



universität  
wien

# DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„Virtual Screening using a homology model of the  
outward occluded conformation of the creatine  
transporter (SLC6A8)“

verfasst von / submitted by

Sarah Mayr

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of  
Magistra der Pharmazie (Mag.pharm.)

Wien, 2021 / Vienna, 2021

Studienkennzahl lt. Studienblatt /  
degree programme code as it appears on  
the student record sheet:

UA 449

Studienrichtung lt. Studienblatt /  
degree programme as it appears on  
the student record sheet:

Diplomstudium Pharmazie

Betreut von / Supervisor:

Univ.-Prof. Mag. Dr. Gerhard Ecker



## Acknowledgements

In view of the fact of the pandemic of covid 19, I was forced to change both the topic and the "where" of the thesis at short notice, in view of the already planned Erasmus semester in France. In a very short time, Prof. Dr. Ecker gave me the opportunity to become part of his research group. The demanding and ambitious work that some team members have been doing for many years has won my fully respect. I thank Prof. Dr. Ecker for allowing me to be part of this research group, which has broadened my perspective and has got an important part of my student career that I would not want to miss.

I would like to thank my co-supervisor, Dr Claire Colas, for her infinite patience and instructions, with which she lead a true novice in computer science to a result, that is evident in this work. Her vast knowledge of SLC proteins and the knowledge of molecular modelling left me stunning to this day. In my eyes, she is a forward looking woman, who lives research to the fullest, which impresses me.

In addition, I would like thank the entire team for all the open ears and always helping hands, when needed. Unfortunately, this extraordinary circumstances made it impossible to get to know each other better.

Finally, I thank my family and especially my mother for their endless support throughout my studies.

---

## Table of Content

<b>1</b>	<b>Introduction .....</b>	<b>1</b>
1.1	<i>The system of energy supply .....</i>	1
1.2	<i>Creatine Synthesis and uptake .....</i>	1
1.3	<i>Function of Creatine.....</i>	2
1.4	<i>Creatine Deficiency Syndrome.....</i>	2
1.4.1	Clinical features.....	3
1.5	<i>Solute carrier transporters (SLCs) .....</i>	4
1.5.1	LeuT-like fold.....	4
1.6	<i>SLC6 Transporters.....</i>	6
1.7	<i>Creatine Transporter (SLC6A8).....</i>	7
1.7.1	The homology model of SLC6A8.....	7
<b>2</b>	<b>Aim of the Diploma Thesis.....</b>	<b>11</b>
<b>3</b>	<b>Materials and Methods .....</b>	<b>12</b>
3.1	<i>Databases .....</i>	12
3.1.1	The Protein Data Bank.....	12
3.1.2	Drugbank.....	12
3.1.3	Enamine database.....	12
3.2	<i>Vizualization in PyMOL.....</i>	13
3.3	<i>Structure based virtual screening .....</i>	13
3.3.1	The search algorithm.....	14
3.3.2	Scoring functions.....	14
3.4	<i>Molecular docking in Maestro.....</i>	16
3.4.1	The calculation steps of Glide .....	16
3.4.2	Scoring Function in Maestro .....	18
3.4.3	Settings.....	19
3.4.4	Method.....	21
3.5	<i>Pharmacophores.....</i>	21
3.5.1	Structure based pharmacophores in LigandScout.....	22
3.5.2	Pharmacophore validation.....	24
3.5.3	Methods.....	24
3.6	<i>Screening schemes.....</i>	27

3.6.1	Pharmacophore filtering method.....	27
3.6.2	Scheme 1.....	27
3.6.3	Scheme 2.....	28
3.7	<i>Automatization in KNIME</i> .....	28
3.7.1	Workflows for scheme 1.....	30
<b>4</b>	<b>Results and Discussion .....</b>	<b>32</b>
4.1	<i>Molecular docking with Drugbank</i> .....	32
4.2	<i>Molecular docking with Enamine</i> .....	32
4.3	<i>Structure Based Pharmacophores and their validations</i> .....	32
4.3.1	Sbph1.....	33
4.3.2	Sbph 2.....	34
4.3.3	Sbph 3.....	35
4.4	<i>Scheme 1</i> .....	37
4.4.1	Sbph 1.....	37
4.4.2	Sbph 2.....	38
4.4.3	Sbph 3.....	39
4.5	<i>Scheme 2</i> .....	40
4.5.1	Sbph1.....	40
4.5.2	Sbph2.....	40
4.5.3	Sbph3.....	42
<b>5</b>	<b>Outlook and Conclusion .....</b>	<b>43</b>
<b>6</b>	<b>References .....</b>	<b>45</b>
<b>7</b>	<b>Table of Figures .....</b>	<b>49</b>
<b>8</b>	<b>Abstract.....</b>	<b>51</b>
<b>9</b>	<b>Zusammenfassung.....</b>	<b>52</b>



# 1 Introduction

## 1.1 The system of energy supply

The so-called creatine-phosphocreatine-creatine kinase system (Cr/PCr/CK) plays a major role in energy supply. This system cooperates with the ATP/ADP substrates. As these compounds are responsible for metabolic processes the creatine system is not allowed to fail. If even only one part of this system does not work, major illnesses occur.

To be more specific, in neurons, the ATP hydrolysis to ADP can be increased within seconds, during which the ATP level is able to intracellularly stay the same at all times, this is commonly known as the stability paradoxon. In the very beginning, the creatine needs to get into the cell via the creatine transporter. Once arrived, the Cr/PCr acts as a shuttle molecule between sites of ATP production (glycolysis, mitochondrial oxidative phosphorylation) and sites of ATP consumption/ hydrolysis (ATPases). In the mitochondria, creatine is loaded with a phosphate group of ATP before the resulting PCr arrives in the cytosol, where in turn the phosphate group is transferred to ADP. Afterwards, the unloaded creatine is then transported back into the mitochondria to be charged via ATP again. The Cr/PCr molecules are relatively small compared to ATP/ADP, but they have a higher diffusion coefficient and their superior concentrations in cytosol differ: The concentrations of ATP and ADP (ATP: 3-5mM; ADP 20-40  $\mu$ M) are less than the ones of Cr and PCr (Cr: 5-10 mM; PCr: 20-35 mM).

The creatine kinases are responsible for the transfer of phosphate groups regulated primarily via compartmentation, enabling the functional coupling of the CK reaction to various cellular ATPases. There are five existing subunits of creatine kinases: the brain type (CK-B), the muscle type (CK-M), the hetero-dimeric heart type (MB-CK), and two mitochondrial creatine kinases (mt-CK) which are homo-octamers expressed in the sarcomere or ubiquitous.<sup>1,2</sup>

## 1.2 Creatine Synthesis and uptake

The components of the endogenous creatine synthesis are glycine, methionine and arginine. Arginine provides the amidino group and methionine provides the methyl group. In the first step, the enzyme AGAT (arginine glycine amidino transferase) catalyzes the attachment of the amidino group to the glycine. We receive ornithine and guanidinoacetate (GAA). Next, the methyl group of S-adenosylmethionine (SAM) gets attached to GAA and it occurs creatine and S-adenosylhomocysteine (SAH) catalyzed

by the GAMT enzyme (guanidinoacetate N-methyl transferase).<sup>1</sup> The de novo creatine synthesis and the externally supplied creatine influence each other<sup>3</sup>. Exogenous creatine represses the endogenous synthesis by AGAT activity and expression<sup>4,5</sup>. Arginine and glycine induce creatine synthesis, methionine doesn't have a ratio limiting role<sup>2</sup>.

Food items like meat and fish and to a lesser extent dairy lead to a exogenous creatine uptake<sup>6</sup>.

### 1.3 Function of Creatine

As mentioned further above, creatine has an important role in energy supply and acts as an energy storage and cytosolic buffer in an interplay with the ATP/ADP compounds. Furthermore, creatine appears to have anti-oxidative and anti-apoptotic effects<sup>7,8,9</sup> and acts as an osmolyte.<sup>10,2</sup>

The important role of creatine in nutritional supplementation in sports medicine is of particular interest. The neuromodulatory function<sup>11</sup> as well as the therapeutic agent in psychiatric disorders seem to be important indications<sup>12</sup>. For example, creatine is used in cases like mitochondrial encephalopathy, strokes, or neurodegenerative and muscular disorders.<sup>13,7,8</sup>

### 1.4 Creatine Deficiency Syndrome

A man of 70 kilograms has a total amount of 120 gram creatine in his body. More than 90 % is situated in skeletal muscle, the rest of the creatine amount rests in tissues like heart, brain, retina and spermatozoa. Low levels of creatine can be found in the liver and kidneys, where the creatine synthesis takes place. The Creatine and Phosphocreatine convert themselves non-enzymatically and spontaneously to creatinine and gets excreted via urine. Of all the Cr and PCr we have in our body, we lose approximately 1,7% per day.<sup>3</sup>

The creatine deficiency syndrome therefore tends to appear due to an underlying gene defect of the enzymes AGAT/GAMT or the creatine transporter (CreaT).<sup>1</sup> The CreaT deficiency was discovered in 2001 in the context of a cerebral creatine deficiency. Mutations in the CreaT were found in up to 2% of X-linked with intellectual disability, whereas AGAT-D and GAMT-D are caused by autosomal recessive conditions.<sup>14</sup>



There have been several endeavors to cure creatine deficiency. However, in contrast to the AGAT and GAMT deficiency the CreaT deficiency cannot be treated well with creatine supplementation, since creatine needs to cross biological membranes (plasma membrane and blood-brain barrier) and depends on a transporter.<sup>2</sup>

Another attempt to treat the syndrome, that is not related to this work, is the idea of modulating creatine or other molecules to cross the membranes independently of the transporter. A promising example of such an attempt is the DAC - Acetyl-creatine-ethyl ester, which is showing hopeful results with a high lipophilicity and accumulation in the cytosol.<sup>15</sup>

### 1.4.1 Clinical features

In a collaborative study of phenotype and genotype in 101 male patients, clinical features were obtained. Intellectual disability is the hallmark of CreaT deficiency and leads to speech delay in 100 % of cases, behavioral abnormalities (in 85% of cases) and seizures (in 59% of cases). Moreover, the intellectual disability progresses with age. There is only one case where intellectual disability was considered mild. Nonetheless, patients often have a cheerful being. Behavioral abnormalities mostly consist of hyperactivity, attention deficit and autistic tendencies. Seizures, on the other hand, can be easily controlled, and febrile induced seizures are mostly infrequent. Even so, few patients can suffer severe refractory epilepsy and status epilepticus.

The motor development in patients with creatine deficiency is only slightly delayed and often appears in the form of unstable, stiff walking, with elevated arms. Extrapyraxidal movement disorders were, for example, reported as abnormal hand movements. Muscle weakness seems to be a rare problem though.

In addition, gastrointestinal complaints are relatively frequent and are apparent during feeding, or as frequent vomiting, leading to severe constipation and ileus, which in turn may necessitate surgery.

The bladder voiding dysfunction or instability occurs more frequently with increasing age. Cardiac symptoms weren't presented in patients that often and if cardiac symptoms were present, they showed rather mildly.

Retinal anomalies can occur, but were also not reported very often. The physical appearance of people suffering creatine transporter deficiency can differ from healthy people. They can have a broader/more prominent forehead as well as myopathic facies. But the most apparent physical feature is the slender build and poorly developed muscular mass.<sup>2</sup>

## 1.5 Solute carrier transporters (SLCs)

The solute carrier transporters (SLC) exist right beside the ABC and GPCR proteins, which present the so-called superfamilies. The SLCs count 456 known human transporters grouped in 65 families.<sup>16</sup> They are diverse in structure and have dissimilar folds. SLC transporters with similar folds are evolutionarily related to each other. The most common ones within the SLC superfamily are the “Major Facilitator Superfamily” (MFS) and the LeuT-like fold. To compare diverse folds, the MFS has 12 TM helices with 6 TM inverted pseudo-repeats, whereas the core of the LeuT-like fold contains 10 TM with 5 TM pseudo repeats.

SLC transporters can have similar substrates despite their diverse folds. Those displaying the same fold or family, however, are able to differ in substrate specificity, ion stoichiometry or energy coupling mechanism.<sup>17</sup> Diverse folds which exist in diverse SLC families are also used as representatives for other SLC families with a similar fold as observed in the X-Ray structures of their closest homologs<sup>18</sup>

Energy coupling mechanisms enable various options for movement across the membrane, including ion channels, secondary active transporters and transporters which do not have transport capability themselves, but interact with other SLC members forming heterodimers. This dynamic process, transporting substrates via secondary active transport, is defined as the “alternating access model”. The alternating access model has three different types of mechanisms, which are named “the rocker switch”, “the gated pore” and “the elevator”.

SLC transporters with the same fold, the serotonin, norepinephrine as well as the dopamine transporter (belonging to the MAT family of the GABA subfamily in SLC6 transporters) have the same conserved binding site as LeuT and show their diversity in transported substrates. MAT transporters transfer monoamines compared to amino acids transported via LeuT, which are chemically distinct.<sup>17, 18</sup>

### 1.5.1 LeuT-like fold

The crystal structure of LeuT was resolved in 2005.<sup>2</sup> As mentioned above, the core of the LeuT-like fold comprises 10 TM with 5 TM pseudo repeats. The first two TMs of each of the repeats are bundle domains and therefore moveable. The others are dedicated as scaffold domains, which remain static. The bundle domain ensures the capture and release of substrates. Therefore the LeuT-like fold uses the gating pore mechanism, also known as the rocking bundle mechanism. In general, the LeuT-like fold tends to

remain the same in the core region, but can vary concerning the rest of its TMs. Therefore TMs 11 to 14 are possible. Consequently, the number of co-transported ions varies as well.

Examples for the Leu-T like fold are the SLC5 family containing the Na<sup>+</sup>/glucose transporters and the SLC6 family containing the Na/Cl dependent neurotransmitter transporters.<sup>17</sup>

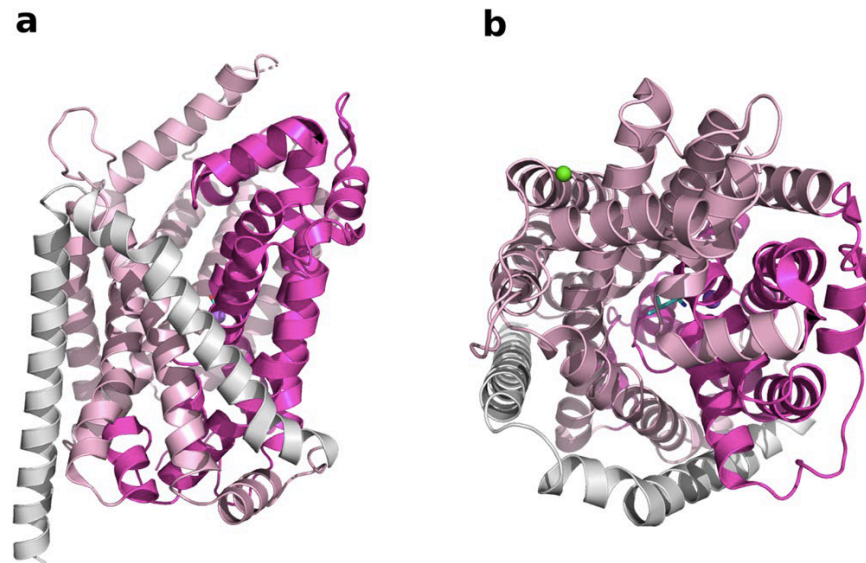


Figure 1: 3D structure of LeuT; the scaffold domain (light pink) and transport domain (dark pink); a (= side view) and b(= top view), the green and purple spheres represents the Na<sup>+</sup> and Cl<sup>-</sup> ions.<sup>19</sup>

#### 1.5.1.1 Binding site of the LeuT

The most studied binding site of LeuT is located approximately halfway across the bilayer between helices 1, 3, 6, 8 and 10. The binding site has 3 subpockets A, B and C. Specifically A consists of TM1b, 6 and 8 and has therefore a critical role in binding of the amine moiety of substrates. Highly conserved residues such as Q98 and Y95 in hSERT pertain to subpocket A. Subpocket B (TM3,8) is responsible for the hydrophobic interactions with aromatic moieties of the ligand. Subpocket C contains aromatic residues and is important for the shape of the binding site. F335 and F341 in hSERT belong to subpocket C. The corresponding area of the subpockets A and C has a ligand anchoring function.

