MeetEU Project - Team Heidelberg - Team 1 – Identification and Enhancement of novel Sars-CoV-2 NSP13 Helicase Inhibitors

Linda Blaier, Paul Brunner, Selina Ernst, Valerie Segatz, and Chloé Weiler

February 2024

1 Abstract

Abbreviations

MD	Molecular dynamics
NSP13	Non-structural protein 13
RTC	Replication transcription complex
SAscore	Synthetic accessibility score
ssRNA	Single-stranded RNA
ZBD	Zinc binding domain
SARS-CoV-2	severe acute respiratory syndrome corona virus 2
FDA	food and drug administration
RAS	renin-angiotensin system
COVID-19	coronavirus disease 2019
SARS-CoV-2	SARSI

2 Introduction

Even though the development of vaccines against SARS-CoV-2 was successful during the recent pandemic, the amount of FDA approuved drugs for the therapy of COVID-19 is still limited to Paxlovid, Veklury, Olumiant, and Actemra (FDA 2023). Improving the landscape of drugs available for treating COVID-19 would be particularly beneficial for people who are at risk of severe illness, as vaccines may not fully prevent infections. The goal of this year's Meet-EU project is to develop a pipeline to identify possible inhibtiors against the SARS-CoV-2 helicase also known as **NSP13!** (**NSP13!**). There are two main reasons as to why this protein is a promising drug target. For one, it is highly conserved among corona viruses, which means that the virus is unlikely to develop resistances against drugs targeting **NSP13!** through rapid mutations in the viral genome (Spratt et al. 2021). On the other hand, **NSP13!** together with other non-structural proteins forms the **RTC!** (**RTC!**), which is essential for viral RNA synthesis (**Malone˙2022**). Therefore, inhibiting **NSP13!** would severely hinder the spread of the virus inside the host. The protein consists of five domains, namely the **ZBD!** (**ZBD!**), the stalk domain, as well as 1A, 2A and 1B. The latter three make up the catalytic centre of the protein, where RNA and ATP bind (**NSP13˙basics**).
Computer-aided structure-based drug discovery can be followed to identify possible inhibitors of the **NSP13!** helicase. These can then be further investigated in wet-lab settings. This process involves several steps: (1) Identification of possible binding sites, (2) high-trhoughput screening of ligands for how well they bind the respective pocket, followed by (3) the evaluation of the binding pathways, the kinetics, and thermodynamics (**Sledz˙2018**). Hereby, focussing this screening on well documented or already FDA approved compounds is very attractive, as this drug repurposing potentially shortents the development period and therefore also the development costs (**Pushpakom˙2019**). **ssRNA!** (**ssRNA!**), **SARS-CoV-2!** (**SARS-CoV-2!**)

2.1 Identification of Consensus Binding Pocket

In drug discovery, the initial step is to investigate the protein structure in order to analyse potential binding sites. These are cavities on the surface or interior of the protein with suitable properties to bind a ligand. The functionality of a binding pocket is determined by its shape and location, but also by the amino acid residues which define its pyhsicochemical characteristics (Stank et al. 2016). Different experimental and theoretical procedures exist to analyse the druggability of such binding pockets. In this work, we combined three different *in silico* tools, each following a different algorithm. Fpocket (Le Guilloux et al. 2009) utilises a geometry-based algorithm based on Voronoi tesselation and sequential clustering to determine potential binding sites. We also implemented P2Rank (Krivák and Hoksza 2018; Jendele et al. 2019; Jakubec et al. 2022), which is based on a machine-learning algorithm. P2Rank assigns structural, physicochemical, and evolutionary features to points on the solvent-accessible surface of a protein. From this information, the machine-learning model is built and used to predict and rank potential ligand binding sites. Lastly, FTMAP (Brenke et al. 2009) was used to validate the binding pocket found with the previously mentioned approaches. FTMAP uses docking results of sixteen small molecules differing in polarity, shape, and size to identify binding hot spots with a fast Fourier transform correlation. The most favourable docked confirmations are determined through energy minimisation and clustering processes. Finally, the results of all three tools were combined to identify a consensus binding pocket of the NSP13 helicase. The resulting coordinates of the consensus binding pocket were then used for molecular docking simulations.

