the Shank1:GKAP complex.

3.0.2 Procedures developed for evaluation of agent-based simulations

The distribution of data is not uniform in that there are brain regions for which data are not available from all patients. For this reason, each experiment must be treated separately, and I have sought to identify the main differences between the results.

Homomultimeric protein associates I created a separate algorithm for its identification. This was necessary in order to obtain a comprehensive picture of how the given proteins capable of multimerization, and how the complexes with up to 100 proteins were glued together, even in the case of larger complexes.

When finding protein associates consisting of the same protein, it is first necessary to determine which protein can multimerize in a chain-like manner. In my test system, e.g. the Shank3 protein has such a property due to its characteristic association with the SAM domain described in the literature [15].

These proteins can be searched for in the protein complexes based on their unique identifier (ID) and the interactions in which both interacting proteins correspond to the given protein can be identified. Chains built from proteins as units can then be identified.

Determining a metric indicating the added information content of the simulations

One of the goals of the tests I carried out was to examine how the distribution of the resulting protein complexes – as the output of the simulations – depends on the available protein frequencies – which appear as input in the simulations. One of the general aspects of the question is whether the output follows trivially from the inputs, or whether the relationship between the two is so complicated that without simulation it is not possible to estimate the output with great certainty for a given input. To decide this, I examined whether the input and output datasets give the same pattern for each region (input dataset).

The K-means clustering [16] procedure creates groups from similar data. In vector spaces, this property means that closer points are placed in the same cluster. I then created a labeled vector for each cluster. Each element of the 524 long vector represents a studied region, and the coordinate is 0 if the data is not in the cluster and one if the data is in the cluster (so-called one-hot representation). We use it to determine how well each cluster displays the same grouping of each region in the case of two different clusterings, for example based on input and output data.

If the cross-spacing of the input and output clusters is large enough, the aggregated simulation may contain additional information that is not visible in the inputs themselves. The procedure can be used both using a smaller six-dimensional output complex vector space defined on the basis of interaction abundance, and a multidimensional output complex vector space defined on the basis of the abundance of individual complexes.

Identification of the complex with the highest information content

During the evaluation of the data, the need arose to determine the complex whose frequency carries the most information, i.e. the one that contributes the most to the differences between the obtained protein complex distributions. To determine this, I defined a metric based on principal component analysis.

I represented the brain regions as points in a multidimensional space, where each coordinate represents the frequency of a specific protein complex. The protein complex distribution of a given brain region is defined as a linear combination as follows:

$$c \in \mathrm{R}^n, \quad c = \sum_{i=1}^{n} \alpha_i p_i \tag{3.3}$$

where α_i is i. frequency (number of copies) of complex (p_i).

Each complex has a characteristic frequency profile in different regions: the complexes with the most varied abundance show the most striking differences between regions. Thus, the most variable complexes determine the main components of the system.

The variance projected onto each complex is calculated by multiplying the variance represented by each principal component vector by the contribution of the respective complex as a base vector to the given principal component, and then summing them up for each principal component. The obtained relevance indicates how decisive the frequency of the given complex is in determining the differences between the individual regions.

$$r \in \mathrm{R}^n, \quad r = \sum_{i=1}^{n} \frac{\lambda_i}{\sum_{j=1}^{n} |v_{ij}|} v_i \tag{3.4}$$

Where r is an n dimensional vector containing the relevance of all complexes, λ_i is i. relevance of eigenvector, v_{ij} to j. coordinate is i. for eigenvector (v_i).

Comparison of protein complex frequencies

The paired T-test is a classic statistical test that is usually chosen when there is only one change - in my case, e.g. the examined Shank1 R743H mutation - may be present in the system, and the question is how the change affected the average value of the given variable. The null hypothesis: the average abundances (wild type and mutant) are the same at a given significance level. Thus, the alternative hypothesis: the average abundances of a complex differ significantly. I performed the T-test for a specific complex as described in the methods.

Given that the abundance of each complex is averaged separately during the simulation, the T-test, which hypothesis can be accepted, only applies to the abundance of the given complex. In order to be able to say whether a given region has changed in general, we have to perform the T-test on all complexes, and their weighted average indicates which hypothesis we can accept on average for the given region. During averaging, we use the known importance of each complex, which is calculated based on the relevance of each complex. Feature importance is the weight of averaging the p-values calculated with the T-test for all complexes in a given region.

$$\text{average p-value of a region} = \sum_{i=0}^{39} \text{complex\_relevance}_i \cdot \text{p-value}_i \tag{3.5}$$

3.0.3 Analysis and raison d'être of simulations

The difference between the input and output distances of a given pair of simulations showed exactly which brain regions, i.e. the relative position of the simulations, changed compared to the input positions (Fig. 3.1). **Thus, I was able to highlight the regions in which the**

most added information could appear. These regions are: H376.VIII.51_STC id:340, H376.VI.50_V1C id:286, H376.IX.51_MFC id:238. The ID shows how many times I performed the simulations , has a role in some figures. The code before the underline represents the subject from which the brain slice is derived, while the three-character code after the underline refers to the specific brain region. STC: posterior (caudal) superior temporal cortex (area 22c), V1C: primary visual cortex (striate cortex, area V1/17), MFC: anterior (rostral) cingulate (medial prefrontal) cortex.

I compared the ratios of the abundance of the input proteins with the ratios of the output complexes, or with the relative frequency of the pairwise protein-protein interactions found in them.

