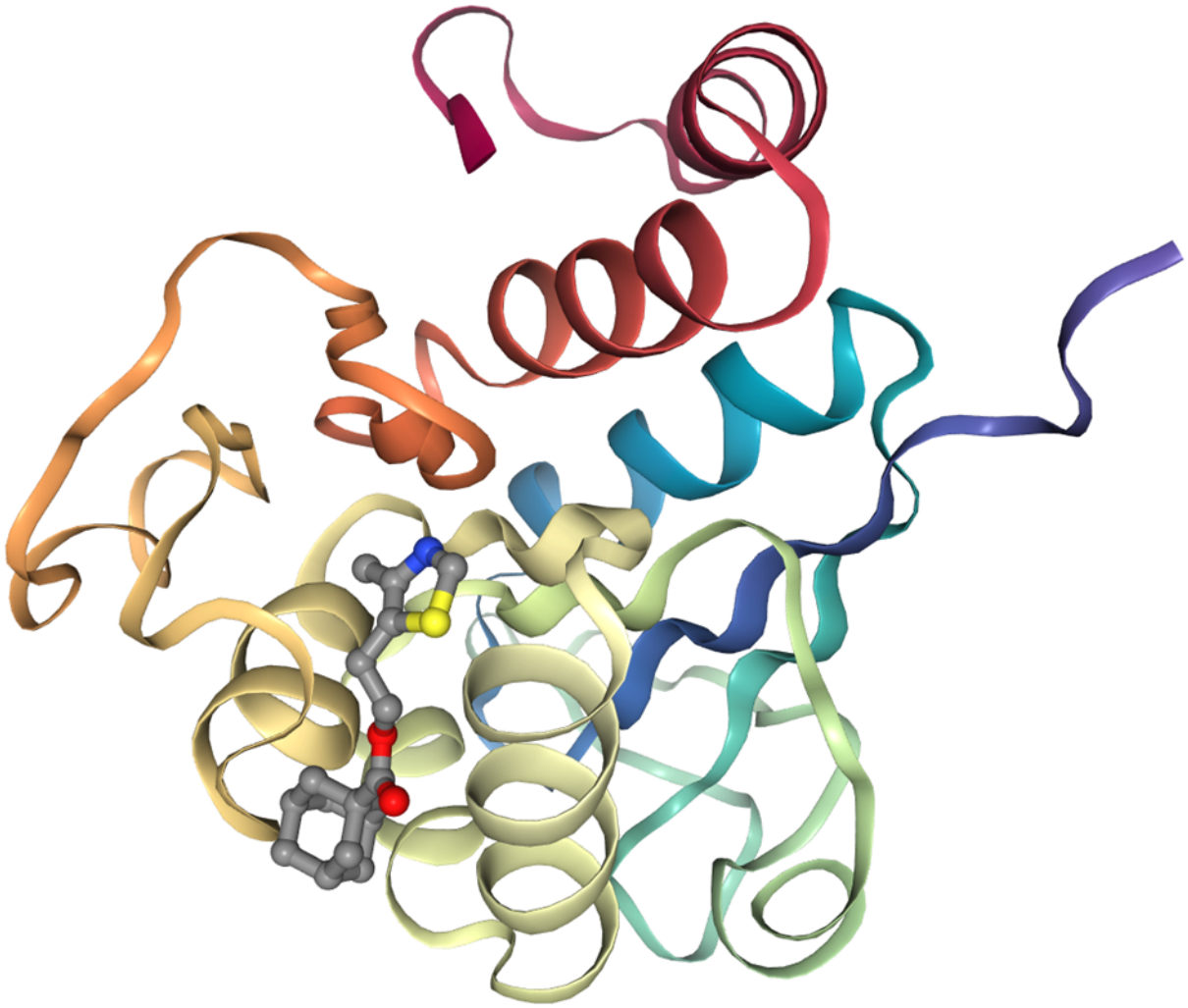
Protein-Liganden-Docking

Bioinformatik gegen Malaria & Co.

Kevin Kretz, German Esaulkov, Leander Schäfer

16. Januar 2022



1 Kurzfassung

Tropenkrankheiten stellen in ihren Verbreitungsgebieten eine extreme Bedrohung für die dortige Bevölkerung dar. Gemessen an ihrer Bedeutung - Malaria ist z.B. mit 200 Millionen Fällen pro Jahr eine der häufigsten Infektionskrankheit der Welt - erhalten sie in nicht betroffenen Industrieländern nur wenig Aufmerksamkeit in den Medien, in Form von Forschungsprojekten und in den Entwicklungsabteilungen von Pharmafirmen.

Aufgrund der Veröffentlichung des AlphaFold2-Papers im Juli 2021 und der gleichzeitig veröffentlichten Datenbank von dreidimensionalen Proteinmodellen sowie dem UseGalaxy-Server der Universität Freiburg haben wir gute Voraussetzungen bekommen, um mithilfe von Protein-Liganden-Docking nach möglichen Wirkstoffen gegen Tropenkrankheiten zu suchen.

2 Inhaltsverzeichnis

Inhaltsverzeichnis

3 Einleitung

Tropenkrankheiten richten in Entwicklungsländern verheerende Schäden an und fordern immer noch viele Opfer. So ist z.B. Malaria mit 200 Millionen Fällen pro Jahr eine der häufigsten Infektionskrankheit der Welt. Trotz der enormen Probleme, die Tropenkrankheiten wie z.B. Malaria, die Afrikanische Schlafkrankheit und die Chagas-Krankheit bereiten, schenken Industrieländer und die Pharmakonzerne diesen Bedrohungen nur wenig Beachtung.[24] Die Entwicklung von Medikamenten gegen diese weit verbreiteten Krankheiten erscheint den Konzernen als nicht lukrativ genug, obwohl der Bedarf dafür vorhanden ist, und wird dementsprechend vernachlässigt. Doch mittlerweile gibt es Vereinigungen, die gegen diese reale Gefahr kämpfen.[22] Wir möchten uns diesen, soweit es uns möglich ist, anschließen. Nun versuchen wir bei unserem Projekt mithilfe von Bioinformatik (weitere) helfende Wirkstoffe zu finden und unter Umständen auch eine neue Entdeckung zu machen. Selbstverständlich wird es uns nicht möglich sein, auf diese Weise ein fertiges Medikament entwickeln, dennoch hoffen wir, eine Grundlage für weitere Forschung schaffen zu können.

Krankheiten Doch warum haben wir ausgerechnet diese Krankheiten gewählt? Malaria, auch als Sumpf- oder Tropenfieber bekannt, ist die häufigste Tropenkrankheit. Sie ist vor allem in den tropischen Regionen Afrikas anzutreffen, aber auch in Südostasien und in den nördlichen Teilen Südamerikas. Der wichtigste Überträger der Krankheit ist die weibliche Anophelesmücke. Malaria kann Symptome wie Fieber, Erbrechen, Gelbsucht und Krämpfe hervorrufen. Vor allem die von uns gewählte Variante der Falciparum-Malaria ruft schwere Symptome wie Lähmung oder Koma hervor. Über Lungen- oder Nierenversagen führt die Krankheit zum Tod.

Abbildung 1: Die Anophelesmücke beim Blutsaugen (CDC/James Gathany)

Ebenfalls eine durch ein Insekt verbreitete Krankheit ist die *Afrikanische Trypanosomiasis*, oder auch Afrikanische Schlafkrankheit, deren Erreger *Trypanosoma brucei* durch die Tsetse-Fliege übertragen wird. Man kann aktuell mit ca. 500.000 Betroffenen in Afrika rechnen.[25] Über Fieber, Gliederschmerzen, Lymphknotenschwellung und Anämie führt die Krankheit zum namengebenden Dämmerzustand und anschließend zum Tod.