Some examples of various amino acids in different transporter families, which are representing the substrate specificity individual cases, are as follows: The carboxylate moiety of GAT substrates interacts with Gly which is substituted with Asp in hSERT. In

the creatine transporter the C144 in TM3 associated to the subpocket B is important for the ligand binding, which is substituted with a glycine in this position in GATs. In TM8, a cysteine (GABAs) or an aspartate (TauT) is substituted with hydrophobic residues like I172 and G442 in hSERT. Finally, the  $\pi$ -helix in TM10 seems to be a specific feature to the GAT subgroup.<sup>18,19</sup>

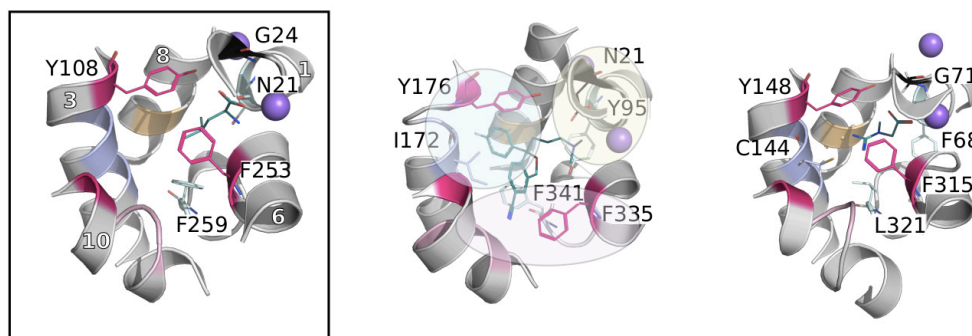


Figure 2: On the left hand side the binding site of LeuT (PDB ID 2A65) is depicted followed by hSERT (PDB ID 5I73) in the middle and a homology model of CreaT on the right hand side. The LeuT picture displays the numbers of the TM helices as well. hSERT in the middle shows 3 sub pockets displayed in yellow (A), blue (B) and pink (C) spheres.<sup>18</sup>

## 1.6 SLC6 Transporters

Due to the sequence similarity and the substrate specificity, the SLC6 family is classified into four subgroups: The monoamine transporter (MAT), the GABA transporter (GAT), the amino acid I (AA) and the amino acid II (AA) subgroups. The MAT subgroup, which is composed of NET (SLC6A2), SERT (SLC6A4) and DAT (SLC6A3) shows 20-30% sequence identity with the LeuT-like fold. The main inhibitory neurotransmitter GABA gets transported via GAT1 (SLC6A1), GAT3 (SLC6A11), BGT1 (SLC6A12) and GAT2 (SLC6A13). Furthermore the TauT (SLC6A6) transporter of the osmolytes taurine and betaine and the CreaT (SLC6A8) transporter belong to the GABA subfamily as well. The amino acid I subfamily contains the glycine transporters (GlyT1 (SLC6A9), GlyT2 (SLC6A5)), which are mainly expressed in the brain and spinal cord due to its inhibitory function, as well as the PROT and the ATB<sup>0,+</sup>. The amino acid II subfamily possesses amino acid transporters responsible for the amino acid homeostasis.<sup>18</sup>

As already mentioned, the SLC6 subfamily belongs to the secondary active transporter and operates as a symporter. These transporters use the electrochemical difference of Na<sup>+</sup> as energy source for transporting substrates.<sup>19</sup> Within the SLC6 subfamily, the transported ions one substrate molecule can vary from 2 to 3 Na<sup>+</sup>-ions to 1 to 2 Cl<sup>-</sup>-ions.

The GAT family has a sequence identity of 50-90% to each other, which makes it even more difficult to discover the residues defining the substrate specificity.<sup>19</sup> But has been found to be a general trend in SLC transporters is that conserved residues in TM 1 and 6 are responsible for the anchorage of substrates and variability in residues in TM3,8 and 10 confers in substrate specificity.<sup>18</sup>

## 1.7 Creatine Transporter (SLC6A8)

The SLC6A8 gene was first sequenced in humans in the 1990s and was located on section Xq28 containing 13 exons. The protein consists of 635 amino acids with a weight of 70,5 kDa. The transporter belongs to the GABA subgroup is closely related to the taurine transporter (52%) and shares a homology with GABA/betaine transporters of 48 – 50%.<sup>2</sup>

The core of the transporter is constituted as described in section 1.5.1.. The creatine transporter consists of 12 TM and the TMs are connected by loops. The N- and C-termini are situated on the cytoplasmic side. The transporter alternates between outward open and inward open conformations during the transport process.<sup>19,20</sup>

### 1.7.1 The homology model of SLC6A8

The molecular mechanisms of the creatine transporter were studied via a homology model using the prokaryotic LeuT transporter (PDB ID: 2A65) and the human serotonin transporter (hSERT) as templates. The hSERT was selected because of its high sequence identity (44%) with CreaT (Uniprot P48029, SLC6A8\_HUMAN). The sequence identity of LeuT to CreaT accounts for 21%. Both transporters have a similar predicted fold. The selected hSERT PDB-ID is in the outward open conformation. The selected PDB-ID of LeuT is in the outward occluded conformation, which represents the model we will present and are working with in this paper. Furthermore, in the multiple sequence alignment, an additional amino acid (S479) present in all GATs has emerged, and is located in the binding site in TM10 which results in a  $\pi$ -helix. A  $\pi$ -helix has on average 7 amino acids per turn and a looser hydrogen bonding connectivity of  $i \rightarrow i+5$ . Because of the diverse location of the additional AA in hSERT and LeuT this feature is one of the important characteristics for CreaT specificity within the subfamily of SLC6.

The other key feature of CreaT seems to be the residue C144, located on TM3 which is involved in substrate recognition with the very conserved tyrosine (Y148 in CreaT and Y176 in hSERT) as well located on TM3. The importance of this feature has been

described further above in 1.5.1.1.. Mutation studies on C144 to serine, alanine or leucine showed that mutants with long and hydrophobic side chains in this position decreased the substrate affinity.

Functional key residues are located in conserved positions as mentioned in section 1.5.1.1.. Residues like Y148 and F315 form the hydrophobic lid encompassing the binding site. D474 and R28 constitute the extracellular gate. The outward occluded conformation has a binding site volume of 117Å<sup>3</sup>, whereas the outward open conformation has a volume of 349Å<sup>3</sup>. The difference is caused by the tilting of the two broken TM 1 and 6 helices on the extracellular side, as well as the residue Y148 acting as a hydrophobic extracellular flap effecting a smaller volume. <sup>19</sup>

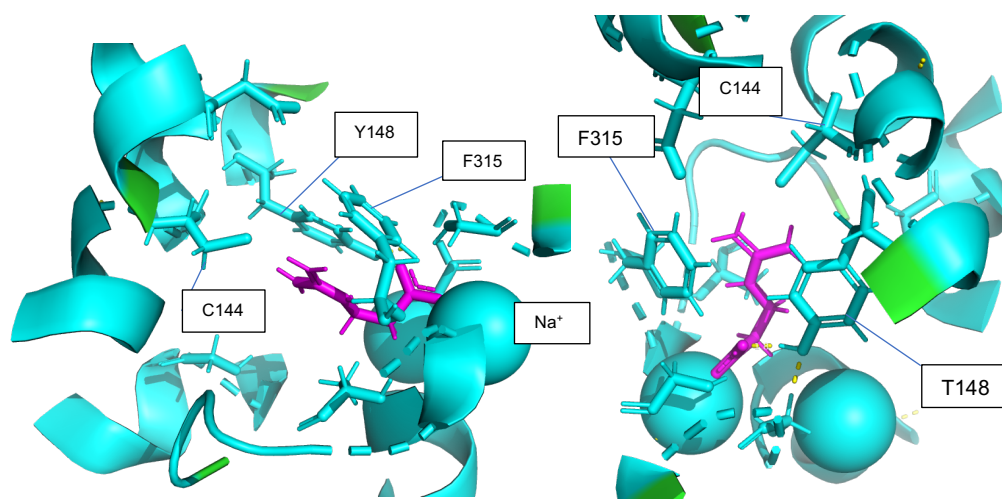


Figure 3: The figure on the left hand side shows the binding pocket in the outward occluded conformation with b-GPA (violet) from the side view. The right figure represents the same binding pocket from the top view. The symmetric order of the b-GPA to the Y148 forming a cation-pi-interaction with the guanidino group and the kink of the carboxylic moiety interacting with the glycines is very visible and gives a first insight into what the binding mode should look like and what some of the possible shapes are. Both were created in PyMol.

#### 1.7.1.1 Substrate properties of SLC6A8

The carboxylate moiety interacting with the backbone of G71 and G73 as well as with the hydroxyl group of Y148 forms polar interactions. The carboxylate group interacts with Na<sup>+</sup> as well. The guanidino moiety forms a salt bridge with the deprotonated C144<sup>21</sup> and a  $\pi$ - $\pi$  interactions with either Y148 or F315.

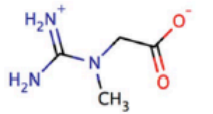
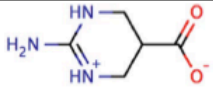
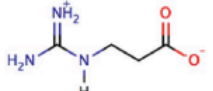
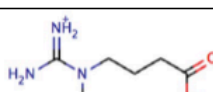
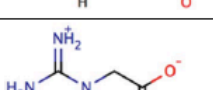
CreaT ligands <sup>a</sup>	2D Sketch <sup>b</sup>	Number of carbon atoms of the linker <sup>c</sup>	Length (Å) <sup>d</sup>	IC <sub>50</sub> <sup>e</sup>	Glide score from IFD <sup>f</sup>	MMGBSA <sup>g</sup> (kcal/mol)
Creatine		1	3.8	Km = 0.2 mM	-5.43	-24.68
ATPCA		2	4.3	66 μM	-7.411	-47.02
Beta-Guanidinopropionate (Beta-GPA)		2	5	44 μM	-6.86	-43.68
Gamma-Guanidinobutyric acid (Gamma-GBA)		3	5.8	697.9 μM	-6.58	-45.57
Guanidinoacetate (GAA)		1	3.7	712 μM	-5.24	-25.79

Figure 4: Table of all physiological compounds interacting with the creatine transporter

19

The table shows all compounds having a carboxylate and a guanidine group within a distance of 1-3 carbons chain length. By means of the values of IC<sub>50</sub> the substrates can be divided into inhibitors and substrates. ATPCA and beta-Guanidinopropionate (b-GPA) have a carbon length of two and are potential inhibitors with IC<sub>50</sub> values of 66 μM and 44 μM. Similar to the guanidinoacetate (GAA), creatine has a carbon length of one but creatine with a Km of 200 μM differs from GAA with a Km of 712 μM.

Taking into consideration that the MM/GBSA is likely to be a good scoring function to determine binding affinities, the scoring function should always be compared with IC<sub>50</sub> values nonetheless. Referring to the table above, gamma-GBA has a really good MM/GBSA scoring value and could have been counted as ATPCA and Beta-GPA as a potential inhibitor. Contemplating the IC<sub>50</sub> values, gamma-GBA is not a good inhibitor, probably because of the exceptionally long carbon chain. GAA and Creatine have similar MM/GBSA values, but differ in IC<sub>50</sub> values also. In all cases the reason is likely to be the carbon length. GAA is too small and can only accept weak interactions with the binding residues and Gamma-GBA is too large, which is only considered by the IC<sub>50</sub> value not by the MM/GBSA. Another reason might be a binding pocket of the conformation of the homology model that is too strict. The gamma-GBA is therefore slightly constricted and would bind much better to a slightly more open conformation.

In conclusion, the hypothesis was established of a carbon linker of approximately 4.5-5Å to gain a hydrogen bond with C144 on the guanidine moiety and a hydrogen bond with G71 or G73 on the carboxylate group. Most likely, it requires a dipole moment. Features which do not seem effective to their activity are subtle differences in geometry and flexibility. ATPCA is slightly smaller and less flexible than b-GPA.<sup>19</sup>



## 2 Aim of the Diploma Thesis

SLCs bear an essential role in absorption, distribution, metabolism and elimination of therapeutic drugs, although the importance, diversity and complexity of SLCs remained understudied until recently. The difficulty in detecting SLCs lies in the lack of resolved 3D structures due to the very hard expression and purification of membrane proteins maintaining their native states. Until this day, about 100 structures of human SLCs have been resolved representing only 25 unique proteins.

Thankfully, several structures of prokaryotic homologs have already been unveiled,<sup>16</sup> so that the creation of a homology model of the creatine transporter has been accomplished in previous work by Colas, C. et al. in the Pharmacoinformatics Research Group. For further investigation, the homology model of LeuT and CreaT was used, which has a lower sequence identity to the leucine transporter (21%). Even though this model is more likely to depict ligand interactions, as the outward occluded conformation is more compact, whereas both models<sup>19</sup> show similar quality.

Five endogenous ligands with known IC<sub>50</sub> values have already been tested on this outward occluded model of the creatine transporter, which lead to hypothesis about specific ligand properties for SLC6A8 interactions<sup>19</sup>. In the process of validating the protein model, this hypothesis shall be validated simultaneously.

In order to achieve this goal, the first step and starting point is a literature review of published compounds with corresponding IC<sub>50</sub> values. Next, these compounds should ideally be compared to the known ligands in docking experiments, in order to get more information on the ligand properties and binding modes, which would help to adjust the model.

Since this approach was not feasible, due to missing data in literature, the validation had to be carried out in an alternative fashion. The new objective was to find reliable structures in databases that can achieve good IC<sub>50</sub> values in the laboratory and would thereby be a first indication for a validation of the model<sup>22,23</sup>. The retrieval of new compounds should be achieved by using various software packages with different computational methods. On the one hand the “docking based screening” and on the other hand the “pharmacophore based screening” of two different databases will be executed.

Two different schemes were created to match the data obtained from the different experiments. The schemes are then compared with each other.

## 3 Materials and Methods

### 3.1 Databases

#### 3.1.1 The Protein Data Bank

The PDB organization provides the three dimensional structure of proteins, nucleic acids and complex molecules concluded in PDB-files supplied to the global community, though mostly used by the academic community. The Protein Data bank has existed since 1971 and is growing constantly. In 2018, it saw additions of more than 31 new structures per day.<sup>24</sup> Most structures are determined via experimental methods like X-ray crystallography, NMR spectroscopy or cryo-electron microscopy.

