DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„Homology Modelling of Shaggy-like Kinase α in *Arabidopsis thaliana* (ASKα)“

verfasst von / submitted by

Michaela Ukowitz

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magistra der Pharmazie (Mag.pharm.)

Wien, 2019 / Vienna, 2019

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 449
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Diplomstudium Pharmazie
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Thierry Langer
Mitbetreut von / Co-Supervisor:	Dipl.-Ing. Dr. Thomas Seidel

Acknowledgements

I would like to take this opportunity to thank my professor Dr. Thierry Langer for giving me the chance to work in his research group. Working on this interesting topic allowed me to get a taste of an area that I have not been confronted with too often during my studies.
In this context I also want to thank Dr. Claudia Jonak. She brought this part of her project at AIT to university and provided me with very valuable information about this topic.

Another thank you goes to Dr. Marcus Wieder, who, despite the changed circumstances at short notice, was always available for any kind of questions and could always answer them patiently with his enormous knowledge. He made me work very independently but was still always available to help if there were any kind of problems. Even with the time difference, he regularly took time for telephone conferences to support and to be kept up to date so that his local absence did not entail any disadvantage for my work.

A big thank you to Mag. pharm. Eva Hellsberg who took so much time for me and shared her knowledge about homology modelling with me. I'm very grateful to you, for spending your already short time on my problems and questions, always having an open ear and patiently answering thousand questions.

Besides, I would like to thank Dr. Thomas Seidel and the great working group. All the people are so kind and helpful and welcomed me very nicely into their group. At the lab meetings they gave me good ideas and were always available without hesitation for any computer problems I had to struggle with. I am very happy to have worked on my diploma thesis in this research group and had a lot of fun working there.

Finally, I would like to thank my family and friends who supported me both, financially and mentally, throughout my time at university.

List of abbreviations

ASK	Shaggy-like kinases in *Arabidopsis thaliana*
CDPK	Calcium-dependent protein kinase
DOPE	Discrete optimized protein energy
EM	Electron microscopy
GL2	Glabra 2
GSK3	Glycogen synthase kinase 3
G6PD6	Glucose-6-phosphate dehydrogenase-6
MAPK	Mitogen-activated protein kinase
MSA	Multiple sequence alignment
NADPH	Nicotinamide dinucleotide phosphate
NMR	Nuclear magnetic resonance
OPPP	Oxidative pentose phosphate pathway
PAMP	Pathogen-associated molecular pattern
PDB	Protein data bank
PRR	Pattern-recognition receptor
PTI	Pattern-triggered immunity
RMS	Root mean square
ROS	Reactive oxygen species
TREE	Threonine, arginine, glutamic acid, glutamic acid
TTG1	Transparent testa glabra 1
TT2	Transparent Testa 2

Table of contents

Abstract

Plants are often exposed to stress. On the one hand, there is biotic stress caused by bacteria, viruses, fungi or insects. On the other hand, abiotic stress is caused by changing environmental conditions such as drought, heat, frost or high salinity of the soil. These stress factors lead to enormous annual yield losses, which is why research into the mechanisms triggered by stress in plants is of high interest.

Therefore, there is interest to develop new methods to improve stress tolerance in plants, which would bring both, economic benefits and ensure sufficient food supply to the population.

Several studies attribute the kinase ASKα an important role in stress modulation. The activated kinase increases stress tolerance of plants, which is why it is important to find out more about this protein. A three-dimensional structure of this kinase, which has not yet been determined experimentally, is important.

The aim of this work is to model the 3D structure of ASKα and to examine it for possible binding pockets for small molecules. First, a good template was needed, which was identified as GSK3β due to the high sequence identity. The PDB was searched for the most suitable structure of GSK3β. With this template, homology models were calculated with the software *Modeller.* The calculated models were filtered by different validation criteria (molpdf, DOPE-score). For the most appropriate ones, Ramachandran plots were generated with Procheck. The plots were analysed in detail until a final model could be selected. For the first time, we can provide a 3D structure of ASKα in form of a thoroughly validated homology model. This model allows us to gain deeper insights into ASKα's function and allows further predictions about potential ligand bindings. Therefore, the model was examined with *Fpocket* for promising binding pockets.

Zusammenfassung

Pflanzen sind kontinuierlich Stress ausgesetzt. Auf der einen Seite ist es biotischer Stress, der unter anderem von Bakterien, Viren, Pilzen oder Insekten ausgelöst wird und auf der anderen Seite abiotischer Stress, der durch wechselnde Umweltbedingungen wie Trockenheit, Hitze, Frost oder hohen Salzgehalt der Böden verursacht wird. Diese Stressfaktoren führen jährlich zu enormen Ernteverlusten, weshalb die Wissenschaft gefragt ist, die Mechanismen, die durch Stress in Pflanzen ausgelöst werden, genauer zu erforschen. Das Ziel ist es, neue Methoden zu entwickeln, welche die Stresstoleranz in Pflanzen verbessern, was sowohl wirtschaftliche Vorteile bringen, als auch die ausreichende Versorgung der Bevölkerung mit Nahrung gewährleisten würde.

Mehrere Studien sprechen der Kinase ASKα eine wichtige Rolle bei der Stressmodulation zu. Die aktivierte Kinase erhöht die Stresstoleranz von Pflanzen, weshalb es wichtig ist, mehr über dieses Protein herauszufinden. Dafür ist eine dreidimensionale Struktur dieser Kinase von Bedeutung, die bis dato aber noch nicht experimentell ermittelt wurde.

Das Ziel dieser Arbeit war es, die 3D Struktur von ASKα zu modellieren und diese auf mögliche Bindetaschen für niedermolekulare Verbindungen zu untersuchen. Zuerst wurde eine geeignete Vorlage benötigt, als welche GSK3β aufgrund der hohen Sequenzidentität identifiziert wurde. Die PDB wurde nach der geeignetsten Struktur von GSK3β durchsucht. Mit dieser Vorlage wurden dann Homologiemodelle mit der Software *Modeller* berechnet. Die kalkulierten Modelle wurden mittels verschiedener Validierungskriterien (molpdf, DOPE-score) gefiltert und von den passendsten wurden mittels *Procheck* Ramachandran Plots erstellt. Die Plots wurden analysiert und ein finales Modell ausgewählt. Dieses Modell wurde mit *Fpocket* auf mögliche Bindungstaschen untersucht.

1. Introduction

1.1. Stress Response in Plants

The human population is growing, therefore more and more people need to be supplied with food. Due to decreasing resources, science is forced to learn more about the mechanisms how plants react to stress in order to develop methods to positively change plant productivity
Yield loss caused by stress is about 65%-87% compared to optimal growth conditions. Thus, practices that improve crop would make a tremendous contribution to food supply and also have financial benefits.[1]

Plants are exposed to various conditions during their growth. They might be exposed to biotic stress from other organisms or abiotic stress like heat, cold, frost, drought or salinity. Environmental stress can damage the structure of cells and disrupt important physiological activities. Membranes could lose their organization, but also proteins may denaturate or lose their functionality. Therefore, often high amounts of ROS (reactive oxygen species) are produced causing oxidative damage. The consequences of these circumstances are problems in plant growth and development, reduced functionality, or in worst case, plant death.[1]

Different types of plants vary a lot in their needs. When environmental conditions are perfect for one kind of plant, they can cause enormous stress for another species. Just as different plant species can be stressed in different ways, they can also react differently to stress.[2] Fig. 1 shows that many criteria have an effect on the plants response to stress. There are numerous aspects that influence whether a plant responses with susceptibility or resistance, such as duration, severity, number of exposures or the combination of stresses as characteristics of stress. Genotype, developmental stage and the involved organ or tissue affect the reaction as plant characteristics. When the actions of the plant are successful, it survives the problematic conditions, otherwise the result might be the death of the plant.[1]

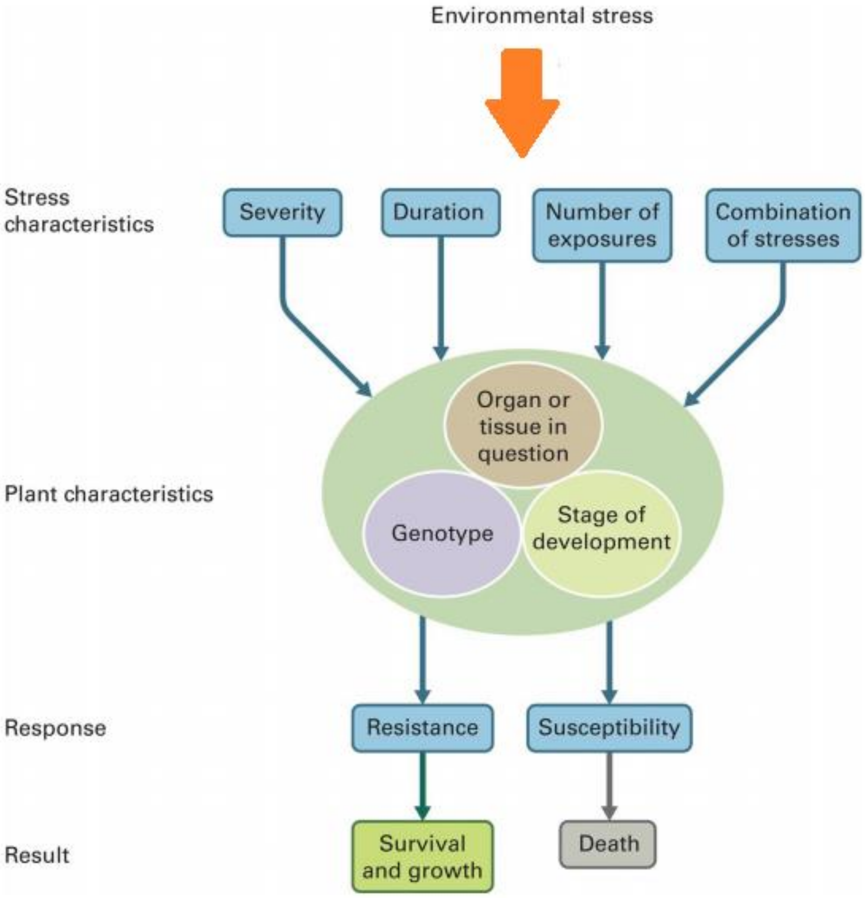


Fig. 1: Depending on different factors of environmental stress and the plant characteristics, the response and the result to the stress impact differs.[1]

To deal with stress, plants have two different strategies, which are stress avoidance and stress tolerance (Fig. 2).[1] If frost is stressful for a plant, stress avoidance would prevent ice formation by, for example, osmoregulation. Within the framework of stress tolerance, however, the plant would survive intracellular ice formation through stress proteins or altered membrane lipids.[3] Stress avoidance refers to mechanisms that delay or prohibit the consequences of stress. The adjustments are permanent and are passed on to the next generations. Stress tolerance is a direct reaction to conditions abnormal to a plant, like very low temperatures in summer. In contrast to stress avoidance, these adaptions are only temporary.[1]

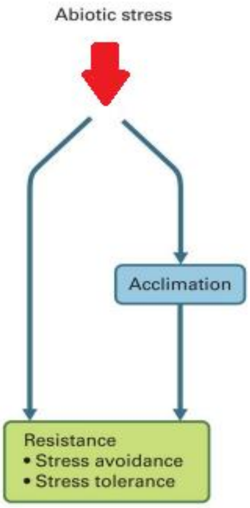


Fig. 2: Changed plant physiology to adapt to environmental changes.[1]

Besides the environmental factors mentioned above representing abiotic stress, also biotic stress caused by bacteria, fungi, viruses or insects exists.[4]

1.2. GSK3/Shaggy-like Kinases

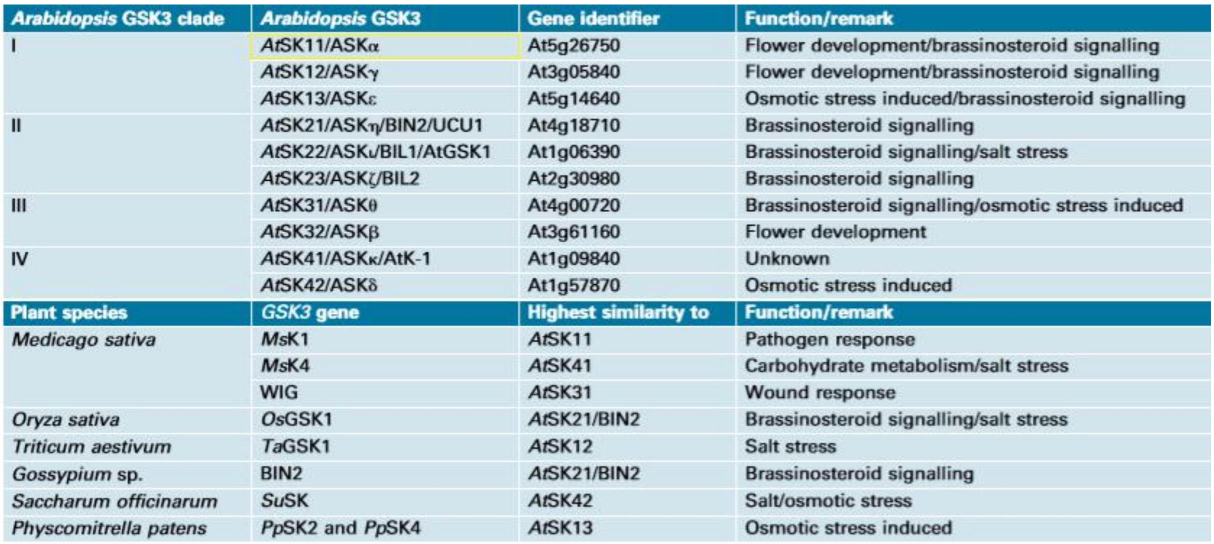
GSK3 (Glycogen synthase kinase 3) is a serine/threonine kinase that phosphorylates the enzyme glycogen synthase in the insulin signalling pathway. Mammalian GSK3 for instance plays an important role in developmental processes and in the animal Wnt signalling pathway, where the inactivation of GSK3 by Wnt subsequently leads to beta catenin no longer being degraded in the proteasome. In the nucleus β-catenin can build up and regulates the target gene expression. Furthermore, its functionality contributes to widespread diseases.[5]

GSK3 has become a target of high interest for the treatment of severe human diseases like Alzheimer’s disease, Diabetes type II, bipolar disorders, neurodegenerative pathologies or chronic inflammatory diseases.[6] There are two isoforms of that enzyme in mammalians, encoded by the genes GSK3α and GSK3β, which have a conserved kinase domain, but differ in their N- and C domains.[5]

Before substrates get phosphorylated by GSK3, another kinase must phosphorylate them to get the substrates into a proper configuration, so GSK3 can phosphorylate them. This process is also called “prime phosphorylation”. For the activation of GSK3β itself, Y216 gets phosphorylated (Fig. 4).[5]

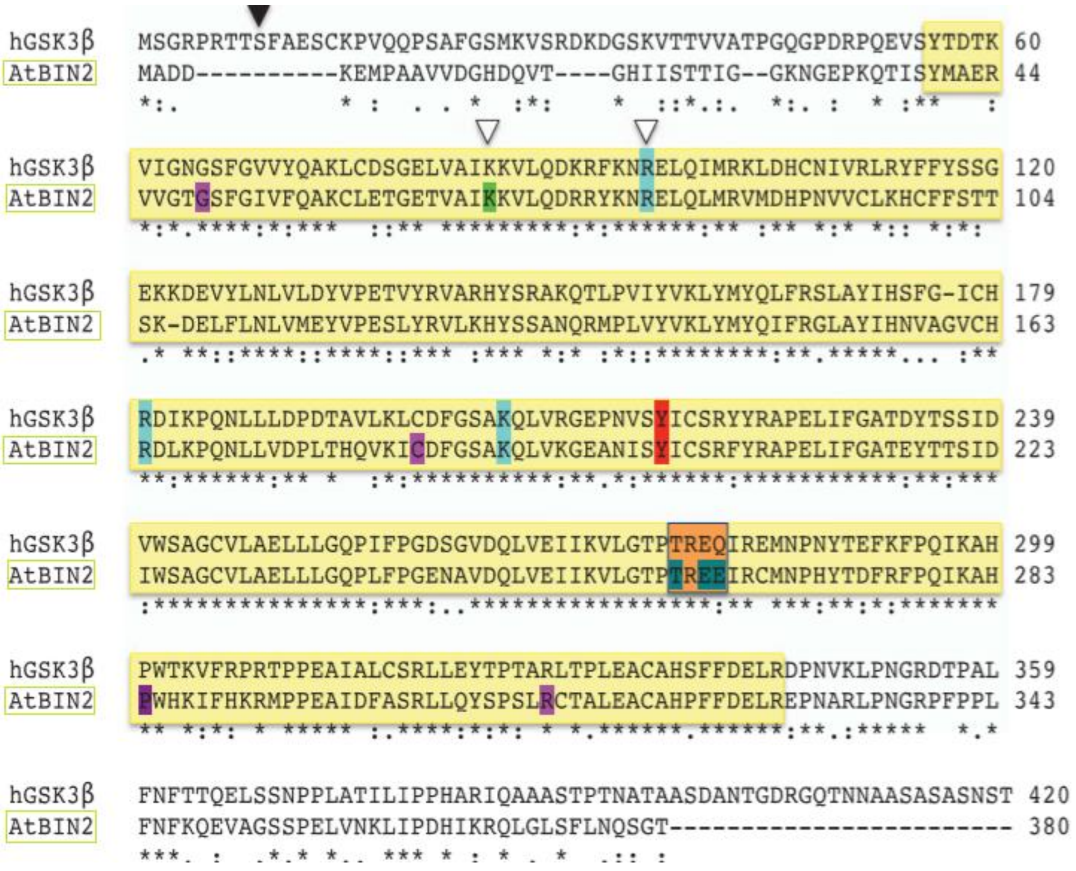
Homologues of that kinase have been found in all different kinds of eukaryotes. In contrast to the ones in animals, kinases in land plants are encoded by large multigene families, where the members have a relatively high sequence similarity.[5]

In *Arabidopsis thaliana*, a plant of the family Cruciferae[7], there are 10 homologues of GSK3. The names of these kinases can confuse a little bit. They are called Shaggy-like kinases, SK, ASK or AtSK. In Fig. 3, all 10 kinases are listed with their possible nomenclature as well as their group affiliation, gene identifier and function.[5]