Als dritte Krankheit betrachten wir die Chagas-Krankheit, auch *Südamerikanische Trypanosomiasis* genannt. Der Erreger *Trypanosoma cruci*, ein Verwandter der *Trypanosoma brucei*, wird durch den Kot verschiedener Raubwanzen, aber vor allem von *Triatoma infestans*, übertragen.[27] Aktuell gibt es ca. 18 Millionen Erkrankte, wobei jedes Jahr 50.000 dazukommen. Die Zahl der Todesfälle beträgt jährlich um die 15.000. Die Krankheit verursacht Ödeme, chronisches Herzversagen und das Absterben von Nervenzellen im Darm. Dies führt teilweise zum Tod durch Darmverschluss oder Darmdurchbruch.

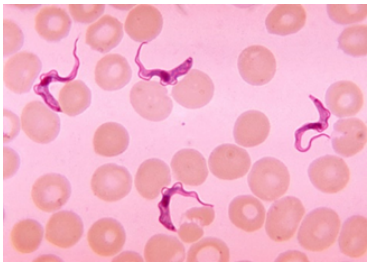


Abbildung 2: Trypanosoma, die Erreger der Schlafkrankheit (CDC/Dr. Myron G. Schultz)

Sowohl die Chagas-Krankheit als auch die *Afrikanische Trypanosomiasis* sind von der Weltgesundheitsorganisation als *Neglected Tropical Diseases* (NTD´s), also als `Vernachlässigte Tropenkrankheiten`, eingestuft. Die Auswirkungen dieser Vernachlässigung sind in den betroffenen Ländern deutlich zu spüren, weshalb also schnellstmöglich Mittel gegen diese Seuchen gefunden und entwickelt werden sollten.

Verfahren Wir möchten mithilfe von Protein-Liganden-Docking, einer Molecular Modelling-Technik, wobei mittels informatischer Methoden versucht wird, herauszufinden, wie Liganden mit Proteinen an welcher Stelle binden, Wirkstoffe für potenzielle Medikamente finden. Die Bindung eines Liganden an ein Protein kann die Struktur und die Funktion desselben beeinflussen bzw. verändern.[23] Die Bindungsart ist eine chemische Bindung wie eine Ionenbindung, Wasserstoffbrückenbindung oder Van-der-Waals-Kräfte. Man benötigt dementsprechend Daten über den Liganden und den Rezeptor des Proteins.[38]

Dieses Verfahren hat eine große pharmazeutische Bedeutung, da man „in silico“ mögliche Moleküle finden kann, die vitale Proteine des Krankheitserregers außer Kraft setzen können, ohne dass alle Kandidaten im Labor untersucht werden müssen. Dadurch können sehr schnell sehr viele Wirkstoffkandidaten kostengünstig getestet werden.

Auch wir möchten dieses bioinformatische Verfahren anwenden.

Werkzeuge Da wir persönlich keine Laborforschungsmittel oder Supercomputer besitzen, möchten wir öffentlich zugängliche Ressourcen verwenden.

Dabei haben wir uns für die Galaxy-Plattform entschieden.

Galaxy ist eine browserbasierte Plattform für Computational Science, mit Fokus auf Biologie, die das Nutzen bioinformatischer Tools mit überschaubaren Programmierkenntnissen möglich macht. Die Abläufe werden zusammen mit den Daten in sogenannten *Histories* gespeichert, die auch veröffentlicht werden können.[17, 26]

Abbildung 3: Logo von Galaxy Europe

Wir haben die öffentlich zugängliche, europäische Galaxy-Instanz (`https://usegalaxy.eu`) verwendet, da uns ihr großes Cluster genügend Rechenleistung für unser Projekt zur Verfügung stellt. Sie wird maßgeblich an der Uni Freiburg entwickelt und betrieben. Dort wollen wir, Protein-Liganden-Docking zu simulieren und auf diese Weise wirkungsvolle Moleküle gegen die Krankheiten zu finden.

Mit Galaxy haben wir nun also schon eine gute Grundlage zur Erforschung und Durchführung des Protein-Liganden-Dockings, jedoch fehlen uns die Daten über den Wirkstoff, also den Liganden, und das Organell, sprich ein vitales Protein, welches im Erreger außer Kraft gesetzt werden soll, sodass er stirbt. Man benötigt also Datenbanken mit den entsprechenden Sequenzen bzw. Strukturen. Lange Zeit war das Bestimmen der räumlichen Struktur von Proteinen eins der großen Probleme in der Bioinformatik. Durch die von DeepMind entwickelte AlphaFold-Software und die daraus resultierenden Daten in der AlphaFold Protein Structure Database hat sich hier die Lage stark verbessert, was zu genaueren Docking-Vorhersagen führt. Die AlphaFold Protein Structure Database führt Proteinstrukturen auf, die von einer KI auf Grundlage der Aminosäurensequenz ermittelt bzw. vorhergesagt wurden.[13] Dabei sind diese Vorhersagen sehr präzise. Hat man dann eine mögliche Struktur für das Protein gefunden, kann man nach realen Chemikalien in der EMBL-EBI-Datenbank, bzw. der von EMBL-EBI betriebenen Datenbank für bioaktive Moleküle ChEMBL suchen.[7, 9, 10, 30] Das *European Molecular Biology Laboratory Bioinformatics Institute* beherbergt die größte öffentlich zugängliche biologische Datenbank und bietet gleichzeitig auch bioinformatische Dienste für Forschende aus aller Welt an. Nur mithilfe von Galaxy, AlphaFold und EMBL-EBI bzw. ChEMBL können wir unser Projekt durchführen.

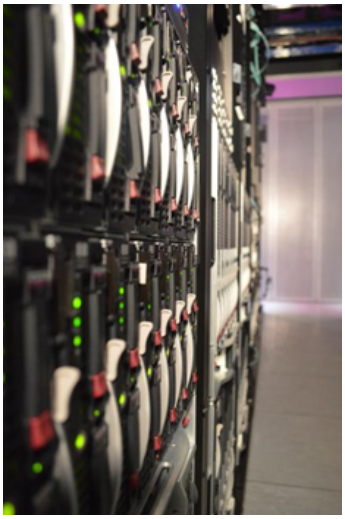


Abbildung 4: Datencenter des EMBL-EBI (EMBL-EBI)

Wie genau das abläuft wird im weiteren Verlauf erläutert.

4 Vorgehensweise, Materialien und Methode

4.1 Proteine und natürliche Liganden finden

Zunächst gilt es, Proteine zu finden, welche für wichtige Funktionen der Erreger verantwortlich sind. Hierzu betrieben wir Internetrecherchen, bei denen wir nach folgenden Kriterien geeignete Proteine auswählten:

- Die Struktur des Proteins im gefalteten Zustand ist bekannt, bzw. wird von AlphaFold mit großer Sicherheit vorhergesagt.
- Das Protein ist wichtig für die Lebensfähigkeit des Organismus.
- Der natürliche Ligand des Proteins ist bekannt.
- Das Protein ist nicht zu groß.

4.1.1 Malaria

Protein Unsere Forschung mit dem Malaria-Erreger *Plasmodium Falciparum* baut auf einer im *Malaria Journal* veröffentlichten Studie aus dem August des Jahres 2021 auf, in der man potenzielle Zielproteine für eine Behandlung von Malaria identifizierte. Wir wählten das Protein *Protein DJ-1* (`C6KTB1_PLAF7`; UniProt: `C6KTB1`), da es ein kleines Protein ist, für welches AlphaFold eine sehr genaue Strukturvorhersage liefert.[1] Wir mussten dann auf AlphaFold nur die UniProt-ID eingeben und konnten die PDB-Datei der Modellierung des Proteins herunterladen.

Ligand Nun suchten wir den natürlichen Liganden des Proteins. Nach einer Weile fanden wir in der ChEBI-Molekül datenbank ein Molekül namens `5-(2-hydroxyethyl)-4-methylthiazole` mit der ChEBI-ID `17957`.[11] Auf der Seite war eine MOL-Datei verfügbar. Wie wir zum SMILES-String gelangten, wird später beschrieben.