When downloading files from the PDB databank, the resolution of the Protein is one important component. The higher the resolution, the more accurate the amino acids of the protein. The template of the homology model used in this project (LeuT; PDB ID 2A65) has a resolution of 1,65Å, which is of high quality.<sup>19</sup>

#### 3.1.2 Drugbank

DrugBank is a Web-based, open source, bioinformatic/cheminformatics resource combining data on drugs and drug target databases with comprehensive data about drug action. It has primarily been developed to facilitate drug targeting and drug discovery. Hence, DrugBank is applicable for drug docking and screening but also for drug metabolism prediction, drug interaction prediction and general pharmaceutical education.<sup>25</sup> It therefore became a very popular resource for medicinal chemists, bioinformaticians, pharmacists, physicians and cheminformaticians. The database was first released in 2006 and is being updated every 6 to 12 months. Since 2006, DrugBank has rapidly evolved and has released its fifth edition in 2018. The database encompasses over 11,900 drug entries including 2538 small molecules, 1670 biotechnology drugs (protein/peptide) approved by the FDA and nearly 6000 investigational drugs.<sup>24 26</sup>

#### 3.1.3 Enamine database

The Enamine REAL DataBase (RDB) exists since 1991 and covers 29,000,000 compounds for virtual screening, and over 10,000,000 are Rule-of-5 compliant. All virtual compounds are feasible to be synthesized for in vitro testing when having a hit identification.<sup>27 26</sup>

## 3.2 Vizualization in PyMOL

PyMOL is an open source, python-based program for the visualization of 3D structures commercialized by Schödinger Inc. and is suitable for diverse operating systems. This is the most important program for nearly every technique in CADD, such as molecular preparation, homology modeling, protein preparation, lead design and molecular dynamics.

In our case, PyMOL was used for gaining a deeper insight into the protein. When opening a pdb file in PyMOL, the terminal is able to screen the whole amino acids and yet (is everything but water or protein residues, ligands and metal ions are included<sup>28</sup>), in order to check the completeness of the protein. This step is particularly important in cases when a crystal structure will be downloaded from the PDB-Databank. Furthermore, PyMOL generates high-resolution images and has many options to illustrate the properties of the protein in various options and forms and is therefore essential for creating pictures for publications.<sup>29</sup>

## 3.3 Structure based virtual screening

A major progress in reducing cost and time in the drug development process is the integration of computer aided drug design (CADD). In this context, structure based virtual screening (SBVS)<sup>26</sup> is one of the most promising in silico methods. The Scoring functions of a software for the evaluation of forces of non-covalent interactions between a ligand and its molecular target are the most sensitive part and are the critical component in SBVS which are responsible for success or failure. Since different software packages use different algorithms, different results can be obtained although the same input has been used. For the achieving of accurate SBVS results, docking protocols are essential, which are composed of two main components: in the first place the search algorithm and in the second place the scoring function.

Search algorithms are used to search for the most viable ligand conformations in combination with the most realistic position of the ligand in a systematic way. In case of rigid docking the search algorithm explores different positions of ligands using translational and rotational degrees of freedom. Another variant of docking is the flexible docking, where the conformational degrees of freedom to translations and rotations are added.<sup>24 30</sup>

### 3.3.1 The search algorithm

Algorithms that consider ligand flexibility can be distinguished into three types, the systematic, stochastic and deterministic approach. One or more algorithms can be used in one software.

#### 3.3.1.1 Systematic algorithm

Glide is attributed to the systematic algorithm, which means that the degree of freedom of a molecule will be exploited, which results in further thoughts of an increase of evaluation needed to be performed by the algorithm requiring more time for its execution. Reducing the time for executing, termination criteria are needed. Solutions which are already known as impossible won't be tried out.<sup>31 24</sup>

#### 3.3.1.2 Stochastic search algorithm

The operating principle of the stochastic search algorithm is the random change in the spatial conformation of the ligand, changing one system degree of freedom at a time. This approach leads to the exploration of several credible conformations. Since in this case the uncertainty of converging a good solution is given, several independent executions of stochastic algorithms are performed, normally. Monte Carlo (MC methods used by Glide)<sup>31</sup>, MOE and genetic algorithms used by GOLD and AutoDock4 are based on the stochastic search algorithm.<sup>31 24</sup>

#### 3.3.1.3 Deterministic algorithm

The deterministic algorithm calculates the next pose dependent on its initial pose. The weak points in this algorithm are local minima where the algorithm can be trapped in. For crossing this barrier, the simulation temperature can be increased. Examples therefore are energy minimization methods, as well as molecular dynamics simulations (MD).<sup>24</sup>

### 3.3.2 Scoring functions

Using scoring functions can pursue three different goals: the prediction of the binding affinity of a protein and a ligand; the identification of the ligand binding site (or allosteric site) as well as the conformation of the ligand and its target and last but not least the lead optimization.<sup>24</sup>

In general, the scoring function is calculated by the binding constant ( $K_d$ ) and the Gibbs free energy ( $\Delta G_L$ ) due to the formation of the ligand-receptor complex. The physical-chemical operations need to be evaluated, which are intermolecular interactions,

desolvation and entropic effects. The greater the evaluation based on the number of utilized parameters, the greater the accuracy of the scoring function.<sup>22</sup> Therefore most authors divide scoring functions into three types: the force field (FF), empirical and knowledge-based. Further scoring functions have been described as: machine-learning-based<sup>32</sup> and hybrid methods.<sup>24,33</sup>

#### 3.3.2.1 The force field scoring function

The force field scoring function, also known as “physics-based”<sup>34</sup> is based on experimental data in accordance with molecular mechanics. The scoring function is based on intermolecular interactions, non-bonded and bonded like van der Waals, electrostatic forces and bond stretching/bending/ torsional force interactions. But due to the lack of physical models to gain more accurate functions, the force field scoring functions have limitations in estimating entropic contributions. In addition, the estimation of the desolvation energy is neither accurate nor really included.<sup>24 22</sup>

#### 3.3.2.2 The empirical scoring function

Secondly, the empirical scoring function weights the binding free energy on structural parameters by adjusting them to experimentally determined binding constants (affinities) of a known set of protein ligand complexes. The prediction of the values of some variables is based on a linear regression. Subsequently the equation term will be adjusted by the weight constants generated by the empirical function used as coefficients. Hydrogen bonding, ionic bonding, non-polar interactions, entropic effects and desolvation effects are terms of the function. To sum up, the function is as good as the accuracy of the used data for developing the training model.<sup>24 22 35 36</sup>

#### 3.3.2.3 The knowledge based scoring function

The knowledge based scoring function is based on the estimation of the binding affinity by summing the binding interactions of atoms of a protein and the atoms of the molecular target. Statistical observations performed on large databases are required. This method derives from the intermolecular interactions occurring near certain types of atoms or functional groups and moreover occurring more frequently to contribute favorably to the binding affinity more likely. The score gives a sum of the score of all individual interactions.

As every scoring function has their virtues and limitations, the combination has been employed in obtaining a consensus scoring.<sup>24 22</sup>

## 3.4 Molecular docking in Maestro

Molecular docking, representing one of the SBVS techniques emerged in the 1980s, and showed to be a very promising method. Due to improved techniques, increased computational power, greater access in structural data and target molecules it became widely used just in the 1990s.<sup>24</sup>

In this project, the graphical user interface of Maestro makes the use of the Glide package of the Schrödinger software version 20-2 easier. The project can be ran from the command line as well.<sup>37</sup>

The Glide (Grid-based Ligand Docking with Energetics)-program first tests the spatial fit of the ligand to the binding site and trains the complementarity of ligand receptor interactions via the grid-based method similar to the empirical ChemScore function. Secondly, when compounds have passed this initial filter, the final stage of the algorithm is completed. The evaluation and minimization of the grid approximation to the OPLS nonbonded ligand-receptor interaction energy takes place. GlideScore is a multi-ligand scoring function and used to score the poses in the energy-minimized stage.<sup>37</sup> Glide is using empirical and force field based terms.<sup>30</sup>

### 3.4.1 The calculation steps of Glide

The conformational search is kind of a heuristic approach which eliminates unsuitable conformations (long-range internal hydrogen bonds).

Each ligand has a core region and some number of rotamer groups, whereas each rotamer group is attached to the core, but has no additional rotatable bond. Methyl-, amino-, ammonium groups and groups with terminated hydrogens are not considered because of their little significance. The core plus all enumerated, possible rotamer conformations will be docked as a single project. Glide can also pre-compute sets of conformations and internally generated conformations offer the greatest values. The shape and properties of the protein are embedded in a grid by several different sets of fields.

#### 3.4.1.1 The location and orientation of the compound

The search for possible locations and orientations on the active site of the protein begins with the selection of "site points". This site points are located on a grid covering the active-site region and are equally spaced in 2Å. The selection of the "site points" is made by calculating the distances from the site points to the receptor surface via a series of pre-specified directions which are sorted into distance ranges ("bins") of a width of 1Å.

Similarly, the distances from the ligand center defined as the midpoint of the two most widely separated atoms will be measured to the ligand surface. The distances will be also sorted into bins of width 1Å. Afterwards the distance ranges of the ligand center to the ligand surface will be compared to the distance ranges of the site points to the receptor surface. If the match of the ligand center and the site point is not good enough, the site point will be skipped.

#### 3.4.1.2 The placement of atoms

The atoms lie between a specified distance drawn on a line between the most widely separated atoms, the ligand diameter. A pre-specified selection of possible conformations will be collected. If there are too many steric clashes, the orientation will be skipped. Afterwards, the rotation around the ligand center will be included. Moreover, a subset test will be performed in scoring all interactions of atoms making hydrogen bonds or ligand-metal interactions with the receptor.

To combine these approaches for the right placement of atoms a similar scoring function to ChemScore is used. As ChemScore itself, this algorithm assesses favorable hydrophobic, hydrogen bonding and metal-ligation interactions, as well as steric clashes. In this stage, atoms are moved plus/minus 1Å in x,y or z direction to decrease the large 2Å jumps in the site-point/ligand-center position. Due to this characteristic, it is also called as “greedy scoring” and the best greedy scoring pose undergoes a “refinement” procedure. The turn is on the ligand, which can move rigidly as a whole plus/minus 1Å.

#### 3.4.1.3 The energy minimization

The aim of this step (only a small number of refined poses accomplishes the third stage) is the reduction of large energy and gradient terms resulting from too-close interatomic contacts. That's where the energy minimization begins in “softening” the pre-computed OPLS van der Waals and electrostatic grids. The final energy minimization is based on a full-scale OPLS non-bonded energy surface, which consists of rigid body translations and rotations in the case of docking external generated conformations. Internally generated conformations also include torsional motion about the core and end-group rotatable bonds. The very final improvement in energy minimization lies in the attempt of subjecting the top-ranked poses to a sampling procedure in which alternative local minima of core and rotamer-group torsion angles are examined.

#### 3.4.1.4 The re-scoring

The used re-scoring function is the GlideScore<sup>38</sup> based on the ChemScore, but includes a steric-clash term, penalizes electrostatic mismatches, amide-twists, calculates

hydrophobic enclosure terms and excluded volumes plus some modifications on terms already known in ChemScore.

### 3.4.2 Scoring Function in Maestro

Component	Description
vdW	Van der Waals energy. This term is calculated with reduced net ionic charges on groups with formal charges, such as metals, carboxylates, and guanidiniums.
Coul	Coulomb energy. This term is calculated with reduced net ionic charges on groups with formal charges, such as metals, carboxylates, and guanidiniums.
Lipo	Lipophilic term, which is a pairwise term in SP but is derived from the hydrophobic grid potential for XP. Rewards favorable hydrophobic interactions.
HBond	Hydrogen-bonding term. This term is separated into differently weighted components that depend on whether the donor and acceptor are neutral, one is neutral and the other is charged, or both are charged.
Metal	Metal-binding term. Only the interactions with anionic or highly polar acceptor atoms are included. If the net metal charge in the apo protein is positive, the preference for anionic or polar ligands is included; if the net charge is zero, the preference is suppressed.
Rewards	Rewards and penalties for various features, such as buried polar groups, hydrophobic enclosure, correlated hydrogen bonds, amide twists, and so on. This category covers all terms other than those explicitly mentioned.
RotB	Penalty for freezing rotatable bonds.
Site	Polar interactions in the active site. Polar but non-hydrogen-bonding atoms in a hydrophobic region are rewarded.

Figure 5: This table shows all terms used for the calculation of the GScore.<sup>37</sup>

Beside the GlideScore, whose terms are listed in the table above, further scoring functions are the “docking score” and the “Emodel score”. The docking score is calculated as the GlideScore, but is supplemented with Epik state penalties and strain corrections. Epik is a program in Schrödinger which processes large collections of input structures and estimates state penalties for the assumption that all structures are in solution. Epik can be chosen in the LigPrep panel, which will be mentioned later on.

The Emodel score is used for estimating the best pose of all poses that have been tried out for one structure.<sup>37 28</sup>

In this case, the GlideScore is the only scoring function we are using, because of the utilization of the Ionizer in the LigPrep panel instead of Epik and no strain corrections have been set, the docking score and the GlideScore are identical. Furthermore, as the setting was chosen for gaining one single conformation of each structure, the Emodel Score is not usable as well.



### 3.4.3 Settings

To gain accurate results in a docking run, pre-steps need to be done. The protein preparation and ligand preparation are one of these precautions, as well as the receptor grid generation without a ligand docking job is not possible.<sup>37 39</sup>

#### 3.4.3.1 Protein Preparation

Using the “Protein preparation wizard” panel a PDB file is assumed to start in general. The content of a PDB file can be different, whether the Protein comes from an experimental method offering a crystal structure or directly from a computer model, like the homology model, which is already processed.

In each case, the PDB-file needs to be prepared before doing a docking for gaining accurate results. Missing hydrogen atoms can be added, hydrogen bonds can be optimized, and atomic clashes can be removed. In addition, other operations can be performed which are not possible to run in the x-ray crystal structure refinement processes. This information gets added to the PDB file by using the protein preparation wizard.<sup>37</sup>

Default options were chosen in this work whereby the pH-range for generating het states using Epik was optimized from pH 7,0 +/- 2,0 to pH 7,0 +/- 0,5, because the pH value in blood is 7,4. After processing it is important to look if this action has not attached an H-atom in position C144. This was not the case so further investigations were necessary and the minimized prepared protein was able to work on with.

#### 3.4.3.2 Ligand preparation

As the ligands come from the databases, they need to be adjusted to fit in the target protein. In general, Glide can only modify the torsional internal coordinates of the ligand during the docking process; all the other parameters (bond lengths and bond angles) need to be optimized previously. Valences will be filled up in adding their hydrogens and the protonation state for physiological conditions will be adjusted.<sup>37</sup>

In this project, the Smiles codes of Drugbank and Enamine will be uploaded in the LigPrep panel. The OPLS3e force field<sup>40</sup> is checked by default options, whereas the pH range needs to be adjusted like in the protein preparation part to physiological conditions by using the Ionizer (not Epik) allows to generate all possible ligand protonation states in this pH-range. The desalt option is checked by default as well as the “tautomerizer”. As the input structure is wanted the tautomerizer is unchecked. The last step in this panel is to check the stereoisomers. Three options are eligible, as the third option

“generate all combinations” is not relevant for huge databases. The option “retain specified chiralities” takes the chirality from the input file (SD/ Maestro format) and the option “determine chiralities from 3D structure” takes the chirality information from the 3D geometry (the Smiles code) and ignores the input file.<sup>28</sup> Subsequently, the option of determine chiralities from 3D structure can only generate one structure with the given stereoisomer. In consequence, the number of output ligands (SDF format) do not need to be adjusted.

### 3.4.3.3 The receptor grid generation

As described in 3.4.1., the shape and properties of the receptor are represented on a grid by some different sets of fields. Different grids are required for different conformations, but for example different hydroxyl conformations can be handled with a single grid.