Arabidopsis GSK3 clade	*Arabidopsis* GSK3	Gene identifier	Function/remark
I	AtSK11/ASKα	At5g26750	Flower development/brassinosteroid signalling
	AtSK12/ASKγ	At3g05840	Flower development/brassinosteroid signalling
	AtSK13/ASKε	At5g14640	Osmotic stress induced/brassinosteroid signalling
II	AtSK21/ASKη/BIN2/UCU1	At4g18710	Brassinosteroid signalling
	AtSK22/ASKι/BIL1/AtGSK1	At1g06390	Brassinosteroid signalling/salt stress
	AtSK23/ASKζ/BIL2	At2g30980	Brassinosteroid signalling
III	AtSK31/ASKθ	At4g00720	Brassinosteroid signalling/osmotic stress induced
	AtSK32/ASKβ	At3g61160	Flower development
IV	AtSK41/ASKκ/AtK-1	At1g09840	Unknown
	AtSK42/ASKδ	At1g57870	Osmotic stress induced
Plant species	***GSK3* gene**	**Highest similarity to**	**Function/remark**
Medicago sativa	*Ms*K1	AtSK11	Pathogen response
	*Ms*K4	AtSK41	Carbohydrate metabolism/salt stress
	WIG	AtSK31	Wound response
Oryza sativa	*Os*GSK1	AtSK21/BIN2	Brassinosteroid signalling/salt stress
Triticum aestivum	*Ta*GSK1	AtSK12	Salt stress
Gossypium sp.	BIN2	AtSK21/BIN2	Brassinosteroid signalling
Saccharum officinarum	*Su*SK	AtSK42	Salt/osmotic stress
Physcomitrella patens	*Pp*SK2 and *Pp*SK4	AtSK13	Osmotic stress induced

Fig. 3: List of the possible nomenclature of GSK3 like kinases in Arabidopsis thaliana.[5]



```
hGSK3β   MSGRPRTTSFAESCKPVQQPSAFGSMKVSRDKDGSKVTTVVATPGQGPDRPQEVSYTDTK  60
AtBIN2   MADD----------KEMPAAVVDGHDQVT----GHIISTTIG--GKNGEPKQTISYMAER  44
         *:.          * :  . . *  :*:    *  ::*.:.  *:. :  * :**   :

hGSK3β   VIGNGSFGVVYQAKLCDSGELVAIKKVLQDKRFKNRELQIMRKLDHCNIVRLRYFFYSSG 120
AtBIN2   VVGTGSFGIVFQAKCLETGETVAIKKVLQDRRYKNRELQLMRVMDHPNVVCLKHCFFSTT 104
         *:*.****:*:***  ::** *********:*:******:** :** *:* *:: *:*:

hGSK3β   EKKDEVYLNLVLDYVPETVYRVARHYSRAKQTLPVIYVKLYMYQLFRSLAYIHSFG-ICH 179
AtBIN2   SK-DELFLNLVMEYVPESLYRVLKHYSSANQRMPLVYVKLYMYQIFRGLAYIHNVAGVCH 163
         .* **::****::****::*** :*** *:* :*::********:**.*****... :**

hGSK3β   RDIKPQNLLLDPDTAVLKLCDFGSAKQLVRGEPNVSYICSRYYRAPELIFGATDYTSSID 239
AtBIN2   RDLKPQNLLVDPLTHQVKICDFGSAKQLVKGEANISYICSRFYRAPELIFGATEYTTSID 223
         **:******:** *  :*:***********:**.*:******:************:**:***

hGSK3β   VWSAGCVLAELLLGQPIFPGDSGVDQLVEIIKVLGTPTREQIREMNPNYTEFKFPQIKAH 299
AtBIN2   IWSAGCVLAELLLGQPLFPGENAVDQLVEIIKVLGTPTREEIRCMNPHYTDFRFPQIKAH 283
         :***************:***:..*****************:** ***:**:*:*******

hGSK3β   PWTKVFRPRTPPEAIALCSRLLEYTPTARLTPLEACAHSFFDELRDPNVKLPNGRDTPAL 359
AtBIN2   PWHKIFHKRMPPEAIDFASRLLQYSPSLRCTALEACAHPFFDELREPNARLPNGRPFPPL 343
         ** *:*: * ***** :.****:*:*: * *.******.******:**.:*****  *.*

hGSK3β   FNFTTQELSSNPPLATILIPPHARIQAAASTPTNATAASDANTGDRGQTNNAASASASNST 420
AtBIN2   FNFKQEVAGSSPELVNKLIPDHIKRQLGLSFLNQSGT------------------------ 380
         ***. :  .*.* *.. *** * : * . *  .:: :
```

Fig. 4: Alignment of a human GSK3β with ASK4 where the tyrosine that needs to be phosphorylated for the activity of the kinase is coloured red, and the conserved TREE(amino acids threonine, arginine, glutamic acid, glutamic acid) region is shown in orange. The part of the sequence boxed in yellow is the kinase domain.[5]

1.2.1. ASKα as a Positive Modulator of Salt Tolerance

High soil salt concentrations are a stress factor for plants and have a negative effect on crop yield. It leads to a deficit of water as well as ion concentration imbalances, which can cause destabilization of enzymes or membranes, a reduced supply of nutrients and an increased production of ROS. The production of ROS is a universal feature in the aerobic metabolism. Low levels are necessary for an optimal physiological function, whereas high levels caused by an external damage lead to oxidative damage and eventually in the end to cell death.[8]

Glucose-6-phosphate dehydrogenase (G6PD) is an enzyme catalysing a step in the oxidative pentose phosphate pathway (OPPP) that gets phosphorylated by ASKα. During the OPPP, NADPH (Nicotinamide dinucleotide phosphate) is provided for reductive biosynthesis and for the maintenance of the redox state in cells. The activity of G6PD is present in the cytosol and plastids of plants and increases e.g. under conditions of high salinity.[8]

To show the importance of ASKα to salt stress tolerance, knockout ASKα and ASKα overexpressing mutants were investigated. Under conditions of high salinity, the ASKα overexpressing mutant turned out to grow better than the knockout mutant. The salt tolerance during early seedling development was higher for mutants with overexpressed ASKα. Also a significant increase in germination efficiency and root length has been documented for these mutants.[8]

The activity of G6PD provides reducing power, which is important for the detoxification of ROS. In mutated plants with knockout ASKα under conditions of high salinity, the activity of G6PD has decreased, resulting in higher levels of H_2O_2. In contrast, plants overexpressing ASKα show lower levels of H_2O_2 due to a higher activity of G6PD. That indicates that ASKα is responsible for regulating the production and accumulation of ROS induced by stress (Fig. 5).[8]

The phosphorylation of Y229 is very important for the activation of ASKα, but under conditions of high salinity also other pathways of activation are involved. The underlying mechanisms need to be further investigated in future research.[8]

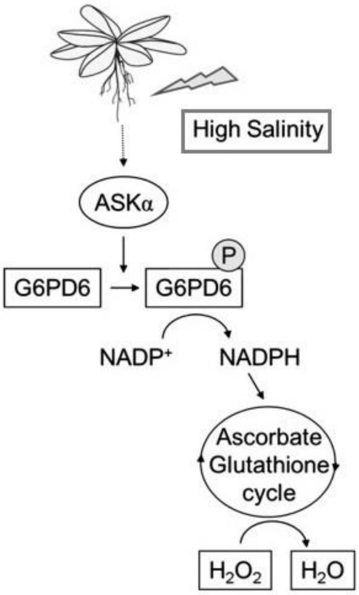


Fig. 5: Under activation of high salinity, activated ASKα phosphorylates G6PD6, which consequently provides NADPH for the reduction of H_2O_2 to H_2O.[8]

1.2.2. Contribution of ASKα to Pattern-triggered Immunity

In case plants are exposed to pathogenic microbes, pattern recognition receptors (PRR) on plants recognise conserved pathogen- associated molecular patterns (PAMPs) of the largest variety of pathogens. After PAMP binding to PRR, pattern triggered immunity (PTI) is initiated over a signalling cascade. Events such as an apoplastic burst of ROS, the influx of Ca^{2+} into the cytosol, the activation of Ca^{2+} -dependent or mitogen-activated protein kinases (CDPKs or MAPKs), the transfer of callose to the cell wall or transcriptional modifications are possible answers of PTI in order to protect from the pathogens.[9]

The burst of ROS is a conserved reaction in defending pathogens because of its antimicrobial features as well as the ability to cross-link compounds of cell walls to prevent the entrance of pathogenic microbes.[10]

To produce ROS, NADPH oxidase needs NADPH, which is generated in the oxidative pentose phosphate pathway by G6PD. ASKα phosphorylates T467 for the activity of the dehydrogenase.[9]

Possible pathogenic microbes are bacteria. A recent study reports, that the activity of ASKα was induced within 15 minutes after seedlings of *A. thaliana* have been exposed to flg22, an epitope of flagellin with 22 amino acids (Fig. 6).[9] A mutated plant with enhanced ASKα activity produced more ROS compared to one with knockout ASKα. In contrast, there is no significant difference between the two mutations concerning MAPK activation, showing that either ASKα works independently from or downstream of MAPK activation. Also, activated ASKα is important for hindering pathogen entry by strengthening cell wall integrity.

The induction of ASKα was also detected in other molecular patterns such as fungal chitin.

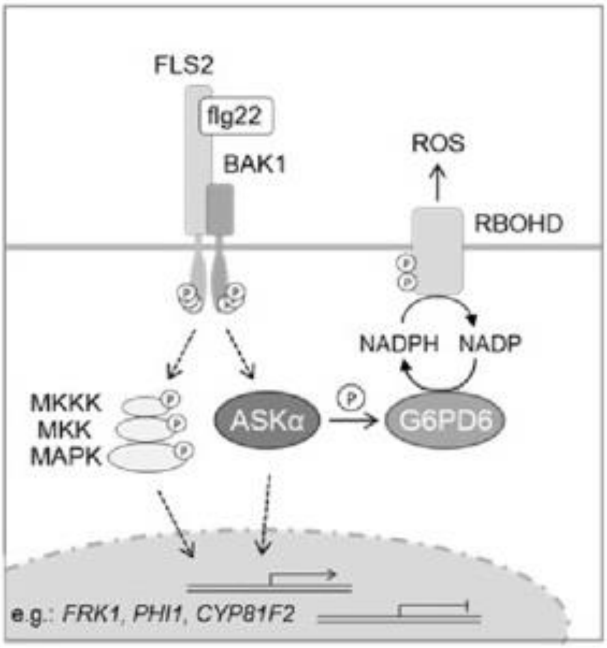


Fig. 6: Mechanism after flg22 binding;
FLS2: Flagellin Sensing2 (receptor kinase)
BAK1: BRI1-associated kinase1[9]

1.2.3. Regulation of Carbon Partitioning in *Arabidopsis* seeds

The development of seeds relies on the presence of nutrients like a carbon source (e.g. sucrose) by the parent plant. Mature seeds in *Arabidopsis thaliana* store their reserve compounds like storage proteins and oil as triacylglycerols in the zygotic embryo. The storage of reserve compounds is in competition with other metabolic pathways that also use sucrose as a nutrient. Sucrose for example is also used to produce seed coat mucilage and pigments.[11]

Transparent Testa Glabra 1 (TTG1) is an WD-40 repeat protein[12] that is part of the biosynthesis of flavonoids as well as the development of seed coat mucilage and pigments.[11]

ASKα and ASKγ are both kinases of group one GSK3-like kinases in *Arabidopsis thaliana*. They phosphorylate TTG1 and thus lead the metabolic pathway into the biosynthesis of fatty acids in the embryo while preventing the synthesis of flavonoid pigments and mucilage for the seed coat (Fig. 7).[11]

GL2 (Glabra2) activity increases when TTG1 interacts with TT2 (Transparent Testa 2) during seed development, which results in leading the carbon source into the production of seed mucilage and pigments. In contrast, ASKα and ASKγ phosphorylate TTG1 at S215. Phosphorylated TTG1 doesn't interact with TT2, resulting in a lower GL2 expression level and the movement of the metabolic pathway to the biosynthesis of fatty acids.[11]

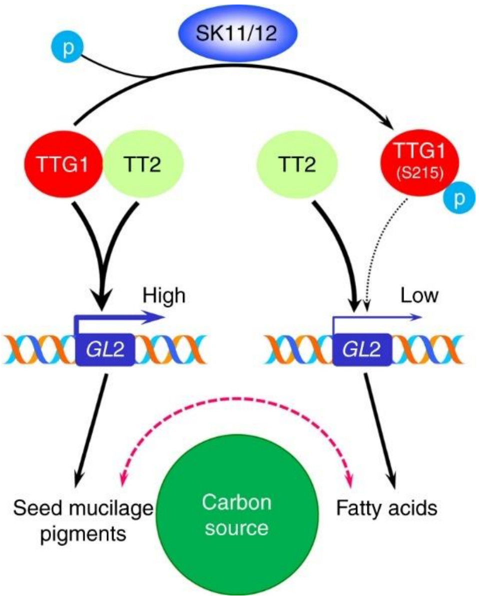


Fig. 7: Distribution of the carbon source between the production of seed mucilage/pigments and fatty acids.[11] Phosphorylated TTG1 leads to higher fatty acid production.

Notably, a mutation in the TREE (residues threonine, arginine, glutamic acid, glutamic acid at 290-293) region from glutamic acid (E) to lysine (K) at position 292 causes gain of function. The mutated residue lead to a higher protein stability and thus higher levels of fatty acids in seeds.[11]

1.3. Protein Structure Elucidation

To design compounds or to activate proteins like ASKα, it is necessary to know the three-dimensional structure of proteins. There are different ways to get these structures. Experimental ways are X-ray diffraction where the protein needs to be crystallized or NMR (nuclear magnetic resonance) analysis for proteins in solution.[13] Furthermore, the cryo-EM (cryo electron microscopy) method is steadily increasing in importance.[14]

Today many resolved macromolecular structures can be found in the PDB (Protein Data Bank).[15] However, structures of some macromolecules cannot be determined experimentally, because they are just too large for NMR analysis or cannot be crystallized for X-Ray diffraction. In this case, *in silico* methods are a good alternative for predicting proteins structure with an accuracy comparable to the best results that were achieved experimentally.[16]

1.3.1. Homology Modelling

The first step in homology modelling is finding possible templates which are related or whose sequence is similar to the target sequence (Fig. 8). Proteins with high sequence identities are likely to fold the same way, even though they are not related.[16] Proteins with the same function also tend to be folded the same, even if the sequence identity is not very high. Therefore, proteins with the same functionality but less sequence identity can also be very good templates. Beside the sequence identity, other parameters like the solvent or the pH should be taken into account when selecting the template.[17]

The next step is to align the sequence of the target to the selected templates. For the alignment, various programs can be used. When the sequence identity of target and template is over 40%, the alignment is mostly correct. With a sequence identity of less than 30%, one also speaks of the twilight zone. Around 30% sequence identity, only around 80% of the amino acids are matched correctly. It is very important to invest enough time and effort in alignments. If the alignment is wrong, no available modelling program can calculate a valid model. Alignments can be improved by applying methods such as the incorporation of structural information or the creation of models from several alignments. Of these created models, one then evaluates the alignments on the basis of the assessment scores for models and thus improves the alignment in an iterative process.[17] Aligning regions in a sequence with a low sequence identity can become difficult. In that case multiple sequence alignments are useful. Gaps in alignments should, if possible, not be in or close to functionally important regions.

Insertions and deletions have to be avoided whenever possible. Insertions are additional amino acids in the model, deletions are when amino acids are missing in the model. In this case, multiple sequence alignments are also very advantageous.[16]

When the optimal alignment has been found, model building can start. This step is executed automatically by a program. In case the template has one or more regions that are poorly resolved, "good" regions of different templates can be used together.[16]

After the models have been calculated, model optimization and evaluation steps are performed. There is no model without errors. Errors depend on two factors. First, errors occur in the models when there are errors in the template. Second, the sequence identity plays a role. An identity higher than 90% leads to models that can be compared faithfully to 3D structures determined by X-ray. If the identity is lower than 25%, the model probably contains a lot of errors. When the sequence identity is between 50% and 90%, there are rms errors of about 1,5Å as well as probably larger local errors.[16]

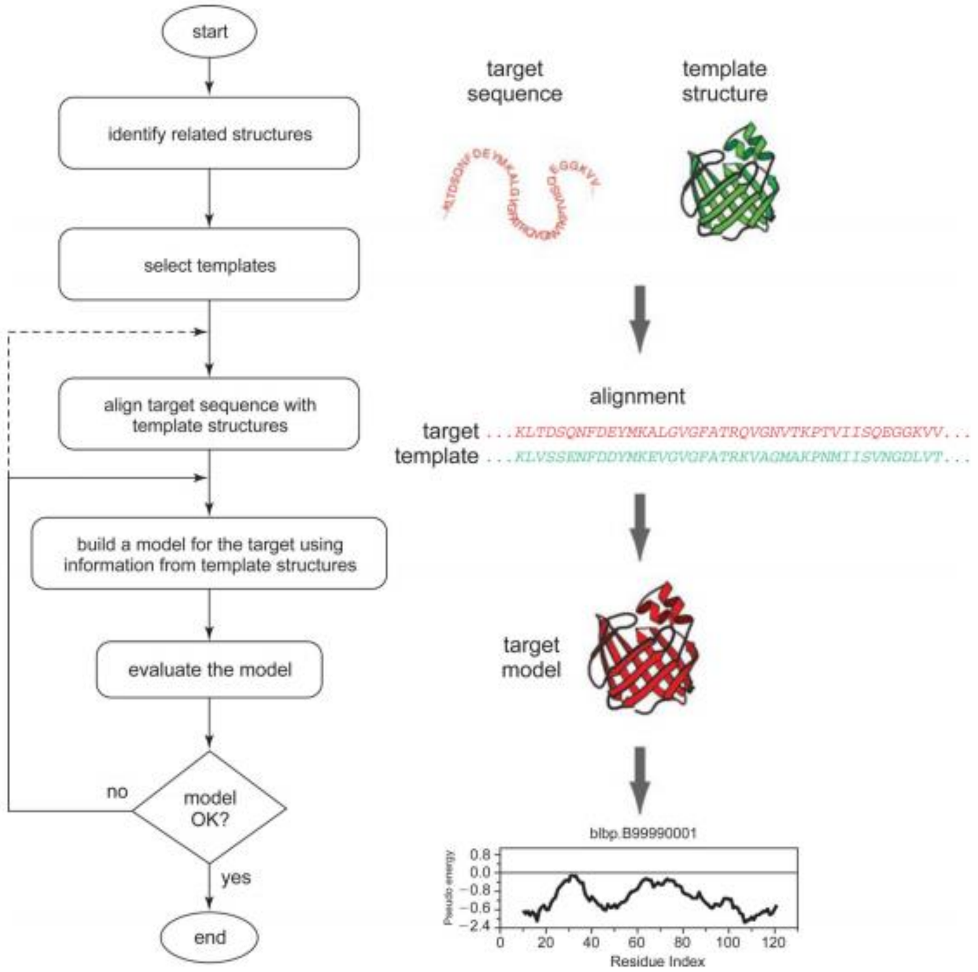

start
identify related structures
select templates
align target sequence with template structures
build a model for the target using information from template structures
evaluate the model
model OK?
no
yes
end
target sequence
template structure
alignment
target ...KLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVV...
template ...KLVSSENFDDYMKEVGVGFATRKVAGMAKPNMIISVNGDLVT...
target model
blbp.B99990001
Pseudo energy
0.8
0.0
−0.8
−0.6
−2.4
0
20
40
60
80
100
120
Residue Index

Fig. 8: Workflow in homology modelling.[18]

2. Aim of The Thesis

As clearly demonstrated above, an activated ASKα is of high importance for stress tolerance. Since the three-dimensional structure of that kinase has not been resolved yet, the aim of the thesis is to generate a 3D model of the protein ASKα in *Arabidopsis thaliana* by homology modelling. The final model is necessary to find possible binding pockets for small molecules which may activate the kinase.