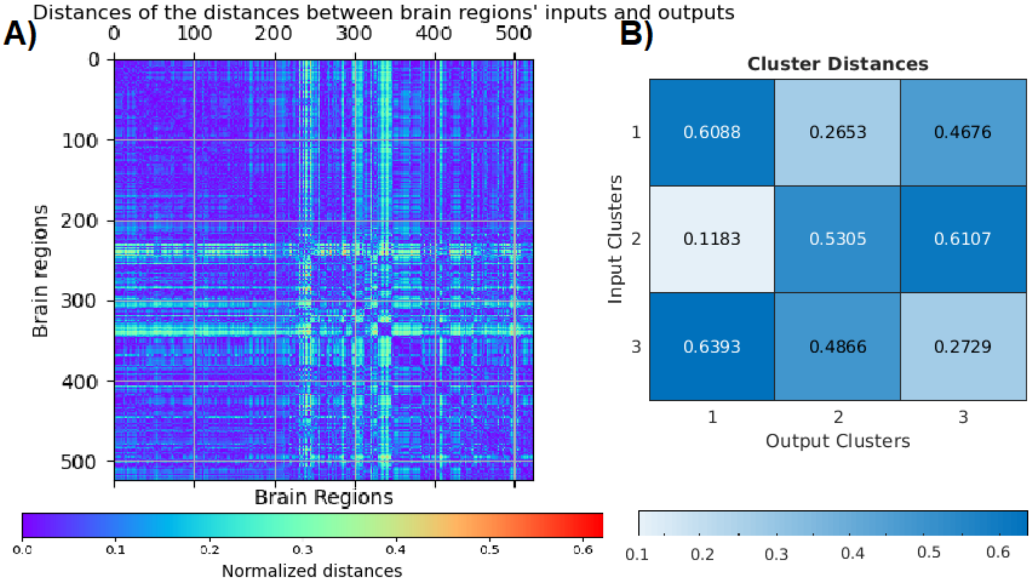


Figure 3.1: **The relative displacement of the regions in relation to each other on the heat map and the result of the clustering.** I placed the individual regions in a given vector space based on both input and output. Thus, the relative position of the regions shows how similar two regions are to each other, in the case of input in terms of the amount of proteins, in the case of output in terms of the amount of specific complexes. By displacement we mean that two regions have moved closer or further away from each other in the vector space created on the basis of the output compared to their input position. In many cases, we can hardly see any change, because most parts are blue, however, there are pixels that fall into red. During clustering, it is not clear which cluster should be matched with which, so I matched the closest clusters to each other and looked for the change ratio between them.

The performed simulations showed that the frequency of the emerging complexes does not only depend on the initial protein amount, but also has a significant effect on possible bonds. The frequency of a complex containing a certain protein depends more on the frequency of the protein's neighbors than on the frequency of the given protein itself.

The presented examples of regions with a similar set of proteins, but containing complexes in a different distribution, suggest that the local synthesis and degradation of selected proteins can lead to redistribution of protein complexes to a degree that can significantly change the "identity" of the synapse. The presence of local mRNA translation in dendritic spines, which produce PSD proteins among others, is well known and has been linked to several neuronal processes such as late-phase LTP [17]. Similar complex distributions can be achieved with different combinations of the constituent proteins, which provides the opportunity to reach functionally similar states.

The AMPAR/PSD-95/SYNGAP complex is the most informative

Principal component analysis I identified which complexes are the most informative in terms of distinguishing brain regions. The first two main components are both wild type and mutant case, it covers 44% and 24% of the total variance of the outputs, respectively. The first main axis is dominated by the AMPAR/PSD-95 (id:12) complex, while the second is dominated by the PSD-95/SynGAP (id:8) complex.

Overall, the most informative complex - with the largest contribution in terms of all principal components and the variance explained by them - is AMPAR/PSD-95/SynGAP (id:5) with 19% contribution. In comparison, the average complex significance is very small, approximately 4.48e-06, and the median is 9.98e-08. The reason for the low rate is the large number of possible complexes - so the probability rate inherently shares a large set of elementary events. The fact that the AMPAR/PSD-95/SynGAP complex significantly influences the regions is consistent with the information found in the literature, SynGAP binding regulates the frequency of AMPAR binding [18].

3.0.4 Testing the effect of hypomorphic mutation

The specific effect of mutations can be enigmatic, especially when they affect proteins present in many different tissues. This is especially true for cell types in which even the major partners and interactions are the same. The diversity of neurons in terms of different amounts of postsynaptic proteins offers a unique opportunity to explore the effect of specific mutations in a complex, yet simplified multicomponent system with the same set of building blocks.

In the case of the simulations run to analyze the hypomorphic Shank1 mutation, we run simulations using the same initial protein frequencies for both the wild-type and mutant cases. In the case of the mutant, I assumed a homozygous state, so either only wild-type or only mutant protein was present in the system.

The principal component analysis of the simulation results of the wild-type and mutant scenarios shows a very similar overall picture. The two PCA plots can be directly compared since the axes are the same even in the two independent PCA outputs. The data points corresponding to the wild-type and mutant cases differ only minimally from each other. The average distance between wild-type and mutant regions is 1.5 ±0.8 units.

I note that PCA usually does not separate the given brain regions (fig 3.2), however, in the case of these simulations, the cerebellar cortex-type regions are well separated from the others.

My results indicate that the total protein complex distribution is primarily determined by the availability of individual proteins, and the presence of the examined hypomorphic mutation does not cause significant global effects. This is consistent with the system retaining its overall functionality. In order to analyze the effect of the mutation in more detail, we examined the abundance of each protein complex.

