Joshua Gabbay
Rohs Lab
12/6/24

Molecular Docking and Convolutional Neural Network Classification of Three BRAF V600E Conformations

Abstract:

The BRAF V600E mutation, a key oncogenic driver in multiple cancers, is necessary to investigate for the development of novel therapeutic inhibitors to address current drug resistance and improve patient outcomes. This study integrates molecular docking, clustering, and machine learning to identify potential inhibitors for the BRAF V600E receptor across its three active conformations. Using AutoDock QVina, docking simulations identified high-affinity binders from a dataset of ChEMBL molecules and a large, randomized set of ZINC molecules. Dimensionality reduction with UMAP and clustering via KMeans facilitated structural analysis, while a convolutional neural network (CNN) classified molecules as binders for specific conformations or non-binders. High potential binders were isolated, and the results underscored the potential of combining molecular docking and machine learning to accelerate the discovery of targeted cancer drugs.

Introduction:

BRAF is a serine/threonine protein kinase that, as a dimer, plays a pivotal role in regulating cell growth through the MAPK/ERK signaling pathway. Mutations in BRAF are associated with a range of cancers, with the V600E mutation being one of the most clinically significant.[1] This mutation, which replaces valine (Val) with glutamate (Glu) at position 600, disrupts the hydrophobic interactions that stabilize the inactive conformation of BRAF. Consequently, BRAF V600E becomes active as a monomer, bypassing normal regulatory mechanisms and driving unregulated cell growth.[2] This makes it a critical target for therapeutic intervention in cancers such as melanoma, thyroid, ovarian, and colorectal cancers, where the mutation is highly prevalent.[3]

BRAF V600E has three conformations that can be bound to, depending on where two key structures lie: the alpha-C helix and the DFG motif. The three conformations represent three structural make-ups: alpha-C helix in/DFG out, alpha-C helix out/DFG in, and alpha-C helix in/DFG in.[4] These conformations are critical for understanding the receptor's binding preferences and potential inhibitor interactions. It is important to explore all these different

[1] Hanrahan, A.J., Chen, Z., Rosen, N. *et al.* BRAF — a tumour-agnostic drug target with lineage-specific dependencies. *Nat Rev Clin Oncol* 21, 224–247 (2024). https://doi.org/10.1038/s41571-023-00852-0

[2] Śmiech, M., Leszczyński, P., Kono, H., Wardell, C., & Taniguchi, H. (2020, November 12). *Emerging BRAF mutations in cancer progression and their possible effects on transcriptional networks*. Genes. https://pmc.ncbi.nlm.nih.gov/articles/PMC7697059/

[3] Kowalewski, A., Durślewicz, J., Zdrenka, M., Grzanka, D., & Szylberg, Ł. (2020, August). *Clinical relevance of BRAF V600E mutation status in brain tumors with a focus on a novel management algorithm*. Targeted oncology. https://pmc.ncbi.nlm.nih.gov/articles/PMC7434793/

[4] Cotto-Rios, X.M., Agianian, B., Gitego, N. *et al.* Inhibitors of BRAF dimers using an allosteric site. *Nat Commun* 11, 4370 (2020). https://doi.org/10.1038/s41467-020-18123-2

conformations because of their structural differences – even minute changes can influence whether a molecule would have a high chance of binding.

Current FDA-approved therapies, including Vemurafenib, Dabrafenib, and Encorafenib, effectively target the monomeric form of BRAF V600E and have shown remarkable efficacy in treating metastatic melanoma.[5] However, resistance mechanisms, such as RAF dimerization and reactivation of ERK signaling, limit the long-term success of these treatments. These resistance mechanisms are essentially a paradoxical activation – when an inhibited monomeric BRAF V600E forms a dimer, the drug has reduced affinity for the second BRAF monomer, leading to reactivation of downstream pathways.[6] These challenges highlight the need for new approaches to identify inhibitors capable of targeting both monomeric and dimeric forms of BRAF, potentially overcoming resistance and improving therapeutic outcomes.

The integration of multiple computational tools creates a robust and versatile pipeline that combines the strengths of docking simulations, clustering methods, and machine learning models. By applying this innovative computational approach, this study seeks to identify and categorize promising BRAF V600E inhibitors efficiently and systematically. These insights not only inform future experimental validation efforts but also advance the broader understanding of how machine learning and computational chemistry can be harnessed to address pressing challenges in drug discovery.

Materials and Methods:

In this study, advanced computational tools are used to accelerate the discovery of potential inhibitors for the BRAF V600E receptor. These tools address key challenges in drug discovery by enabling efficient and high-throughput screening, clustering, and classification of drug-like molecules. Each tool serves a specific and critical purpose in the pipeline.

To conduct computational molecular docking, Autodock QVina is a tool used to predict the binding affinities of drug-like molecules to the BRAF V600E receptor. By simulating the binding interactions, docking provides valuable preliminary insights into which molecules are likely to form stable and high-affinity complexes with the receptor. This step has been foundational in narrowing down large chemical libraries for promising candidates.