3. Methods

3.1. Visualization software

During work in progress, some decisions had to be made and approaches had to be considered. To achieve this, the protein had to be carefully examined. There are several freely available programs for this, their application can be learned fast due to a simple graphical user interface. The programs used for this work are as follows.

3.1.1. PyMOL

PyMOL[19] is a free-access visualization software maintained and contributed by Schrödinger to graphically display biomolecules such as proteins. A graphical user interface is provided to simplify the handling, but it can also be controlled by a text interface. During this project, it has been used several times, especially in the steps of template selection, pocket detection and model adaption. The program allows to detect non-resolved regions in a possible template and proximity to important regions of the protein. Furthermore, each step that had modified the structure of the model, was checked for correctness with PyMOL.[19] [20]

The version 2.2.0 has been used. The software can be downloaded at https://pymol.org/2/ and for a temporally unlimited usage, a licence needs to be requested.

3.1.2. VMD

VMD[21] (Visual molecular dynamics) is a software for modelling, visualising and analysing biological systems like proteins or nucleic acids. It can display any number of structures and provides a lot of tools for e.g colouring or displaying a proteins structure. Another feature and a popular field of application is the animation and analyse of the trajectory of MD simulations.[22] A graphical user interface and a text interface are provided for program control.[21] In regard of this work, VMD was applied for displaying missing side chains in the template.

The version 1.9.3 was used and is available for download by https://www.ks.uiuc.edu/Research/vmd/.

3.2. Template selection

To build a homology model, both sequence and structure of the template are needed to model the unknown 3D structure of a target protein. The target sequence is available at UniProt, an online source for protein sequences.[23]

The template is supposed to have a similar function to the target as well as a highest possible sequence identity. A high sequence identity increases the possibility of creating a model of high quality which is very close to the natural structure.

Publicly available options can be found in the Protein Data Bank (PDB) at https://www.rcsb.org. The PDB is an archive for structural data of biological macromolecules, which is continuously growing.[15] In case of ASKα, GSK3β is the most appropriate and related template.[5] For this reason, the PDB has been searched for the most suitable GSK3β structure.

3.3. Sequence Alignment

Sequence alignments are intended to determine and illustrate the similarity between sequences of, for example, amino acids.[24] They are important tools to understand the nature of related groups of amino acid sequences.[25] The sequences are superimposed in a way that as many amino acids as possible of the different sequences match. There are multiple sequence alignments (MSA), in which three or more amino acid sequences are compared, or pairwise sequence alignments, in which only two sequences are overlaid. Both methods calculate the global sequence identity and give the opportunity to compare the local alignment quality. Similarity can be defined in different ways, considering mainly amino acid properties.

There are plenty of tools available using different algorithms for aligning sequences. A few have been tested to see the differences between the various programs. The ones important for the results are described in the following.

3.3.1. PROMALS3D

PROMALS3D[26] (PROfile Multiple Alignment with predicted Local Structures and 3D constraints) is an online tool used for protein sequence and structure alignments. It can identify homologous proteins with known structure of a given amino acid sequence. The program is an optimized version of a naïve sequence alignment method because it includes the information of the three-dimensional structure in the alignment. The input is

the target amino acid sequence and the PDB-ID of the template. The tool then considers structural constraints and combines it with sequence alignments.[26]
The output is an alignment in three different formats: A coloured alignment with predicted secondary structure and conservation information, a CLUSTAL format alignment and a FASTA format, which is required for the further operation steps.[27]

The webpage is freely available via http://prodata.swmed.edu/promals3d/promals3d.php and was used with default settings.

3.3.2. Jalview

Sequence alignments can also be modified manually. Jalview was developed to simplify editing, viewing and analysing of alignments.[28] Even the best alignment programs cannot optimally arrange the sequences on top of each other, if the sequence identity of different proteins is very low. Human skills are essential here, because the scientist has more biological background information available than the programs, which only refer to the sequence and possibly also to structural information.[25]

In regard of this project, Jalview was used for a better visualization of similarities between target and template according to their amino acid properties. In addition, it was applied to cut off the terminal sequence residues identified by the PROMALS3D alignment. The N- and C-termini are excluded from the modelling process because there are no templates of GSK3β available including these residues. Missing termini occur frequently in resolved protein structures due to their high flexibility.

The *Launch Jalview Desktop* version used is freely available via http://www.jalview.org.

3.3.3. Clustal Omega

Clustal Omega is an aligning program for multiple sequence alignments, which can align from three up to 4000 sequences.[29,30] The input are FASTA sequences and the output provides you with a demonstrative alignment as well as a percent identity matrix.

The version 12.1 has been used for the alignment of the ten kinases in *A. thaliana* to see the differences in the sequence of the proteins and for the template selection.

It is available via https://www.ebi.ac.uk/Tools/msa/clustalo/ and has been used with default settings.

3.3.4. EMBOSS Needle

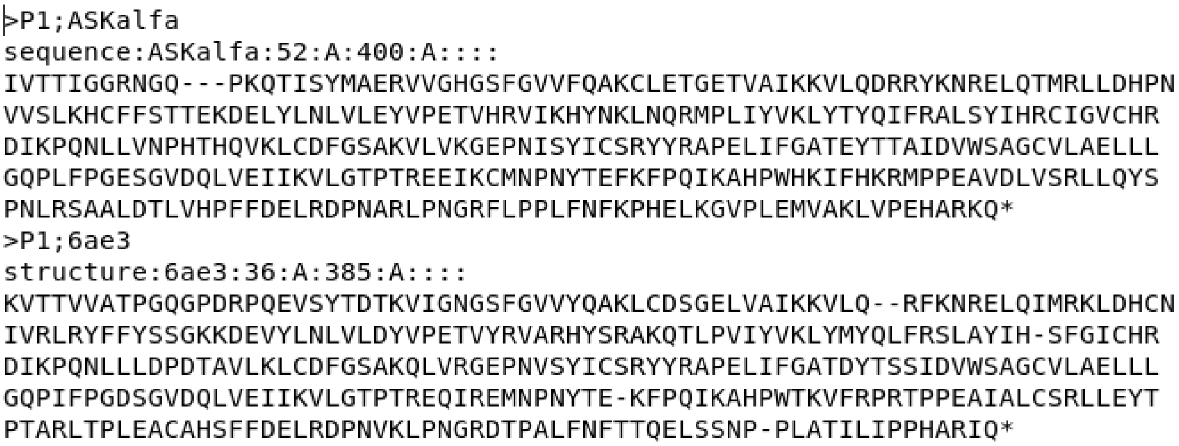
EMBOSS Needle uses the Needleman-Wunsch algorithm and provides an optimal global alignment of two sequences.[31] The method used is called dynamic programming.[32] The algorithm determines the optimal similarity score, which is a measure for the similarity of two sequences. The higher this score is, the more similar these two sequences are. The score is the sum of the matched amino acids. For each gap opened or enlarged, a gap penalty is deducted.[31] As input serve the two sequences of the proteins of interest in FASTA format. The output is the alignment with useful values like the sequence identity and similarity or the number of introduced gaps.

It has been used for various alignments during the project, especially to compare the results to the output of PROMALS3D.

The tool is available via https://www.ebi.ac.uk/Tools/psa/emboss_needle/ and has been used with default options.

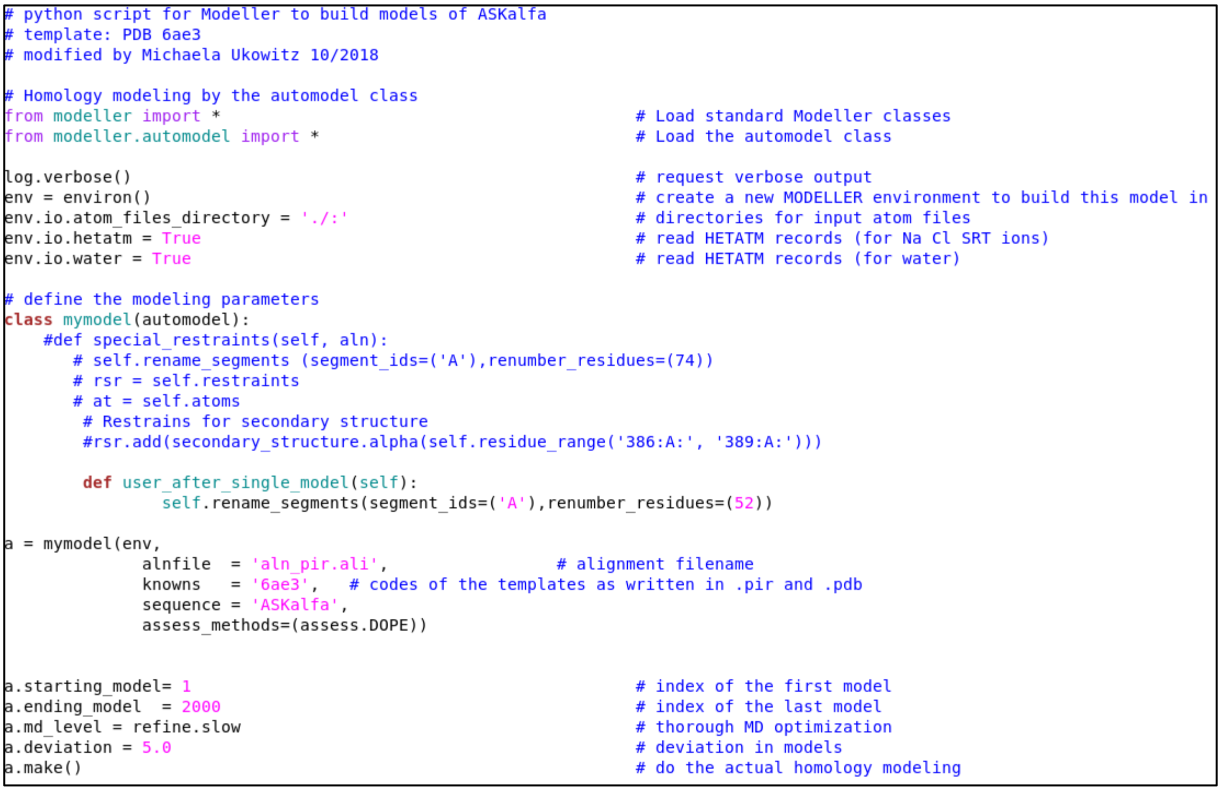
3.4. Homology Modelling with Modeller

Modeller is a program for homology or comparative modelling of 3D structures, which calculates models with all non-hydrogen atoms by the satisfaction of spatial restrictions.[33] The input needed is an alignment file of the target and the template in the PIR (.ali) format as well as the template's PDB file. The first line of the alignment file consists of the sequence code in the format ">P1;code". The second line has ten fields, separated by colons, in which information about the structure file, if available, is written. The numbers in the second line of each sequence indicate at which amino acid of the sequence of chain A the alignment starts and ends. The next lines contain the sequence, the end must be marked with a "*" (Fig. 9). To start the calculations with Modeller, a python script is required containing all the commands to be executed (Fig. 10).



```
>P1;ASKalfa
sequence:ASKalfa:52:A:400:A::::
IVTTIGGRNGQ---PKQTISYMAERVVGHGSFGVVFQAKCLETGETVAIKKVLQDRRYKNRELQTMRLLDHPN
VVSLKHCFFSTTEKDELYLNLVLEYVPETVHRVIKHYNKLNQRMPLIYVKLYTYQIFRALSYIHRCIGVCHR
DIKPQNLLVNPHTHQVKLCDFGSAKVLVKGEPNISYICSRYYRAPELIFGATEYTTAIDVWSAGCVLAELLL
GQPLFPGESGVDQLVEIIKVLGTPTREEIKCMNPNYTEFKFPQIKAHPWHKIFHKRMPPEAVDLVSRLLQYS
PNLRSAALDTLVHPFFDELRDPNARLPNGRFLPPLFNFKPHELKGVPLEMVAKLVPEHARKQ*
>P1;6ae3
structure:6ae3:36:A:385:A::::
KVTTVVATPGQGPDRPQEVSYTDTKVIGNGSFGVVYQAKLCDSGELVAIKKVLQ--RFKNRELQIMRKLDHCN
IVRLRYFFYSSGKKDEVYLNLVLDYVPETVYRVARHYSRAKQTLPVIYVKLYMYQLFRSLAYIH-SFGICHR
DIKPQNLLLDPDTAVLKLCDFGSAKQLVRGEPNVSYICSRYYRAPELIFGATDYTSSIDVWSAGCVLAELLL
GQPIFPGDSGVDQLVEIIKVLGTPTREQIREMNPNYTE-KFPQIKAHPWTKVFRPRTPPEAIALCSRLLEYT
PTARLTPLEACAHSFFDELRDPNVKLPNGRDTPALFNFTTQELSSNP-PLATILIPPHARIQ*
```

Fig. 9: Input alignment in">P1;code" format.



```
# python script for Modeller to build models of ASKalfa
# template: PDB 6ae3
# modified by Michaela Ukowitz 10/2018

# Homology modeling by the automodel class
from modeller import *                                          # Load standard Modeller classes
from modeller.automodel import *                                # Load the automodel class

log.verbose()                                                   # request verbose output
env = environ()                                                 # create a new MODELLER environment to build this model in
env.io.atom_files_directory = './:'                             # directories for input atom files
env.io.hetatm = True                                            # read HETATM records (for Na Cl SRT ions)
env.io.water = True                                             # read HETATM records (for water)

# define the modeling parameters
class mymodel(automodel):
    #def special_restraints(self, aln):
       # self.rename_segments (segment_ids=('A'),renumber_residues=(74))
       # rsr = self.restraints
       # at = self.atoms
        # Restrains for secondary structure
        #rsr.add(secondary_structure.alpha(self.residue_range('386:A:', '389:A:')))

        def user_after_single_model(self):
                self.rename_segments(segment_ids=('A'),renumber_residues=(52))

a = mymodel(env,
              alnfile  = 'aln_pir.ali',                # alignment filename
              knowns   = '6ae3',   # codes of the templates as written in .pir and .pdb
              sequence = 'ASKalfa',
              assess_methods=(assess.DOPE))


a.starting_model= 1                                             # index of the first model
a.ending_model  = 2000                                          # index of the last model
a.md_level = refine.slow                                        # thorough MD optimization
a.deviation = 5.0                                               # deviation in models
a.make()                                                        # do the actual homology modeling
```

Fig. 10: Used Python script to run Modeller

The version Modeller 9.20 was used with default settings. 2000 models including two scores which are implemented in Modeller, the molpdf and DOPE score, were calculated. The output, which was then used for further work, were the PDB files of the individual models with the corresponding scores as well as a log file in which all calculations were recorded.

Modeller is a very complex program. Instructions, tutorials and answers to various questions, as well as ready-made scripts can be found at https://salilab.org/modeller/.

3.5. Validation of The Model

Modeller calculates several models based on statistical factors. To limit the choice of models, different scores and validation methods are used.

3.5.1. Molpdf

The molpdf (molecular probability density function) score is an objective function which is automatically calculated in Modeller.[31,34] It is the standard scoring function for model assessment and sums up all the restraint violations,[35] so the lower the score, the better the model. The score is no absolute measure, therefore it can only be used to rank models from the same alignment (Formula 1).[34]

$$\boxed{F = F(R) = F_{symm} + \sum_{i} c_i\,(f_i, p_i)}$$

Formula 1: Molpdf score

$\boldsymbol{F_{symm}}$= optional symmetry term

R= Cartesian coordinates of all atoms

c= restraint

f= geometric feature of a molecule

p= parameters

Modeller minimizes the objective function F regarding Cartesian coordinates of about 10000 atoms, which are 3D points, that form a system. The form of restraint **c** includes a quadratic function, cosine, a weighted sum of a few Gaussian functions, Coulomb law, Lennard-Jones potential, cubic splines, and some other simple functions. The geometric features **f,** for example, contain a distance, an angle, a dihedral angle, a pair of dihedral angles between two, three, four atoms and eight atoms as well as an atom density which is the number of atoms around the central atom.[36]

3.5.2. DOPE Score

The DOPE (Discrete Optimized Protein Energy) score is a statistical potential to assess the three-dimensional structures of proteins in homology modelling. It is incorporated in the Modeller software.[37] For the score to be calculated, the command must be included in the input script.