4.1.2 Afrikanische Schlafkrankheit

Protein Dieser Organismus verfügt über das Hitzeschockprotein 83, das die Reifung, den Strukturerhalt und die ordnungsgemäße Regulierung spezifischer Zielproteine fördert, die z.B. an der Kontrolle des Zellzyklus und der Signaltransduktion beteiligt sind.

Ligand Durch Recherche im Internet fanden wir heraus, dass der natürliche Ligand des HSP83 folgender ist: `CHEMBL561224`. Auf seinem ChEMBL-Eintrag stand sein PDBe-Eintrag verlinkt (PDBe steht für Protein Data Bank in Europe). Von dort kamen wir auf den betreffenden PDBeChem-Eintrag, wo der SMILES-String `FC(F)(F)c2nn (c1c2C (=O)CC(C1)(C)C)c4ccc(C(=O)N)c(NC3CCC(O)CC3)c4` anzutreffen war.

4.1.3 Chagas-Krankheit

Protein Im Fall der Chagas-Krankheit entschieden wir uns, einen etwas anderen Ansatz zu verfolgen. Wie auch die Afrikanische Schlafkrankheit, wird die Chagas-Krankheit durch einen Flagellaten aus der Gattung der Trypanosomen verursacht, genauer durch *Trypanosoma bruci*.[5] Trypanosomen nutzen Ergosterin statt Cholesterin als Baustein ihrer Zellmembranen.[31] Die Blockierung der Ergosterin-Biosynthese stellt also eine mögliche Behandlung von Trypanosomen-Erkrankungen dar. Also wollten wir das Ergosterin angreifen, es binden. Als zu bindendes Protein wählten wir *Sterol 14-alpha demethylase* (UniProt-ID `Q7Z1V1`) aus, das zur Biosynsthese von Ergosterol bei *Trypanosoma cruzi* notwendig ist.

Ligand Um passende Liganden zu finden, gingen wir auf den Wikipedia-Artikel von Ergosterol und von dort aus auf die aufgelisteten Artikel der Inhibitoren (Azole). Diese waren *Fluconazol*, *Miconazol*, *Itraconazol*, *Clotrimazol*, *Myclobutanil* und *Lanosterol*.[32, 33, 34, 35, 37] Von ihnen kopierten wir die SMILES-Strings, erstellten eine Compound library und fuhren wie gehabt fort.

4.2 Docking mit Galaxy

4.2.1 Protein- und Ligandenstruktur eingeben

Auf der Galaxy-Website kann man eine neue Datenanalyse in Form einer sogenannten History erstellen. Darin stehen über das Web-Interface Tools zur Verfügung, die man nacheinander ausführen kann. Es handelt sich um graphisches Programmieren, wobei auch die Daten direkt in der History gespeichert sind. Histories können veröffentlicht oder geteilt werden. Als Leitfaden haben wir das Tutorial *Protein-ligand-docking* verwendet.[4]

Nachdem eine History erstellt war, luden wir in der History als erstes über den `Upload-Data`-Button die PDB -Datei von *Protein DJ-1* von unserem lokalen PC hoch.

Dann brauchten wir eine Compound library, in welcher Moleküle aufgelistet sein sollten,

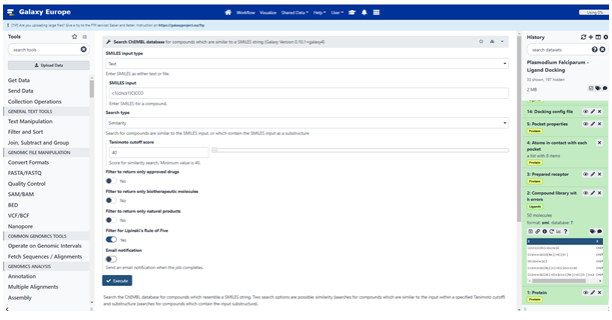


Abbildung 5: GUI von Galaxy Europe

die dem Molekül des natürlichen Liganden vom *Protein DJ-1* ähnelten. Diese erstellten wir, indem wir das auf Galaxy bereitstehende Tool `Search ChEMBL database` nutzten.[2] Dieses Tool ermöglicht es, die ChEMBL-Datenbank mithilfe des Tanimoto-Algorithmus nach Molekülen zu durchsuchen, die ähnlich zu einem Molekül sind, das man über einen einzugebenden SMILES-String übermittelt. Wir gaben also den SMILES-String `(c1(c(ncs1)C)CCO)` des natürlichen Liganden `(5-(2-hydroxyethyl)-4-methylthiazole)` als SMILES-Input ein. Diesen bekamen wir, indem wir die Mol-Datei von *(5-(2-hydroxyethyl)-4-methylthiazole* in einer Galaxy-History hochluden und mit Compound conversion, einem auf Open Babel basierenden Dateiformat-Konvertierungstool, in eine SMILES-Datei umwandelten.[3]

In dem Tool konnte man auswählen, wie hoch der *Tanimoto cutoff score* sein soll. Dieser Score entscheidet, wie hoch die Ähnlichkeit zwischen einem Liganden aus der Datenbank und dem eingegeben Vergleichsliganden sein muss, um in die Compound library aufgenommen zu werden. Ein Bereich von 40% (niedrige Ähnlichkeit ist ausreichend) bis 100% steht zur Verfügung. Wir wählten 40%. Außerdem haben wir ausgewählt, dass der *Filter for Lipinski's Rule of Five* angewendet werden soll. Die Rule of Five ist eine Faustregel, die Moleküle auf ihre orale Bioverfügbarkeit überprüft, also auf die Frage, ob ein Molekül prinzipiell als oral einzunehmendes Arzneimittel geeignet wäre.[28]

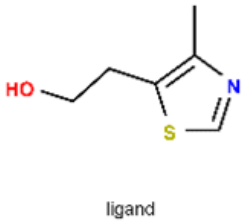


Abbildung 6: Strukturformel natürlicher Ligand

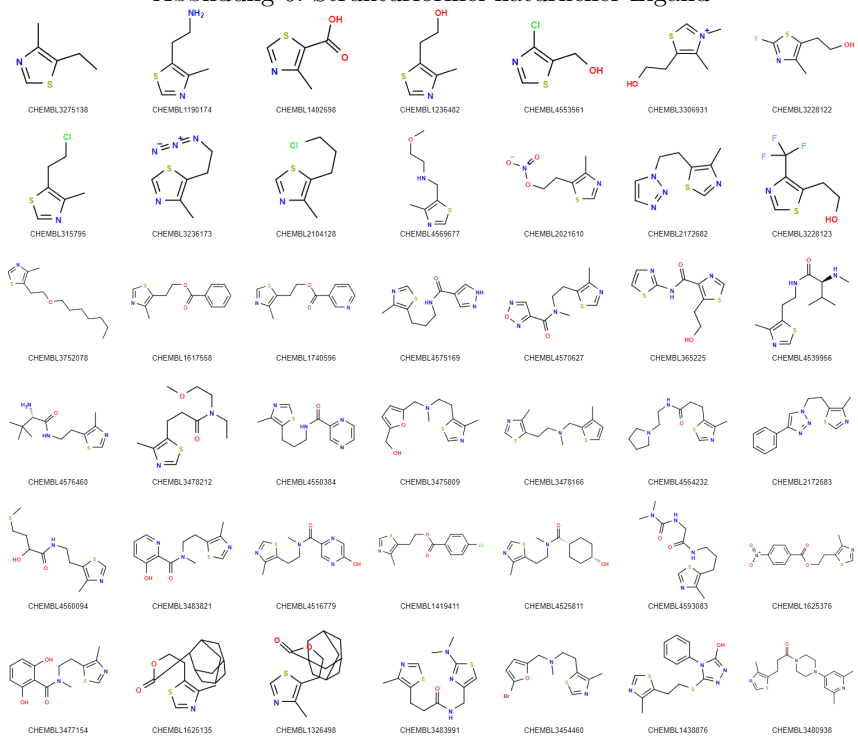


Abbildung 7: Strukturformeln Compound library

Nach dem Ausführen dieses Befehls hatten wir nun also eine Compound library im Format einer SMILES-Liste, die aus 50 Molekülen bestand, die dem natürlichen Liganden des *Protein DJ-1* ähnlich sind, also auch prinzipiell stabile Bindungen erzielen sollten, und außerdem als Medikamente geeignet wären. Die Datei sieht folgendermaßen aus: In jeder Zeile ist ein Molekül aufgelistet, in der ersten Spalte steht der SMILES-String, in der zweiten der Titel, also die ChEMBL-ID des jeweiligen Moleküls.