Additionally, high-dimensional data, which is common when dealing with drug-like molecules, can be visualized and analyzed using UMAP. This dimensionality reduction technique simplifies complex data, allowing for clearer visualization of molecular relationships and structural similarities. KMeans clustering can then be applied to group molecules based on

[5] Leonardi, Giulia & Candido, Saverio & Carbone, Maurizio & Raiti, Fabio & Colaianni, Valeria & Garozzo, Sebastiano & Cinà, Diana & McCubrey, James & Libra, Massimo. (2012). BRAF mutations in papillary thyroid carcinoma and emerging targeted therapies (Review). Molecular medicine reports. 6. 687-94. 10.3892/mmr.2012.1016.

[6] Degirmenci, Ufuk & Wang, Mei & Hu, Jiancheng. (2020). Targeting Aberrant RAS/RAF/MEK/ERK Signaling for Cancer Therapy. Cells. 9. 10.3390/cells9010198.

shared features, revealing distinct clusters of potential inhibitors. These insights can be critical for prioritizing molecules for further development.

Finally, using a Convolutional Neural Network (CNN), which was originally developed for tasks like image recognition, can be utilized on molecular data. By detecting spatial and feature-based patterns in molecular data, CNNs excel in identifying key structural motifs associated with binding potential. This machine learning approach adds a layer of predictive power to identifying high and low potential inhibitors.

The following pipeline is specifically designed to analyze the BRAF V600E receptor in three distinct conformations – based on PDB structures 6p7g, 4xv2, and 4e26 – that reflect variations in the alpha-C helix and DFG motif positions: alpha-C helix in/DFG out, alpha-C helix out/DFG in, and alpha-C helix in/DFG, respectively.

Data preprocessing:

PDB structures for the BRAF V600E receptor (6p7g, 4xv2, and 4e26) were downloaded and cleaned to remove water molecules, ions, and other non-relevant entities. Missing residues in the PDB structures were modeled and filled to ensure complete binding pockets using the homology modeling tool MODELLER. SMILES strings of ChEMBL compounds were converted into 3D molecular structures, including their stereoisomers and multiple conformers to ensure the best possible docking pose. Molecules larger than a predefined threshold (based on QVina limits) were excluded to avoid issues in the docking pipeline. Each 3D structure was optimized using the Merck Molecular Force Field to minimize energy and stabilize the conformers. Conformers of each molecule were then clustered using root-mean-square-deviation to identify representative structures for each cluster. These representative conformers were selected for docking to reduce computational runtime while still preserving diversity.

Docking:

Molecular docking simulations were performed using AutoDock QVina. Docking scores predicted the binding affinity between compounds and the BRAF V600E receptor. The docking simulations were run in parallel using multiprocessing for efficiency. The pipeline was repeated for each of the three receptor conformations. Docking scores of ChEMBL molecules were compared across conformations to identify which conformational state provided the best fit. This analysis helped establish whether a molecule had potential to bind and, if so, a baseline for predicting preferred binding conformations.

ZINC Molecule Screening:

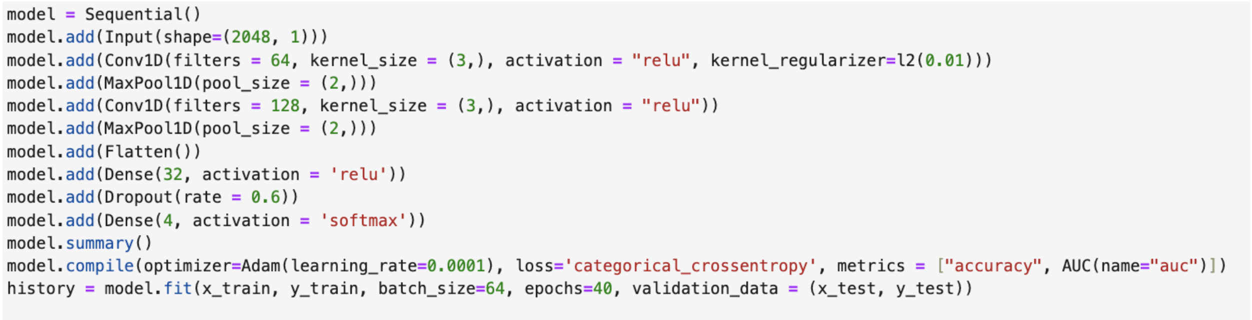
A subset of ZINC database molecules was converted from SMILES to 3D structures, repeating the process outlined above for ChEMBL molecules. Docking simulations were run for each ZINC molecule using the same pipeline as above. Docking results were compared to known good binders from ChEMBL to assess the pipeline’s ability to recover high-quality hits.