The idea is that the native structure of a protein under native conditions has the lowest free energy. The DOPE-score is based on an improved reference state corresponding to non-interacting atoms in a homogeneous sphere, where the radius depends on a native sample structure. It therefore considers the finite and spherical shape of the native structure.[37]
The assessment method was determined by evaluating 1472 crystal structures and interpreting the DOPE score to recognize the model that comes closest to the native structure. Compared to other scoring functions, the DOPE score performed best in differentiating good from bad models.[37]

As the DOPE score was obtained by non-membrane proteins, it is theoretically only reliable for soluble proteins.[38]

3.5.3. Procheck

Procheck is a program that assesses the geometric quality of a protein or the model of a protein. It investigates the general geometry as well as the residue-by-residue geometry.[39]

To analyse more than one model at once, the installation of the program is necessary. The link and the instructions to start the download can be found here https://www.ebi.ac.uk/thornton-srv/software/PROCHECK/download.html.
In case you just want to check a single pdb file, PDBsum http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=index.html is available with no need of a download.

The only input the program needs are the pdb files of all models where plots should be generated. As an output, the plots as well as a residue-by-residue listing was created. [39]

For selecting the best model, Ramachandran plots were analysed. A Ramachandran plot shows the distribution of the combinations of the two dihedral angles Psi and Phi in a protein backbone. Procheck thus provides statistics, where it differs between residues in *most favoured, additional allowed, generously allowed* and *disallowed* regions. Based on the percentage of residues in the respective region, a conclusion about how good the quality of a protein is, can be made.[39]

3.6. Model refinement

Modeller calculates the models without hydrogens, because they are not included in the template. For a complete model, these must be added. In addition, other subtle changes and adaptations to a model or preparations for further studies may be necessary occasionally.

3.6.1. CHARMM-GUI

CHARMM-GUI is a website that intends to simplify and standardize the usage of simulation techniques using the CHARMM (Chemistry at Harvard Macromolecular Mechanics). It provides different tools for molecular dynamic calculations of macromolecules such as proteins and prepares input files for several programs, e.g. CHARMM[40], OpenMM[41], Gromacs[42], etc. Furthermore, CHARMM-GUI offers several versions of the CHARMM force field.[43]

The webservice CHARMM-GUI has been invented to supply a graphical user interface (GUI) and is available at http://www.charmm-gui.org/. It is a more intuitive way for preparing input files as well as molecular systems for CHARMM and other programs.[43]

The following steps were performed with the PDB Reader.[44] *Terminal group patching* with ACE (acetylation) for the N- and CTER (standard) for the C-terminus selected, *preserve hydrogen coordinates* and *phosphorylation* of Y229. The phosphorylation of tyrosine is necessary because it is essential for the activity of the kinase. To mutate E292 all mentioned steps plus the additional *mutation* step were executed. The mutation step is very important to compare the native protein to the mutated one at the end.

3.7. Pocket Detection with Fpocket

Fpocket is an algorithm to predict pockets in proteins. The input is a simple pdb file. As an output, several files are generated.[45] Of special interest is the ".pml" file to visualize the pockets in PyMOL and a text file with a list of different scores for each pocket.

The number of alpha spheres is a normalized value in Fpocket and indicates the size of a pocket accordingly.[46]

The druggability score uses a logistic function with a value between 0 and 1 given to a possible binding pocket. It represents the likelihood of small, drug-like molecules to bind in that pocket. A low value indicates that it is highly unlikely that small molecules will bind to it. If the number is higher than the threshold 0.5, a small molecule probably can bind.[45,46]

The version used was Fpocket2 and is available at http://fpocket.sourceforge.net/ with the download instructions as well as the user manual.

4. Results and discussion

4.1. Template selection

To get the best ASKα homology model possible, a template with a high similarity to the target is needed. Therefore, when looking for suitable templates in the PDB, various factors were considered.

GSK3β is a kinase that plays a role in different pathways and diseases. Homologues have been found in all kinds of eukaryotes, like the land plant *Arabidopsis thaliana*.[5] The similarity and the conserved kinase domain made GSK3β the chosen protein, for which the PDB was searched.

Generally, a template of high quality is very important. Any structure created by homology modelling is only as good as the experimental data used. If already the template has a bad resolution, the created homology model may be not very meaningful.

Since ASKα is a protein with 405 residues[47], the template should consist of about the same number of amino acids. Also, we preferably wanted a template with resolved N- and C-termini. GSK3β has 420 residues[48], the available structures all had approximately 350 residues resolved. So, none of the possibilities in the PDB includes the termini of the protein as parts of its structure. This is the reason why we could not use the amount of resolved residues or resolved ends as a priority selection criterion.

There were two main parameters for the selection of the template. First, only the available structures with a high resolution were considered. Within those with the highest resolution, the determining factor was, that important regions of the kinase are resolved and that there are no gaps close to these regions.
As important regions, Y229, that gets phosphorylated, as well as the residues 290-293, were considered. That region from 290-293 is also called TREE region, because it consists of the residues threonine (T), arginine (R) and two glutamic acids (E).This region is of special interest, because E292 is going to get mutated to K292. The reason for this is the assumption that this mutation increases the stability of the protein.[5]

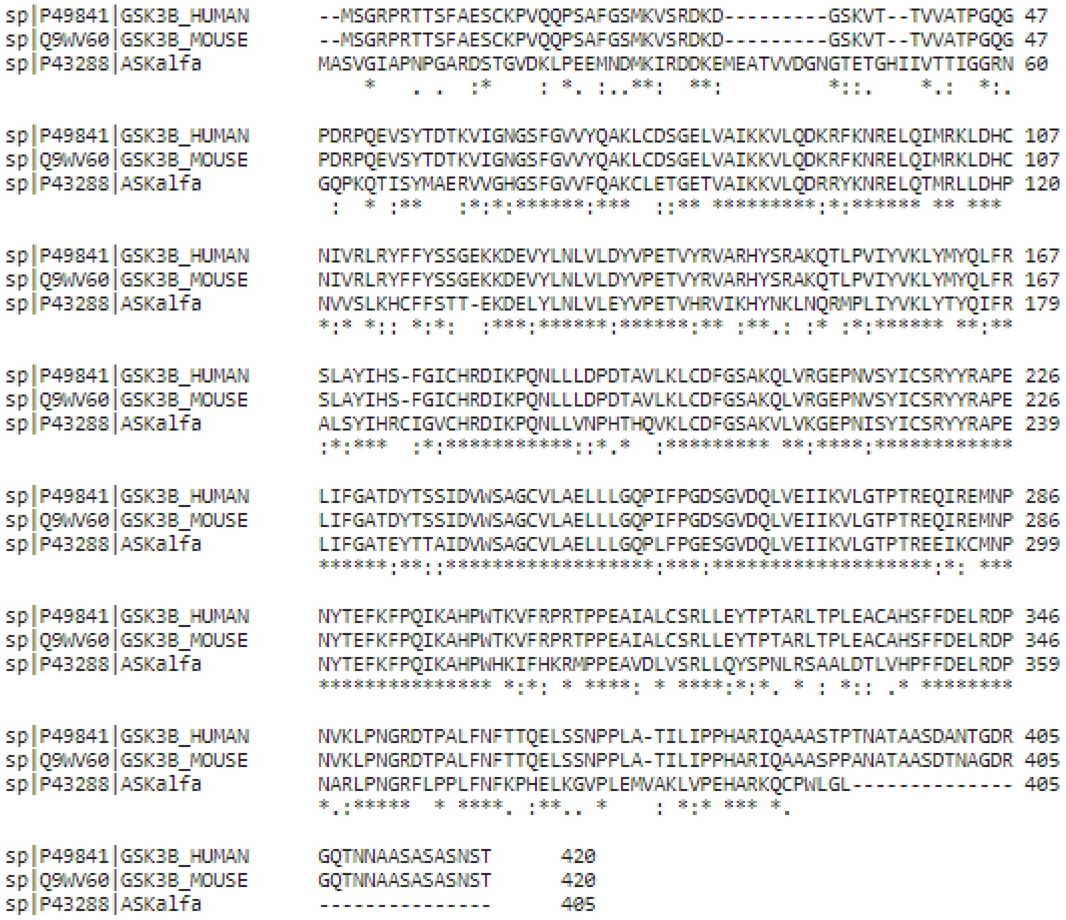
Table 1 shows the structures of GSK3β with the highest resolution. 1Q5K and 4AFJ are human GSK3β, whereas 6AE3 is present in *Mus musculus*. The sequence identity between the human GSK3β, mouse GSK3β and ASKα was checked to be sure the protein doesn't differ between these two species. The sequence identity between the human and the mouse kinase is 99,05 % and human and mouse kinase have each a sequence homology of 61,54 % with the plant kinase. Since there is no difference, both kinases could be used. (Fig. 11,Fig. 12).

PDB ID	**Resolution (Å)**	**Termini resolved**	**Mutation**	**Ligand**	**Residues resolved (without termini)**	**Reference**
1Q5K	1,94	no	0	TMU	349	(2003) J.Biol.Chem. **278**: 45937-45945
4AFJ	1,98	no	0	SJJ	351	(2018) Biochem Biophys Res Commun **504** 519-524
6AE3	2,14	no	0	Morin	348	(2012) Bioorg Med Chem Lett **22** 1989

Table 1: The three structures of GSK3β with the highest resolution.

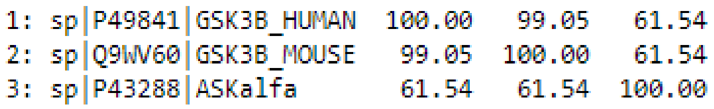
TMU= N-(4-Methoxybenzyl)-N'-(5-Nitro-1,3-Thiazol-2-yl)urea

SJJ= 5-(4-Methoxyphenyl)-N-(Pyridin-4-ylmethyl)-1,3-oxyzole-4-carboxamide



```
sp|P49841|GSK3B_HUMAN      --MSGRPRTTSFAESCKPVQQPSAFGSMKVSRDKD---------GSKVT--TVVATPGQG 47
sp|Q9WV60|GSK3B_MOUSE      --MSGRPRTTSFAESCKPVQQPSAFGSMKVSRDKD---------GSKVT--TVVATPGQG 47
sp|P43288|ASKalfa          MASVGIAPNPGARDSTGVDKLPEEMNDMKIRDDKEMEATVVDGNGTETGHIIVTTIGGRN 60
                               *   . .  :*    : *. :..**:  **:         *::.    *.:  *:.

sp|P49841|GSK3B_HUMAN      PDRPQEVSYTDTKVIGNGSFGVVYQAKLCDSGELVAIKKVLQDKRFKNRELQIMRKLDHC 107
sp|Q9WV60|GSK3B_MOUSE      PDRPQEVSYTDTKVIGNGSFGVVYQAKLCDSGELVAIKKVLQDKRFKNRELQIMRKLDHC 107
sp|P43288|ASKalfa          GQPKQTISYMAERVVGHGSFGVVFQAKCLETGETVAIKKVLQDRRYKNRELQTMRLLDHP 120
                            :  * :**   :*:*:******:***  ::** *********:*:****** ** ***

sp|P49841|GSK3B_HUMAN      NIVRLRYFFYSSGEKKDEVYLNLVLDYVPETVYRVARHYSRAKQTLPVIYVKLYMYQLFR 167
sp|Q9WV60|GSK3B_MOUSE      NIVRLRYFFYSSGEKKDEVYLNLVLDYVPETVYRVARHYSRAKQTLPVIYVKLYMYQLFR 167
sp|P43288|ASKalfa          NVVSLKHCFFSTT-EKDELYLNLVLEYVPETVHRVIKHYNKLNQRMPLIYVKLYTYQIFR 179
                           *:* *:: *:*:  :***:******:******:** :**.: :* :*:****** **:**

sp|P49841|GSK3B_HUMAN      SLAYIHS-FGICHRDIKPQNLLLDPDTAVLKLCDFGSAKQLVRGEPNVSYICSRYYRAPE 226
sp|Q9WV60|GSK3B_MOUSE      SLAYIHS-FGICHRDIKPQNLLLDPDTAVLKLCDFGSAKQLVRGEPNVSYICSRYYRAPE 226
sp|P43288|ASKalfa          ALSYIHRCIGVCHRDIKPQNLLVNPHTHQVKLCDFGSAKVLVKGEPNISYICSRYYRAPE 239
                           :*:***  :*:***********::*.*  :********* **:****:************

sp|P49841|GSK3B_HUMAN      LIFGATDYTSSIDVWSAGCVLAELLLGQPIFPGDSGVDQLVEIIKVLGTPTREQIREMNP 286
sp|Q9WV60|GSK3B_MOUSE      LIFGATDYTSSIDVWSAGCVLAELLLGQPIFPGDSGVDQLVEIIKVLGTPTREQIREMNP 286
sp|P43288|ASKalfa          LIFGATEYTTAIDVWSAGCVLAELLLGQPLFPGESGVDQLVEIIKVLGTPTREEIKCMNP 299
                           ******:**::******************:***:*******************:*: ***

sp|P49841|GSK3B_HUMAN      NYTEFKFPQIKAHPWTKVFRPRTPPEAIALCSRLLEYTPTARLTPLEACAHSFFDELRDP 346
sp|Q9WV60|GSK3B_MOUSE      NYTEFKFPQIKAHPWTKVFRPRTPPEAIALCSRLLEYTPTARLTPLEACAHSFFDELRDP 346
sp|P43288|ASKalfa          NYTEFKFPQIKAHPWHKIFHKRMPPEAVDLVSRLLQYSPNLRSAALDTLVHPFFDELRDP 359
                           *************** *:*: * ****: * ****:*:*. * : *:: .* ********

sp|P49841|GSK3B_HUMAN      NVKLPNGRDTPALFNFTTQELSSNPPLA-TILIPPHARIQAAASTPTNATAASDANTGDR 405
sp|Q9WV60|GSK3B_MOUSE      NVKLPNGRDTPALFNFTTQELSSNPPLA-TILIPPHARIQAAASPPANATAASDTNAGDR 405
sp|P43288|ASKalfa          NARLPNGRFLPPLFNFKPHELKGVPLEMVAKLVPEHARKQCPWLGL-------------- 405
                           *.:*****  * ****. :**.. *    : *:* *** *.

sp|P49841|GSK3B_HUMAN      GQTNNAASASASNST      420
sp|Q9WV60|GSK3B_MOUSE      GQTNNAASASASNST      420
sp|P43288|ASKalfa          ---------------      405
```

Fig. 11: Alignment of human/mouse GSK3β and ASKα. At the beginning of each line is the uniprot[23] identifier and the name of the protein whose sequence is in the respective line. The symbol below stands for the conservation of the according residues: the asterisk (*) stands for fully conserved residues, the colon (:) for conservation between residues that are similar in their properties and the period (.) for conservation between residues with slightly similar properties. At the end of each line, the residue number of the last amino acid in this line is given.[49]



```
1: sp|P49841|GSK3B_HUMAN  100.00   99.05   61.54
2: sp|Q9WV60|GSK3B_MOUSE   99.05  100.00   61.54
3: sp|P43288|ASKalfa       61.54   61.54  100.00
```

Fig. 12:Sequence identity matrix of the alignment in Fig. 11.

All the available structures have gaps. After intensively analysing all aspects of the ones with the highest resolution, especially the location of gaps, 6AE3 turned out to be the most obvious choice, although it is not the one with the very best resolution. The occupancy for each residue is 1,00, so there are no alternate rotamers for side chains.[50] The gaps are neither close to the phosphorylation location nor the TREE- region.

To validate experimentally determined structures, there are geometric quality criteria. In Fig. 13, residues are coloured depending on the number of these criteria in which they have at least one outlier. Green means there is no such criterion, yellow that there is one criterion, orange two and in red ones there are three or more with outliers. If there is a red dot above the residue, it poorly fits into the electron density.[51]

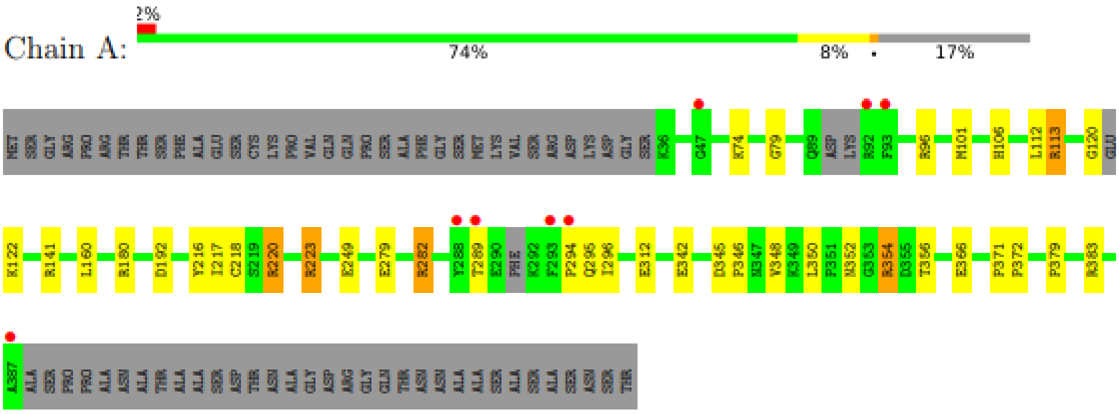


Fig. 13: Residue-property plot of chain A 6AE3.[51],[52–54] *See main text for explanation of the figure.*

The final choice of the template was chain A of 6AE3 (Fig. 14).[51] The reason for this selection was the actuality of the structure and it is one of the available structures with the highest resolution.

The missing ends of the template and model are no problem. The tyrosine that gets phosphorylated and the TREE-region are resolved, which are the important parts of the protein for further studies. With the high resolution as well as sequence identity, the main factors for a good model are given.