The panel is divided into 5 tabs, called “Receptor”, “Site”, “Constraints”, “Rotatable Groups” and “Excluded volumes”. Since in the outward occluded conformation of the homology model b-GPA is already depicted, the ligand will be excluded with the “receptor grid generation” panel, the position will be determined and the size of the active site will be shown by the receptor grid. The receptor does not need to be uploaded in the panel, it gets recognized from the workspace. In the “Receptor” tab, default options are retained. As centroid of the grid, the workspace ligand is chosen and the size of the screened ligands should be similar to the Workspace ligand, whereas this option can be checked in the “Site” tab. In the “Constraints” tab, positional constraints or H-bond constraints can be set. In this case, an hydrogen bond constraint to G71 was set up, because of its difference in hSERT to the GATs and the creatine transporter, as already mentioned in 1.5.1.1..

The rest of the panel was left by default.<sup>37 28 41</sup>

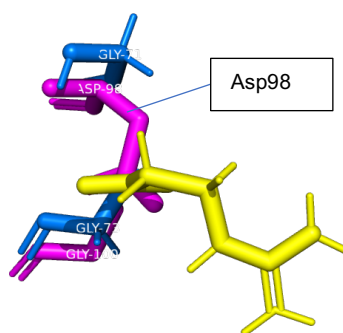


Figure 6: b-GPA (yellow); residues of hSERT (violet); residues of CreaT (blue); This image shows the interference of the Asp with the ligand, caused by the size and the charge of the amino acid. Asp displays an important difference to Gly71.

### 3.4.4 Method

#### 3.4.4.1 Ligand docking

For the utilization of the Glide ligand docking panel, the previously prepared databases and the protein are necessary.

Six different tabs give an overview on this panel: “Ligands”, “Settings”, “Core”, “Constrains”, “Torsional Constrains” and “Output”. First, the receptor grid file as well as the generated ligand file available in SDF-format need to be uploaded. Default options were used in this tab. The job was executed in the standard precision<sup>42</sup> and all other options left by default in the “Settings” tab. The core tab was skipped, and the set constraint was novated from the uploaded file of the receptor grid generation, therefore it does not need to be characterized anymore. After skipping the “torsional constraints” tab the “output” tab was adjusted. Retrieving ligands in the “Ligand pose” means the receptor is excluded and the results are stored in an sdf file. The file includes one pose per ligand. The post docking minimization is set as default options. Huge databases need many CPUs, subsequently, running the job on the local host would last a very long time and switching over to another host is recommended.

The output file contains every structure from the file generated in LigPrep retrievable in the single best conformation and inserted in the workspace into the protein structure, where nonbonded interactions can be reviewed. The GlideScore was calculated for each ligand, for assessing the docking compared to the other ligands. The Scores are denoted in kcal/mol and should approximate the binding free energy.<sup>37</sup>

## 3.5 Pharmacophores

A pharmacophore is a concept of describing pharmacon-drug interactions, which means that relevant (steric and electronic) chemical ligand features are arranged in a 3 dimensional space.<sup>43</sup> The pharmacophore should describe the chemical properties of a molecule, which are necessary for its biological activity.<sup>44</sup> This technique is used with well-acceptance for high-throughput virtual screening and is indispensable for drug development.

There are two possible ways to create a pharmacophore: Either by analysis of a known ligand-protein complex, or by using as a starting point a set of ligands supposed to bind in the same area within a target. The structure-based pharmacophore derives relevant chemical features from the known complex, whereas a ligand-based pharmacophore is searching for a maximum common set of chemical features.<sup>43 45</sup>

### 3.5.1 Structure based pharmacophores in LigandScout

LigandScout by Inte:Ligand has 4 different perspectives, called the “structure-based”, the “ligand-based”, the “alignment” and the “screening”. The “structure-based view” is used for the creation of structure-based pharmacophores and binding site analysis. In the “ligand-based view” the creation of ligand-based pharmacophores is possible. The analysis of common characteristics and the alignment of pharmacophores gaining merged characteristics can be done in the “alignment view”. In the “screening view” pharmacophore models can be optimized by changing features and is used to perform screenings with an inserted database of compounds against the created pharmacophore.

To screen a structure-based pharmacophore, only two perspectives are necessary: the “Structure-based” and the “Screening”.

First of all, the upload of a pdb file into the structure-based perspective needs to be done. The PDB file can be uploaded by using the 4-letter code or a downloaded pdb-file can be opened via the “file” menu “open” button. In this case the pdb file contains a ligand already existing in the binding site. Therefore, once the protein is uploaded, the binding site is marked with a yellow square. Clicking on this square, the view zooms into the binding site already containing the ligand. It is important to do the “Minimize MMFF94 Energy of Core Molecule and Side Chains” in the Molecule menu. By pressing the button “Show Binding Affinity Surface” in the same menu, the binding affinity of the b-GPA to the CreaT in our complex will be estimated. With the energy minimization, the binding energy of the complex can be optimized, whereby often more energy minimization runs are necessary. Afterwards the creation of an initial pharmacophore can be made by pressing the “Create pharmacophore” button in the “pharmacophore” menu. Adding an excluding volume coat to the pharmacophore, representing the shape of the active site, is important to prevent steric clashes and acts as a filter for too big ligands as a consequence. Via the “copy to other perspective widget”, the pharmacophore can be transferred to the “screening” perspective, where further investigations on the pharmacophore can be made.



Figure 7: Possible feature adjustments in the screening perspective by clicking on the “Pharmacophore” button in the menu bar. <sup>46</sup>

In the “Screening” perspective, the perfect pharmacophore for virtual screening needs to be created, before doing the virtual screening with huge databases. The table above shows all possibilities that can be used. <sup>46</sup>

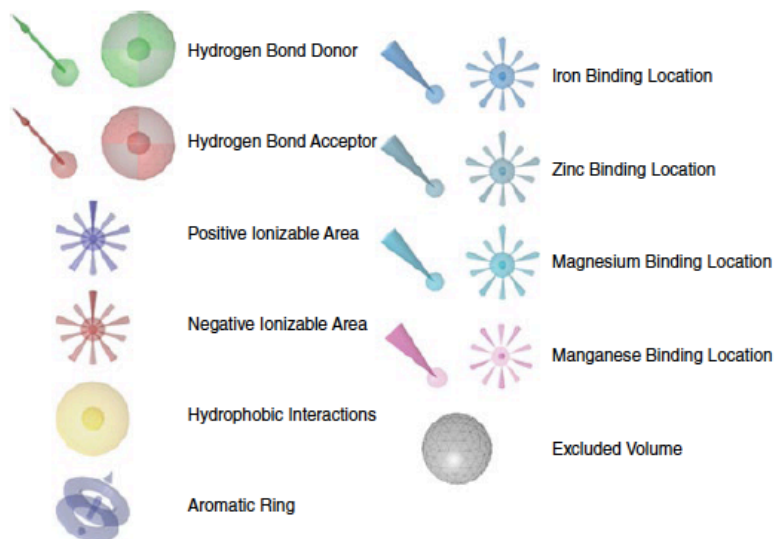


Figure 8: Features displayed in LigandScout as symbols. <sup>47</sup>

## 3.5.2 Pharmacophore validation

To create a good pharmacophore, a validation with a set of actives compared with a set of decoys needs to be made. If the hit list shows mainly active ligands, the pharmacophore covers active molecules well. If the pharmacophore cannot differentiate between actives and inactives, the model needs to be improved.<sup>46 24</sup>

### 3.5.2.1 Decoys

Decoys are generated from random molecular modifications of a structure related to a true active compound. Moreover, they have physically similar properties (molecular mass, number of rotatable bonds, logP, ..) but are different in structure, which is supposed to render the compound inactive.

Different databases, which are proposed to generate decoys are available, such as the Zinc database or the DUD-E database. Specifically the DUD-E database generates 50 different decoys per active compound, as used in this case. Because of the similarity of the physical properties in decoys, it is a good validation of the reliability of the used program.<sup>24,48</sup>

## 3.5.3 Methods

### 3.5.3.1 Pharmacophore validation in LigandScout

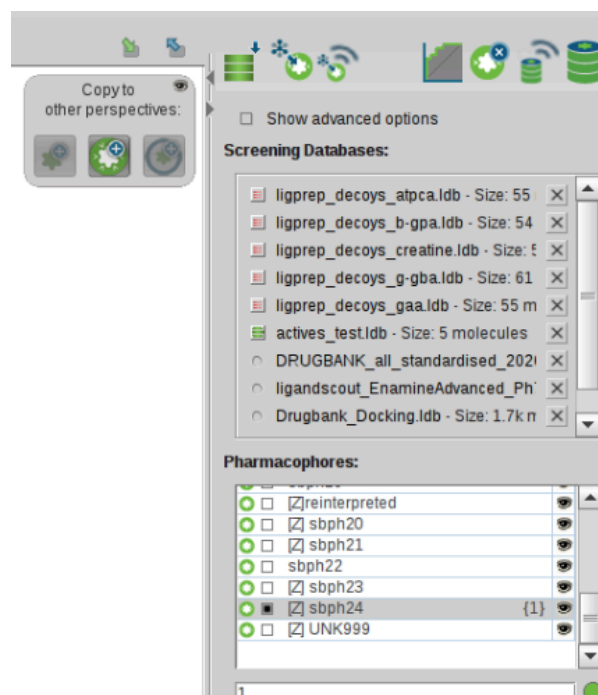


Figure 9: Screening panel<sup>49</sup>

The validation takes place in the “Screening” perspective. In using the very right button (Figur 6), the so-called “Create Screening Database” the already prepared ligands get converted into an ldb-file. Ligprep of Maestro was used for the ligand preparation using the same settings as in the “Ligprep” part of the database preparations. The arisen sdf files get uploaded in the “Create Screening Database” panel and an output file is defined. After clicking on “next”, the conformer generation takes place. Three options are available: to take over the conformation from the input, “iCon Fast” and “iCon Best”. “Fast” can be chosen for a high throughput and “Best” for high quality conformations, which we decided to use. Clicking on “next”, the ldb file will be created.

With the “Load Database” Icon on the very left side of the screening panel (figure 6), the ldb-file gets uploaded into the “Screening Databases” list. As shown in the picture above, the files in the list have green or red boxes on the left side of the file name or nothing. A red box marks a decoy file and a green box an active file, which is important for calculating the ROC. If no box is visible on the left side, the file will not be screened by the pharmacophore.

Since the DUD-E database creates a decoy file separately for every active, five ldb files were created. In every decoy file the related active compound is included. A separate active file was created, for which the actives from the decoy files were used merging into one.

By pressing the “Perform Screening” Icon (the second from the left side) the validation of the pharmacophore is executed.<sup>46 49</sup>

### 3.5.3.2 ROC

A small number of actives (TPCs) are sufficient for the calculation of an AUC-ROC. The ROC curve plots the distribution of “true positive compounds” and “false positive compounds” on a graph, where the sensitivity (% of selected ligands) of a virtual screening experiment in comparison to its specificity (% of selected decoys) is analysed. Subsequently, the larger the Area under the curve, the more TPCs will be discovered instead of FPCs. Excellent AUC values (0,90 -1,00), good AUC values (0,80 – 0,90), fair values (0,70 – 0,80), poor values (0,60 – 0,70) and failure (0,50 – 0,60) can be differentiated.

One disadvantage of this validation tool is the missing ranking of the best compounds for use in in vitro experiments. The Boltzmann-Enhanced Discrimination of ROC ranks active compounds more accurate and can be used instead. But the AUC-ROC and the BEDROC correlates very well in case of virtual screening simulations.<sup>24</sup>

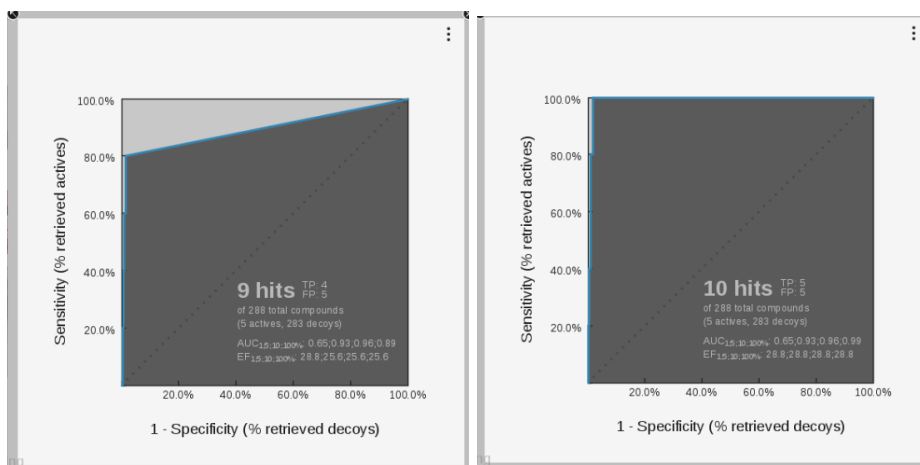


Figure 10: ROC of the original Ph4 on the left hand side, ROC of the modulated Ph4 on the right hand side. The median indicates the border to insufficient dedication of the pharmacophore of true positives (TP) and false positives (FP).<sup>49</sup>

### 3.5.3.3 Virtual screening

The virtual screening is relatively similar to the validation process. Once a pharmacophore shows a good validation, the databases are ready to screen with the model. Ensure that just one file in the “screening databases list “ (Figure 9) is selected with a green box on the left side of the file. The screening hits are listed in the library view, with all properties and the “pharmacophore fit score”.

The screening can be started in the “structure-based perspective” as well, by transferring the refined pharmacophore into “the structure-based perspective” again, where the selection “screening against an external library” is possible. The advantage of this option is the possibility of changing the settings for the screening run.<sup>49 46</sup>

### 3.5.3.4 Scoring Functions

The pharmacophore fit score considers only the feature RMS deviation and the pharmacophoric features. The “gaussian shape similarity score” includes steric properties and calculates gaussian functions to approximate the atom spheres. The binding affinity score calculates the binding affinity between the protein and the ligands in the library.<sup>46</sup>



## 3.6 Screening schemes

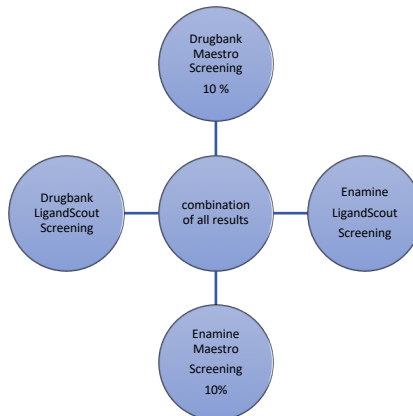
### 3.6.1 Pharmacophore filtering method

In this case, the filtering method with a structure-based pharmacophore was used. Poses which do not fully fill the binding site or which leave unpaired hydrogen bond acceptors or donors, should get filtered out. The advantage of this method compared with others is the possibility of quickly tested, compared and re-adjusted protein-ligand interactions with a single docking simulation run. In addition, Glide provides a “pose-filter” Python script which can filter certain poses by trying out if they fulfill some certain interactions or not. Despite that, a pharmacophore program can define filters of greater flexibility and has more options.

In a certain study of Megan L. Peach and Marc C. Nicklaus<sup>42</sup>, it has shown that using the pharmacophore filtering method as a post-processing filter, the advantages of both the docking and pharmacophore screenings can be exploited.<sup>50</sup>

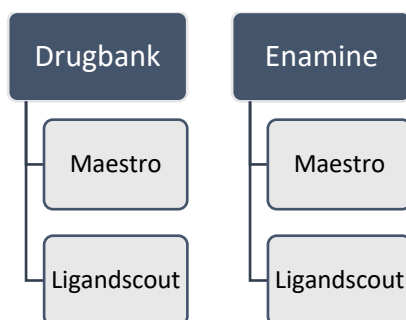
For our hit identification we have created two different schemes:

### 3.6.2 Scheme 1



In the first scheme, the initial screening of the whole databases (Drugbank and Enamine) with the two software packages of Schrödinger and Inte:Ligand were executed. For the obtained results, KNIME workflows were built to combine the hit lists to one solution. The KNIME workflow will be explained in “Workflows for scheme 1”. Afterwards the solutions of these workflows will be discussed in “Scheme 1”.