Die Abbildungen zeigen die Strukturformeln der Compound library und des natürlichen Liganden. Alle Moleküle aus der Compound library haben als Titel ihre ChEMBL-ID. Ähnlichkeiten sind zum Teil klar erkennbar.

4.2.2 Dateien für das Docking vorbereiten

Nun konvertierten wir das Protein, also den Rezeptor, in eine PDBQT-Datei (wir nannten sie *Prepared receptor*), die für das Docking benötigt wurde.

Dann nutzten wir das `fpocket`-Tool, um überhaupt die Stelle auszumachen, an der der Ligand an das Protein docken sollte.[16] Fpocket ist die wohl beste Möglichkeit, Bindungsstellen für kleine Moleküle

zu ermitteln. Es ist ein Open-Source-Programm, welches mithilfe der Alpha-Sphären-Theorie arbeitet. Input war die PDB-Datei des Proteins, Output war zum einen eine Datei namens `Pocket properties`, in der die verschiedenen Eigenschaften der Taschen aufgelistet waren. Am wichtigsten war ihr Gesamtscore, nach dem wurden sie auch sortiert. Pocket1 war mit einem Score von rund 0,4 die beste.

Zum anderen kam eine Dateiliste heraus, die alle acht Taschen im PDB-Format umfasste, also die Atome in Kontakt mit jeder Tasche.

Danach konnten wir mit dem Tool `Calculate the box parameters using RDKit` eine Konfigurationsdatei für das Docking erstellen, die als Input pocket1 hatte und unter anderem die Größe des für das Docking benötigten Bereichs beinhaltet, also x-, y- und z-Puffer.[15] Das Tool nutzt RDKit, eine Sammlung an Programmen für Chemieinformatik und maschinelles Lernen.[14] Die Datei hieß `Docking config file`.

Daraufhin mussten wir nur noch die Compound library mit dem Tool `Compound conversion` in eine SDF-Datei namens `Prepared ligands with errors` (Weshalb „with errors“, wird später erklärt.) umwandeln, und das Docking konnte beginnen.

4.2.3 Docking

Als Docking-Programm verwendeten wir AutoDock Vina, welches bei Galaxy über das Tool `Vina Docking` zur Verfügung steht.[20, 21] AutoDock Vina ist ein Open-Source-Programm, das von `The Scripps Research Institute` entwickelt wurde.[29] In Tests, in denen es mit anderen Molecular-Docking- Simulationsprogrammen verglichen wurde, schnitt es gut bis sehr gut ab und findet dementsprechend die größte Verwendung.

AutoDock Vina legt um die pocket ein virtuelles dreidimensionales Gitter mit sehr kleinen Abständen im Angström-Bereich zwischen den einzelnen Gitterzeilen. An jeder Gitterkreuzung berechnet AutoDock für verschiedene räumliche Orientierungen einen Score, also pro Punkt und Orientierung (Pose). Bei der Scoring-Funktion werden sowohl intramolekulare als auch intermolekulare Bindungs- und Abstoßungskräfte bewertet. Auf diese Weise kann man am Ende die beste Pose eines Liganden zu einer pocket finden.

Als Input verwendeten wir für den Rezeptor die PDBQT-Datei `Prepared receptor`, für die Liganden die SDF-Datei `Prepared ligands with errors` und als Box-Konfiguration die TXT-Datei `Docking config file`. Den pH-Wert setzten wir auf 7.4, den pH-Wert von menschlichem Blut.

Das Ergebnis, eine SDF-Dateiliste, war jedoch als Fehler markiert, da acht der 50 SMILES-Strings nicht kompatibel waren. In der Fehlermeldung sahen wir nach, welche Liganden Probleme machten. Dort waren sieben Liganden aufgelistet, ein weiterer Ligand wurde zwar nicht in der Fehlermeldung aufgeführt, wurde jedoch auch nicht ausgeführt und war dementsprechend auch nicht nutzbar. Wir luden uns daraufhin die Compound library auf unseren PC herunter, öffneten sie mit einem Editor und löschten alle Problem-Liganden (es waren die Liganden Nummer 6, 9, 12, 20, 27, 30, 39 und 45) manuell. Die Datei, die nun also noch 42 Liganden umfasste, luden wir wieder in der Galaxy-History unter dem Namen `Compound library without errors` hoch. Dort konvertierten wir sie wir gehabt in eine SDF-Datei und wiederholten das Docking mit denselben Inputs (selbstverständlich abgesehen von den Liganden). Nun war das Ergebnis eine funktionierende Dateiliste.

4.3 Auswertung der Daten

4.3.1 Compund library aufbereiten

Compound library-Erweiterung Jetzt erstellten wir eine SMI-Datei des natürlichen Liganden, indem wir den schon am Anfang verwendeten SMI-String mit dem Titel `ligand` über den Button `Upload Data` und die Eingabemethode `Paste/Fetch data` mit der Einstellung `Convert spaces to tabs` eingaben. Die Datei nannten wir `Natural ligand`.

Dann nutzten wir das Tool `Concatenate datasets`, um die Datasets `Natural ligand` und `Compound library without errors` zu einer SMI-Datei mit 43 Liganden zusammenzuführen. Dieses Dataset nannten wir `Labelled compound library`.

Fingerprints und Clustering Mit dem Tool `Molecule to fingerprint` erstellten wir aus der Labelled compound library eine Open Babel FP2 fingerprints-Datei.[8, 18] Molekulare Fingerprints kodieren Molekülstrukturen in Bits. Mit ihnen kann man Moleküle auf ihre Ähnlichkeit untersuchen.

Danach konnten wir mit der Fingerprints-Datei als Input die Liganden mit dem Tool `Taylor Butina clustering` in Cluster einsortieren.[6] Der Taylor-Butina Algorithmus erstellt Cluster, in denen Moleküle

aufgelistet sind, die zu dem zentralen Molekül des Clusters eine bestimmte Ähnlichkeit aufweisen. Den Ähnlichkeits-Schwellenwert kann man im Tool angeben, wir haben 0,8 verwendet. Das Taylor-Butina Clustering ist (im Gegensatz zum nächsten Clustertool, welches wir verwenden werden) nicht hierarchisch, es gibt also keine Über- und Untercluster.

Dann erstellten wir noch mit dem Tool `NxN clustering` ein Cluster-Dendogramm. Jenes wurde erzeugt, indem das Tool eine Selbstähnlichkeitsmatrix erstellte, um die Ähnlichkeit zu ermitteln.

4.3.2 Ergebnisse des Dockings auswerten

Veranschaulichung Um die Ergebnisse des Dockings, die bisher noch in unterschiedlichen SDF-Dateien lagen, übersichtlicher zu machen, verwendeten wir das Tool `Extract values from an SD-file`. Input war die im Docking-Schritt erstellte Liste aus SDF-Dateien, Output eine Liste aus Tabular-Dateien, die die Informationen anders, übersichtlicher, darstellten.

Nun wäre *eine* Datei mit allen Ergebnissen praktisch. Um dies zu realisieren kam das Tool `Collapse Collection` zum Einsatz. Dieses reihte alle vom gerade benutzen Tool erstellten Dateien in einer Tabular-Datei hintereinander auf. Diese Datei sortierten wir mit dem Tool `Sort Dataset` nach dem Docking-Score.

Dann hatten wir eine übersichtliche Datei, in der alle Ergebnisse des Dockings, also für jeden Liganden jede Torsion einzeln, nach Bindungsstabilität sortiert waren.