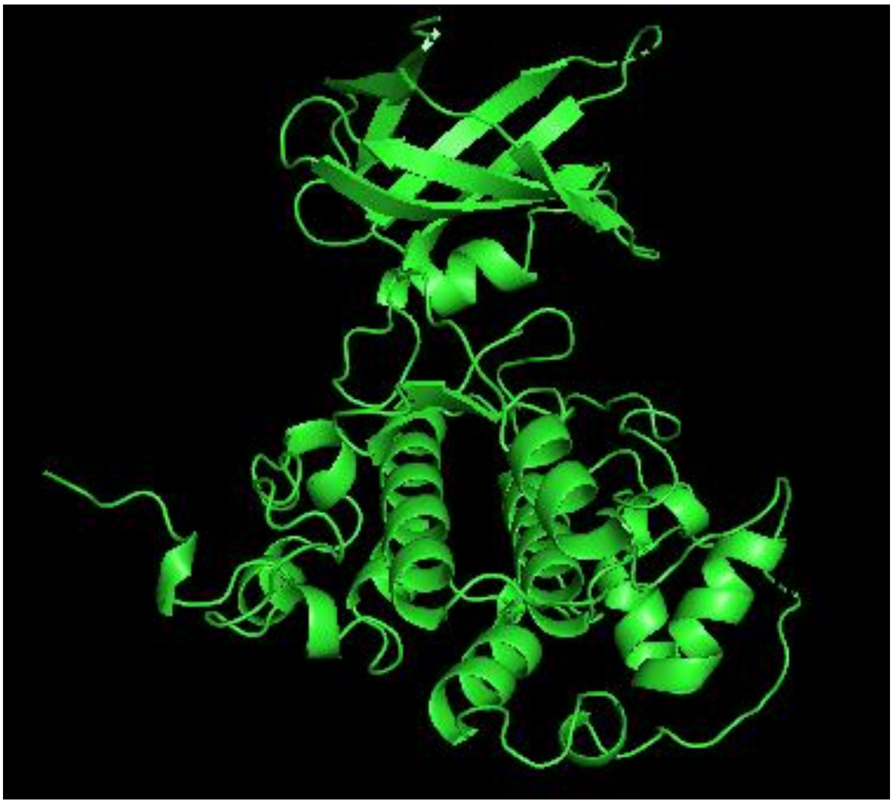


Fig. 14: Structure of a mouse GSK3β (PDB-ID 6AE3).

4.2. Sequence Alignment

4.2.1. PROMALS3D

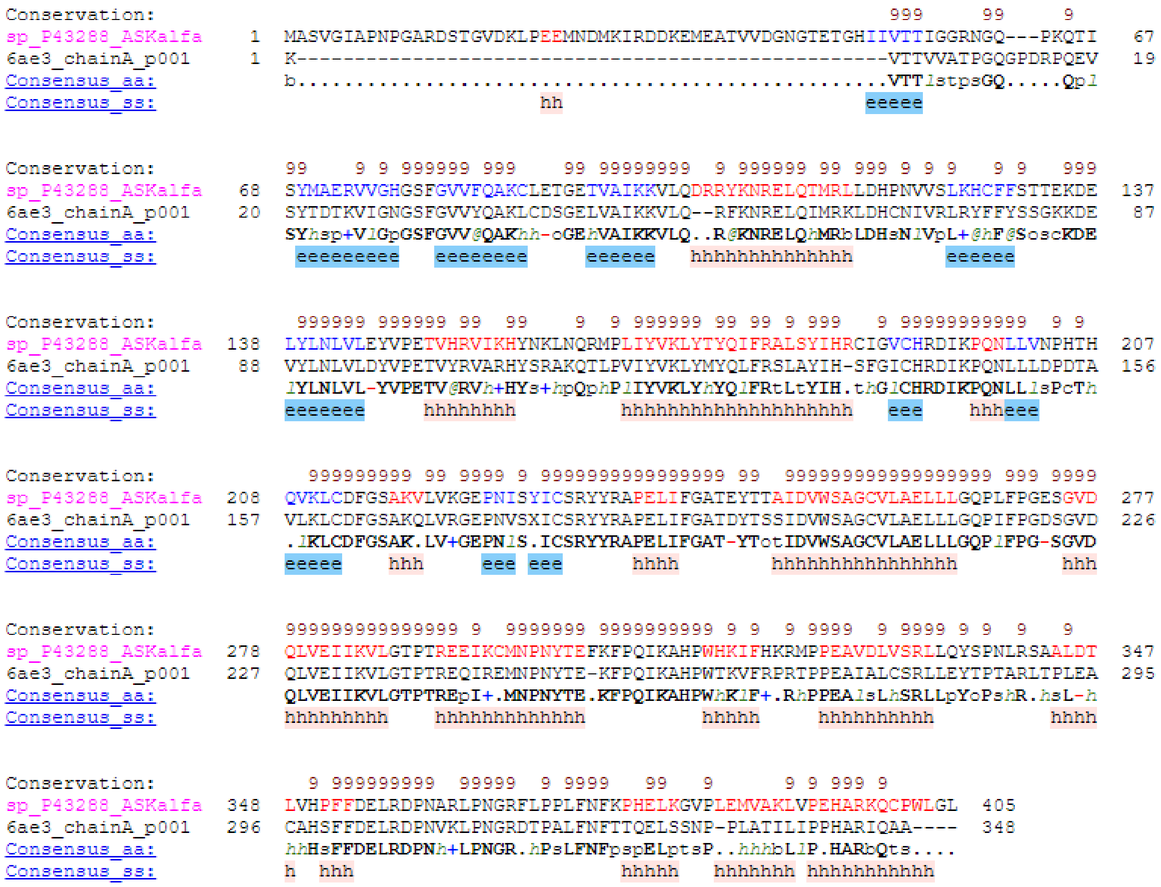
After choosing the template, the sequences of the target and the template have been aligned with Promals3d.[26] Fig. 15 shows an alignment, where the blue amino acids are part of a beta strand, the red ones of an alpha-helix. The "9" above the sequence is present, when the residue of both input proteins sequences match. At the beginning and end of each line the residue number of the first and last amino acid is shown. The Consensus_aa (consensus amino acid sequence) symbols are:

- conserved amino acids= bold and uppercase letters
- aliphatic (I, V, L)= l
- hydrophobic (W, F, Y, M, L, I, V, A, C, T, H)= h
- aromatic (Y, H, W, F)= @
- alcohol (S, T)= o
- polar residues (D, E, H, K, N, Q, R, S, T)= p
- tiny (A, G, C, S)= t
- small (A, G, C, S, V, N, D, T, P)= s
- bulky residues (E, F, I, K, L, M, Q, R, W, Y)= b

- charged (D, E, K, R, H)= c
- positively charged (K, R, H)= +
- negatively charged (D, E)= -

The Consensus_ss (consensus predicted secondary structure) at the bottom shows “h” for an alpha-helix and “e” for a beta-strand.[55]

The alignment has later been used in Jalview to cut off the termini.



```
Conservation:                                                                      999     99     9
sp_P43288_ASKalfa       1  MASVGIAPNPGARDSTGVDKLPEEMNDMKIRDDKEMEATVVDGNGTETGHIIVTTIGGRNGQ---PKQTI    67
6ae3_chainA_p001        1  K---------------------------------------------------VTTVVATPGQGPDRPQEV    19
Consensus_aa:              b...................................................VTTlstpsGQ.....Qpl
Consensus_ss:                                    hh                          eeeee

Conservation:              99    9 9 999999 999    99 99999999  9 999999 99 999 9 9 9   9 9   999
sp_P43288_ASKalfa      68  SYMAERVVGHGSFGVVFQAKCLETGETVAIKKVLQDRRYKNRELQTMRLLDHPNVVSLKHCFFSTTEKDE   137
6ae3_chainA_p001       20  SYTDTKVIGNGSFGVVYQAKLCDSGELVAIKKVLQ--RFKNRELQIMRKLDHCNIVRLRYFFYSSGKKDE    87
Consensus_aa:              SYhsp+VlGpGSFGVV@QAKhh-oGEhVAIKKVLQ..R@KNRELQhMRbLDHsNlVpL+@hF@SoscKDE
Consensus_ss:               eeeeeeeee   eeeeeeee     eeeeee   hhhhhhhhhhhhhh        eeeeee

Conservation:               999999 999999 99  99     9  9 999999 99 99 9 999   9 99999999999  9 9
sp_P43288_ASKalfa     138  LYLNLVLEYVPETVHRVIKHYNKLNQRMPLIYVKLYTYQIFRALSYIHRCIGVCHRDIKPQNLLVNPHTH   207
6ae3_chainA_p001       88  VYLNLVLDYVPETVYRVARHYSRAKQTLPVIYVKLYMYQLFRSLAYIH-SFGICHRDIKPQNLLLDPDTA   156
Consensus_aa:              lYLNLVL-YVPETV@RVh+HYs+hpQphPlIYVKLYhYQlFRtLtYIH.thGlCHRDIKPQNLLlsPcTh
Consensus_ss:              eeeeeee     hhhhhhhh         hhhhhhhhhhhhhhhhhhhh   eee    hhheee

Conservation:                999999999 99 9999 9 9999999999999999 99  999999999999999999 999 9999
sp_P43288_ASKalfa     208  QVKLCDFGSAKVLVKGEPNISYICSRYYRAPELIFGATEYTTAIDVWSAGCVLAELLLGQPLFPGESGVD   277
6ae3_chainA_p001      157  VLKLCDFGSAKQLVRGEPNVSXICSRYYRAPELIFGATDYTSSIDVWSAGCVLAELLLGQPIFPGDSGVD   226
Consensus_aa:              .lKLCDFGSAK.LV+GEPNlS.ICSRYYRAPELIFGAT-YTotIDVWSAGCVLAELLLGQPlFPG-SGVD
Consensus_ss:              eeeee    hhh     eee eee      hhhh        hhhhhhhhhhhhhhhh         hhh

Conservation:              999999999999999 9  9999999 9999999999 9 9  9 9999  9 9999 9 9  9   9
sp_P43288_ASKalfa     278  QLVEIIKVLGTPTREEIKCMNPNYTEFKFPQIKAHPWHKIFHKRMPPEAVDLVSRLLQYSPNLRSAALDT   347
6ae3_chainA_p001      227  QLVEIIKVLGTPTREQIREMNPNYTE-KFPQIKAHPWTKVFRPRTPPEAIALCSRLLEYTPTARLTPLEA   295
Consensus_aa:              QLVEIIKVLGTPTREpI+.MNPNYTE.KFPQIKAHPWhKlF+.RhPPEAlsLhSRLLpYoPshR.hsL-h
Consensus_ss:              hhhhhhhhh    hhhhhhhhhhhhh          hhhhh     hhhhhhhhhh          hhhh

Conservation:                9 999999999  99999  9 9999   99   9      9 9 999 9
sp_P43288_ASKalfa     348  LVHPFFDELRDPNARLPNGRFLPPLFNFKPHELKGVPLEMVAKLVPEHARKQCPWLGL  405
6ae3_chainA_p001      296  CAHSFFDELRDPNVKLPNGRDTPALFNFTTQELSSNP-PLATILIPPHARIQAA----  348
Consensus_aa:              hhHsFFDELRDPNh+LPNGR.hPsLFNFpspELptsP..hhhbLlP.HARbQts....
Consensus_ss:              h  hhh                       hhhhh   hhhhhhh hhhhhhhhhhh
```

Fig. 15: Coloured alignment in PROMALS3D.

There are four gaps in the template in the alignment of Fig. 15. Three of them happen because the template is not resolved in these parts (Q54-R55, H135-S136, E252-K253). The gap P332-P333 is caused by an insertion in the template, because there is one more residue in the target sequence. Insertions in a secondary structure, here an alpha helix, are not optimal. In this case, this helix is very far away from the important places in the protein such as the TREE-region (Fig. 16). Furthermore, the insertion is only at the edge of an alpha helix close to a loop region which is not considered a problem and has no effect on the model or later results in this project. It should be noted however, that this insertion might influence the helical geometry and should be taken into account in potential future studies.

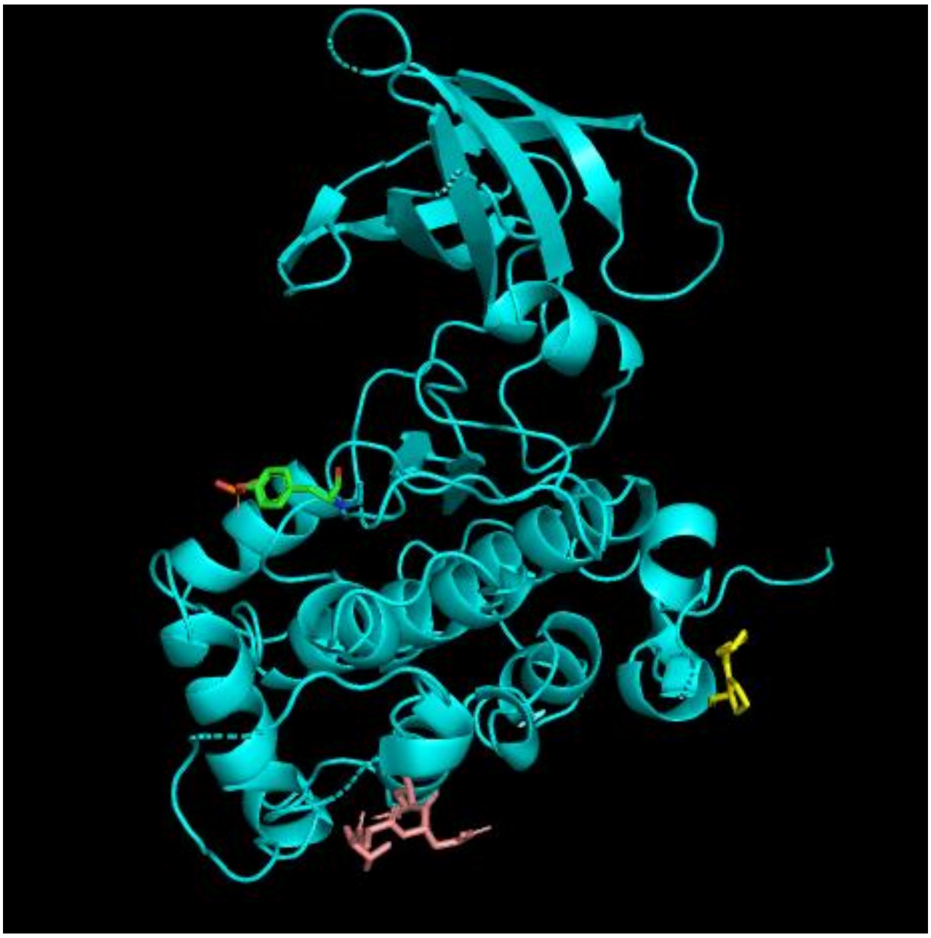


Fig. 16: 6ae3 chain A with the TREE region coloured in pink, the phosphorylated Y216 coloured in green and the region with the gap in yellow.

4.2.2. Jalview

In the absence of a template for the termini of the kinase, they were cut off with Jalview.[28] The alignment now starts with residue 53 of ASKα and ends with 400 (Fig. 17).

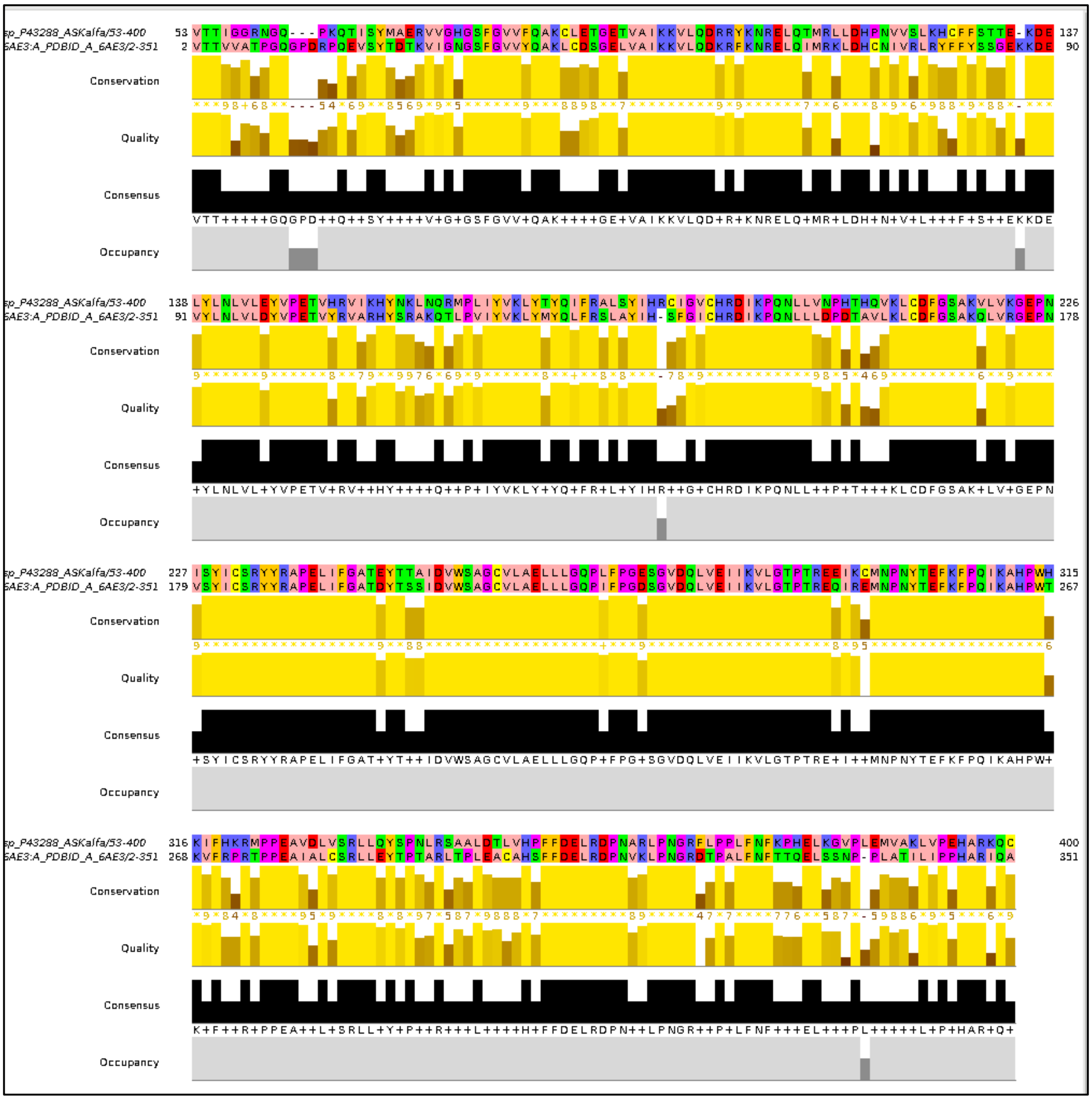


Fig. 17: Alignment of ASKα and GSK3β in Jalview[28] without termini. The conservation shows conserved residues as high yellow bars[56], the quality indicates the likeliness of finding a mutation that might be present.[57]The consensus bars show the percentage of the modal residue of each column.[58] The occupancy visualises gaps in the alignment.