### 3.6.3 Scheme 2



In the second scheme, the pharmacophore screening in LigandScout should function as a filter executed with the screening results of the docking screening. No further combinations of results will be necessary in this approach.

## 3.7 Automatization in KNIME

As a leader in open resources of data mining tools, the KNIME software (Konstanz Information Miner) scripted in Java has been established and is extensible for further plugins. The main goal of using KNIME is to generate workflows, where the data will be uploaded in an input node. The data will be processed through other nodes and afterwards stored in a new format in an output file.<sup>51,52</sup>

In this work, Knime 4.2.2 was used.

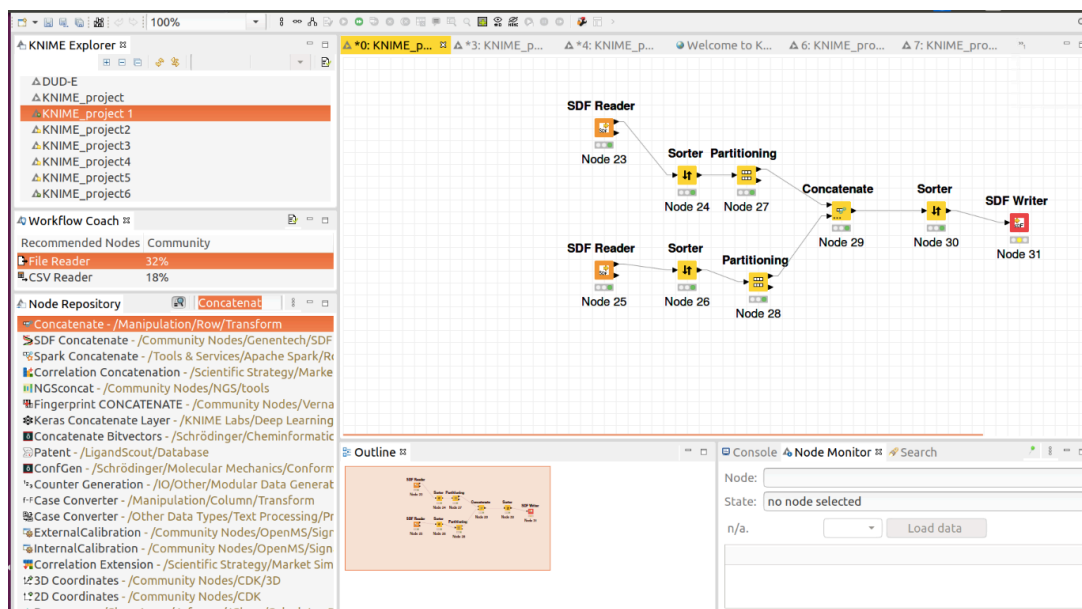


Figure 11: Interface of Knime

KNIME is offering a clear interface with different views: For opening already existing workflows on the local machine or existing workflows from the KNIME server, the KNIME

explorer represents each. The “workflow coach” recommends nodes appropriate to the previous node and the “Node description” is giving an explanation to the wanted node. In the “Node Repository” the search for nodes can be done in the search bar and by clicking double time on the requested node, the node will be inserted into the plaid workflow building view. In the “outline” view, the position which is represented in the workflow building view can be customized, which is relevant for huge workflows that cannot be shown on the whole in the workflow building view.

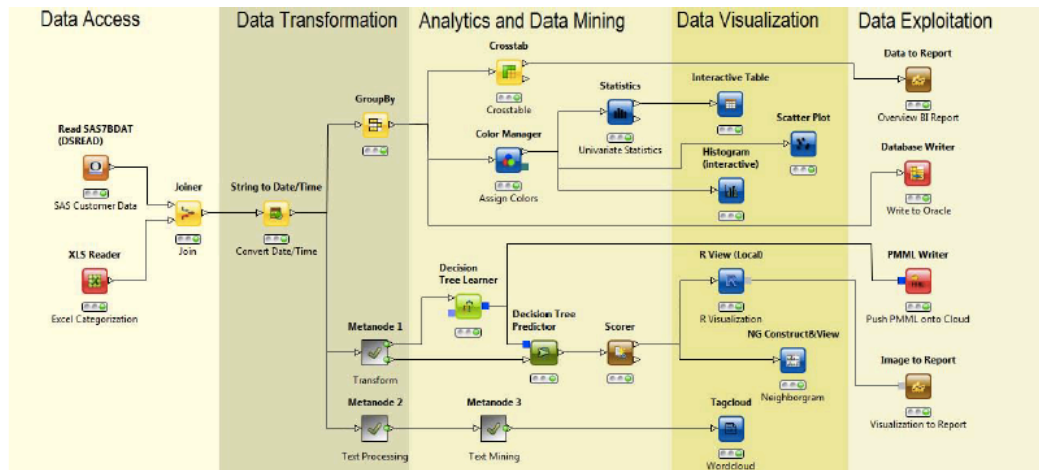


Figure 12: The various steps in the data mining process by building a Knime workflow.<sup>51</sup>

Knime provides diverse reader nodes for all possible data formats which need to be inserted in the data access part of a workflow. The “data transformation” part handles standard preprocessing functions like manipulating row and columns plus filtering them, concatenation and joining, binning, merging, transforming and row grouping as well as transformation. It is important to pre-process the data into a suitable form before mining. The next part is called analytics and data mining. In this sequence, the pattern recognition algorithm, the building, evaluating and interpreting of models is included. Finally, the last two steps, the “data visualization” and “data exploitation” is summarized as “knowledge deployment”.<sup>51</sup>

Building a workflow node by node, the data will be processed depending on the previous node. Therefore, every node has an input port and an output port, with which the nodes will be connected with each other. The configuration of a node can be done with a right click and choosing the “configure” option. Afterwards the node can be executed and checked, by opening the table of the individual node/step. If there is some incorrectness, the “Knime Console” field shows warnings as well as error messages for the solution of the problem.<sup>53</sup>

### 3.7.1 Workflows for scheme 1

#### 3.7.1.1 Combining Databases

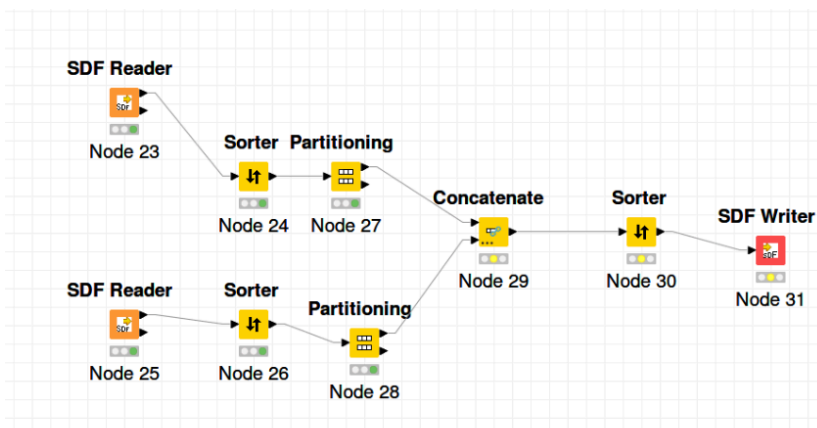


Figure 13: Knime Workflow; the combination of the docking-based screenings.

In the first SDF reader the docking results of Drugbank got uploaded. The “Sorter” was used to arrange the docked poses along their ascending values and with the “partitioning” node the first 10 % were extracted. The same procedure Enamine was running through. The “Concatenate” node combined the top 10 % of each database. To sort the concatenated table again along their ascending docking score, the “sorter” was once again inserted before writing the new sdf file with the “SDF Writer” node.

#### 3.7.1.2 Combining Screenings

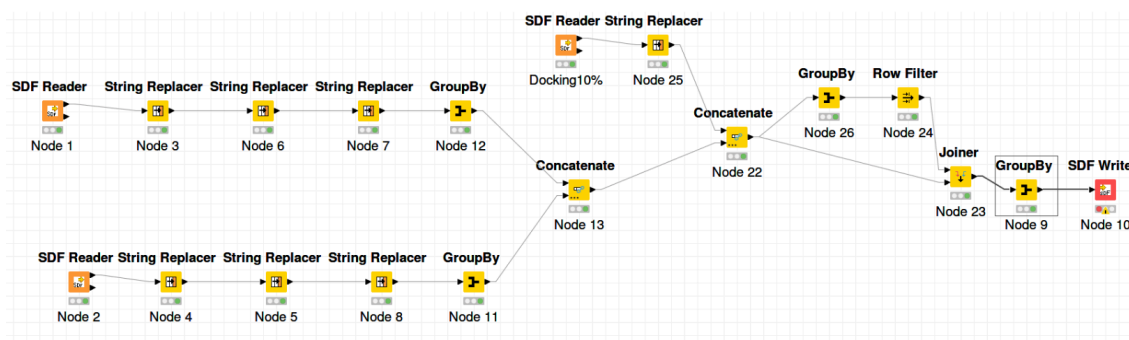


Figure 14: Knime Workflow in order to combine all screening results.

The aim of this workflow is the combination of the docking results with the pharmacophore screening results. For this purpose, the screening results of the pharmacophores will be uploaded in the “SDF-reader”. The names of the molecules are shown in the column called “s\_lp\_variant”. As in the part of the ligand preparation, the

same inserted ligand got different charges, ligands with the same "s\_lp\_variant" name were listed several times in the output list by adding a dash and numbers like "-1" or "-2". The "string replacer" nodes are deleting these attachments. After executing the "GroupBy" node, all molecules in the list are listed one time. The pharmacophore screening results do not offer that many output molecules, so that the partitioning of 10 % of the list is not necessary. The concatenation of the two databases followed with the best 10 % of the docking results was executed. In the next nodes, the molecules which were listed two times in the list, were counted, meaning just molecules were taken in the final "SDF Writing" node detected in both, the docking screening and the pharmacophore screening.

## 4 Results and Discussion

### 4.1 Molecular docking with Drugbank

In the initial dataset 9662 structures were presented. In doing the screening in Maestro, 7841 structures were filtered out. The glide score is ranked from -10.549 to +3.352.

### 4.2 Molecular docking with Enamine

The Enamine library is much bigger than the Drugbank database and initially had 582095 structures. In this case, 560045 structures were filtered out and the hit list shows 22050 molecules in a range of the glide score of -10.188 to +10.866.

### 4.3 Structure Based Pharmacophores and their validations

Depending on which binding affinity score the uploaded complex (b-GPA is included) has, after executing the “minimize MMFF94 Energy of Core Molecule and Side Chains” button, pharmacophores vary.

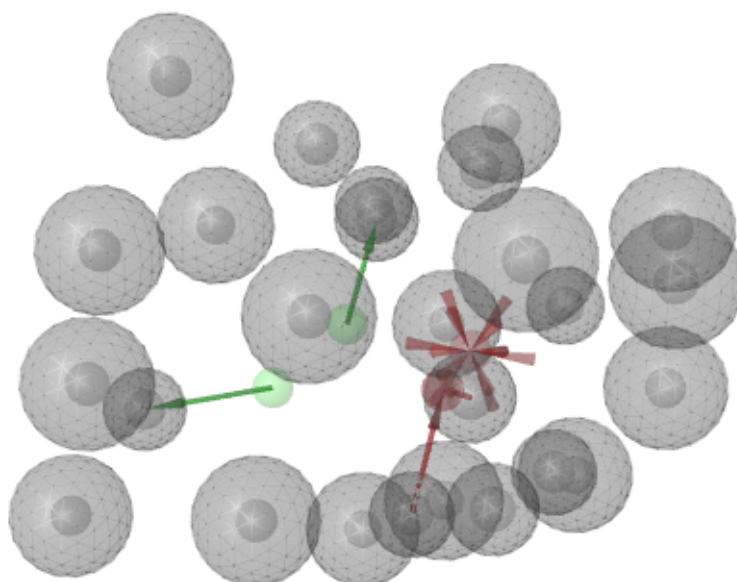


Figure 15: Pharmacophore in the original form with the excluded volume coat.

The gained pharmacophore, directly deriving from the structure-based perspective of the complex with a predicted binding affinity score of -2,83 of b-GPA, is already very

good. It shows two hydrogen-bond donors, two hydrogen-bond acceptors and one negative ionization feature surrounded by an excluded volume coat. The very left hydrogen bond donor points in the direction of the C144 and the other one is pointing in the direction of the F315 interacting with the backbone of the AA. The hydrogen bond acceptors are pointing in the direction of G73 and Y148. The negative ionization feature in the middle of all hydrogen bond acceptors interacts with the Na<sup>+</sup> which acts as a symporter for the substrates, as well.

Hits for Query »[Z] UNK999« Hitrate: 3.12% (9 of 288)   Filter: + + +					
	Mark	Name	#	Matching Features	Pharmacophore-Fit Score
1	<input type="checkbox"/>	b-gpaP10000005	8	■ ■ ■ ■ ■ ■ ■ ■	57.46
2	<input type="checkbox"/>	b-gpaP10000005	1	■ ■ ■ ■ ■ ■ ■ ■	57.46
3	<input type="checkbox"/>	creatineP10000009	7	■ ■ ■ ■ ■ ■ ■ ■	57.22
4	<input type="checkbox"/>	creatineP10000009	2	■ ■ ■ ■ ■ ■ ■ ■	57.22
5	<input type="checkbox"/>	g-gbaP10000002	9	■ ■ ■ ■ ■ ■ ■ ■ ■	56.52
6	<input type="checkbox"/>	g-gbaP10000002	3	■ ■ ■ ■ ■ ■ ■ ■	56.52
7	<input type="checkbox"/>	C33765216P50031618	5	■ ■ ■ ■ ■ ■ ■ ■	56.26
8	<input type="checkbox"/>	glycociamineP10000007	6	■ ■ ■ ■ ■ ■ ■ ■	55.72
9	<input type="checkbox"/>	glycociamineP10000007	4	■ ■ ■ ■ ■ ■ ■ ■	55.72

LigandScout (C) 1999-2021 Inte.Ligand GmbH

Figure 16: Validation hit library of a pharmacophore.

As the validation should register every active compound two times (ones from the active file and one from a decoy file), 4 actives were matched plus one decoy. Since all active compounds are very similar, a false positive rate of 20 % would not be tolerable.

#### 4.3.1 Sbph1

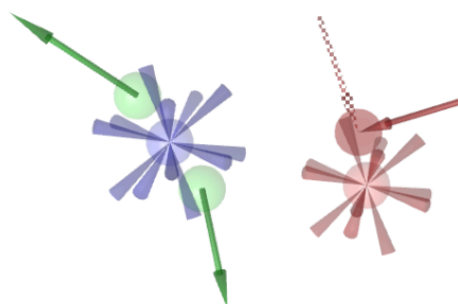


Figure 17: Pharmacophore 1 without the excluded volume coat.

Starting from the original template, the feature optimizations representing the hydrogen bond donors started with the duplication of themselves. These were changed into a positive ionizable feature which got interpolated to one feature in the middle of the original position of the two features. Subsequently one hydrogen bond acceptor feature was set to optional.