Von den fünf Liganden, die die stabilsten Bindungen erzielt hatten, konvertierten wir danach noch die SDF-Dateien des Dockings mit `Compound conversion` in PDB-Dateien, um mit dem NGL-Viewer 3D-Visualisierungen von ihnen vornehmen zu können, das ist direkt in Galaxy möglich.[19]

Auf der Website von NGL konnten wir auch direkt die PDB-Datei des Proteins und die gerade erstellte PDB-Datei des Liganden hochladen, und so beide in der beim Docking bewerteten Position visualisiert betrachten.

Substrukturen-Abgleich mit Python Nun wollten wir ChEMBL nach Molekülen durchsuchen, die Moleküle aus unserer `Compound library without errors` als Substrukturen besitzen und die bereits als Medikament verwendet werden oder sich im Prozess der Medikamentenentwicklung befinden. Da ein solcher Abgleich manuell enorm Zeitintensiv gewesen wäre, automatisierten wir diesen mit Python.

Das Repository ist unter `https://github.com/theKevinKretz/Protein-Ligand-Docking` auf GitHub veröffentlicht.

Die ausführbare Datei ist die `main.py`.

Die Funktion `main()` ist die Hauptfunktion. Sie ruft für jede der SMI-Dateien im Inputverzeichnis `Input smi-files/` die Funktion `run_for_disease` auf. Die Funktion `run_for_disease` ruft wiederum die notwendigen Funktionen auf, um die jeweilige SMI-Datei zu bearbeiten. Hierzu wird für jeden SMILES-String in der SMI-Datei über die Funktion `substructure_search()` eine Anfrage zu einer Substrukturen-Suche an die API der ChEMBL-Datenbank gesendet. Die Antwort des Servers wird durch die Funktion `store_file()` gespeichert. Nun wird für jedes Ergebnis der Substrukturen-Suche der Wert `max_phase` bestimmt. Wenn er größer als Null ist, werden die ChEMBL-ID des Originals aus der SMI-Datei, die des Ergebnisses der Substrukturen-Suche, sowie der `max_phase`-Wert in der Dictionary-Variablen `everything` gespeichert. Der Inhalt der Variablen wird in der Datei `output.json` gespeichert und durch die Funktion `make_look_good` aufbereitet und in der Konsole ausgegeben.

Von den gefundenen Ergebnissen suchten wir uns passende SMILES-Strings von Wikipedia, luden diese dann als SMI-Dateien hoch, präparierten sie für das Docking und dockten sie. Die Docking-Datei konvertierten wir wieder in eine Tabular-Datei. Danach konvertierten wir die SDF-Datei des gedockten Liganden in PDB-Dateien, erstellten aus der SMI-Datei des Medikamenten-Liganden und der bereits bestehenden `Labelled compound library` eine `Labelled compound library with drugs` und visualisierten deren Strukturformeln.

Hier die Links zu den Malaria-Histories (SMILES-Vorbereitung und eigentliche History):

`https://usegalaxy.eu/u/leander_schaefer/h/5-2-hydroxyethyl-4-methylthiazole-natural-ligand-of-protein-dj-1`

`https://usegalaxy.eu/u/leander_schaefer/h/protein-dj-1---ligand-docking`

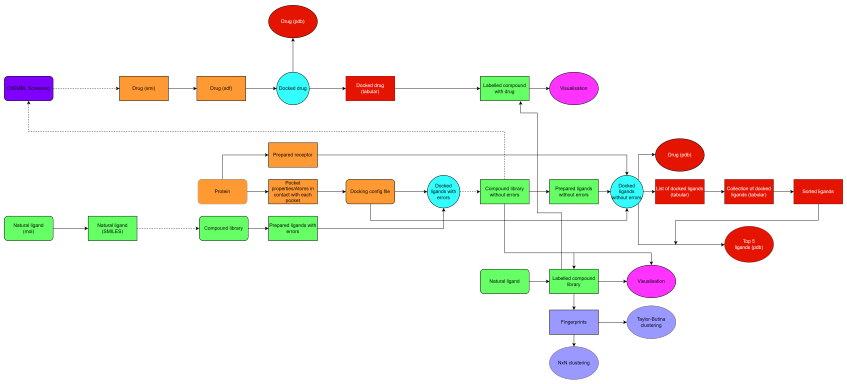


Abbildung 8: Flowchart: Malaria

4.4 Andere Krankheiten

4.4.1 Afrikanische Schlafkrankheit

Nach einem sehr ähnlichen Verfahren wie bei Malaria behandelten wir die Afrikanische Schlafkrankheit, genauer das für sie wichtige Hitzeschockprotein 83. Unterschiede lagen darin, dass das HSP 83 wesentlich mehr Taschen als das *Protein DJ-1* von *P. Falciparum* besaß (57). Außerdem umfasste die Compound library 52 Moleküle, welche von Anfang an alle funktionstüchtig waren, der bei Malaria notwendige Schritt des manuellen Löschens von Problem-Liganden, erneute Aufbereitung und erneutes Docking entfielen somit.

Hier der Link zur Afrikanischen Schlafkrankheits-History:
`https://usegalaxy.eu/u/leander_schaefer/h/hsp-83---ligand-docking-1`

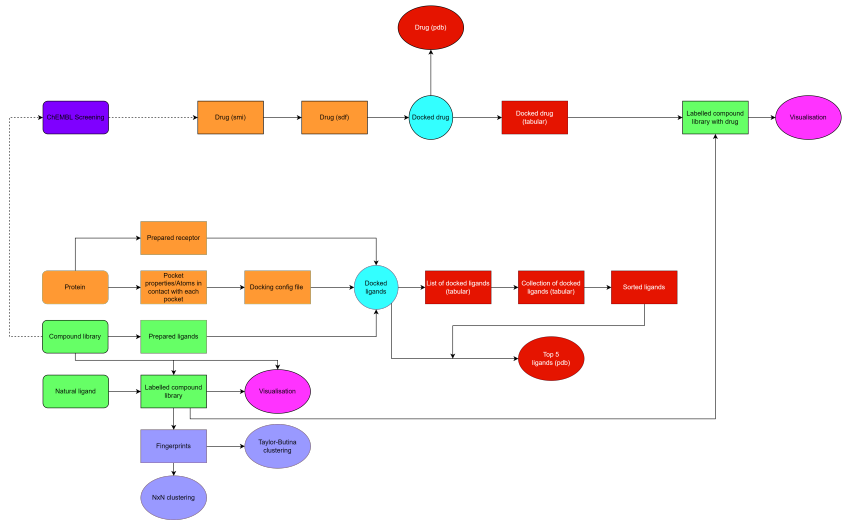


Abbildung 9: Flowchart: Afrikanische Schlafkrankheit

4.4.2 Chagas-Krankheit

Auch bei der Chagas-Krankheit verwendeten wir, nachdem wir das Protein und die Compound library erstellt hatten, das bekannte Verfahren. Das Protein *Sterol 14-alpha demethylase)* besitzt 37 pockets.

Das Python-Script setzen wir hier nicht ein, da alle Liganden bereits als Medikamente in anderen Zusammenhängen eingesetzt werden. Uns ging es in diesem Fall einzig darum, herauszufinden, welche der Azolverbindungen am besten an die erst über Alphafold verfügbar gewordene Struktur des *Trypanosoma cruzi*-Proteins passt.