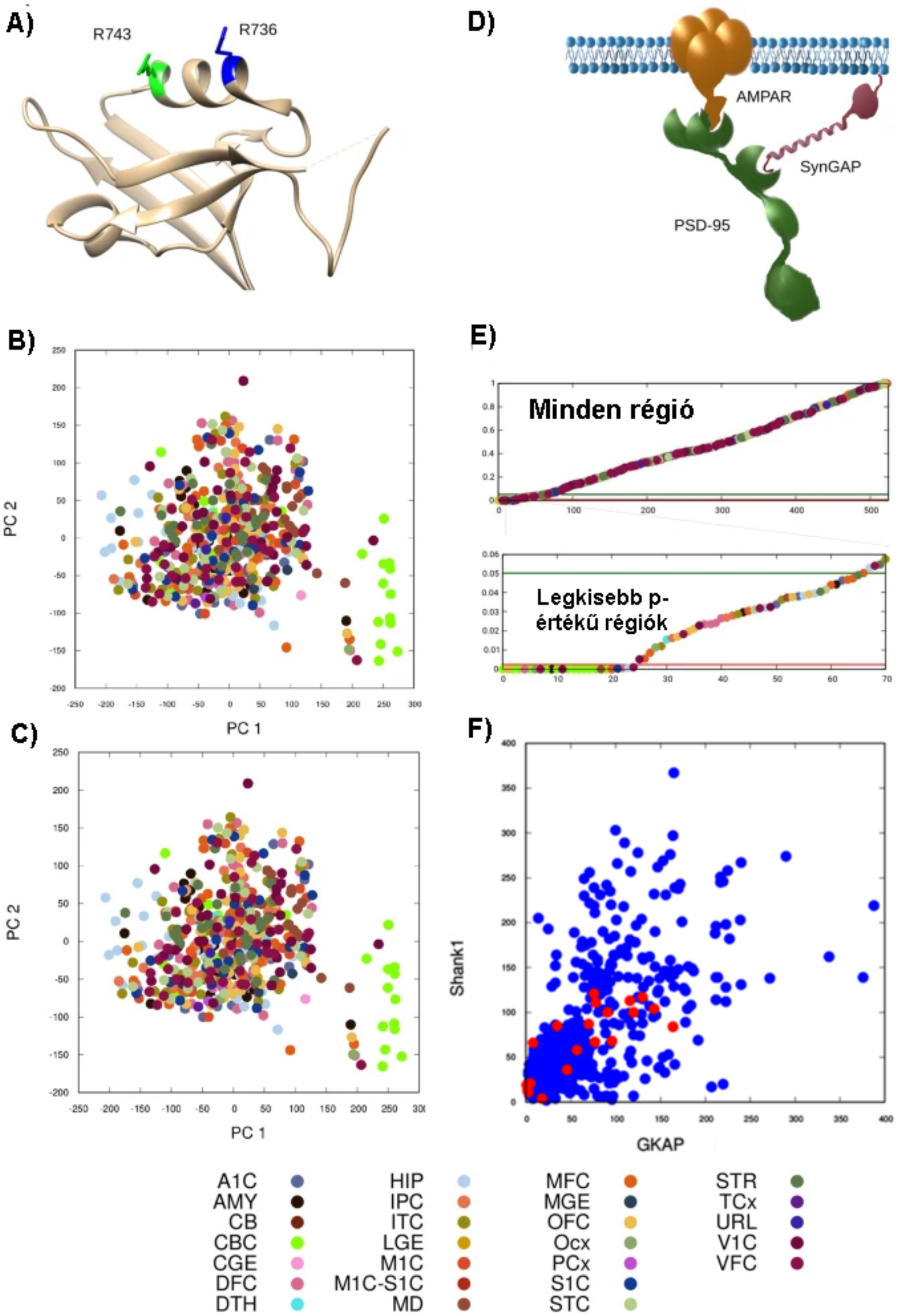


Figure 3.2: **Effect of mutation on the simulation: A)** The position of the selected mutation (R743H, green) and a similar mutation reported in ASD (R736Q, blue) on the ribbon representation of the Shank1 PDZ domain (PDB ID 6YWZ) . Both arginines are located on helix $\alpha2$ flanking the ligand-binding groove. Principal component analysis of the obtained protein complex distributions for the **B)** wild-type and **C)** mutant scenarios. Different colors represent different brain regions according to the key at the bottom. **D)** Schematic representation of the most informative complex according to PCA (AMPAR/PSD-95/SynGAP). **E)** *p*-values describing the change in mutationsorted for the most informative complex. Lines indicate the 0.05 and the corrected 0.01 significance levels. Colors by brain region according to the key at the bottom. **F)** Abundance of the two proteins affected by the mutation, Shank1 and GKAP, in the input data sets. Red circles indicate data sets where the abundance of the most informative complex changed significantly in the output. The figure was published in my article [19].

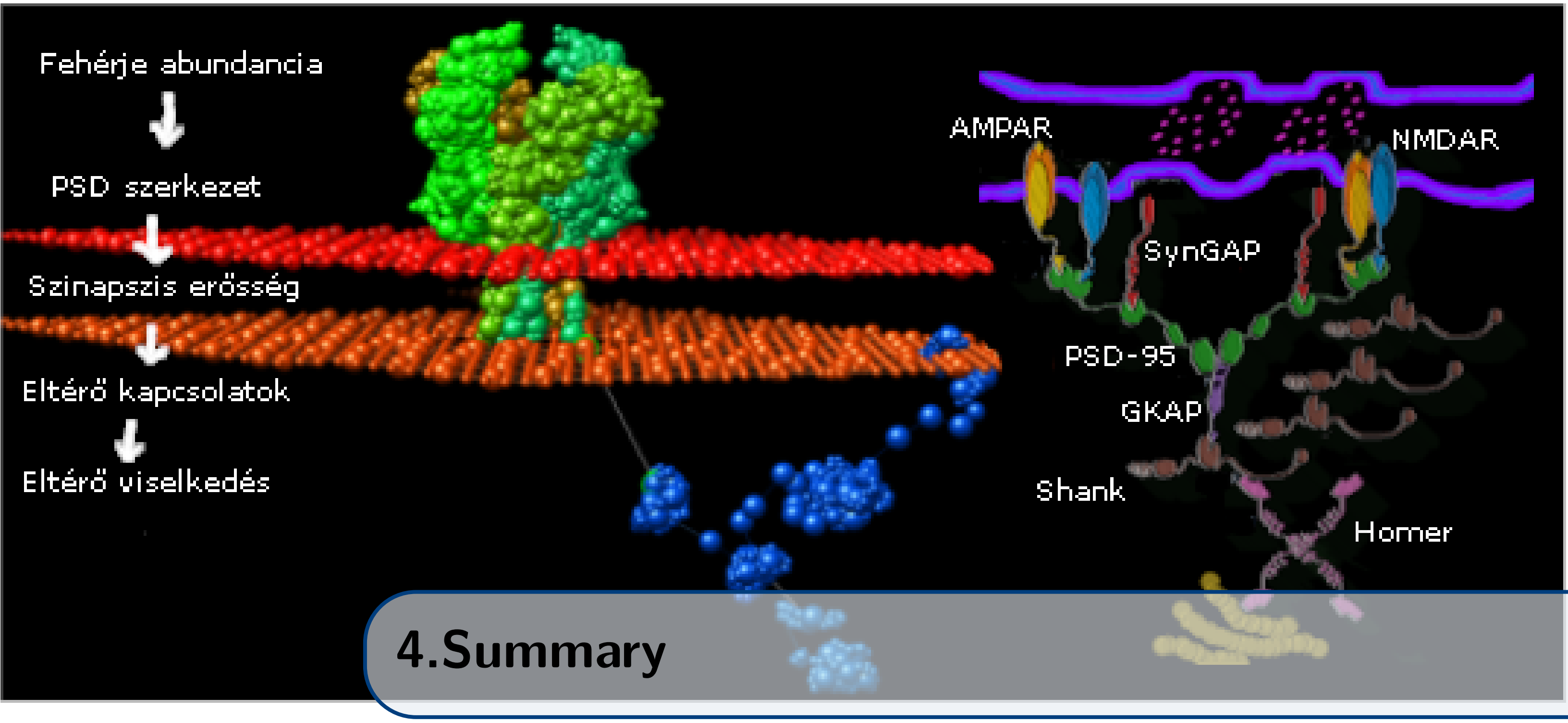
3.0.5 Three-dimensional examination of complexes

The procedure consists of two parts running in parallel, which can later be integrated into a common model - in this way, the process also connects the work of the Faculty's two research groups. One direction is the large-scale simulation of the formation of protein complexes with the Gillespie algorithm, the other direction of the procedure is based on spatial structures and sequence annotations and takes into account the three-dimensional nature of individual proteins as components. In this case, we create a string of beads model from the proteins, thereby viewing it as an object with a real extension. I have come to the conclusion that combining the two directions is not trivial, but can be combined with knowledge from each other. I showed that the space requirements of nanoclusters can be estimated and that the treatment of disordered/unknown protein regions is a complex process that greatly affects the model.