4.3. Homology Modelling

After deleting the termini, the protein had to be renumbered, starting with residue 52. Also, in the PDB-file 6AE3 Y216 is phosphorylated. Modeller cannot execute the script due to this phosphorylation. For this reason, amino acid 216 was edited into an unphosphorylated Y216.

The number of models to be calculated is an individual decision. After calculating the models of ASKα, a script was run through all the models, looking for the minimum molpdf score (Appendix 7.2). Then minimum molpdf scores collected were plotted against the number of models calulated. The score values need to converge, so one can be sure that the lowest scores have been found already.

As mentioned before, we wanted to calculate a sufficient number of models to get the most similar models to the template with the highest probability possible. According to the moldpdf score, most similar means that the models include the lowest number of restraint violations Modeller can calculate. For a sufficient quantity of models, no smaller molpdf should be found anymore. In the case of this project, there was still a lower molpdf found after 500 (Fig. 18) and 1000 models (Fig. 19), so 2000 models were calculated. Within 2000 models (Fig. 20), the molpdf score converged and no lower score was found in more than 1000 models, so we decided this number is enough to include the best models Modeller is able to generate.

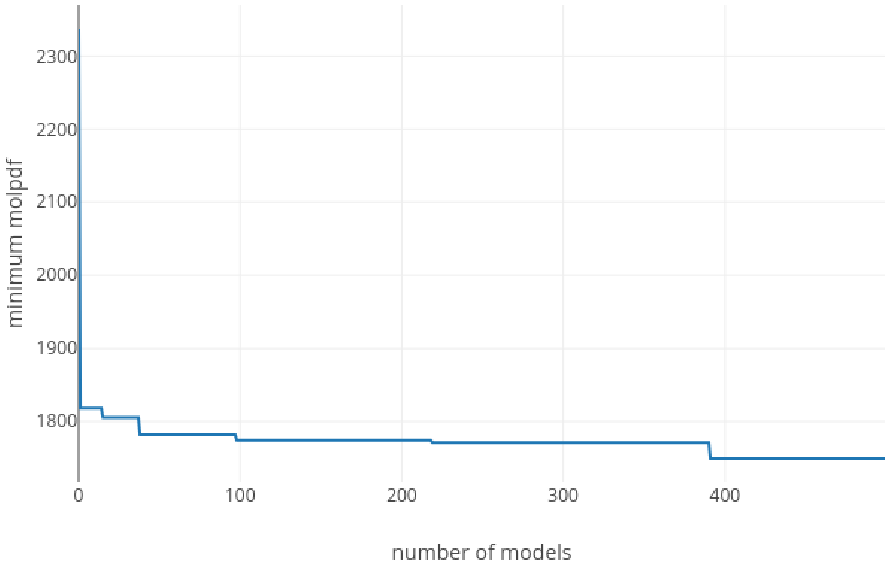


Fig. 18: Lowest moldpf scores after 500 models calculated.

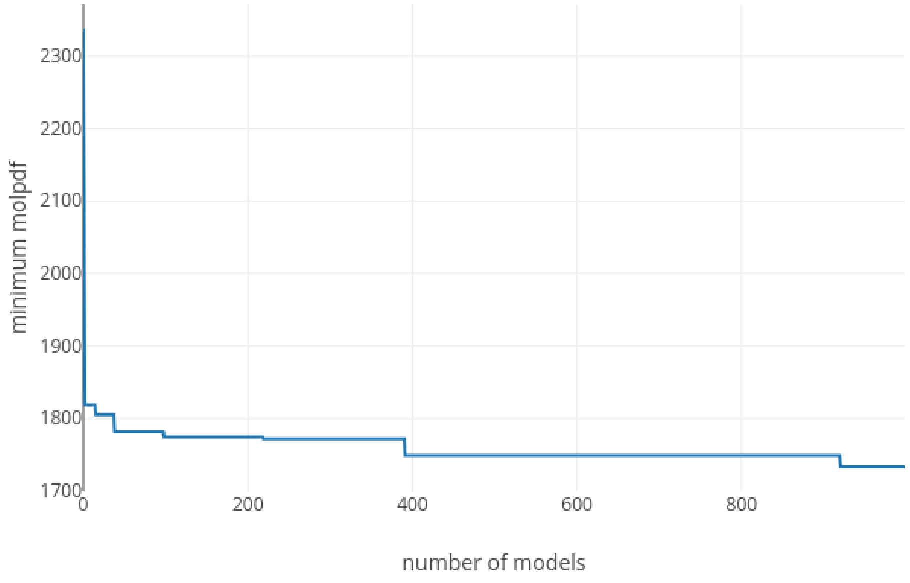


Fig. 19: Lowest moldpf scores after 1000 models calculated.

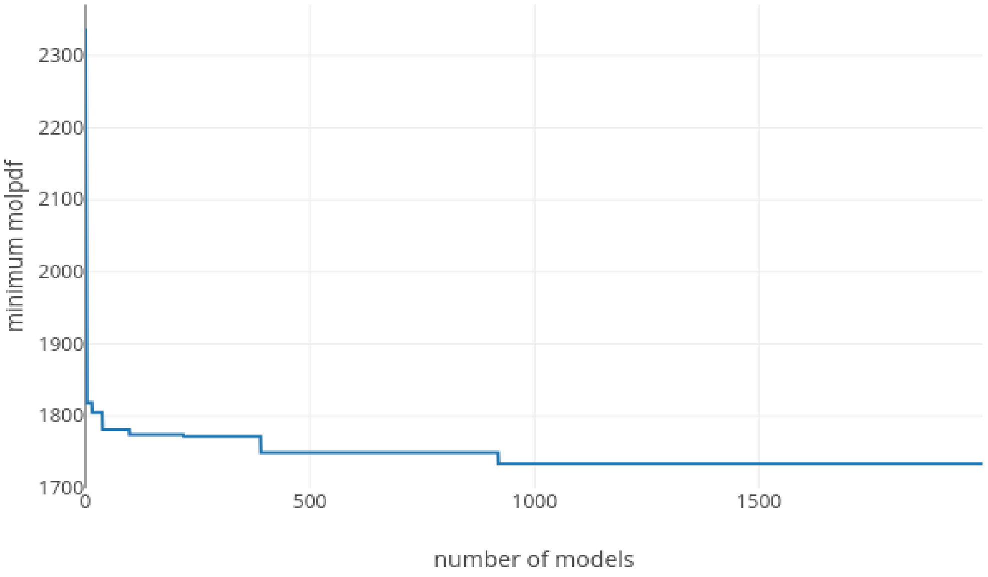
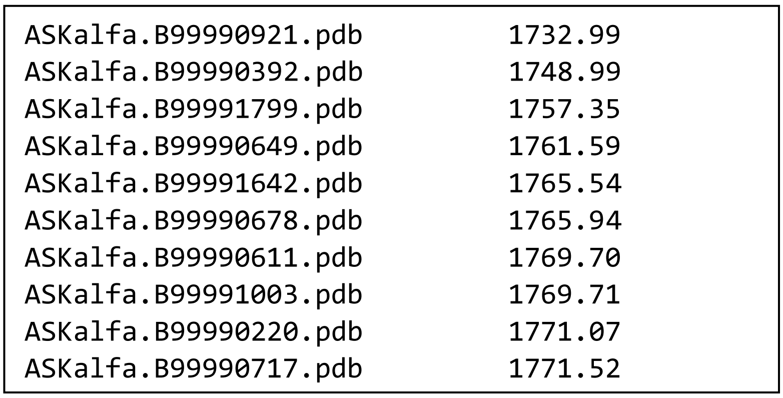


Fig. 20: Lowest moldpf scores after 2000 models calculated.

4.4. Model validation

4.4.1. Molpdf Score

Modeller automatically calculates the molpdf score for each model. A list of all model names as well as the according molpdf scores can be found in the log file after running the model calculation. The list was ranked by ascending molpdf scores. We decided to continue with the best 5%, meaning keeping the 100 models with the lowest molpdf scores. The range from best to worst molpdf score for the first 100 is 81,31. The top ten of them are shown in Fig. 21.



```
ASKalfa.B99990921.pdb         1732.99
ASKalfa.B99990392.pdb         1748.99
ASKalfa.B99991799.pdb         1757.35
ASKalfa.B99990649.pdb         1761.59
ASKalfa.B99991642.pdb         1765.54
ASKalfa.B99990678.pdb         1765.94
ASKalfa.B99990611.pdb         1769.70
ASKalfa.B99991003.pdb         1769.71
ASKalfa.B99990220.pdb         1771.07
ASKalfa.B99990717.pdb         1771.52
```

Fig. 21: Top ten models with the lowest molpdf score.

4.4.2. DOPE Score

The best 5% of the 2000 models were ranked by increasing DOPE scores. To restrict this number, the models with the lowest DOPE score were selected for further validation. To decide how many models to keep for the continuing evaluation, the mean value as well as the standard deviation of the DOPE score were calculated. Then we subtracted the standard deviation from the mean value and retained all models below this value as the best remaining 18 models (Table 2).

$\bar{x}$= -41382, 9356

s= 135,6206

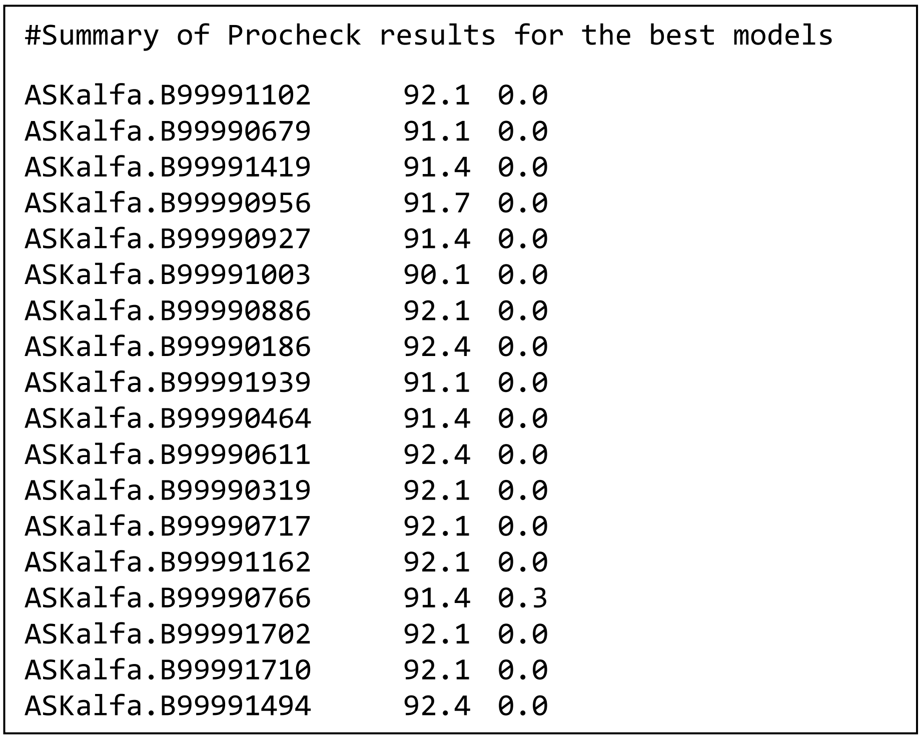
Model name	DOPE score
ASKalfa.B99991102.pdb	-41722.42
ASKalfa.B99990679.pdb	-41719.68
ASKalfa.B99991419.pdb	-41672.03
ASKalfa.B99990956.pdb	-41643.13
ASKalfa.B99990927.pdb	-41627.19

ASKalfa.B99991003.pdb	-41617.24
ASKalfa.B99990886.pdb	-41609.37
ASKalfa.B99990186.pdb	-41607.56
ASKalfa.B99991939.pdb	-41602.27
ASKalfa.B99990464.pdb	-41594.37
ASKalfa.B99990611.pdb	-41586.93
ASKalfa.B99990319.pdb	-41566.13
ASKalfa.B99990717.pdb	-41554.04
ASKalfa.B99991162.pdb	-41546.59
ASKalfa.B99990766.pdb	-41534.17
ASKalfa.B99991702.pdb	-41524.22
ASKalfa.B99991710.pdb	-41522.81
ASKalfa.B99991494.pdb	-41517.72

Table 2: Models with the lowest DOPE score that were kept for further evaluation.

4.4.3. Ramachandran Plots

Ramachandran plots were created for each of the remaining 18 models (Fig. 22). The values for non-proline and non-glycine residues in most favoured, additionally, generously and disallowed regions are all similar and extremely good, which is why the choice was made between many very good models. Also, the values of the Ramachandran plot of the template 6ae3, which consists of four chains, are in the same range as the calculated models (91,9% residues in most favoured regions, 7,9% in additional allowed regions, 0,2% in generously allowed regions and 0,0% in disallowed regions).



```
#Summary of Procheck results for the best models

ASKalfa.B99991102     92.1  0.0
ASKalfa.B99990679     91.1  0.0
ASKalfa.B99991419     91.4  0.0
ASKalfa.B99990956     91.7  0.0
ASKalfa.B99990927     91.4  0.0
ASKalfa.B99991003     90.1  0.0
ASKalfa.B99990886     92.1  0.0
ASKalfa.B99990186     92.4  0.0
ASKalfa.B99991939     91.1  0.0
ASKalfa.B99990464     91.4  0.0
ASKalfa.B99990611     92.4  0.0
ASKalfa.B99990319     92.1  0.0
ASKalfa.B99990717     92.1  0.0
ASKalfa.B99991162     92.1  0.0
ASKalfa.B99990766     91.4  0.3
ASKalfa.B99991702     92.1  0.0
ASKalfa.B99991710     92.1  0.0
ASKalfa.B99991494     92.4  0.0
```

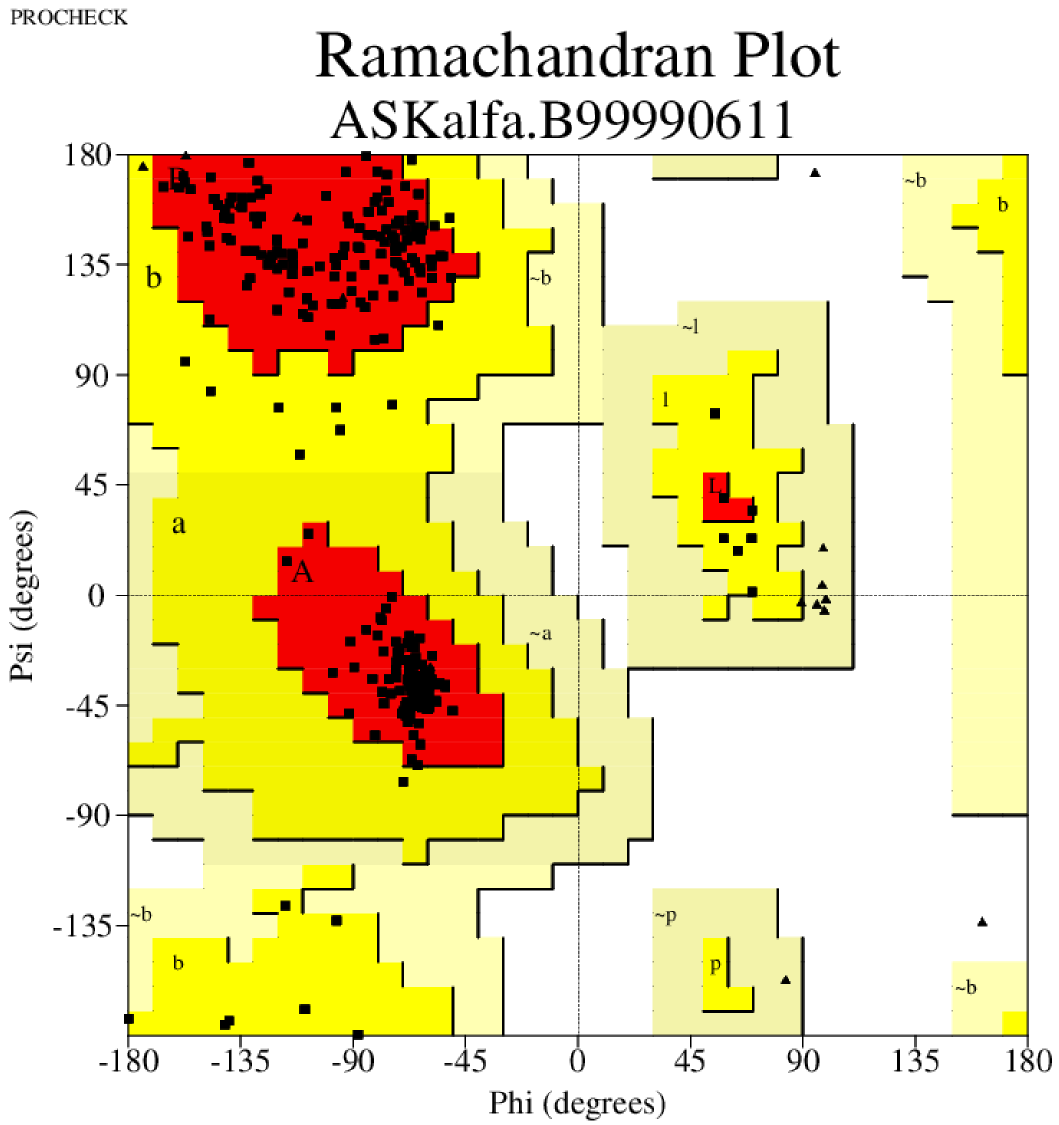
Fig. 22: Procheck results with model name, residues in most favoured and residues in disallowed regions.

Of all plots, the selection was limited to three models, which were those with the highest number of residues in the most favoured regions (Table 3).