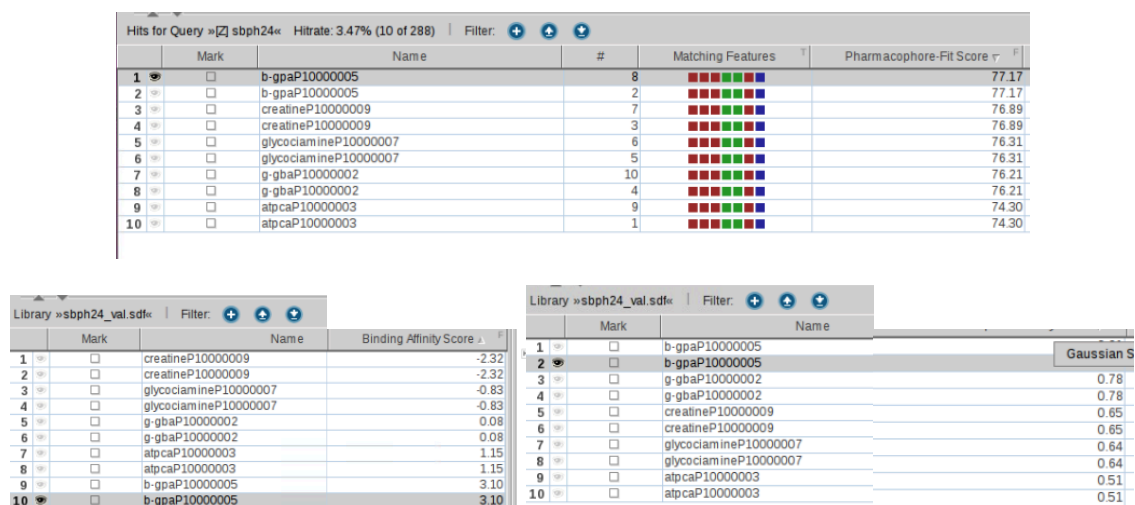


Figure 18: Validation of the optimized pharmacophore 1.

After optimization, all actives match the pharmacophore. For making a hypothesis the “gaussian shape similarity score” includes sterically conditions in addition to the matched features. As the pharmacophore was earned from the b-GPA, it is important that b-GPA is ranked as the “most suitable”.

The binding affinity score of b-GPA in the complex directly after inserting the homology model into the “structure based” view is -2.83, whereas b-GPA achieves 3.10 in this validation.

### 4.3.2 Sbph 2

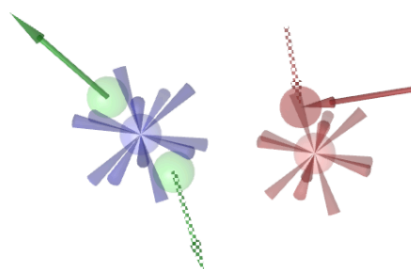


Figure 19: Pharmacophore 2 without the excluded volume coat.

The pharmacophore 2 looks very similar to pharmacophore 1, but received a much different validation. The binding affinity score of b-GPA to the binding site was set to a minimum of -5,24 after doing minimization steps. Further improvements like generating



the “positive ionizable feature” (in the same way as described in pharmacophore 1) were set, as well as the one of the hydrogen bond acceptor features was set to optional.

Hits for Query »[Z] UNK999« Hitrate: 3.47% (10 of 288)   Filter: + -						
	Mark	Name	#	Matching Features	Pharmacophore-Fit Score $r^2$	
1	<input type="checkbox"/>	g-gbaP1000002	9	■ ■ ■ ■ ■ ■ ■ ■ ■ ■	67.63	
2	<input type="checkbox"/>	g-gbaP1000002	4	■ ■ ■ ■ ■ ■ ■ ■ ■ ■	67.63	
3	<input type="checkbox"/>	atpcaP1000003	8	■ ■ ■ ■ ■ ■ ■ ■ ■ ■	64.64	
4	<input type="checkbox"/>	atpcaP1000003	1	■ ■ ■ ■ ■ ■ ■ ■ ■ ■	64.64	
5	<input type="checkbox"/>	b-gpaP1000005	10	■ ■ ■ ■ ■ ■ ■ ■ ■ ■	57.63	
6	<input type="checkbox"/>	b-gpaP1000005	2	■ ■ ■ ■ ■ ■ ■ ■ ■ ■	57.63	
7	<input type="checkbox"/>	glycociamineP1000007	7	■ ■ ■ ■ ■ ■ ■ ■ ■ ■	56.59	
8	<input type="checkbox"/>	glycociamineP1000007	5	■ ■ ■ ■ ■ ■ ■ ■ ■ ■	56.59	
9	<input type="checkbox"/>	creatineP1000009	6	■ ■ ■ ■ ■ ■ ■ ■ ■ ■	56.56	
10	<input type="checkbox"/>	creatineP1000009	3	■ ■ ■ ■ ■ ■ ■ ■ ■ ■	56.56	

Library »sbph27_val.sdf«   Filter: + -			
	Mark	Name	Binding Affinity Score
1	<input type="checkbox"/>	g-gbaP1000002	-4.38
2	<input checked="" type="checkbox"/>	g-gbaP1000002	-4.38
3	<input type="checkbox"/>	atpcaP1000003	-3.47
4	<input type="checkbox"/>	atpcaP1000003	-3.47
5	<input type="checkbox"/>	b-gpaP1000005	-1.51
6	<input type="checkbox"/>	b-gpaP1000005	-1.51
7	<input type="checkbox"/>	creatineP1000009	-0.41
8	<input type="checkbox"/>	creatineP1000009	-0.41
9	<input type="checkbox"/>	glycociamineP1000007	2.60
10	<input type="checkbox"/>	glycociamineP1000007	2.60

Library »sbph27_val.sdf«   Filter: + -			
	Mark	Name	Gaussian Shape Similarity Score $r^2$
1	<input type="checkbox"/>	g-gbaP1000002	0.71
2	<input checked="" type="checkbox"/>	g-gbaP1000002	0.71
3	<input type="checkbox"/>	glycociamineP1000007	0.65
4	<input type="checkbox"/>	glycociamineP1000007	0.65
5	<input type="checkbox"/>	b-gpaP1000005	0.65
6	<input type="checkbox"/>	b-gpaP1000005	0.65
7	<input type="checkbox"/>	creatineP1000009	0.59
8	<input type="checkbox"/>	creatineP1000009	0.59
9	<input type="checkbox"/>	atpcaP1000003	0.55
10	<input type="checkbox"/>	atpcaP1000003	0.55

Figure 20: Validation of pharmacophore 2.

The binding affinity score of the b-GPA (-1.51) in this case seems closer to the binding affinity score of -2,83. It is striking that the binding affinity score ranks actives along their sizes. Furthermore, the binding affinity should be in the negative range, whereas glycocyamine was calculated positive, which also gained the worst  $IC_{50}$  value as depicted in chapter 1.7.1.1. But the “binding affinity score” does not correlate with the “Gaussian shape similarity score” at all.

Interestingly, this pharmacophore of all tested pharmacophores ranks the ATPCA in the second place looking at the “pharmacophore-fit score”.

### 4.3.3 Sbph 3

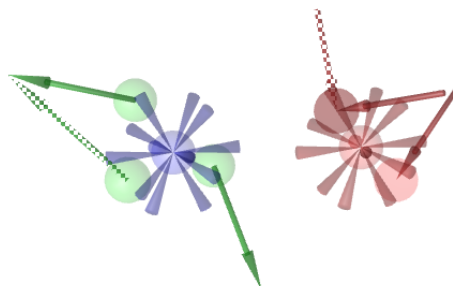


Figure 21: Pharmacophore 3 without the excluded volume coat.

In pharmacophore number three, the minimized complex has a binding affinity score of -5,09 and is representing more features in its original state. For the first time, the hydrogen bond donor and acceptor features are countable to three, but the “positive ionization feature” was not included from the beginning. For gaining the positive ionizable feature in this case, all three h-bond donors were duplicated and changed into the positive ionizable feature. All three features were interpolated, so that the feature is in the middle of all h-bond donor features. The one hydrogen bond acceptor feature pointing to the G71 is added, besides one feature was set to optional.

Mark	Name	#	Matching Features	Pharmacophore-Fit Score
1	g-gbaP10000002	8	■ ■ ■ ■ ■ ■ ■ ■	85.53
2	g-gbaP10000002	3	■ ■ ■ ■ ■ ■ ■ ■	85.53
3	b-gpaP10000005	6	■ ■ ■ ■ ■ ■ ■ ■	76.48
4	b-gpaP10000005	1	■ ■ ■ ■ ■ ■ ■ ■	76.48
5	creatineP10000009	7	■ ■ ■ ■ ■ ■ ■ ■	76.40
6	creatineP10000009	2	■ ■ ■ ■ ■ ■ ■ ■	76.40
7	glycociamineP10000007	5	■ ■ ■ ■ ■ ■ ■ ■	75.90
8	glycociamineP10000007	4	■ ■ ■ ■ ■ ■ ■ ■	75.90

Mark	Name	Binding Affinity Score
1	creatineP10000009	-5.87
2	creatineP10000009	-5.87
3	b-gpaP10000005	-2.66
4	b-gpaP10000005	-2.66
5	glycociamineP10000007	1.74
6	glycociamineP10000007	1.74
7	g-gbaP10000002	3.57
8	g-gbaP10000002	3.57

Mark	Name	Gaussian Shape Similarity Score
1	b-gpaP10000005	0.79
2	b-gpaP10000005	0.79
3	creatineP10000009	0.71
4	creatineP10000009	0.71
5	glycociamineP10000007	0.71
6	glycociamineP10000007	0.71
7	g-gbaP10000002	0.59
8	g-gbaP10000002	0.59

Figure 22: Validation of pharmacophore 3.

This pharmacophore is matching only 4 features, whereas ATPCA is missing. The pharmacophore fit score is gaining the highest values of all pharmacophores and the ranking of the “binding affinity score” and the “gaussian shape similarity score” seems to favor creatine and b-GPA. The “binding affinity score” shows positive values for too small (glycocyamine) and too big (g-GBA) ligands. Furthermore, b-GPA is ranked in the first place again concerning the “Gaussian shape similarity score”.

Since the validation of the pharmacophores are all relatively similar and all of them seem to be good, but not great, it is difficult to pick out one. Due to its limitation of having just five known ligands which are all very similar, it is difficult to impossible to create a pharmacophore which is not too strict for gaining unexpected results. Hence the screening was performed with each of them.

## 4.4 Scheme 1

Since only one docking screening result exists, the arrangement is organized by the different pharmacophores. Every result needs to be checked manually in Maestro, where non-bonded interactions can be displayed. The yellow dashed line shows hydrogen bond interactions, pi-pi stacking interactions are shown in blue lines. The green line is representing a cation-pi interaction. Clashes are differentiated to bad and ugly interactions whereas the bad interaction is shown in orange lines and the ugly interaction means red line.

By creating the ldb-files of the databases for the insertion in LigandScout, the iCONBest was selected for the conformer generation, when the "LigPrep" part was already executed in its settings as described in its chapter. The top 10 % of the docking results are 2387 structures which calculate the KNIME workflow.

### 4.4.1 Sbph 1

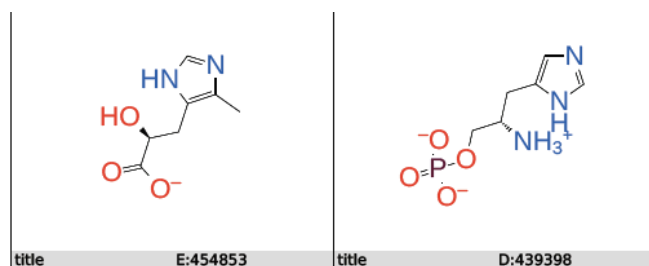


Figure 23: Screening results of pharmacophore 1.

To explain the first structure: The methyl rest of the imidazole ring of E:454853 is showing 4 bad interactions, two with the C144 and two with the Y148 because of too short distances. Further three bad interactions make the carboxylic group of the structure, where two of them make bad interactions with the amide group of the backbone of G71 and the other one interacts badly with the oxygen atom of the hydroxyl group of the T148. It is clearly caused by a too tight position. In the same time, three hydrogen bond interactions take place from the carboxyl group.

The other structure coming from DrugBank is showing even more bad interactions, which suggests again that the size of the molecule is too big.

## 4.4.2 Sbph 2

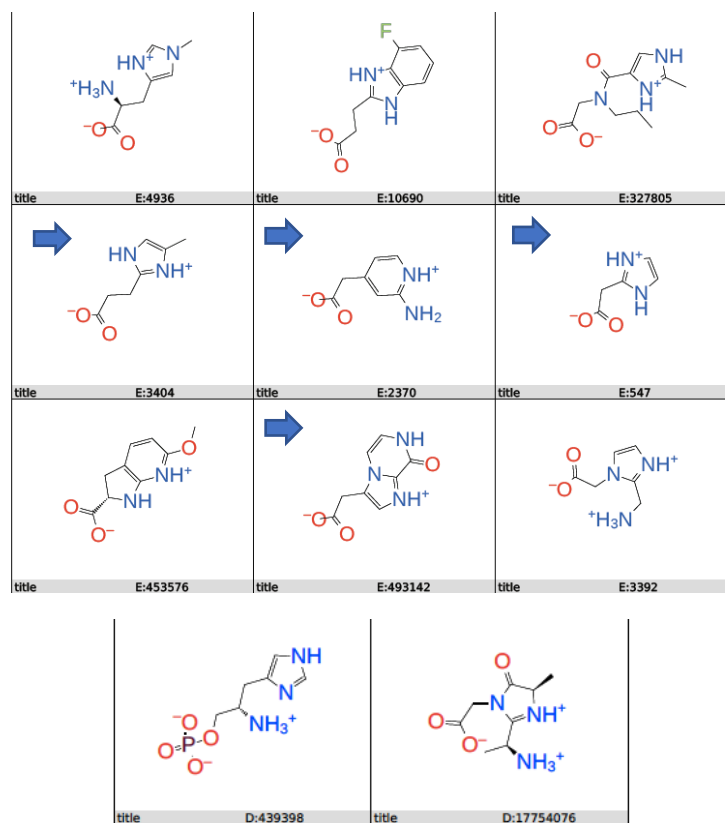


Figure 24: Screening results of pharmacophore 2.

In this pharmacophore, a few structures seem very useful, by looking through them manually. Out of this content the structure of E:493142 sticks out as it is the only structure having no bad interactions in Maestro. Since structures were prepared in the “ligprep” panel at the beginning of the whole procedure, the structures are listed two times mostly or even more often caused by the different charges which are possible by doing the setting of a pH of 7.0 plus/minus 0.5. So in this case, the structure with the GlideScore of -8,425 makes steric clashes compared to the structure with a charged oxygen on the carboxylic group and a GlideScore of -7.447. The imidazole ring in this picture is charged, as it is connected with the higher docking score. The imidazole ring with the lower docking score doesn't have any charges and forms a pi-pi-interaction with F315. On the carboxylic moiety the same hydrogen bond interactions with G71 and G73 take place, as this is the case in the carboxylic moieties of the known ligands described in 1.7.1.1..

In addition, structure E:3404 having the imidazole ring in a distance of 2 carbons from the carboxylic moiety looks very promising as well. The imidazole ring in its uncharged form interacts with the T148 and the F315 by forming a cation-pi-interaction<sup>54</sup> respectively. One hydrogen bond interaction with the backbone of F68 takes place. On the carboxylic group of the structure, two bad interactions are displayed with the nitrogen in the backbone of G73 with which hydrogen bond interactions take place as well. The charged structure has a GlideScore value of -7.498 whereas the uncharged structure has a GlideScore of -6.849, which is obtaining better interactions.