Hier der Link zur Chagas-History:
`https://usegalaxy.eu/u/leander_schaefer/h/sterol-14-alpha-demethylase---ligand-docking-1`

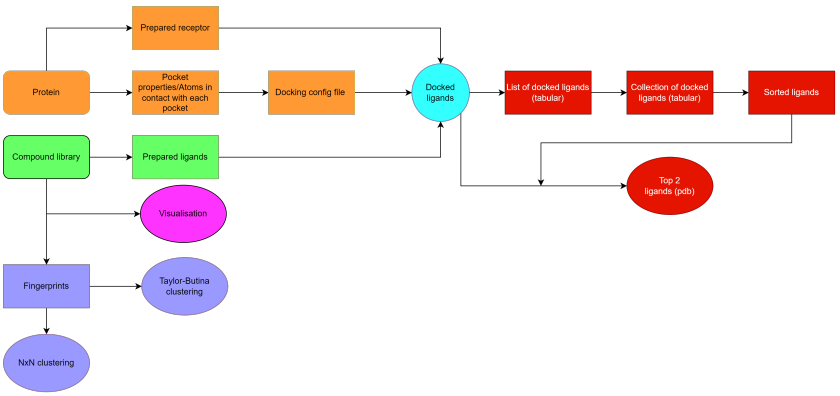


Abbildung 10: Flowchart: Chagas-Krankheit

5 Ergebnisse

5.1 Malaria

5.1.1 Liganden-Clustering

Das Taylor-Butina Clustering ergab, dass im Fall von Malaria fünf Cluster bestehen. Der größte Cluster umfasste fünf Liganden, der zentrale Ligand war `CHEMBL3275138`, die „Cluster-Mitglieder“ waren `CHEMBL315795`, der natürliche Ligand, `CHEMBL1236482` und `CHEMBL1190174`. Des Weiteren existierten drei Cluster mit jeweils drei Liganden und ein kleiner, aus zwei Liganden bestehender Cluster.

Das hierarchische Dendogramm des NxN-Clusterings sah folgendermaßen aus (*ligand* ist nach wie vor der natürliche Ligand).

Je höher die Zahl auf der y-Achse, desto geringer ist die Ähnlichkeit. Die Ergebnisse sind denen des NxN-Clusterings ähnlich.

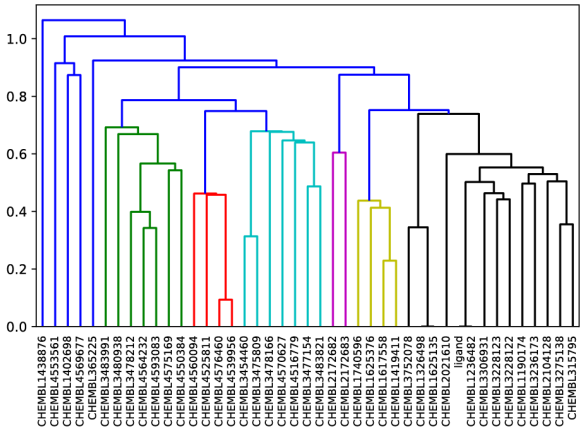


Abbildung 11: NxN-Clustering

5.1.2 Docking-Ergebnisse

Dies ist die Tabular-Datei, die die Docking-Ergebnisse Torsion für Torsion, nach Score sortiert auflistet, hier die besten zehn Ergebnisse (SMILES-Strings aus Gründen der Übersichtlichkeit ausgenommen):

Index	MODEL	RMSD_LB	RMSD_UB	SCORE	SDFMoleculeName	TORSDO
0	1.0	0.0	0.0	-5.3	CHEMBL1326498	F 5
0	1.0	0.0	0.0	-5.3	CHEMBL1625135	F 5
0	1.0	0.0	0.0	-5.2	CHEMBL3480938	F 4
0	1.0	0.0	0.0	-5.1	CHEMBL2172683	F 4
0	1.0	0.0	0.0	-5.1	CHEMBL3477154	F 6
1	2.0	0.972	2.104	-5.1	CHEMBL1625135	F 5
1	2.0	3.111	9.891	-5.1	CHEMBL3480938	F 4
2	3.0	2.624	4.043	-5.0	CHEMBL3480938	F 4
0	1.0	0.0	0.0	-4.9	CHEMBL1419411	F 5
1	2.0	2.01	2.62	-4.9	CHEMBL1326498	F 5

Demnach sind die Liganden, die die fünf stabilsten Bindungen erzielt haben, die Liganden Nummer 7: `CHEMBL1326498`, 38: `CHEMBL1625135`, 19: `CHEMBL3480938`, 36: `CHEMBL2172683` und 42: `CHEMBL3477154`. Dies ist eine NGL-3D-Visualisierungen von Ligand 7.

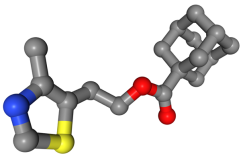


Abbildung 12: CHEMBL1326498

Hier die Visualisierung von `CHEMBL1326498` und *Protein DJ-1* in ihrer stabilsten Bindung:

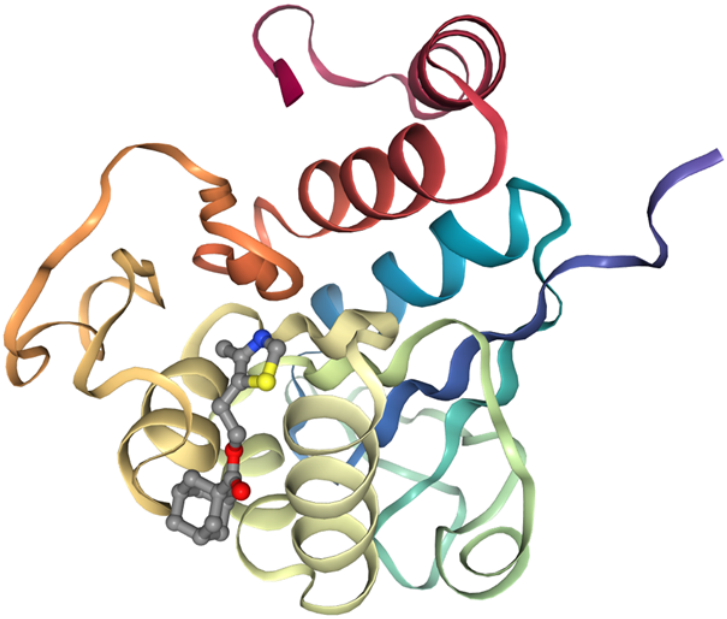


Abbildung 13: Protein DJ-1 mit gedocktem CHEMBL1326498

Das Ergebnis des Python-Scripts bei Malaria war, dass ein Molekül aus unserer Compound library without errors, `CHEMBL315795` oder Colmethiazol unter dem Handelsnamen *Heminevrin* bereits unter anderem gegen Schlafstörungen eingesetzt wird. Beim Docking hatte es mit einem Score von -2,7 aber bestenfalls ein Ergebnis im unteren Bereich.

Der Ligand `CHEMBL1190174` aus unserer Compound library without errors ist ein Teil, also eine Substruktur, von `CHEMBL301265` (*Pramipexole*), dass unter den Handelsnamen *Mirapexin*, *Neliprax*, *Oprymea* und *Pipexus* schon unter anderem gegen Parkinson Verwendung findet.[36] Das Ergebnis des Pramipexole-Dockings war, dass Pramipexole mit einem Score von -4,1 ziemlich gut bindet.