2.2 Molecular Docking

Molecular Docking programs are used to evaluate binding affinities between a potential drug candidate and the target protein. A key aspect of this task is the prediction of the ligand position, orientation, and conformation. Search-based methods approach this task by continuously modifying the ligand pose, while estimating its quality or likelihood (score) and stochastically trying to infer the global optimum of the scoring function. Among the most widely used tools are AutoDock Vina (**Trott.2010**) and Glide (**Halgren.2004**), which mainly differ in their scoring functions. However, such search-based methods are computationally expensive. Therefore, in order to be able to screen large datasets, search-based methods are generally restricted to a previously defined binding pocket (**Corso.2022**). Consequently, potential other binding sites of a ligand are not assessed. Machine learning-based blind docking approaches try to address that problem by stochastically predicting binding pocket and ligand pose based on learned characteristics and aligning them. The most promising results are achieved by using Diffdock (**Corso.2022**), a generative model which applies a reverse diffusion process to the docking paradigm. In this manner, Diffdock iteratively transforms an uninformed noisy distribution over ligand poses defined by the degrees of freedom involved in docking (position, turns around its centre of mass, and twists of torsion angles) into a learned model distribution (**Corso.2022**). Corso *et al.* thereby describe this process as a progressive refinement of random ligand poses via updates of their translations, rotations and torsion angles.

2.3 clustering of compounds

In an effort to gain a better understanding of the compounds that emerged from our pipeline, our top scoring compounds from our docking simulation were clustered with those of the teams we were paired with ("Sorbonne5" and "Warsaw1"). Direct comparison of the compounds is made challenging by the fact that different binding sites were chosen by the three teams. Nevertheless, this clustering is still interesting, as it may provide an insight into the structural similarity of compounds that can potentially inhibit the **NSP13!** helicase Spratt et al. 2021.

2.4 Estimation of Toxicity and Synthetic Accessibility

In addition to determining the activity of novel drug candidates on the therapeutic target, predicition of toxic effects is an indispensible step in drug design to be able to assess the predicted risk vs. benefit ratio of the potential drug **roncaglioni2013silico**. As conventional *in vivo* animal tests are time-consuming, expensive, and ethically controversial, researchers nowadays tend to favour *in silico* methods as they are significantly cheaper and faster than wet expirements and they allow for simultaneous evaluation of large numbers of potential drug candidates **raies2016silico**; **roncaglioni2013silico**. Therefore, *in silico* toxicity tests are routinely integrated into the early stages of drug discovery in an attempt to minimise late-stage failures in drug design **dearden2003silico**. Moreover, novel drugs must not only ensure the safety of patients but also have the capability for large-scale synthesis in order to one day be commercially viable. For that reason, determination of the synthetic accessibility, that is the ease of synthesis of a chemical compound, is essential for estimating the feasibility of an active compound as a pharmaceutical **boda2007structure**. Therefore, in our pipeline we followed up the identification of the lead compounds that exhibit optimal binding affinity within the consensus pocket with an evaluation of the general toxicity and synthetic accessibility of these compounds to help estimate the suitability of the compounds as real-life pharmaceuticals against **COVID-19!** (**COVID-19!**).

2.5 Molecular Dynamics Simulation

As the last step of our pipeline, a **MD!** (**MD!**) simulation is conducted using the best-scoring compound as a ligand in the binding pocket of the NSP13 protein. Using GROMACS (Version 2023.3) (Abraham et al. 2015), this enables us to interpret the stability of the protein-ligand interaction, as well as to identify important residues for the interaction. Using a given force-field, a set of equations describing different forces between the atoms and residues in the protein and ligand, the movement of all atoms in the system can be simulated and analysed. However, this is only possible in a very limited timeframe with a small time step size. As this process is rather resource-heavy, it has to be conducted on a cluster with access to a GPU.