Name	Most favoured	Additional allowed	Generously allowed	Unallowed
ASKalfa.B99990186	92,4 %	7,3	0,3 %	0,0 %
ASKalfa.B99990611	92,4 %	7,6 %	0,0 %	0,0 %
ASKalfa.B99991494	92,4 %	7,3 %	0,3 %	0,0 %

Table 3: Top three models with most residues in most favoured regions.

As shown in Fig. 23, most of non-proline and non-glycine residues, illustrated as black dots, are in the most favoured regions coloured in red. The others of these residues are in additionally allowed regions. Two glycine residues, illustrated as black triangles, are in disallowed regions, which are the white regions of the plot.



Plot statistics

Residues in most favoured regions [A,B,L]	280	92.4%
Residues in additional allowed regions [a,b,l,p]	23	7.6%
Residues in generously allowed regions [~a,~b,~l,~p]	0	0.0%
Residues in disallowed regions	0	0.0%
	----	------
Number of non-glycine and non-proline residues	303	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	18	
Number of proline residues	25	

Total number of residues	348	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms
and R-factor no greater than 20%, a good quality model would be expected
to have over 90% in the most favoured regions.

Fig. 23: Ramachandran plot of final selected model

4.5. Selection of Final Model

Since all remaining models are of high quality, the final decision was made based on the data of the residue-distribution in the plots. The selected model has the highest percentage of residues in most favoured and additionally allowed regions. A look at the molpdf and the DOPE scores also showed that this model is one of the top models according to these scores with a molpdf of 1769,71 and a DOPE score of -41586,93. So, the final selection was model number 611 (Fig. 24).

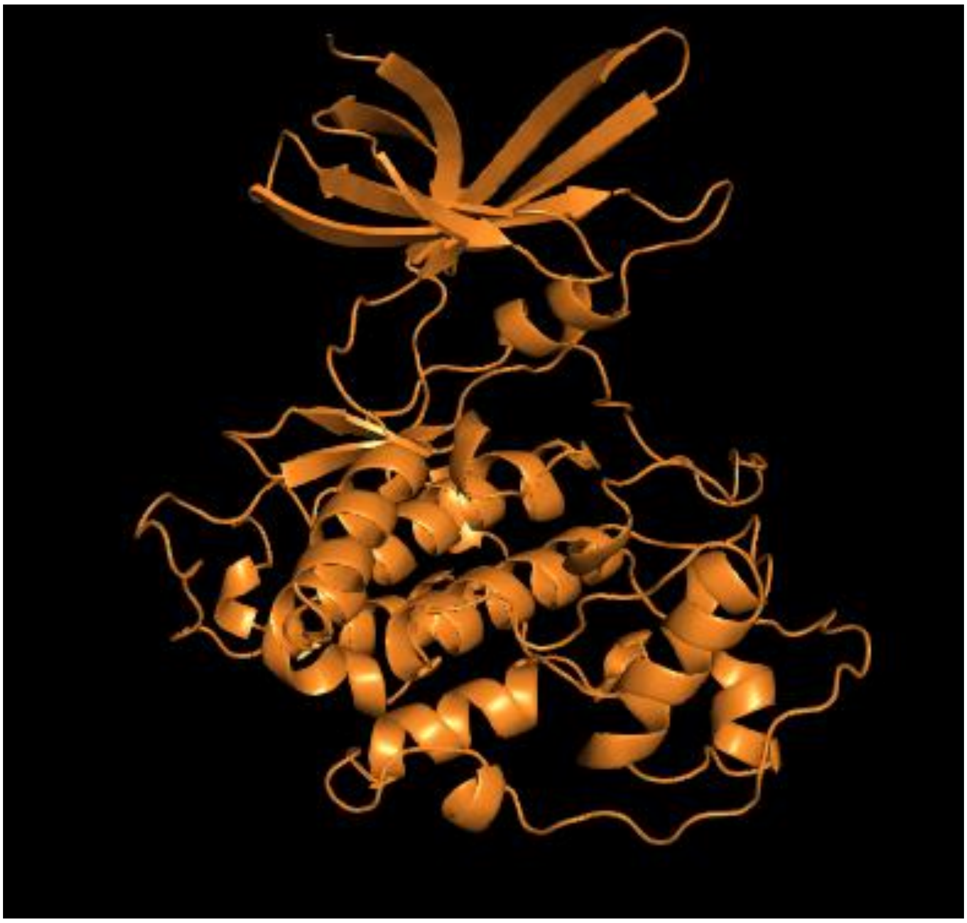
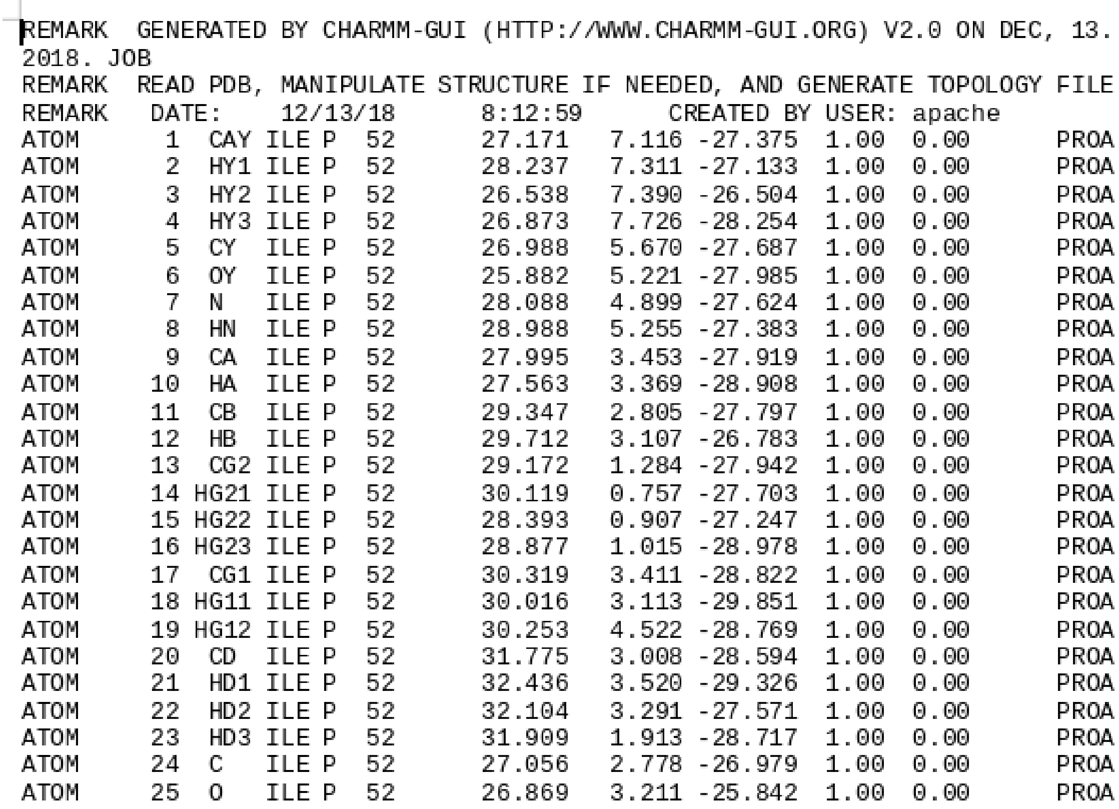


Fig. 24: Homology model of ASKα.

4.6. Model Adaptations

For adaptations of the final model, the PDB Reader in CHARMM-GUI was used. The output of the PDB reader is an updated PDB file with the executed modifications. All hydrogen atoms necessary have been added and Y229, which is important for an activated kinase got phosphorylated. The termini of the protein got capped so there is no chance they can react with any residues nearby (Fig. 25).



```
REMARK  GENERATED BY CHARMM-GUI (HTTP://WWW.CHARMM-GUI.ORG) V2.0 ON DEC, 13.
2018. JOB
REMARK  READ PDB, MANIPULATE STRUCTURE IF NEEDED, AND GENERATE TOPOLOGY FILE
REMARK   DATE:    12/13/18      8:12:59      CREATED BY USER: apache
ATOM      1  CAY ILE P  52      27.171   7.116 -27.375  1.00  0.00      PROA
ATOM      2  HY1 ILE P  52      28.237   7.311 -27.133  1.00  0.00      PROA
ATOM      3  HY2 ILE P  52      26.538   7.390 -26.504  1.00  0.00      PROA
ATOM      4  HY3 ILE P  52      26.873   7.726 -28.254  1.00  0.00      PROA
ATOM      5  CY  ILE P  52      26.988   5.670 -27.687  1.00  0.00      PROA
ATOM      6  OY  ILE P  52      25.882   5.221 -27.985  1.00  0.00      PROA
ATOM      7  N   ILE P  52      28.088   4.899 -27.624  1.00  0.00      PROA
ATOM      8  HN  ILE P  52      28.988   5.255 -27.383  1.00  0.00      PROA
ATOM      9  CA  ILE P  52      27.995   3.453 -27.919  1.00  0.00      PROA
ATOM     10  HA  ILE P  52      27.563   3.369 -28.908  1.00  0.00      PROA
ATOM     11  CB  ILE P  52      29.347   2.805 -27.797  1.00  0.00      PROA
ATOM     12  HB  ILE P  52      29.712   3.107 -26.783  1.00  0.00      PROA
ATOM     13  CG2 ILE P  52      29.172   1.284 -27.942  1.00  0.00      PROA
ATOM     14 HG21 ILE P  52      30.119   0.757 -27.703  1.00  0.00      PROA
ATOM     15 HG22 ILE P  52      28.393   0.907 -27.247  1.00  0.00      PROA
ATOM     16 HG23 ILE P  52      28.877   1.015 -28.978  1.00  0.00      PROA
ATOM     17  CG1 ILE P  52      30.319   3.411 -28.822  1.00  0.00      PROA
ATOM     18 HG11 ILE P  52      30.016   3.113 -29.851  1.00  0.00      PROA
ATOM     19 HG12 ILE P  52      30.253   4.522 -28.769  1.00  0.00      PROA
ATOM     20  CD  ILE P  52      31.775   3.008 -28.594  1.00  0.00      PROA
ATOM     21  HD1 ILE P  52      32.436   3.520 -29.326  1.00  0.00      PROA
ATOM     22  HD2 ILE P  52      32.104   3.291 -27.571  1.00  0.00      PROA
ATOM     23  HD3 ILE P  52      31.909   1.913 -28.717  1.00  0.00      PROA
ATOM     24  C   ILE P  52      27.056   2.778 -26.979  1.00  0.00      PROA
ATOM     25  O   ILE P  52      26.869   3.211 -25.842  1.00  0.00      PROA
```

Fig. 25: Excerpt of the PDB file of the final model with an acetyl group on the N-terminus and the added hydrogens.

In addition, residue 292 was mutated from glutamic acid to lysine to compare the native and mutated kinase.

4.7. Pocket Detection with Fpocket

After finalising the native and mutated protein, both were examined for possible binding pockets with Fpocket. Possible binding pockets are illustrated as spheres in different colours (Fig. 26,Fig. 27). Each pocket is characterised by different scores, which give information about its potential as a binding pocket (Table 4,Table 5).

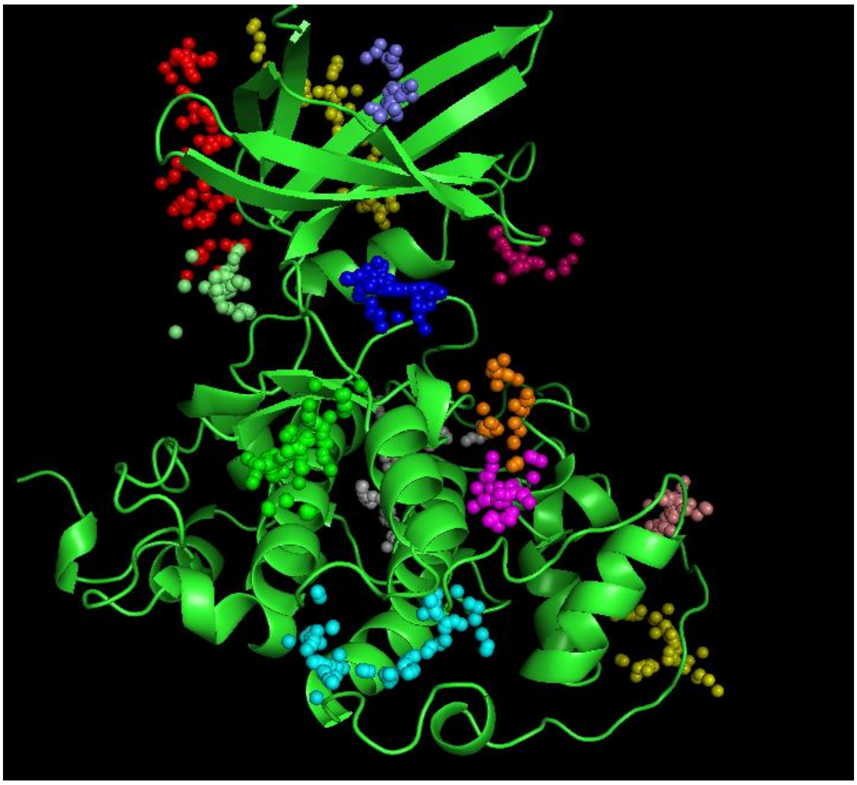


Fig. 26: Native kinase with its potential binding pockets

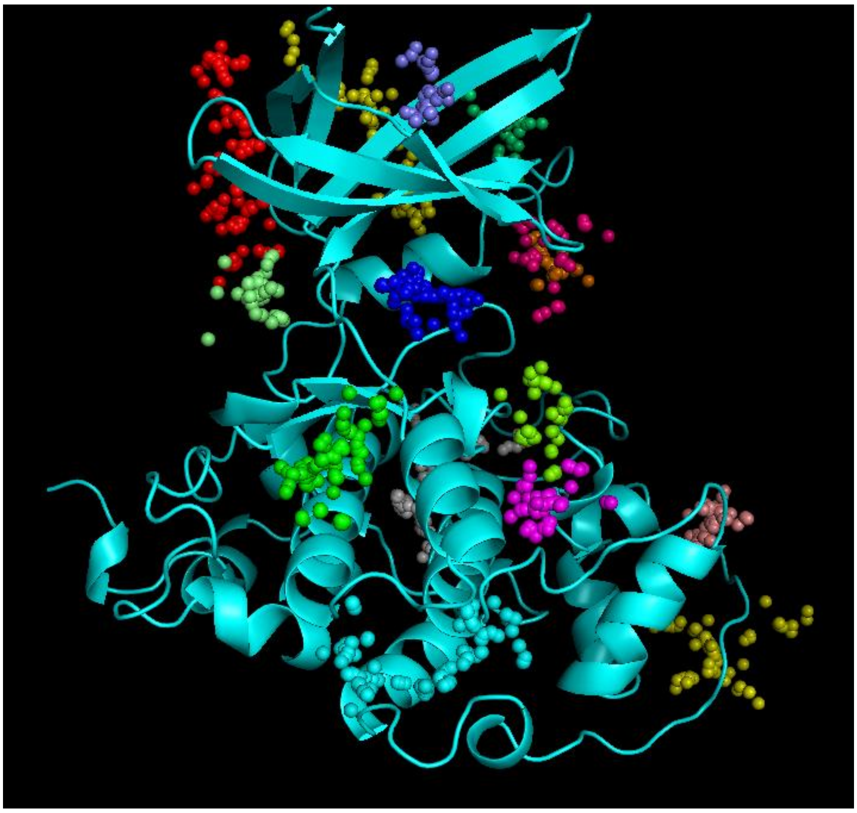


Fig. 27: Kinase including the E292K mutation with its potential binding pockets

Pocket	Score	Druggability score	Number of alpha spheres
1	37.41	0.232	107
2	35.38	0.606	89
3	33.84	0.501	96
4	31.35	0.131	94
5	28.69	0.161	97
6	24.03	0.102	74
7	23.11	0.043	93
8	20.54	0.019	47
9	18.61	0.088	51
10	18.36	0.025	41
11	16.55	0.016	44
12	16.11	0.012	49
13	11.29	0.077	39

Table 4: Scores of the pockets from the native kinase

Pocket	Score	Druggability score	Number of alpha spheres
1	38.33	0.348	115
2	33.97	0.618	90
3	31.97	0.501	96
4	29.06	0.127	91
5	27.38	0.230	99
6	24.44	0.128	85
7	21.3	0.043	93
8	19.8	0.033	49
9	17.62	0.088	51
10	17.57	0.025	41
11	15.7	0.016	44
12	14.13	0.022	41
13	12.32	0.008	37
14	10.54	0.077	39
15	9.262	0.019	49

Table 5: Scores of the pockets from the mutated kinase

In the mutated kinase, two more binding pockets have been found. In both proteins, the druggability score is for only two pockets higher than 0,5 (pocket 2 and 3). Small, drug-like molecules are therefore more likely to bind in these two parts of the kinase. Pocket 2, coloured in dark blue, is in the ATP binding site of the kinase. Therefore, molecules binding here are more likely to be competitive inhibitors than activators.

5. Discussion and Conclusion

As mentioned at the beginning, ASKα plays an important role in stress tolerance. Therefore, this area of research is very promising and of highest importance. Scientific breakthroughs in this field would make a major socio-economic contribution.

A lot of efforts were devoted for finding a suitable template that would deliver the best models possible. For this purpose, a wide variety of alignments were created and the available PDB structures were examined carefully in detail with regard to their resolution and also the unresolved parts. Different approaches were used to generate the final target-template sequence alignment showing a reasonable sequence identity of 61,54%. The models calculated by Modeller were validated carefully according to various criteria until the most suitable model could be selected. After refinement, this model was used to search for binding pockets in the protein.

Conclusively, many well-chosen methods were used and explored to obtain a reliable model of the 3D structure of ASKα. For an exact 3D representation of the protein, however, experimental methods for structure elucidation would of course be important. Due to the high sequence identity of the template to the target, one can be sure that the model determined by homology modelling comes very close to the natural structure.

This work is part of a collaboration with the AIT (Austrian institute of technology), where the stress tolerance of plants is intensively researched and experiments are continuously carried out to understand the activity of ASKα. Based on the homology model of the kinase, different ligands for the protein are currently being investigated as part of a master thesis. These molecules will then be tested experimentally at the AIT.