Some parts of the procedure outlined above are, of course, previously described and currently applied steps, however, as far as I know, the connection of the entire process and the two modeling approaches has not yet been described in this form.

Overall, the innovation value is the combination of the two methods and the joint use of the results of the two research groups.

The first protein for which I started assembling the full 3D structural model was PSD-95. During assembly, I found that it is still difficult to visualize the terminal sections of the protein in the absence of spatial structure, because the rudimentary structural models created based on the protein sequences were not reliable, because the torsion angles differed significantly from the accepted ranges in many places. From the models built using the IMP process, I found that PSD-95 bridges two Kir2.1 molecules between 70-75Åthrough the PDZ tandem. Considering PDZ domains as cylinders, we get that the average length of PDZ domain A is 31Å and its radius is 7Å a **??**. shown in purple in the figure. This may be one of the limitations why distances higher than 35Åin this case, PSD-95 shuttles between the potassium channels [20].



4.Summary

4.1 Justification of my approach

Our model with only seven PSD proteins and no specific spatial organization is definitely a very simplified model far removed from the actual biological complexity of the postsynaptic. Furthermore, for simplicity, we consider in each case a situation that corresponds to a homozygous scenario, i.e. where only the wild-type or only the mutant version of a given protein is present, but both are absent at the same time. Last but not least, we modeled only one well-defined effect of the selected mutation, ignoring possible pleiotropic effects such as changes in the expression levels of other proteins, as e.g. observed in several Shank mutations [21]. Thus, it is not expected that the resulting protein complex distributions can be directly compared to *in vivo* situations. Modeling all these aspects with acceptable accuracy would require much more data than is currently available. However, I argue that our model system, focusing on a well-defined set of core PSD proteins and interactions, is complex enough to capture general aspects of the behavior of elaborate protein networks, while remaining manageable from the point of view of data analysis, since the number of possible protein complexes in the model is not yet extremely high. According to my considerations, the mechanistic connection of genotypes and phenotypes – here by phenotype we mean protein complex distributions that are functionally closer to determining the identity of synapses – where different genotypes lead to similar phenotypes, is only possible through a combination of experimental data and modeling approaches.

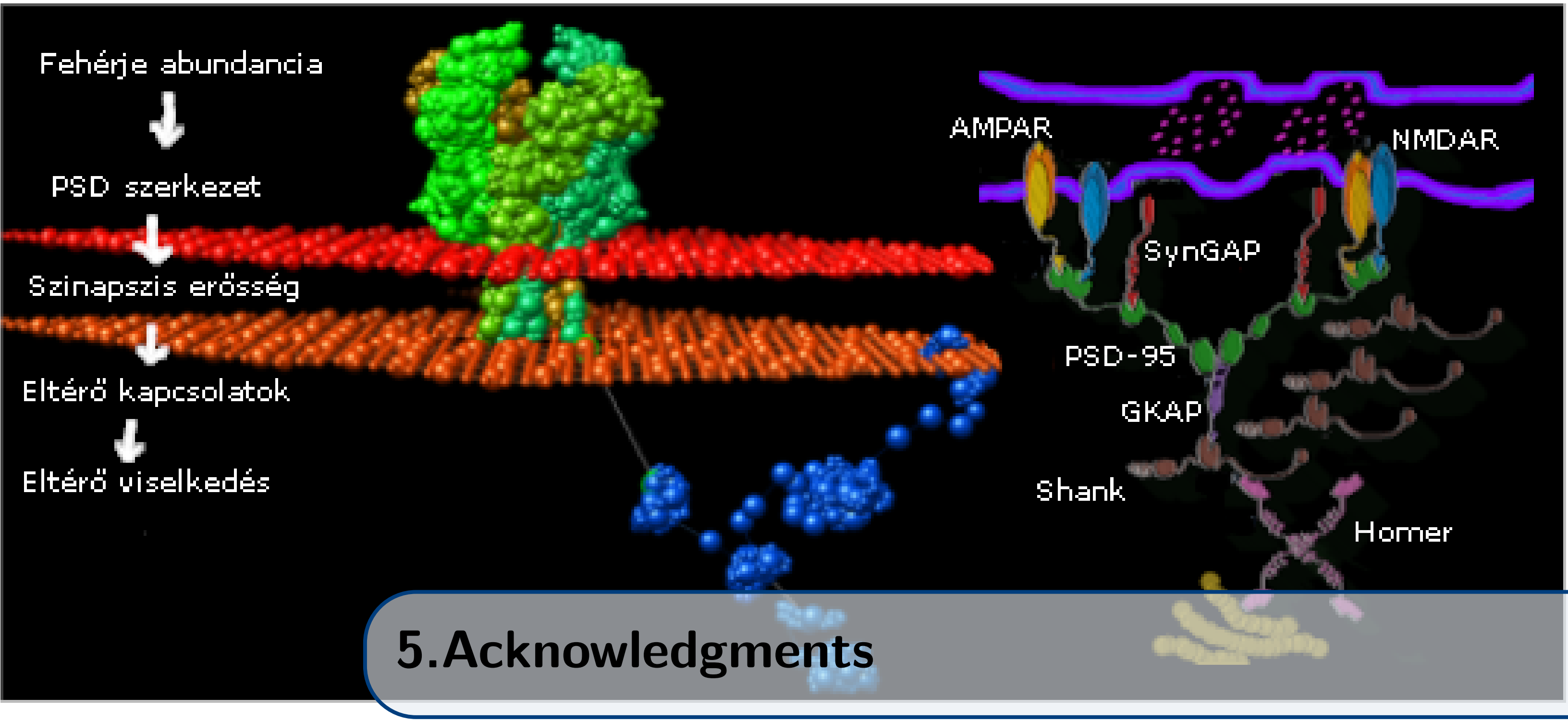
4.2 Models of PSD complexes - how far are we from reality?