3 Material and Methods

3.1 Toxicity and Synthetic Accessibility Prediction using *e*ToxPred

The general toxicity and synthetic accessibility of the each compound was estimated using the machine-learning tool *e*ToxPred (**pu2019toxpred**). The SMILES files of the Top100 compounds from AutoDock Vina (**Trott.2010**) served as input for the pre-trained model.The toxicity predictor was pre-trained on the FDA-approuved and the KEGG-drug datasets whose compounds were considered non-toxic as well as the TOXNET and the T3DB datasets whose compounds were considered toxic using a deep-belief-network based model. This predictor yields a Tox-score between 0 and 1 and in accordance to the paper, all compounds with a Tox-score below 0.58 were deemed non-toxic. The synthetic accessibility was reflected in a synthetic accessibility score (SAscore) which was obtained by training an extra-trees-based classifier on NuBBE, UNPD, FDA-approuved, and DUD-E-active datasets.

Clustering of Compounds Lastly, for a deeper understanding of the potential inhibitors of **NSP13!**, the 10 top scoring compounds from *Glide* as well as the top 10 best-docking ligands of the teams ”Sorebonne5” and ”Warsaw1” were clustered. For one, all 30 compounds were hierarchically clustered based on structural similarity. This was done using the *ChemmineR* package from the *Cheminformatic Toolkit for R* **ChemmineR**. In a first step, the atom pair descriptors for each compound were calculated. These descriptors were then used to calculate a similarity matrix using the Tanimoto coefficient. Substraction of the Tanimoto coefficients from 1 resulted in a distance matrix which was then clustered using the single-linkage method. The resulting heatmap is shown in Figure **??**. Secondly, respective compounds were clustered based on their physico-chemical properties. To this end, the **HBD!** (**HBD!**), the **HBA!** (**HBA!**), the **MW!** (**MW!**), and the **LogP!** (**LogP!**) values were calulcated for each compound using *Open Babel Descriptors* **<empty citation>** To assess the oral activity and drug-like characteristics of the identified compounds, we applied the filtration criteria outlined in Lipinski’s Rule of Five **Lipinski˙1997**. The resulting heatmap of the compounds fulfilling this rule is shown in Figure **??**.

4 Results

4.1 Top 100 Compounds exhibit low toxicity and high synthetic accessibility

After identifying the top 100 best-binding ligands through AutoDock Vina, our subsequent analysis focused on evaluating their practical applicability as potential drugs by considering their predicted toxicity and synthetic accessibility. The resulting SAscore and Tox-Score for each compound were visualised in a scatter plot as seen in Figure **??**.