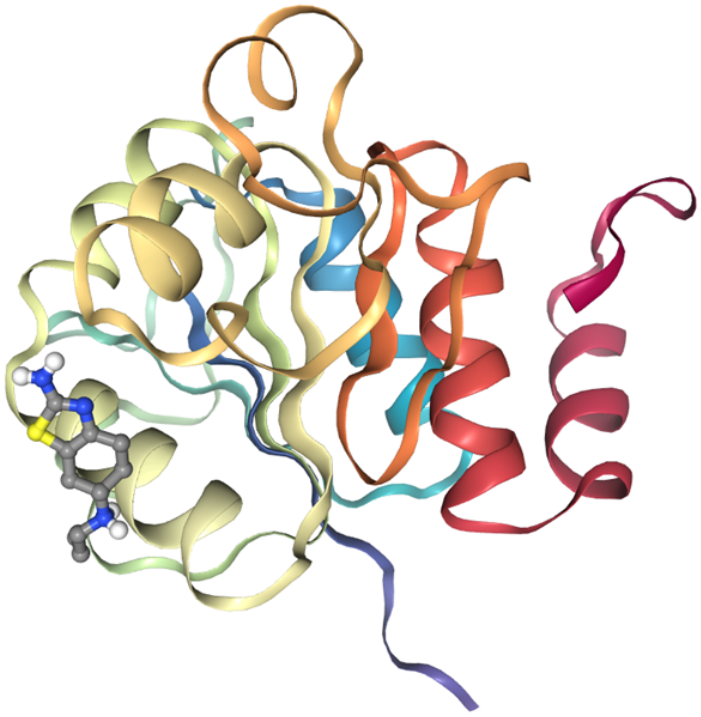


Abbildung 14: Protein DJ-1 mit gedocktem Pramipexole

5.2 Afrikanische Schlafkrankheit

Bei der Afrikanischen Schlafkrankheit erreichten die Liganden 47: `CHEMBL561498` (Score: -11,1), 52: `CHEMBL3235340` (Score: -10,9), 42: `CHEMBL407146` (Score: -10,5), 4: `CHEMBL259487` (Score: -10,4) und 23: `CHEMBL3235351` (Score auch -10,4) die besten Bindungen mit dem Hitzeschockprotein 83.

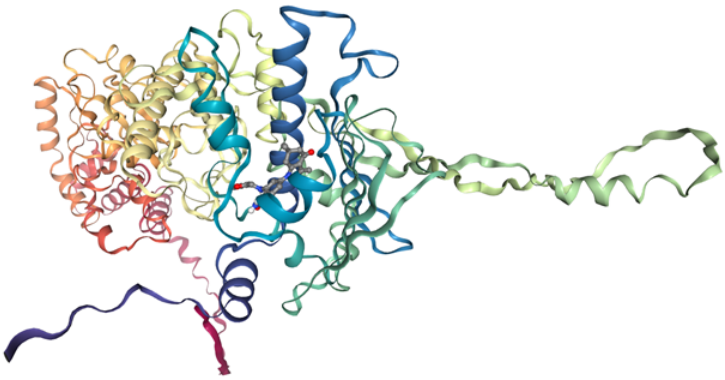


Abbildung 15: HSP83 mit gedocktem CHEMBL561498

Der Lauf des Python-Scripts ergab, dass viele Liganden aus der Compound library Substrukturen von `CHEMBL1195136` waren, welches sich in der Zulassungsphase 2 gegen Neoplasie bei Krebserkrankungen befindet. Den SMILES-String fanden wir auf der vom ChEMBL-Eintrag verlinkten J-Global-Eintrag zum betreffenden Molekül.[12] Es dockte mit einem Score von -10,1 sehr gut.

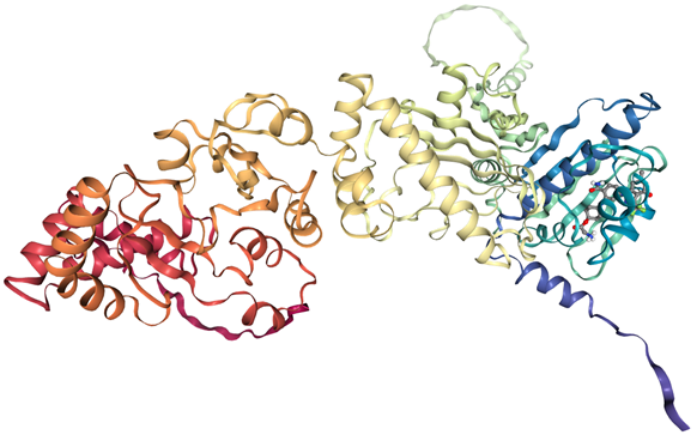


Abbildung 16: HSP83 mit gedocktem CHEMBL1195136

5.3 Chagas-Krankheit

Bei der Chagas-Krankheit erreichte *Itraconazol* mit einem Score von -9.8 einen ähnlich guten Wert für die Bindung an *Sterol 14-alpha demethylase* wie deren natürlicher Ligand *Lanosterol* (-9.7) und stellt damit den vielversprechendsten Inhibitor aus der von uns untersuchten Reihe von Azolverbindungen dar.

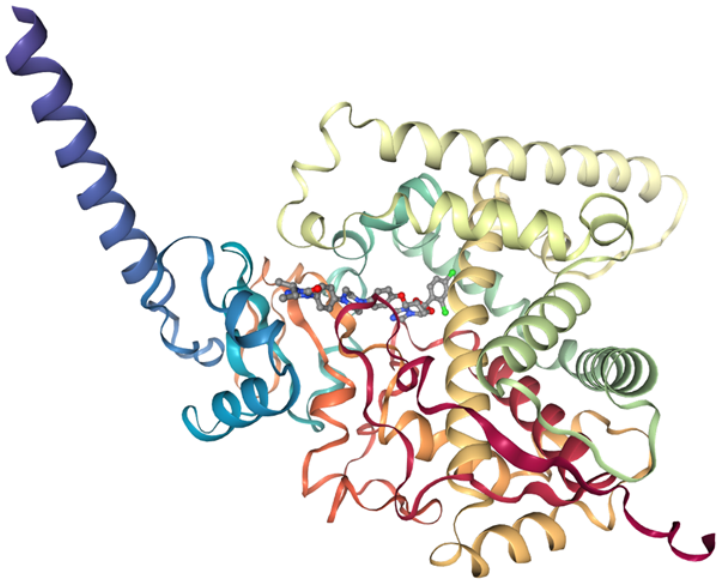


Abbildung 17: Sterol 14-alpha demethylase mit gedocktem Itraconazol

6 Ergebnisdiskussion

Die Suche nach tatsächlich realistischen Liganden hat sich als komplizierter herausgestellt als anfangs erwartet. Erfreulicherweise konnten wir ein Python-Script schreiben, welches uns sehr bei der Arbeit geholfen hat. Dass wir damit auch bereits eingesetzte Arzneimittel gefunden haben, hat uns erstaunt und sehr gefreut.

Mit den vorliegenden Wirkstoffen wäre im Prinzip die Grundlage für ein echtes Heilmittel schon geschaffen. Zur endgültigen Zulassung eines solchen fehlen aber natürlich noch die Tests im Labor und die klinischen Studien für die Wirksamkeit und Unbedenklichkeit bei der Anwendung. Wir haben leider nicht die Ressourcen für solche Studien, weshalb wir diese anderen Institutionen überlassen. Es ist aber dennoch bemerkenswert, welche Möglichkeiten allein die Verwendung öffentlicher Server, Datenbanken und frei verfügbarer Software im Bereich des Medikamenten-Screenings bieten.

Wir hoffen mit unserer Arbeit die Aufmerksamkeit der Öffentlichkeit auf die Krankheiten und die betroffenen Länder zu ziehen und somit einen Beitrag für die Bekämpfung der Tropenkrankheiten zu leisten.

7 Zusammenfassung

Unsere Forschung mit Galaxy und Alphafold hat sich als ergebnisreich erwiesen und dabei haben sich viele Wirkstoffe als vielversprechend erwiesen: Zusammenfassend kann man sagen, dass es uns tatsächlich gelungen ist, Kandidaten für mögliche Medikamente gegen Malaria, die Afrikanische Schlafkrankheit und die Chagas-Krankheit zu finden. So könnte man nun beispielsweise Studien starten, um zu überprüfen, ob das Molekül `CHEMBL561498` tatsächlich als Medikament gegen die Afrikanische Schlafkrankheit wirkt. Noch aufwandsfreier, sowohl zeitlich als auch monetär, wäre es, Studien durchzuführen, in denen man testet, ob beispielsweise Itraconazol, welches bereits als Medikament eingesetzt wird, auch gegen die Chagas-Krankheit wirkt.