6. References

1. Buchanan, B. B., Gruissem, W. & Jones, R. L. Biochemistry & molecular biology of plants. 2441–2443. (2015).

2. Krasensky, J. & Jonak, C. Drought, salt, and temperature stress-induced metabolic rearrangements and regulatory networks. *J. Exp. Bot.* **63**, 1593–1608 (2012).

3. Strasburger, E., Noll, F., Schenck, H. & Schimper, A. F. . *Lehrbuch der Botanik*. 952. (2008).

4. Hussain, S. S., Ali, M., Ahmad, M. & Siddique, K. H. M. Polyamines: natural and engineered abiotic and biotic stress tolerance in plants. *Biotechnol. Adv.* **29**, 300–11 (2011).

5. Saidi, Y., Hearn, T. J. & Coates, J. C. Function and evolution of 'green' GSK3/Shaggy-like kinases. *Trends Plant Sci.* **17**, 39–46 (2012).

6. Palomo, V. *et al.* Exploring the binding sites of glycogen synthase kinase 3. identification and characterization of allosteric modulation cavities. *J. Med. Chem.* **54**, 8461–8470 (2011).

7. Goodman, H. M., Ecker, J. R. & Dean, C. The genome of Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 10831–5 (1995).

8. Dal Santo, S. *et al.* Stress-Induced GSK3 Regulates the Redox Stress Response by Phosphorylating Glucose-6-Phosphate Dehydrogenase in Arabidopsis. *Plant Cell* **24**, 3380–3392 (2012).

9. Stampfl, H., Fritz, M., Dal Santo, S. & Jonak, C. The GSK3/Shaggy-like kinase ASKα contributes to pattern-triggered immunity in Arabidopsis thaliana. *Plant Physiol.* **171**, pp.01741.2015 (2016).

10. Lamb, C. & Dixon, R. A. THE OXIDATIVE BURST IN PLANT DISEASE RESISTANCE. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **48**, 251–275 (1997).

11. Li, C., Zhang, B., Chen, B., Ji, L. & Yu, H. Site-specific phosphorylation of TRANSPARENT TESTA GLABRA1 mediates carbon partitioning in Arabidopsis seeds. *Nat. Commun.* **9**, (2018).

12. Dressel, A. & Hemleben, V. Transparent Testa Glabra 1 (TTG1) and *TTG1* -like genes in *Matthiola incana* R. Br. and related Brassicaceae and mutation in the WD-40 motif. *Plant Biol.* **11**, 204–212 (2009).

13. Müller-Esterl, W. Biochemie. 120–122 (2009).

14. Subramaniam, S., Earl, L. A., Falconieri, V., Milne, J. L. & Egelman, E. H. Resolution advances in cryo-EM enable application to drug discovery. *Curr. Opin. Struct. Biol.* **41**, 194–202 (2016).

15. Berman, H. M. *et al.* The Protein Data Bank Helen. *Nucleic Acids Res.* **28**, 235–242 (2000).

16. Krieger, E., Nabuurs, S. B. & Vriend, G. Chapter 25: homology modeling. *Struct. Bioinforma.* **44**, 507–521 (2003).

17. Saxena, A., Sangwan, R. S. & Mishra, S. Fundamentals of Homology Modeling Steps and Comparison among Important Bioinformatics Tools: An Overview. *Science International* **1**, 237–252 (2013).

18. Eswar, N. *et al.* Comparative Protein Structure Modeling Using Modeller*. Curr Protoc Bioinformatics* (2006).

19. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.

20. The AxPyMOL Molecular Graphics Plugin for PowerPoint, Version 2.0 Schrödinger, LLC.

21. Humphrey, W., Dalke, A. and Schulten, K., 'VMD - Visual Molecular Dynamics', J. Molec. Graphics, 1996, vol. 14, pp. 33-38.

22. What is VMD? Available at: https://www.ks.uiuc.edu/Research/vmd/allversions/what_is_vmd.html. (Accessed: 25th March 2019)

23. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

24. Meng, L., Sun, F., Zhang, X. & Waterman, M. S. Sequence alignment as hypothesis testing. *J. Comput. Biol.* **18**, 677–91 (2011).

25. Clamp, M., Cuff, J., Searle, S. M. & Barton, G. J. The Jalview Java alignment editor. *Bioinforma. Appl. NOTE* **20**, 426–427 (2004).

26. Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: A tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36**, 2295–2300 (2008).

27. PROMALS3D Documentation. Available at: http://prodata.swmed.edu/promals3d/info/promals3d_help.html#output1. (Accessed: 25th March 2019)

28. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).

29. Goujon, M. *et al.* A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.* **38**, W695–W699 (2010).

30. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539–539 (2014).

31. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).

32. Emboss needle. Available at: http://www.bioinformatics.nl/cgi-bin/emboss/help/needle. (Accessed:28th February 2019)

33. Šali, A. & Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **234**, 779–815 (1993).

34. Modeller. Available at: https://salilab.org/modeller/tutorial/basic.html. (Accessed: 14th February 2019)

35. Modeller. Available at: https://salilab.org/modeller/tutorial/cryoem/assess.html. (Accessed: 14th February 2019)

36. Modeller. Available at: https://salilab.org/modeller/9.11/manual/node468.html. (Accessed:14th February 2019)

37. Shen, M. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15**, 2507–2524 (2006).

38. Modeller. Availabele at: https://salilab.org/archives/modeller_usage/2009/msg00202.html. (Accessed: 14th February 2019)

39. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).

40. Brooks, B. R. *et al.* CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **30**, 1545–614 (2009).

41. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Comput. Biol.* **13**, e1005659 (2017).

42. Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* **91**, 43–56 (1995).

43. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29**, 1859–1865 (2008).

44. Jo, S. *et al.* CHARMM-GUI PDB Manipulator for Advanced Modeling and Simulations of Proteins Containing Nonstandard Residues. in 235–265 (2014).

45. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 168 (2009).

46. Le Guilloux, V., Schmidtke, P. & Tufféry, P. *fpocket Users' Manual*. (2008).

47. UniProt. Available at: https://www.uniprot.org/uniprot/P43288. (Accessed:18th February 2019)

48. Uniprot. Available at: https://www.uniprot.org/uniprot/Q9WV60. (Accessed:18th February 2019)

49. Uniprot. Available at: https://www.uniprot.org/help/sequence-alignments. (Accessed:13th March 2019)

50. B-factor data bank (BDB). Available at: http://www.cmbi.ru.nl/bdb/theory/. (Accessed:19th February 2019)

51. Kim, K. *et al.* Crystal structure of GSK3β in complex with the flavonoid, morin. *Biochem. Biophys. Res. Commun.* **504**, 519–524 (2018).

52. Burley, S. K. *et al.* Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res.* **47**, D520–D528 (2019).

53. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Mol. Biol.* **10**, 980–980 (2003).

54. Berman, H., Henrick, K., Nakamura, H. & Markley, J. L. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* **35**, D301–D303 (2007).

55. PROMALS3D. Available at: http://prodata.swmed.edu/promals3d/info/promals3d_help.html#output1. (Accessed:18th January 2019)

56. Jalview. Available at: http://www.jalview.org/version118/documentation.html. (Accessed:13th March 2019)

57. Jalview. Available at: http://www.jalview.org/help/html/calculations/quality.html. (Accessed:13th March 2019)

58. Jalview. Available at: http://www.jalview.org/help/html/calculations/consensus.html. (Accessed: 13th March 2019

7. Appendix

7.1. List templates PDB

PDB-ID	Å	Mutation	Ligand	Paper
6AE3	2,14	0	Morin	(2018) Biochem Biophys Res Commun 504 519-524
6GN1	2,6	0	PIK-75	(2018) Angew Chem Int Ed Engl 57 9970-9975
5KPK	2,4	0	BRD0209	(2018) Sci Transl Med 10
5KPL	2,6	0	BRD0705	(2018) Sci Transl Med 10
5KPM	2,69	0	BRD3731	(2018) Sci Transl Med 10
5T31	2,85	1	75F	(2018) Sci Transl Med 10
6B8J	2,6	0	65C	(2017) J Med Chem 60 8482-8514
5OY4	3,2	0	B4K	(2017) ACS Med Chem Lett 8 1093-1098
5K5N	2,2	1	6QH	(2016) Angew Chem Int Ed Engl 55 9601-9605
5HLN	3,1	0	CHIR99021	(2016) ACS Chem Biol 11 1952-1963
5HLP	2,45	0	BRD3937	(2016) ACS Chem Biol 11 1952-1963
5F94	2,51	0	3UO	(2016) J Med Chem 59 1041-1051
5F95	2,53	0	3UP	(2016) J Med Chem 59 1041-1051
4PTC	2,71	0	2WE	(2015) Bioorg Med Chem Lett 25 1856-1863
4PTE	2,03	0	2WF	(2015) Bioorg Med Chem Lett 25 1856-1863
4PTG	2,36	0	2WG	(2015) Bioorg Med Chem Lett 25 1856-1863
5AIR	2,53	0	LRP6 peptide	(2015) Biodesign 3: 55
4NM0	2,5	0	Axin	(2014) Elife 3 e01998-e01998
4NM3	2,1	0	Axin, ps9 peptide	(2014) Elife 3: e01998-e01998
4NM5	2,3	0	Axin, LRP6 c-motif	(2014) Elife 3 e01998-e01998
4NM7	2,3	0	Axin, LRP6e-motif	(2014) Elife 3 e01998-e01998
4NU1	2,5	0	Axin, ps9 peptide	(2014) Elife 3 e01998-e01998
4IQ6	3,12	0	IQ6	(2013) ACS Med Chem Lett 4 211-215
4J1R	2,7	0	I5R	to be published
4J71	2,31	0	inhibitor 1R	to be published

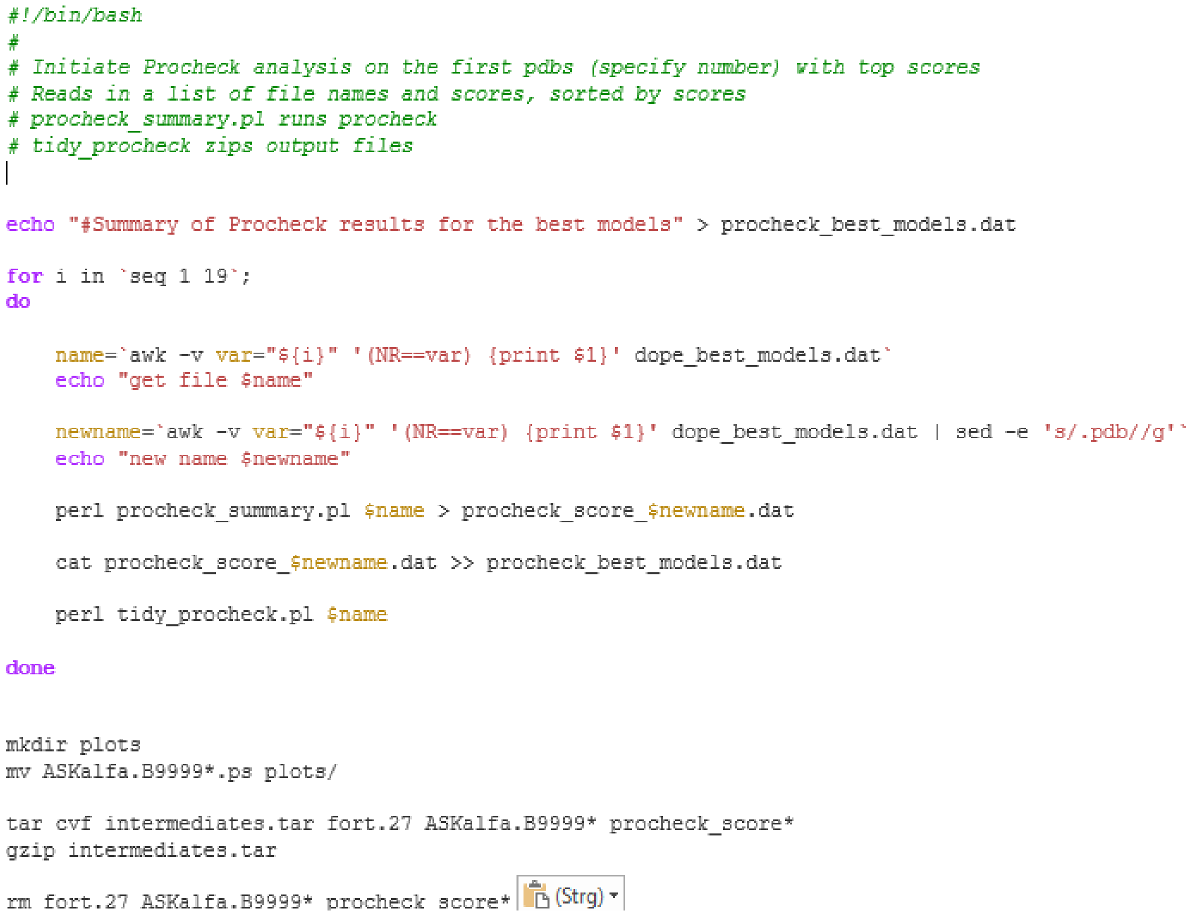
4B7T	2,77	0	Axin, Leucettine L4	(2012) J.Med.Chem. 55: 9312
3ZDI	2,65	0	Axin, Inhibitor 7d	(2013) J.Med.Chem. 56: 264
3SAY	2,23	0	Inhibitor 142	to be published
4ACC	2,21	0	Inhibitor 7YG	(2012) J Med Chem 55 9107-9119
4ACD	2,6	0	Inhibitor GR9	(2012) J. Med. Chem. 55: 9107-9119
4ACG	2,6	0	Inhibitor GLQ	(2012) J. Med. Chem. 55: 9107-9119
4ACH	2,6	0	Inhibitor KDI	(2012) J Med Chem 55 9107-9119
4AFJ	1,98	0	SJJ	(2012) Bioorg Med Chem Lett 22 1989
4DIT	2,6	0	Imidazopyridine inhibitor	to be published
3SD0	2,7	0	Inhibitor TSK,EPE	to be published
3Q3B	2,7	0	55E	(2011) Bioorg.Med.Chem.Lett. 21: 1429-1433
3M1S	3,13	0	Ruthenium pyridocarbazole	(2011) J.Biol.Inorg.Chem. 16: 45-50
3PUP	2,99	0	Ruthenium octasporine	(2011) J.Am.Chem.Soc. 133: 5976-5986
3GB2	2,4	0	Inhibitor G3B	(2009) J Med Chem 52 6270-6286
3L1S	2,9	0	Z92	(2010) Bioorg.Med.Chem.Lett. 20: 1661-1664
3I4B	2,3	0	pyrimidylpyrrole	(2009) J.Med.Chem. 52: 6362-6368
3F7Z	2,4	0	inhibitor 34O	(2009) Bioorg.Med.Chem. 17: 2017-2029
3F88	2,6	0	3HT, 2HT	(2009) Bioorg.Med.Chem. 17: 2017-2029
3DU8	2,2	0	NMS-869553A	(2009) J.Med.Chem. 52: 293-307
2OW3	2,8	0	BIM	(2007) Bioorg.Med.Chem.Lett. 17: 2863-2868
1R0E	2,25	0	DFN, FLC	to be published
1Q5K	1,94	0	TMU	(2003) J.Biol.Chem. 278: 45937-45945
1PYX	2,4	0	AMP-PNP	(2003) J.Mol.Biol. 333: 393-407
1Q3D	2,2	0	Staurosporine	(2003) J.Mol.Biol. 333: 393-407
1Q3W	2,3	0	Alsterpaullone	(2003) J.Mol.Biol. 333: 393-407
1Q41	2,1	0	Indirubin-3'-monoxime	(2003) J.Mol.Biol. 333: 393-407
1Q4L	2,77	0	Inhibitor I-5	(2003) J.Mol.Biol. 333: 393-407

1O9U	2,4	0	Axin	(2003) EMBO J 22 494
1GNG	2,6	0	FRATide peptide	(2001) Structure 9: 1143
1H8F	2,8	0	Hepes	(2001) Cell 105: 721
1I09	2,7	0	-	(2001) Nat.Struct.Mol.Biol. 8: 593-596

7.2. Script get_min_molpdf

```
#!/bin/bash
#
#Use the model number order and pick the minimum molpdf score as you
#go through each model by model number
#Usage: get_min_molpdf.sh name nummodels
#


num=$1 #Number of models
#num=500 #Number of models
name=$2
#name="hSERT_ceo.B9999"
min=1000000 #Initialize 'minimum so far' number

# tidy up output file
rm get_min_molpdf.txt

for i in `seq 1 $1`
do

#Get the file name in order
  if  [ $i -lt 10 ]; then
    newname=${name}000${i}.pdb
    echo "$newname"

    elif [ $i -lt 100 ]; then
    newname=${name}00${i}.pdb
    echo "$newname"



    else
    newname=${name}0${i}.pdb
    echo "$newname"

  fi

  score=`grep OBJECTIVE $newname |awk '{print $6}'`
  echo "score is: $score"

  if (( $(echo "$score < $min" |bc -l) )); then
    min=$score
    echo "minimum so far is: $min"

  else
    echo "minimum is still: $min since file $newname"
  fi

  echo "$newname $min" >> get_min_molpdf.txt

done
```

7.3. Script to organize Procheck output



```
#!/bin/bash
#
# Initiate Procheck analysis on the first pdbs (specify number) with top scores
# Reads in a list of file names and scores, sorted by scores
# procheck_summary.pl runs procheck
# tidy_procheck zips output files

echo "#Summary of Procheck results for the best models" > procheck_best_models.dat

for i in `seq 1 19`;
do

    name=`awk -v var="${i}" '(NR==var) {print $1}' dope_best_models.dat`
    echo "get file $name"

    newname=`awk -v var="${i}" '(NR==var) {print $1}' dope_best_models.dat | sed -e 's/.pdb//g'`
    echo "new name $newname"

    perl procheck_summary.pl $name > procheck_score_$newname.dat

    cat procheck_score_$newname.dat >> procheck_best_models.dat

    perl tidy_procheck.pl $name

done


mkdir plots
mv ASKalfa.B9999*.ps plots/

tar cvf intermediates.tar fort.27 ASKalfa.B9999* procheck_score*
gzip intermediates.tar

rm fort.27 ASKalfa.B9999* procheck_score*
```