The structure E:2370 seems to be very interesting because of the interaction with C144, which would highlight the hypothesis for the specificity of SLC6A8. In addition, a cation-pi interaction and a pi-pi interaction from the pyridine ring of the structure with the F315 is forming out. Two bad interactions have been formed as well. The GlideScore of the charged carboxyl group and the charged pyridine ring has a value of -8.253. The GlideScore of the charged carboxyl group is ranked with a value of -7.593. The interaction profile is relatively similar in this case.

The structure of E:547 is the smallest one in this selection and is very similar to E:3404. This structure has only one carbon in between the carboxylic group and the imidazole ring plus the methyl group attached to the imidazole ring is missing. Depending on which position the structure takes in, it can form a hydrogen bond interaction or a pi-pi interaction with the F315.

All the other structures are displaying more than one steric clash and can be manually excluded.

#### 4.4.3 Sbph 3

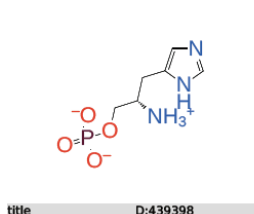


Figure 25: Screening results of pharmacophore 3.

In screening the databases with the pharmacophore 3 and doing the alignment with the docking screening in KNIME, just one structure is represented. Interestingly, this structure was obtained with all the other pharmacophores as well. Despite that, the structure has so many sterically clashes.

## 4.5 Scheme 2

The top 10 % of the docking results got extracted again, because all the other structures did not make sense; like pharmacophore 2 filters out 154 structures by screening the whole docking screening results.

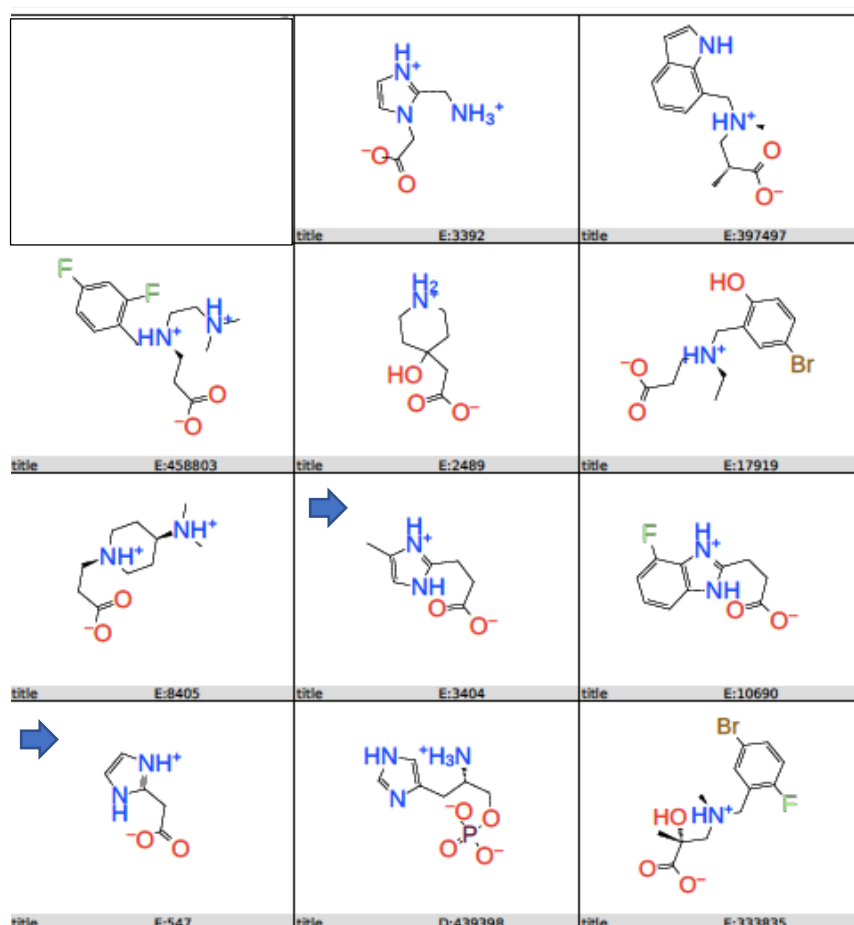
Since in scheme two, the docking results are used for further screenings with pharmacophores, the ligands do not have to get prepared in “LigPrep” again. The sdf-file will be converted into an ldb file, by using the button called “Create screening database” in LigandScout choosing the “iCON\_Best” option. In the drugbank database there are 1790 structures included and in Enamine there are 22002 compounds included. The top 10 %, combining both databases, are 2374 structures, which get screened in this section.

### 4.5.1 Sbph1

This pharmacophore does not get any matches. But in scheme 1 the matches were not reliable anyway, which means that the pharmacophore is not good enough.

### 4.5.2 Sbph2

Pharmacophore two shows 25 hits.



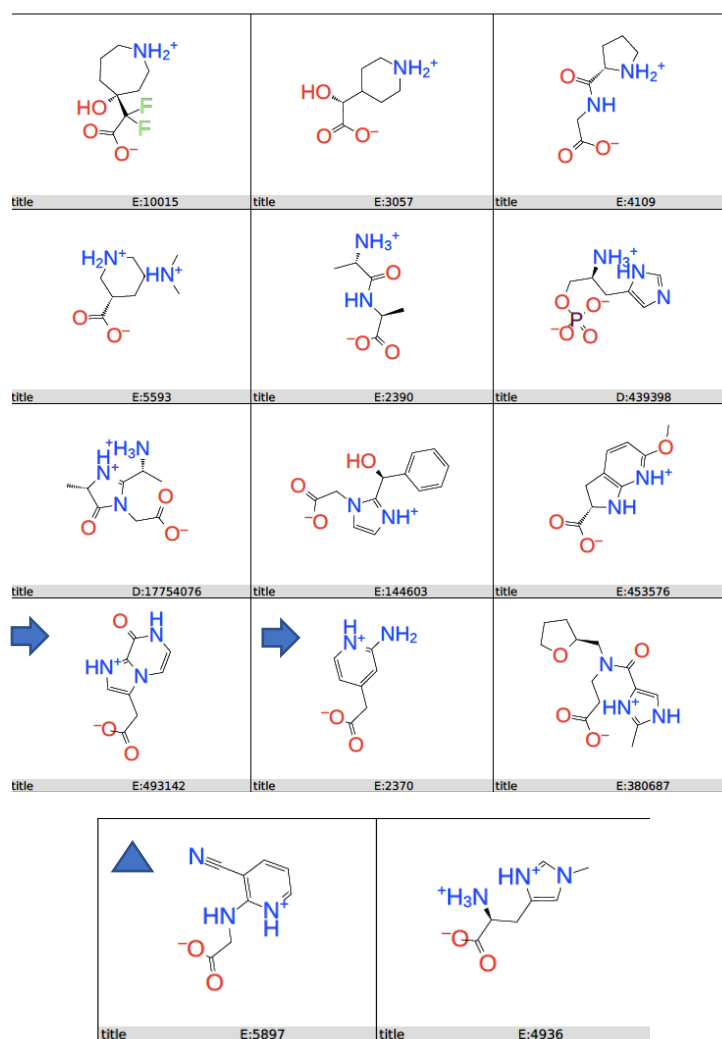


Figure 26: The arrows display the structures retrieved in scheme 1; the triangle indicates the structure, found to be new in scheme 2.

In this phase of finding new structures, all relevant structures found in scheme1 using the pharmacophore 2 were detected one more time in addition with many other structures. Maybe the change of conformation after the ligands had already found their right conformation in the docking process by using the iCON-Best button comes up with a different solution. But this needs to be adjusted, because of the utilization of another force field (MMFF94) in LigandScout than in Maestro (OPLS3e).

The compound E:5897 has in the conformation with the charged pyridine ring no bad interactions and a GlideScore of -7.036. From the amine in the pyridine ring a cation-pi interaction is arising with the T148 as well as a hydrogen bond interaction is forming with the backbone of F68. Furthermore, three hydrogen bond interactions on the carboxylic moiety with G71 and G73 are formed. The uncharged pyridine ring has not the right position for creating a cation-pi interaction.

### 4.5.3 Sbph3

Pharmacophore 3 does not display any matches as well, coming up to the same conclusion as in pharmacophore 1 (4.4.3.).



## 5 Outlook and Conclusion

In this case, validation means finding new structures with which real values can be determined in using in vitro methods, which in turn can be used for further in silico investigations and model adaptations.

The ligand properties are very strictly defined, as described at the beginning of this work. The ligand should have the right size, whereas the size is determined on the carbon number accurately between the positive ionizable and the negative ionizable part of the compounds, as the binding pocket in the outward occluded conformation is that small. Four of the matched compounds have carbon lengths of 1 and one has 2 carbon atoms, concerning the distance of the carboxy termini to the heterocycles. Two compounds contain an imidazole ring, two a pyridine ring and one contains a heterocycle with an A and B ring, whereas A is an imidazole as well. Notably ATPCA is the only known structure having an aromatic ring and is representing a good inhibitory property, which is maybe as well the case in compound E:2370. The carbon length from the carboxy terminus to the positive ionizable nitrogen differ more than in the already known ligands. Carbon lengths from 1 to 4 are present. The ligand must interact with the carboxy terminal with TM8, related to the sub pocket A, which is in all resulted compounds the case.

To conclude, the registered compounds do not really comply with the transporter specificity criteria of the creatine transporter mentioned in the introduction. Only one compound is interacting with residue C144. Further investigations on gaining more data for reliable structures docking experiments can be done with a second constraint to C144. Due to the avoidance of gaining too strict hits, it wasn't set in this case. Due to the lower specificity it is questionable and it would be interesting if the IC<sub>50</sub> values of these structures can be compared to the hypothesis set up by my working group described in the introduction, which is based on very similar ligands.

The different schemes are gaining diverse results, as the number of compounds differ from the calculating part in earning 10 % of the docking screening results. The screening scheme 1 seems more accurate, as less compounds are found with the pharmacophore two. Unfortunately, the other pharmacophores retrieved hits as well, which are unreliable though. The ligands detected in scheme 2 of pharmacophore 2 led to less reliable results (more structures compared to less hits) but extracts one more hit. To sum up, it is a good way to revise and to compare the findings. Despite that, it should be determined why the schemes are gaining diverse results.

Unfortunately, DrugBank does not lead to good results and the results in Enamine are also very “thin”. The size of the bad compounds is a common property and a limiting factor of the pharmacophores, in which direction the pharmacophore needs to be adjusted. Creating a pharmacophore needs much expertise and maybe the use of diverse programs for generating the optimal pharmacophore (for example Catalyst or MOE) is also necessary.

As compared to Maestro, non-bonded interactions are not able to be looked at in LigandScout. Therefore, the ligand pose is better to look at in Maestro of double dedicated ligands in the pose of the docking screening. For this reason, one approach can be the double check of the hits of pharmacophore 2 in performing an induced fit docking before shopping in the ligands for in vitro testing. This approach could be made, if no more docking run with another software will be performed.

In general, a so called consensus docking consisting of diverse scoring functions from different programs, for example a force field based combined with an empirical based scoring function represents another filtering method and in this work a further hedge.<sup>55,56</sup> As individual programs obtain incorrect results, which are mostly random, the combination of different results might be closer to the correct answer.<sup>57</sup> Due to time constraints this approach, which would have been probably due to the ligand similarity described in the introduction the better option, was not pursued.

## 6 References

1. Ndika, J. D. T. Creatine deficiency syndromes: a clinical, molecular and functional approach. ([Uitgever niet vastgesteld], 2014).
2. Kamp, J. M. van de, Mancini, G. M. & Salomons, G. S. X-linked creatine transporter deficiency: clinical aspects and pathophysiology. *J. Inherit. Metab. Dis.* **37**, 715–733 (2014).
3. Wyss, M. & Kaddurah-Daouk, R. Creatine and Creatinine Metabolism. *Physiol. Rev.* **80**, 1107–1213 (2000).
4. McGuire, D. M., Gross, M. D., Van Pilsum, J. F. & Towle, H. C. Repression of rat kidney L-arginine:glycine amidinotransferase synthesis by creatine at a pretranslational level. *J. Biol. Chem.* **259**, 12034–12038 (1984).
5. Edison, E. E., Brosnan, M. E., Meyer, C. & Brosnan, J. T. Creatine synthesis: production of guanidinoacetate by the rat and human kidney in vivo. *Am. J. Physiol.-Ren. Physiol.* **293**, F1799–F1804 (2007).
6. Brosnan, J. T., da Silva, R. P. & Brosnan, M. E. The metabolic burden of creatine synthesis. *Amino Acids* **40**, 1325–1331 (2011).
7. Andres, R. H., Ducray, A. D., Schlattner, U., Wallimann, T. & Widmer, H. R. Functions and effects of creatine in the central nervous system. *Brain Res. Bull.* **76**, 329–343 (2008).
8. Brosnan, J. T. & Brosnan, M. E. Creatine: Endogenous Metabolite, Dietary, and Therapeutic Supplement. *Annu. Rev. Nutr.* **27**, 241–261 (2007).
9. The creatine kinase system and pleiotropic effects of creatine | SpringerLink. <https://link.springer.com/article/10.1007/s00726-011-0877-3>.
10. Alfieri, R. R. *et al.* Creatine as a compatible osmolyte in muscle cells exposed to hypertonic stress. *J. Physiol.* **576**, 391–401 (2006).
11. Almeida, L. S., Salomons, G. S., Hogenboom, F., Jakobs, C. & Schoffemeer, A. N. M. Exocytotic release of creatine in rat brain. *Synapse* **60**, 118–123 (2006).
12. Allen, P. J. Creatine metabolism and psychiatric disorders: Does creatine supplementation have therapeutic value? *Neurosci. Biobehav. Rev.* **36**, 1442–1462 (2012).
13. Evangelidou, A., Vasilaki, K. & Nikolaidis, P. K. and N. Clinical Applications of Creatine Supplementation on Paediatrics. *Current Pharmaceutical Biotechnology* vol. 10 683–690 <https://www.eurekaselect.com/70242/article> (2009).
14. van de Kamp, J. M. *et al.* Genotype-phenotype correlation of contiguous gene deletions of *SLC6A8*, *BCAP31* and *ABCD1*: Gene deletions of *SLC6A8*, *BCAP31* and *ABCD1*. *Clin. Genet.* **87**, 141–

147 (2015).