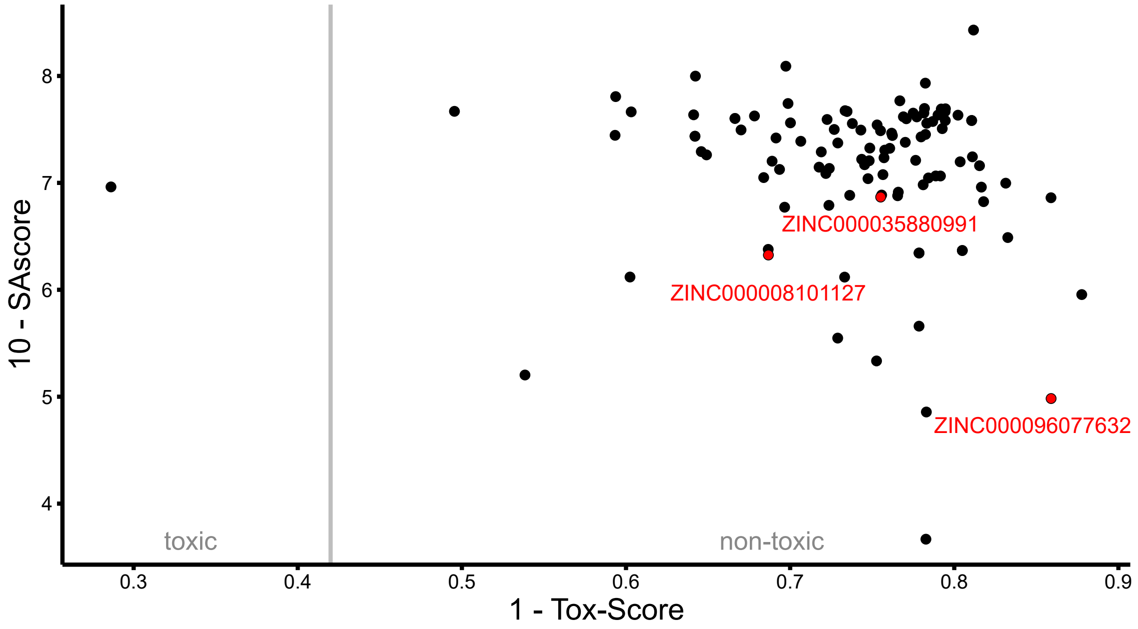

Figure 1: **Scatter plot of predicted toxicity and synthetic accessibility of the 100 best-binding compounds.** The predicted **SAscore!** (**SAscore!**) was plotted against the Tox-Score for the top 100 best scorers from AutoDock Vina. The top 3 scorers from Glide are highlighted in red. The vertical gray line represents the threshold for the toxicity, with the compounds to the right of this line being considered non-toxic.

Of those 100 compounds, 99 presented with a Tox-score below the threshold of toxicity, indicating a low probability of being toxic to humans. The overall median Tox-Score is 0.24, with a mean of 0.26 across all compounds. Across all compounds the median SAscore of 2.69 and a mean of 2.87 which suggests that they are genereally easy to synthesise. The top scorer from Glide, ZINC000096077632, was predicted to have a Tox-Score of 0.14 and an SAscore of 5.02.

5 Discussion and Outlook

5.1 toxicity

The Tox-Score predictor of *e*ToxPred was trained using **FDA!** (**FDA!**)-approved dataset as non-toxic incidences. Consequently, the low mean Tox-Score of the tested compounds aligns with our expectation, considering that the compounds from the ZINC database are derived from an **FDA!**-approved dataset. The single toxic incidence we detected was from the ECBD database which was comprised of **FDA!**-approved and non-**FDA!**-approved molecules. As highlighted by **pu2019toxpred**, natural compounds typically exhibit higher **SAscore!** values compared to synthetic compounds due to their inherent complexity **pu2019toxpred**. The relatively high **SAscore!** of ZINC000096077632 can be explained by the fact that ZINC000096077632 corresponds to angiotensin-(1-7) which is a naturally occuring compound with a crucial role in the **RAS!** (**RAS!**) **santos2014angiotensin**. The analysis of natural compound datasets by **pu2019toxpred** revealed a bimodal distribution in the **SAscore!**, with peaks around 3 and 5. Furthermore the very low Tox-Score of the top scorer can also be explained by the fact that it is a naturally occuring molecule in the human body.

6 Supplementary Material

References

Abraham, M. J., T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl (2015). “GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers”. In: *SoftwareX* 1, pp. 19–25. ISSN: 2352-7110. DOI: `10.1016/j.softx.2015.06.001`.

Berta, D., M. Badaoui, S. A. Martino, P. J. Buigues, A. V. Pisliakov, N. Elghobashi-Meinhardt, G. Wells, S. A. Harris, E. Frezza, and E. Rosta (2021). “Modelling the active SARS-CoV-2 helicase complex as a basis for structure-based inhibitor design”. In: *Chemical Science* 12.40, pp. 13492–13505.