8 Quellen- und Literaturverzeichnis

Literatur

[1] `https://malariajournal.biomedcentral.com/track/pdf/10.1186/s12936-021-03865-1.pdf` *Ali, Fawad Wali, Hira Jan, Saadia Zia, Asad Aslam, Muneeba Ahmad, Imtiaz Afridi, Sahib Khan, Asifullah. (2021). Analysing the essential proteins set of Plasmodium falciparum PF3D7 for novel drug targets identification against malaria. Malaria Journal. 10.1186/s12936-021-03865-1* Abgerufen: 15.12.2021

[2] `https://jcheminf.biomedcentral.com/articles/10.1186/s13321-015-0069-3` *Bajusz, D., Rácz, A. Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?. J Cheminform 7, 20 (2015).* Abgerufen: 08.01.2022

[3] `https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-3-33` *Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., Hutchison, G. R. (2011). Open Babel: An open chemical toolbox. Journal of Cheminformatics, 3(1).* Abgerufen: 03.12.2021

[4] `https://training.galaxyproject.org/training-material/topics/computational-chemistry/tutorials/cheminformatics/tutorial.html` *Simon Bray, 2021 Protein-ligand docking (Galaxy Training Materials).* Abgerufen: 28.11.2021

[5] `https://journals.asm.org/doi/full/10.1128/AAC.42.12.3245` *Frederick S. Buckner, Aaron J. Wilson, Theodore C. White, Wesley C. Van Voorhis, Induction of Resistance to Azole Drugs in Trypanosoma cruzi* Abgerufen:

[6] `https://pubs.acs.org/doi/10.1021/ci9803381` *Butina, D. (1999). Unsupervised Data Base Clustering Based on DaylightFingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. Journal of Chemical Information and Computer Sciences, 39(4), 747–750.* Abgerufen: 28.12.2021

[7] `https://www.ebi.ac.uk/pdbe-srv/pdbechem/chemicalCompound/show/HIE` *EMBL-EBI, HIE : Summary* Abgerufen: 05.01.2021

[8] `https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-5-S1-P36` *Dalke, A. (2013). The FPS fingerprint format and chemfp toolkit. Journal of Cheminformatics, 5(S1).* Abgerufen: 28.12.2021

[9] `https://academic.oup.com/nar/article/43/W1/W612/2467881` *Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., ... Overington, J. P. (2015). ChEMBL web services: streamlining access to drug discovery data and utilities. Nucleic Acids Research, 43(W1), W612–W620.* Abgerufen: 03.12.2021

[10] https://www.ebi.ac.uk/chembl/ *Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños MP, Overington JP, Papadatos G, Smit I, Leach AR. (2017) 'The ChEMBL database in 2017.' Nucleic Acids Res., 45(D1) D945-D954.* Abgerufen: 07.12.2021

[11] https://www.ebi.ac.uk/chebi/init.do *Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Res.* Abgerufen: 07.12.2021

[12] https://jglobal.jst.go.jp/en/detail?JGLOBAL_ID=201007072378971804#%7B%22category%22%3A%227%22%2C%22fields%22%3A%5B%7B%22op%22%3A%22AND%22%2C%22nm%22%3A%22SNID%22%2C%22vals%22%3A%5B%7B%22v%22%3A%22J2.821.849D%22%2C%22m%22%3A1%7D%5D%7D%5D%7D *J-Global, SNX-5422* Abgerufen: 13.01.2022

[13] https://alphafold.com *Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021). Varadi, M et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research (2021).* Abgerufen: 03.12.2021

[14] https://sourceforge.net/p/rdkit/wiki/Home/ *Greg Landrum, RDKit Open-Source Cheminformatics and Machine Learning* Abgerufen: 08.01.2022

[15] http://www.rdkit.org/ *Landrum, G. (n.d.). RDKit: Open-source cheminformatics.* Abgerufen: 06.12.2021

[16] https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-10-168 *Le Guilloux, V., Schmidtke, P. Tuffery, P. Fpocket: An open source platform for ligand pocket detection. BMC Bioinformatics 10, 168 (2009).* Abgerufen: 05.12.2021

[17] https://usegalaxy.eu/ *The authors acknowledge the support of the Freiburg Galaxy Team: Dr. Wolfgang Maier and Björn Grüning, Bioinformatics, University of Freiburg (Germany) funded by the Collaborative Research Centre 992 Medical Epigenetics (DFG grant SFB 992/1 2012) and the German Federal Ministry of Education and Research BMBF grant 031 A538A de.NBI-RBC.* Abgerufen: 27.11.2021

[18] http://openbabel.org/docs/dev/Features/Fingerprints.html *Open Babel, Molecular fingerprints and similarity searching* Abgerufen: 12.01.2022

[19] http://nglviewer.org/ngl/ *AS Rose, AR Bradley, Y Valasatava, JM Duarte, A PrliÄ‡ and PW Rose. NGL viewer: web-based molecular graphics for large complexes. Bioinformatics: bty419, 2018.* Abgerufen: 03.01.2022

[20] https://vina.scripps.edu/ *O. Trott, A. J. Olson, AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading, Journal of Computational Chemistry 31 (2010) 455-461* Abgerufen: 03.12.2021

[21] https://onlinelibrary.wiley.com/doi/10.1002/jcc.21334 *Trott, O., Olson, A. J. (2009). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. Journal of Computational Chemistry, NA-NA.* Abgerufen: 03.12.2021

[22] https://www.who.int/teams/control-of-neglected-tropical-diseases *World Health Organization, „Neglected tropical diseases“* Abgerufen: 06.01.2022

[23] https://en.wikipedia.org/wiki/Protein%E2%80%93ligand_docking *Wikipedia, Protein - Ligand Docking* Abgerufen: 05.01.2022

[24] https://de.wikipedia.org/wiki/Malaria#J%C3%A4hrliche_Opfer_und_Inzidenz *Wikipedia, Malaria* Abgerufen: 05.01.2022

[25] https://de.wikipedia.org/wiki/Afrikanische_Trypanosomiasis *Wikipedia, „Afrikanische Trypanosomiasis“* Abgerufen: 06.01.2022

[26] https://en.wikipedia.org/wiki/Galaxy_(computational_biology) *Wikipedia, Galaxy (computational biology)* Abgerufen: 29.11.2021

[27] https://de.wikipedia.org/wiki/Chagas-Krankheit *Wikipedia, Chagas-Krankheit* Abgerufen: 06.01.2022

[28] https://de.wikipedia.org/wiki/Rule_of_Five *Wikipedia, Rule of Five* Abgerufen: 08.01.2022

[29] https://en.wikipedia.org/wiki/AutoDock *Wikipedia, AutoDock* Abgerufen: 08.01.2022

[30] https://en.wikipedia.org/wiki/ChEMBL *Wikipedia, ChEMBL* Abgerufen: 14.01.2022

[31] https://en.wikipedia.org/wiki/Ergosterol *Wikipedia, Ergosterol* Abgerufen: 09.01.2022

[32] https://en.wikipedia.org/wiki/Fluconazole *Wikipedia, Fluconazole* Abgerufen: 09.01.2022

[33] https://en.wikipedia.org/wiki/Miconazole *Wikipedia, Miconazole* Abgerufen: 09.01.2022

[34] https://en.wikipedia.org/wiki/Clotrimazole *Wikipedia, Clotrimazole* Abgerufen: 09.01.2022

[35] https://en.wikipedia.org/wiki/Myclobutanil *Wikipedia, Myclobutanil* Abgerufen: 09.01.2022

[36] https://en.wikipedia.org/wiki/Pramipexole *Wikipedia, Pramipexole* Abgerufen: 13.01.2022

[37] https://en.wikipedia.org/wiki/Lanosterol *Wikipedia, Lanosterol* Abgerufen: 09.01.2022

[38] https://en.wikipedia.org/wiki/Ligand *Wikipedia, Ligand* Abgerufen: 13.12.2021

9 Unterstützungsleistungen

- Herr Dr. Wolfgang Maier, Bioinformatiker, Technische Fakultät Freiburg: Generelle Unterstützung, Wissenschaftlicher Beirat, Korrekturlesen
- Lino Riepenhausen, Schüler: Korrekturlesen