Protein copy numbers estimated by linear scaling of mRNA expression data do not account for translational and posttranslational regulatory effects [22]. Such effects may result in greater divergence of complexes between brain regions. Ideally, direct protein frequency data would be needed to obtain more realistic simulation results. Such data can be obtained using specific imaging – microscopic – techniques. These imaging techniques show similar ratios to mRNA data as validation, but the dimensions are not exact protein counts, but voxels and intensities [23], and significantly less of this kind of data is available than can be generated from mRNA expression data on a larger scale. A detailed description of PSD organization remains a challenge. It is complicated by its size, the number and variations of its constituent proteins, and above all its variable stoichiometry and dynamic nature. Although high-resolution experimental data on binary complexes are available, they typically include only the interacting domains and segments. Our basic assumption is that simulations can complement experimental data and contribute significantly to the understanding of the nature of PSD. The modeling approach presented here is a first approximation, focusing primarily on the variability of protein frequencies, and considers a system of only seven PSD proteins, each with only a single isoform (or mutant version), thus providing a highly simplified representation of the total PSD subset of Thus, its complexity is far from the actual organization of the PSD. Consequently, my results cannot be translated directly to the real distribution of the complexes that actually appear in the PSD in different neurons. More accurate simulations would require quantitative data on the amounts found directly at the protein level, appropriate binding constants and consideration of the 3D organization of the complexes, as well as their localization within the postsynaptic region. In addition, the dynamic exchange of components, the spatial direction of the addition of new proteins, and the phenomenon of phase separation are all issues that are expected to contribute to the actual distribution of the complexes *in vivo*.

At the same time, my simulations already provide clues about the most relevant relationships between the main PSD proteins and shed light on the changes in the accessibility of the given proteins. That's right though the models do not approach the complexity of the real case, my approach is suitable for giving a comprehensive picture of PSD's main organizing principles.

4.3 Prospects of my research: possible long-term directions and developments

The contents of this thesis can be expanded in many ways, and we would like to expand them. The main driving force behind my research is to get a coherent, comprehensive picture of postsynaptic density, to which the work of the research group also contributes. My work provides a rudimentary insight into this. In this way, as the research group progresses, the results of a given sub-research can be integrated with the help of the developed procedural process. Binding constants can be adjusted, new mutations can be observed, structures can be improved. Observed steric hindrances can be implemented. With internal dynamic information, we can change the availability of domains - we can model many conformations. We can increase the number of proteins. We can examine competitions. Expanding the model is a challenging and continuous journey, with which we can look at this complex protein network from a new and hitherto less understood point of view. We can consider and add many new biological processes to our model in order to get closer to native biological systems - starting from the effects of phosphorylations, protein degradation, and factors affecting external expression levels.



5.Acknowledgments

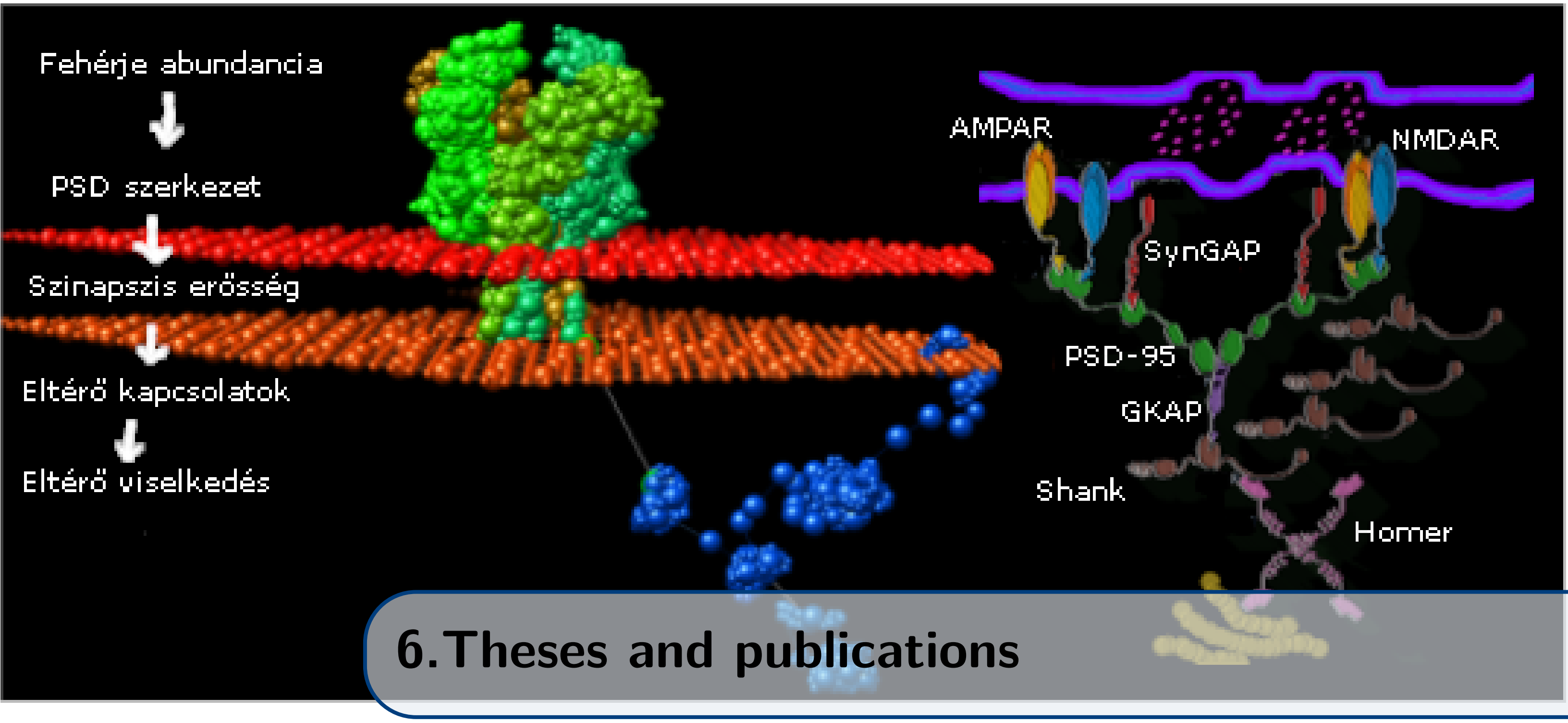
Molecular graphics and analysis were done using UCSF Chimera, developed by San Francisco with NIH grant P41-GM103311. The graphic cover images were created using the STRING database.