15. Adriano, E. *et al.* Di-acetyl creatine ethyl ester, a new creatine derivative for the possible treatment of creatine transporter deficiency. *Neurosci. Lett.* **665**, 217–223 (2018).
16. Colas, C. & Laine, E. Targeting Solute Carrier Transporters through Functional Mapping. *Trends Pharmacol. Sci.* **42**, 3–6 (2021).
17. Colas, C., Ung, P. M.-U. & Schlessinger, A. SLC Transporters: Structure, Function, and Drug Discovery. *MedChemComm* **7**, 1069–1081 (2016).
18. Colas, C. Toward a Systematic Structural and Functional Annotation of Solute Carriers Transporters—Example of the SLC6 and SLC7 Families. *Front. Pharmacol.* **11**, 1229 (2020).
19. Colas, C., Banci, G., Martini, R. & Ecker, G. F. Studies of structural determinants of substrate binding in the Creatine Transporter (CreaT, SLC6A8) using molecular models. *Sci. Rep.* **10**, 6241 (2020).
20. Santacruz, L. & Jacobs, D. O. Structural correlates of the creatine transporter function regulation: the undiscovered country. *Amino Acids* **48**, 2049–2055 (2016).
21. Dodd, J. R. & Christie, D. L. Cysteine 144 in the Third Transmembrane Domain of the Creatine Transporter Is Located Close to a Substrate-binding Site. *J. Biol. Chem.* **276**, 46983–46988 (2001).
22. Ferreira, L. G., Dos Santos, R. N., Oliva, G. & Andricopulo, A. D. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* **20**, 13384–13421 (2015).
23. Fang, Y. Ligand–receptor interaction platforms and their applications for drug discovery. *Expert Opin. Drug Discov.* **7**, 969–988 (2012).
24. Maia, E. H. B., Assis, L. C., de Oliveira, T. A., da Silva, A. M. & Taranto, A. G. Structure-Based Virtual Screening: From Classical to Artificial Intelligence. *Front. Chem.* **8**, (2020).
25. Wishart, D. S. & Wu, A. Using DrugBank for In Silico Drug Exploration and Discovery. *Curr. Protoc. Bioinforma.* **54**, 14.4.1-14.4.31 (2016).
26. Heikamp, K. & Bajorath, J. The Future of Virtual Compound Screening. *Chem. Biol. Drug Des.* **81**, 33–40 (2013).
27. REAL Compound Libraries - Enamine. <https://enamine.net/library-synthesis/real-compounds/real-compound-libraries>.
28. Release 2020-2 | Schrödinger. <https://www.schrodinger.com/releases/release-2020-2>.
29. Yuan, S., Chan, H. C. S. & Hu, Z. Using PyMOL as a platform for computational drug design. *WIREs Comput. Mol. Sci.* **7**, e1298 (2017).
30. Danishuddin, M. & Khan, A. U. Structure based virtual screening to discover putative drug

- candidates: Necessary considerations and successful case studies. *Methods* **71**, 135–145 (2015).
31. Friesner, R. A. *et al.* Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **47**, 1739–1749 (2004).
  32. ashtawy2012.pdf.
  33. Huang, S.-Y., Grinter, S. Z. & Zou, X. Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. *Phys. Chem. Chem. Phys.* **12**, 12899 (2010).
  34. Liu, J. & Wang, R. Classification of Current Scoring Functions. *J. Chem. Inf. Model.* **55**, 475–482 (2015).
  35. Training a Scoring Function for the Alignment of Small Molecules | Journal of Chemical Information and Modeling. <https://pubs.acs.org/doi/10.1021/ci100227h>.
  36. Guedes, I. A., Pereira, F. S. S. & Dardenne, L. E. Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Front. Pharmacol.* **9**, 1089 (2018).
  37. Glide User Manual. 138.
  38. Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V. & Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. 21.
  39. Madhavi Sastry, G., Adzhigirey, M., Day, T., Annabhimoju, R. & Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* **27**, 221–234 (2013).
  40. Roos, K. *et al.* OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. *J. Chem. Theory Comput.* **15**, 1863–1874 (2019).
  41. Jansen, J. M. & Martin, E. J. Target-biased scoring approaches and expert systems in structure-based virtual screening. *Curr. Opin. Chem. Biol.* **8**, 359–364 (2004).
  42. Ramírez, D. & Caballero, J. Is It Reliable to Take the Molecular Docking Top Scoring Position as the Best Solution without Considering Available Structural Data? *Molecules* **23**, 1038 (2018).
  43. Wolber, G. & Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-Bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **45**, 160–169 (2005).
  44. Wermuth, C. G., Ganellin, C. R., Lindberg, P. & Mitscher, L. A. Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). *Pure Appl. Chem.* **70**, 1129–1143 (1998).
  45. Sun, H. Pharmacophore-Based Virtual Screening. *Current Medicinal Chemistry* vol. 15 1018–1024 <https://www.eurekaselect.com/66736/article> (2008).

46. LigandScout User Manual. 143.
47. Seidel, T., Bryant, S. D., Ibis, G., Poli, G. & Langer, T. 3D Pharmacophore Modeling Techniques in Computer-Aided Molecular Design Using LigandScout. in *Tutorials in Chemoinformatics* 279–309 (John Wiley & Sons, Ltd, 2017). doi:10.1002/9781119161110.ch20.
48. Réau, M., Langenfeld, F., Zagury, J.-F., Lagarde, N. & Montes, M. Decoys Selection in Benchmarking Datasets: Overview and Perspectives. *Front. Pharmacol.* **9**, 11 (2018).
49. LigandScout 4.4.6 (Inte:Ligand Software-Entwicklungs und Consulting GmbH).
50. Peach, M. L. & Nicklaus, M. C. Combining docking with pharmacophore filtering for improved virtual screening. *J. Cheminformatics* **1**, 6 (2009).
51. P. Mazanetz, M., J. Marmon, R., B. T. Reisser, C. & Morao, I. Drug Discovery Applications for KNIME: An Open Source Data Mining Platform. *Curr. Top. Med. Chem.* **12**, 1965–1979 (2012).
52. KNIME | Open for Innovation. <https://www.knime.com/>.
53. KNIME Software Overview. *KNIME* <https://www.knime.com/software-overview>.
54. Gallivan, J. P. & Dougherty, D. A. Cation- $\pi$  interactions in structural biology. *Proc Natl Acad Sci USA* **6** (1999).
55. Yang, J.-M., Chen, Y.-F., Shen, T.-W., Kristal, B. S. & Hsu, D. F. Consensus Scoring Criteria for Improving Enrichment in Virtual Screening. *J. Chem. Inf. Model.* **45**, 1134–1146 (2005).
56. Tuccinardi \*, G. P. and T. Consensus Docking in Drug Discovery. *Current Bioactive Compounds* vol. 16 182–190 <https://www.eurekaselect.com/166544/article> (2020).
57. Consensus Docking: Improving the Reliability of Docking in a Virtual Screening Context | Journal of Chemical Information and Modeling. <https://pubs.acs.org/doi/10.1021/ci300399w>.

## 7 Table of Figures

FIGURE 1: 3D STRUCTURE OF LEUT; THE SCAFFOLD DOMAIN (LIGHT PINK) AND TRANSPORT DOMAIN (DARK PINK); A (= SIDE VIEW) AND B(= TOP VIEW), THE GREEN AND PURPLE SPHERES REPRESENTS THE $\text{Na}^+$ AND $\text{Cl}^-$ IONS. <sup>19</sup>	5
FIGURE 2: ON THE LEFT HAND SIDE THE BINDING SITE OF LEUT (PDB ID 2A65) IS DEPICTED FOLLOWED BY HSERT (PDB ID 5I73) IN THE MIDDLE AND A HOMOLOGY MODEL OF CREAT ON THE RIGHT HAND SIDE. THE LEUT PICTURE DISPLAYS THE NUMBERS OF THE TM HELICES AS WELL. HSERT IN THE MIDDLE SHOWS 3 SUB POCKETS DISPLAYED IN YELLOW (A), BLUE (B) AND PINK (C) SPHERES. <sup>18</sup>	6
FIGURE 3: THE FIGURE ON THE LEFT HAND SIDE SHOWS THE BINDING POCKET IN THE OUTWARD OCCLUDED CONFORMATION WITH B-GPA (VIOLET) FROM THE SIDE VIEW. THE RIGHT FIGURE REPRESENTS THE SAME BINDING POCKET FROM THE TOP VIEW. THE SYMMETRIC ORDER OF THE B-GPA TO THE Y148 FORMING A CATION-PI-INTERACTION WITH THE GUANIDINE GROUP AND THE KINK OF THE CARBOXYLIC MOIETY INTERACTING WITH THE GLYCINES IS VERY VISIBLE AND GIVES A FIRST INSIGHT INTO WHAT THE BINDING MODE SHOULD LOOK LIKE AND WHAT SOME OF THE POSSIBLE SHAPES ARE. BOTH WERE CREATED IN PYMOL.	8
FIGURE 4: TABLE OF ALL PHYSIOLOGICAL COMPOUNDS INTERACTING WITH THE CREATINE TRANSPORTER <sup>19</sup>	9
FIGURE 5: THIS TABLE SHOWS ALL TERMS USED FOR THE CALCULATION OF THE GSCORE. <sup>37</sup>	18
FIGURE 6: B-GPA (YELLOW); RESIDUES OF HSERT (VIOLET); RESIDUES OF CREAT (BLUE); THIS IMAGE SHOWS THE INTERFERENCE OF THE ASP WITH THE LIGAND, CAUSED BY THE SIZE AND THE CHARGE OF THE AMINO ACID. ASP DISPLAYS AN IMPORTANT DIFFERENCE TO GLY71.	20
FIGURE 7: POSSIBLE FEATURE ADJUSTMENTS IN THE SCREENING PERSPECTIVE BY CLICKING ON THE "PHARMACOPHORE" BUTTON IN THE MENU BAR. <sup>46</sup>	23
FIGURE 8: FEATURES DISPLAYED IN LIGANDSCOUT AS SYMBOLS. <sup>47</sup>	23
FIGURE 9: SCREENING PANEL <sup>49</sup>	24
FIGURE 10: ROC OF THE ORIGINAL PH4 ON THE LEFT HAND SIDE, ROC OF THE MODULATED PH4 ON THE RIGHT HAND SIDE. THE MEDIAN INDICATES THE BORDER TO INSUFFICIENT DEDICATION OF THE PHARMACOPHORE OF TRUE POSITIVES (TP) AND FALSE POSITIVES (FP). <sup>49</sup>	26
FIGURE 11: INTERFACE OF KNIME	28
FIGURE 12: THE VARIOUS STEPS IN THE DATA MINING PROCESS BY BUILDING A KNIME WORKFLOW. <sup>51</sup>	29
FIGURE 13: KNIME WORKFLOW; THE COMBINATION OF THE DOCKING-BASED SCREENINGS.	30
FIGURE 14: KNIME WORKFLOW IN ORDER TO COMBINE ALL SCREENING RESULTS.	30
FIGURE 15: PHARMACOPHORE IN THE ORIGINAL FORM WITH THE EXCLUDED VOLUME COAT.	32
FIGURE 16: VALIDATION HIT LIBRARY OF A PHARMACOPHORE.	33
FIGURE 17: PHARMACOPHORE 1 WITHOUT THE EXCLUDED VOLUME COAT.	33
FIGURE 18: VALIDATION OF THE OPTIMIZED PHARMACOPHORE 1.	34
FIGURE 19: PHARMACOPHORE 2 WITHOUT THE EXCLUDED VOLUME COAT.	34
FIGURE 20: VALIDATION OF PHARMACOPHORE 2.	35
FIGURE 21: PHARMACOPHORE 3 WITHOUT THE EXCLUDED VOLUME COAT.	35

---

FIGURE 22: VALIDATION OF PHARMACOPHORE 3.	36
FIGURE 23: SCREENING RESULTS OF PHARMACOPHORE 1.	37
FIGURE 24: SCREENING RESULTS OF PHARMACOPHORE 2.	38
FIGURE 25: SCREENING RESULTS OF PHARMACOPHORE 3.	39
FIGURE 26: THE ARROWS DISPLAY THE STRUCTURES RETRIEVED IN SCHEME 1; THE TRIANGLE INDICATES THE STRUCTURE, FOUND TO BE NEW IN SCHEME 2.	41



## 8 Abstract

The creatine transporter is partly responsible for an orphan disease, which occurs with severe disease patterns. Treatment of this disease is possible in cases such as AGAT or GAMT deficiency, depending on the cause of the disease. The third reason for the creatine deficiency syndrome is the creatine transporter deficiency, where supplements are only effective until a limited extent.

In addition, the creatine transporter belongs to the SLC6 proteins and subsequently to the GABA family. An essential task is to discover their differences, due to the high sequence similarity of this family. Therefore, in previous work specifically of our working group a homology model in the outward occluded conformation has been built, based on previous mutation studies in the literature. The aim of this work is to validate the model in order to gain further insights and thus have a model that is accurate enough to be used in the search for new compounds for the treatment of the “orphan disease”. Furthermore, it is important to know the SLC transporter “like the back of your hand” to prevent bad drug interactions, caused by the high sequence similarity of some SLC proteins.

The steps of the validation were defined after finishing the search for literature which was not yet available for this topic. Therefore, the idea was to take one step back and search for interesting compounds in public databases, which can be tested in in-vitro studies for gaining the missing information this way.

The first step in the validation procedure was the execution of a docking-based screening. Drugbank and Enamine were used therefore.

As a second filtering method a pharmacophore was utilized. The pharmacophore was set up in the Inte:Ligand software from the structure based perspective.

Finally, two divers approaches were pursued in this work:

The first scheme was considering the whole databases for each screening. Afterwards the results were compared with a designed workflow in KNIME. In the KNIME workflow, 10 % of the compounds from the docking-based screening and the pharmacophore screening was taken and added together.

In Scheme 2, the pharmacophores were created as a direct filter of the docking-based screening results, using the top 10% of the compounds from the docking-based screening obtained after adding up the databases rather than from each individual.

Overall, the retrieved hit lists were not convincing and additional investigations such as consensus docking seem to be required.

## 9 Zusammenfassung

Der Kreatin-Transporter ist mitverantwortlich für eine „orphan disease“, die mit schweren Krankheitsbildern auftritt. Eine Behandlung dieser Erkrankung ist in Fällen wie AGAT- oder GAMT-Mangel möglich, je nach Ursache. Die dritte Ursache für das Kreatinmangel Syndrom ist die Kreatin-Transporter-Defizienz, wobei Supplemente nur bedingt wirksam sind.

Außerdem gehört der Kreatin-Transporter zu den SLC6-Proteinen und in weiterer Folge zur GABA-Familie. Aufgrund der hohen Sequenzähnlichkeit in dieser Familie liegt eine wesentliche Herausforderung in der Analyse ihrer Unterschiede. Daher wurde in früheren Arbeiten in unserer Arbeitsgruppe ein Homologie-Modell in der nach außen okkludierten Konformation erstellt, das auf früheren Mutationsstudien in der Literatur basiert. Ziel dieser Arbeit ist es, das Modell zu validieren, um weitere Erkenntnisse zu gewinnen und somit ein Modell zu haben, das genau genug ist, um es bei der Suche nach neuen Wirkstoffen für die Behandlung der "Orphan Disease" einzusetzen. Des Weiteren ist es wichtig die SLC-Transporter „wie die eigene Westentasche“ zu kennen, um unerwünschte Arzneimittelinteraktionen zu verhindern, die durch die hohe Sequenzähnlichkeit der SLC-Proteine verursacht werden können.

Die Schritte der Validierung wurden festgelegt, nachdem die Suche nach Literatur, die zu diesem Thema noch nicht verfügbar war, abgeschlossen war. Die Überlegung war daher, einen Schritt zurück zu gehen und in öffentlichen Datenbanken nach interessanten Verbindungen zu suchen, die in in-vitro-Studien getestet werden können, um so die fehlenden Informationen zu gewinnen.

Der erste Schritt im Validierungsverfahren war die Durchführung eines Docking-basierten Screenings. Dazu wurden Drugbank und Enamine verwendet. Als zweite Filtermethode wird ein Pharmakophor verwendet. Das Pharmakophor wird in der Software von Inte:Ligand aus der Struktur basierten Perspektive erstellt.

Schließlich wurden zwei verschiedene Möglichkeiten in dieser Arbeit verfolgt:

Die eine Möglichkeit ist die Betrachtung der gesamten Datenbanken für jedes Screening. Anschließend werden die Ergebnisse mit einem entworfenen Workflow in KNIME verglichen. Im KNIME-Workflow wurden 10 % der Substanzen vom Docking-basierten Screenings und des Pharmakophor-Screenings genommen und addiert.

In Schema 2 wurden die Pharmakophore als direkter Filter der Ergebnisse des Docking-basierten Screenings angelegt, wobei die Top 10 % des Dockings verwendet wurden, die nach Addition der Datenbanken und nicht von jeder einzelnen gewonnen wurden.

Leider konnten die erhaltenen Verbindungen nicht überzeugen und es scheinen noch weitere Untersuchungen wie z.B. das Consensus Docking notwendig zu sein.