Brenke, R., D. Kozakov, G. Y. Chuang, D. Beglov, D. Hall, M. R. Landon, C. Mattos, and S. Vajda (2009). “Fragment-based identification of druggable ’hot spots’ of proteins using Fourier domain correlation techniques”. In: *Bioinformatics* 25.5, pp. 621–7. ISSN: 1367-4803 (Print) 1367-4803. DOI: `10.1093/bioinformatics/btp036`.

Eastman, P., J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande (2017). “OpenMM 7: Rapid development of high performance algorithms for molecular dynamics”. In: *PLOS Computational Biology* 13.7, e1005659. DOI: `10.1371/journal.pcbi.1005659`. URL: `https://doi.org/10.1371/journal.pcbi.1005659`.

FDA (2023). *Know Your Treatment Options for COVID-19*. URL: `https://www.fda.gov/consumers/consumer-updates/know-your-treatment-options-covid-19`.

Jakubec, D., P. Skoda, R. Krivak, M. Novotny, and D. Hoksza (2022). “PrankWeb 3: accelerated ligand-binding site predictions for experimental and modelled protein structures”. In: *Nucleic Acids Research* 50.W1, W593–W597. ISSN: 0305-1048. DOI: `10.1093/nar/gkac389`. URL: `https://doi.org/10.1093/nar/gkac389`.

Jendele, L., R. Krivak, P. Skoda, M. Novotny, and D. Hoksza (2019). “PrankWeb: a web server for ligand binding site prediction and visualization”. In: *Nucleic Acids Research* 47.W1, W345–W349. ISSN: 0305-1048. DOI: `10.1093/nar/gkz424`. URL: `https://doi.org/10.1093/nar/gkz424`.

Krivák, R. and D. Hoksza (2018). “P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure”. In: *Journal of Cheminformatics* 10.1, p. 39. ISSN: 1758-2946. DOI: `10.1186/s13321-018-0285-8`. URL: `https://doi.org/10.1186/s13321-018-0285-8`.

Le Guilloux, V., P. Schmidtke, and P. Tuffery (2009). “Fpocket: An open source platform for ligand pocket detection”. In: *BMC Bioinformatics* 10.1, p. 168. ISSN: 1471-2105. DOI: `10.1186/1471-2105-10-168`. URL: `https://doi.org/10.1186/1471-2105-10-168`.

Newman, J. A., A. Douangamath, S. Yadzani, Y. Yosaatmadja, A. Aimon, J. Brandão-Neto, L. Dunnett, T. Gorrie-stone, R. Skyner, D. Fearon, M. Schapira, F. von Delft, and O. Gileadi (2021). “Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase”. In: *Nature Communications* 12.1, p. 4848. ISSN: 2041-1723. DOI: `10.1038/s41467-021-25166-6`. URL: `https://doi.org/10.1038/s41467-021-25166-6`.

Spiegel, J. O. and J. D. Durrant (2020). “AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization”. In: *Journal of Cheminformatics* 12.1, p. 25. ISSN: 1758-2946. DOI: `10.1186/s13321-020-00429-4`.

Spratt, A. N., F. Gallazzi, T. P. Quinn, C. L. Lorson, A. Sönnerborg, and K. Singh (2021). “Coronavirus helicases: Attractive and unique targets of antiviral drug-development and therapeutic patents”. In: *Expert opinion on therapeutic patents* 31.4, pp. 339–350.

Stank, A., D. B. Kokh, J. C. Fuller, and R. C. Wade (2016). “Protein Binding Pocket Dynamics”. In: *Accounts of Chemical Research* 49.5. doi: 10.1021/acs.accounts.5b00516, pp. 809–815. ISSN:

0001-4842. DOI: 10.1021/acs.accounts.5b00516. URL: https://doi.org/10.1021/acs.accounts.5b00516.