The research was financed by the European Union and co-financed by the European Social Fund under grant number EFOP-3.6.2-162017-00013. The research was carried out with the support of the National Innovation, Research and Development Office grant numbers OTKA 124363, 137947 and TKP2021-EGA-42.

I would like to thank many people for their contributions to my research. To my project managers: Attila Csikász-Nagy and Zoltán Gáspári for their supervision and guidance. Thanks to Kornél Komjáti, who helped me a lot in learning the basics of IMP. I would like to thank Márk Bence Keömley-Horváth for his work and for allowing me to use the Cytocast software created by him.

I would also like to thank Katinka Vida Tivadarné for her tireless work and understanding, which helped me navigate the mysteries of paperwork.

I would like to thank my table company for their support. My grateful thanks extend to my family for their persistence and patience. I especially thank my mother, Ágnes Fekete, for her helpful thoughts and questions.



6. Theses and publications

1. Based on literature data, I prepared a simplified model of the postsynaptic density (PSD) protein network, which contains 7 main proteins and can be used as input for systems biology calculations simulating the distribution of protein complexes. Based on expression data, I examined the seven proteins in a total of 524 different frequency compositions using the Cytocast software. (Miski et al. 2022, [1])
2. I have developed procedures for the comparative analysis of protein complex distributions obtained from simulations of the PSD model. The procedures enable the clear identification of individual protein complexes and the comparison of different simulation results, even in the case of different initial protein frequencies and mutations affecting the partner binding properties. (Miski et al. 2022, [1])
3. I found that the relationship between the distribution of the resulting protein complexes and the frequency of individual proteins is complex and can only be mapped with simulations. Very similar initial protein frequencies can also result in significantly different protein complex distributions, which may have functional significance based on the synaptic theory. (Miski et al. 2022, Miski proc. 2020 [1], [2])
4. I demonstrated that the effect of a hypomorphic mutation, which only slightly weakens a specific protein-protein interaction, significantly depends on the protein frequencies characteristic of the given cells. Therefore, it only causes significant changes in specific cells or cell types. Moreover, it is also characteristic that the complexes showing the greatest changes are not necessarily those involving the interaction partners affected by the mutation. (Miski et al. 2024, [3])
5. I developed a workflow for modeling the possible three-dimensional structures of some large, multi-component protein complexes characteristic of postsynaptic density. (Miski proc. 2019,2020,2022, [4]–[6])

Publications on which the thesis is based

[1] M. Miski, B. M. Keömley-Horváth, D. R. Megyeriné, A. Csikász-Nagy, and Z. Gáspári, "Diversity of synaptic protein complexes as a function of the abundance of their constituent proteins: A modeling approach," *PLOS Computational Biology*, volume 18, number 1, M. Migliore, Ed., e1009758, Jan. 2022. DOI: 10.1371/journal.pcbi.1009758. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1009758 (cited on page 18).

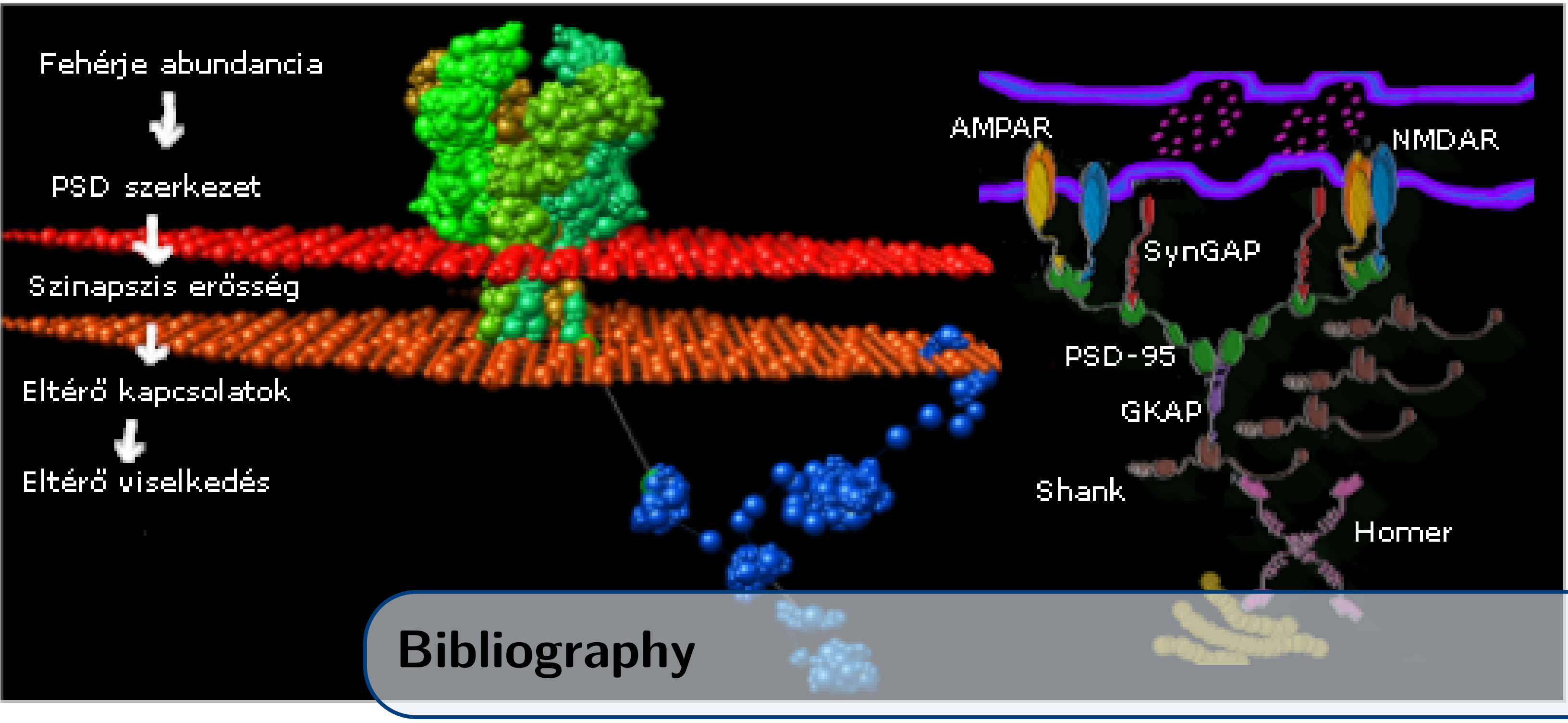
[2] M. Miski and A. Csikász-Nagy, "Analyses of protein-protein interactions in the psd by stochastic simulations," *PHD PROCEEDINGS ANNUAL ISSUES OF THE DOCTORAL SCHOOL FACULTY OF INFORMATION TECHNOLOGY AND BIONICS*, volume 15, pages 88–90, 2020, ISSN: 2064-7271 (cited on page 18).

[3] M. Miski, Á. Weber, K. Fekete-Molnár, B. M. Keömley-Horváth, A. Csikász-Nagy, and Z. Gáspári, "Simulated complexes formed from a set of postsynaptic proteins suggest a localised effect of a hypomorphic shank mutation," *BMC Neuroscience*, volume 25, number 1, Jul. 2024, ISSN: 1471-2202. DOI: 10.1186/s12868-024-00880-1. [Online]. Available: http://dx.doi.org/10.1186/s12868-024-00880-1 (cited on page 18).

[4] M. Miski, K. Kornél, A. Csikász-Nagy, and Z. Gáspári, "Integrative modeling of possible layouts of kir2.1 localized in membranes and connected by psd-95," *JEDLIK LABORATORIES REPORTS*, volume VII, pages 29–32, 2019, ISSN: 2064-3942 (cited on page 18).

[5] M. Miski, "Analyses of protein-protein interactions related to nnos by stochastic simulations," *PHD PROCEEDINGS ANNUAL ISSUES OF THE DOCTORAL SCHOOL FACULTY OF INFORMATION TECHNOLOGY AND BIONICS*, volume vol. 14, pages 59–62, 2019, ISSN: 2064-7271 (cited on page 18).

[6] M. Miski and A. Csikász-Nagy, "Distribution and structure of postsynaptic protein complexes assessed by simulations," *PHD PROCEEDINGS ANNUAL ISSUES OF THE DOCTORAL SCHOOL FACULTY OF INFORMATION TECHNOLOGY AND BIONICS*, volume vol. 16, page 50, 2021, ISSN: 2064-7271 (cited on page 18).



Bibliography

[1] H.-C. Kornau, “Postsynaptic density/architecture at excitatory synapses 73,” in *Reference Module in Neuroscience and Biobehavioral Psychology*, Elsevier, 2017. DOI: 10.1016/b978-0-12-809324-5.02357-9. [Online]. Available: https://doi.org/10.1016/b978-0-12-809324-5.02357-9 (cited on page 4).

[2] H. Jung, S. Kim, J. Ko, and J. W. Um, “Intracellular signaling mechanisms that shape postsynaptic GABAergic synapses,” *Current Opinion in Neurobiology*, volume 81, page 102 728, Aug. 2023. DOI: 10.1016/j.conb.2023.102728. [Online]. Available: https://doi.org/10.1016/j.conb.2023.102728 (cited on page 4).

[3] Y. Yang and J.-J. Liu, “Structural LTP: Signal transduction, actin cytoskeleton reorganization, and membrane remodeling of dendritic spines,” *Current Opinion in Neurobiology*, volume 74, page 102 534, Jun. 2022. DOI: 10.1016/j.conb.2022.102534. [Online]. Available: https://doi.org/10.1016/j.conb.2022.102534 (cited on page 4).

[4] Y. Hayashi, “Molecular mechanism of hippocampal long-term potentiation – towards multiscale understanding of learning and memory,” *Neuroscience Research*, volume 175, pages 3–15, Feb. 2022. DOI: 10.1016/j.neures.2021.08.001. [Online]. Available: https://doi.org/10.1016/j.neures.2021.08.001 (cited on page 4).

[5] S. G. Grant, “The synaptomic theory of behavior and brain disease,” *Cold Spring Harbor Symposia on Quantitative Biology*, volume 83, pages 45–56, 2018. DOI: 10.1101/sqb.2018.83.037887. [Online]. Available: https://doi.org/10.1101/sqb.2018.83.037887 (cited on pages 4, 5).

[6] S. Franklin and T. M. Vondriska, “Genomes, proteomes, and the central dogma,” *Circulation: Cardiovascular Genetics*, volume 4, number 5, pages 576–576, Oct. 2011. DOI: 10.1161/circgenetics.110.957795. [Online]. Available: https://doi.org/10.1161/circgenetics.110.957795 (cited on page 4).

[7] M. Roy, O. Sorokina, N. Skene, C. Simonnet, F. Mazzo, R. Zwart, E. Sher, C. Smith, J. D. Armstrong, and S. G. N. Grant, “Proteomic analysis of postsynaptic proteins in regions of the human neocortex,” *Nature Neuroscience*, volume 21, number 1, pages 130–138, Dec. 2017. DOI: 10.1038/s41593-017-0025-9. [Online]. Available: https://doi.org/10.1038/s41593-017-0025-9 (cited on page 5).

[8] D. Rosenegger, C. Wright, and K. Lukowiak, “A quantitative proteomic analysis of long-term memory,” *Molecular Brain*, volume 3, number 1, Mar. 2010. DOI: 10.1186/1756-6606-3-9. [Online]. Available: https://doi.org/10.1186/1756-6606-3-9 (cited on page 5).

[9] J. Griffiths and S. G. Grant, “Synapse pathology in alzheimer’s disease,” *Seminars in Cell &amp Developmental Biology*, volume 139, pages 13–23, Apr. 2023. DOI: 10.1016/j.semcdb.2022.05.028. [Online]. Available: https://doi.org/10.1016/j.semcdb.2022.05.028 (cited on page 5).

[10] A. A. B. Jamjoom, J. Rhodes, P. J. D. Andrews, and S. G. N. Grant, “The synapse in traumatic brain injury,” *Brain*, volume 144, number 1, pages 18–31, Nov. 2020. DOI: 10.1093/brain/awaa321. [Online]. Available: https://doi.org/10.1093/brain/awaa321 (cited on page 5).

[11] J. H. Kotaleski and K. T. Blackwell, “Modelling the molecular mechanisms of synaptic plasticity using systems biology approaches,” *Nature Reviews Neuroscience*, volume 11, number 4, pages 239–251, Apr. 2010. DOI: 10.1038/nrn2807. [Online]. Available: https://doi.org/10.1038/nrn2807 (cited on page 5).

[12] J. A. Miller, S.-L. Ding, S. M. Sunkin, K. A. Smith, L. Ng, A. Szafer, A. Ebbert, Z. L. Riley, J. J. Royall, K. Aiona, J. M. Arnold, C. Bennet, D. Bertagnolli, K. Brouner, S. Butler, S. Caldejon, A. Carey, C. Cuhaciyan, R. A. Dalley, N. Dee, T. A. Dolbeare, B. A. C. Facer, D. Feng, T. P. Fliss, G. Gee, J. Goldy, L. Gourley, B. W. Gregor, G. Gu, R. E. Howard, J. M. Jochim, C. L. Kuan, C. Lau, C.-K. Lee, F. Lee, T. A. Lemon, P. Lesnar, B. McMurray, N. Mastan, N. Mosqueda, T. Naluai-Cecchini, N.-K. Ngo, J. Nyhus, A. Oldre, E. Olson, J. Parente, P. D. Parker, S. E. Parry, A. Stevens, M. Pletikos, M. Reding, K. Roll, D. Sandman, M. Sarreal, S. Shapouri, N. V. Shapovalova, E. H. Shen, N. Sjoquist, C. R. Slaughterbeck, M. Smith, A. J. Sodt, D. Williams, L. Zöllei, B. Fischl, M. B. Gerstein, D. H. Geschwind, I. A. Glass, M. J. Hawrylycz, R. F. Hevner, H. Huang, A. R. Jones, J. A. Knowles, P. Levitt, J. W. Phillips, N. Šestan, P. Wohnoutka, C. Dang, A. Bernard, J. G. Hohmann, and E. S. Lein, “Transcriptional landscape of the prenatal human brain,” *Nature*, volume 508, number 7495, pages 199–206, Apr. 2014. DOI: 10.1038/nature13185. [Online]. Available: https://doi.org/10.1038/nature13185 (cited on page 7).

[13] M. Giollo, A. J. Martin, I. Walsh, C. Ferrari, and S. C. Tosatto, “NeEMO: A method using residue interaction networks to improve prediction of protein stability upon mutation,” *BMC Genomics*, volume 15, number S4, May 2014. DOI: 10.1186/1471-2164-15-s4-s7. [Online]. Available: https://doi.org/10.1186/1471-2164-15-s4-s7 (cited on page 8).

[14] J. G. Tate, S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, P. Fish, B. Harsha, C. Hathaway, S. C. Jupe, C. Y. Kok, K. Noble, L. Ponting, C. C. Ramshaw, C. E. Rye, H. E. Speedy, R. Stefancsik, S. L. Thompson, S. Wang, S. Ward, P. J. Campbell, and S. A. Forbes, “COSMIC: The catalogue of somatic mutations in cancer,” *Nucleic Acids Research*, volume 47, number D1, pages D941–D947, Oct. 2018. DOI: 10.1093/nar/gky1015. [Online]. Available: https://doi.org/10.1093/nar/gky1015 (cited on page 8).

[15] M. K. Baron, “An architectural framework that may lie at the core of the postsynaptic density,” *Science*, volume 311, number 5760, pages 531–535, Jan. 2006. DOI: 10.1126/science.1118995. [Online]. Available: https://doi.org/10.1126/science.1118995 (cited on page 9).

[16] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A k-means clustering algorithm,” *Applied Statistics*, volume 28, number 1, page 100, 1979. DOI: 10.2307/2346830. [Online]. Available: https://doi.org/10.2307/2346830 (cited on page 9).

[17] C. E. Holt, K. C. Martin, and E. M. Schuman, “Local translation in neurons: Visualization and function,” *Nature Structural & Molecular Biology*, volume 26, number 7, pages 557–566, Jul. 2019. DOI: 10.1038/s41594-019-0263-5. [Online]. Available: https://doi.org/10.1038/s41594-019-0263-5 (cited on page 12).

[18] W. G. Walkup, T. L. Mastro, L. T. Schenker, J. Vielmetter, R. Hu, A. Iancu, M. Reghunathan, B. D. Bannon, and M. B. Kennedy, “A model for regulation by SynGAP-

$$1ofbindingofsynapticproteinstoPDZ-domain'slots'inthepostsynapticdensity$$

,” *eLife*, volume 5, Sep. 2016. DOI: 10.7554/elife.16813. [Online]. Available: https://doi.org/10.7554/elife.16813 (cited on page 12).

[19] M. Miski, Á. Weber, K. Fekete-Molnár, B. M. Keömley-Horváth, A. Csikász-Nagy, and Z. Gáspári, “Simulated complexes formed from a set of postsynaptic proteins suggest a localised effect of a hypomorphic shank mutation,” *BMC Neuroscience*, volume 25, number 1, Jul. 2024, ISSN: 1471-2202. DOI: 10.1186/s12868-024-00880-1. [Online]. Available: http://dx.doi.org/10.1186/s12868-024-00880-1 (cited on page 13).

[20] M. Miski, K. Kornél, A. Csikász-Nagy, and Z. Gáspári, “Integrative modeling of possible layouts of kir2.1 localized in membranes and connected by psd-95,” *JEDLIK LABORATORIES REPORTS*, volume VII, pages 29–32, 2019, ISSN: 2064-3942 (cited on page 14).

[21] A. Ö. Sungur, T. M. Redecker, E. Andres, W. Dürichen, R. K. W. Schwarting, A. del Rey, and M. Wöhr, “Reduced efficacy of d-amphetamine and 3, 4-methylenedioxymethamphetamine in inducing hyperactivity in mice lacking the postsynaptic scaffolding protein SHANK1,” *Frontiers in Molecular Neuroscience*, volume 11, Nov. 2018. DOI: 10.3389/fnmol.2018.00419. [Online]. Available: https://doi.org/10.3389/fnmol.2018.00419 (cited on page 15).

[22] D. Pascovici, J. X. Wu, M. J. McKay, C. Joseph, Z. Noor, K. Kamath, Y. Wu, S. Ranganathan, V. Gupta, and M. Mirzaei, "Clinically relevant post-translational modification analyses—maturing workflows and bioinformatics tools," *International Journal of Molecular Sciences*, volume 20, number 1, page 16, Dec. 2018. DOI: 10.3390/ijms20010016. [Online]. Available: https://doi.org/10.3390/ijms20010016 (cited on page 15).

[23] V. A. Petyuk, W.-J. Qian, M. H. Chin, H. Wang, E. A. Livesay, M. E. Monroe, J. N. Adkins, N. Jaitly, D. J. Anderson, D. G. Camp, D. J. Smith, and R. D. Smith, "Spatial mapping of protein abundances in the mouse brain by voxelation integrated with high-throughput liquid chromatography-mass spectrometry," *Genome Research*, volume 17, number 3, pages 328–336, Feb. 2007. DOI: 10.1101/gr.5799207. [Online]. Available: https://doi.org/10.1101/gr.5799207 (cited on page 16).