Neural Network Data Preprocessing:

Molecular fingerprints were generated for each molecule, converting chemical information into binary feature vectors. Molecules were labeled as good binders or poor binders based on a threshold binding score derived from ChEMBL data. UMAP was then performed to reduce the high-dimensional fingerprint data into 2D space for visualization and clustering. KMeans clustering was applied to the UMAP-reduced data to group molecules into three clusters. The clustering quality was evaluated using silhouette scores for both the available docked molecule labels and UMAP-to-KMeans labels. Resampling techniques were employed to balance the dataset between binders of each conformation, labelled 1, 2, or 3, and poor binders, labelled 0.

Neural Network Generation:

A convolutional neural network was developed for multiclass classification to distinguish between non-binders and binders for each of the three conformations. The CNN architecture and hyperparameters (such as filter size, number of filters, batch size, epochs, learning rate, and regularization techniques) were systematically tuned to optimize performance and avoid overfitting.



```
model = Sequential()
model.add(Input(shape=(2048, 1)))
model.add(Conv1D(filters = 64, kernel_size = (3,), activation = "relu", kernel_regularizer=l2(0.01)))
model.add(MaxPool1D(pool_size = (2,)))
model.add(Conv1D(filters = 128, kernel_size = (3,), activation = "relu"))
model.add(MaxPool1D(pool_size = (2,)))
model.add(Flatten())
model.add(Dense(32, activation = 'relu'))
model.add(Dropout(rate = 0.6))
model.add(Dense(4, activation = 'softmax'))
model.summary()
model.compile(optimizer=Adam(learning_rate=0.0001), loss='categorical_crossentropy', metrics = ["accuracy", AUC(name="auc")])
history = model.fit(x_train, y_train, batch_size=64, epochs=40, validation_data = (x_test, y_test))
```

Neural Network Analysis:

The docking results from the earlier molecular docking simulations for ZINC molecules were obtained, where each molecule was evaluated for its binding affinity to the BRAF V600E receptor (6p7g conformation). The SMILES code for each ZINC molecule was found and converted into molecular fingerprints. These fingerprints were used as input to the CNN, and predicted whether the molecule would bind, and if so, to which conformation. The ZINC molecules that were predicted to bind to the 6p7g class were analyzed in conjunction with their docking score.

This approach integrated docking methods and machine learning predictions, allowing for the identification of ZINC molecules with a high probability of binding to the BRAF V600E alpha-C helix in/DFG out conformation. The process ensured that docking results were validated and supplemented by CNN-based classification, providing a framework for prioritizing molecules for further validation.

Figures:

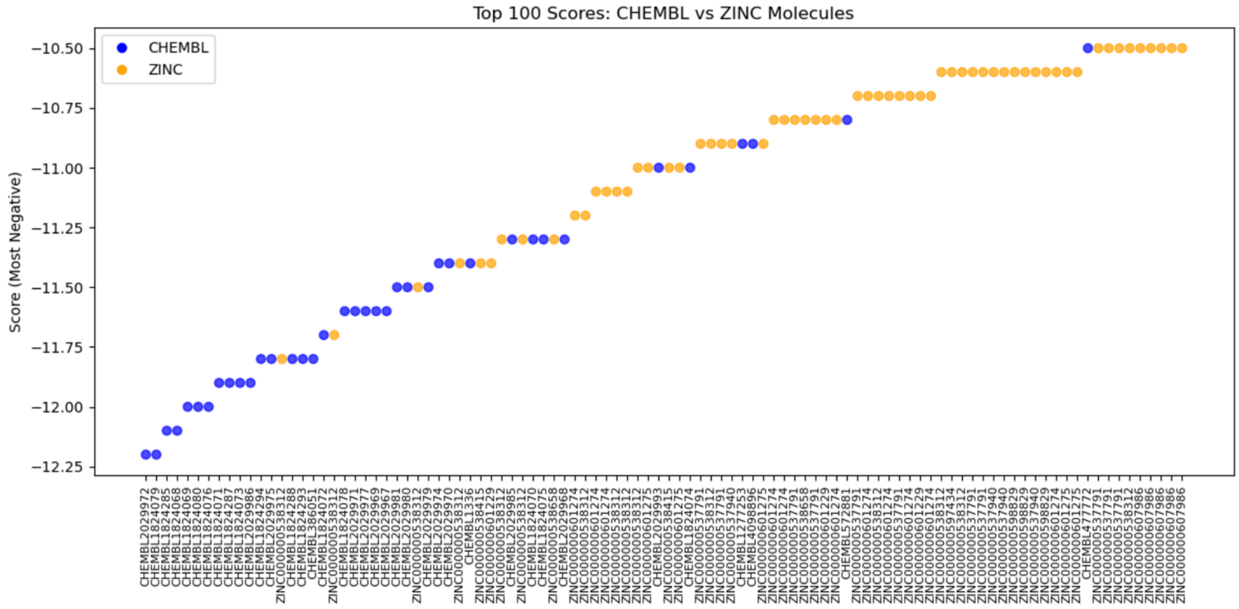


Figure 1. A graph showing the top 100 highest scoring binders amongst the dataset. ChEMBL molecules (known good binders) are colored blue and ZINC molecules are colored yellow.

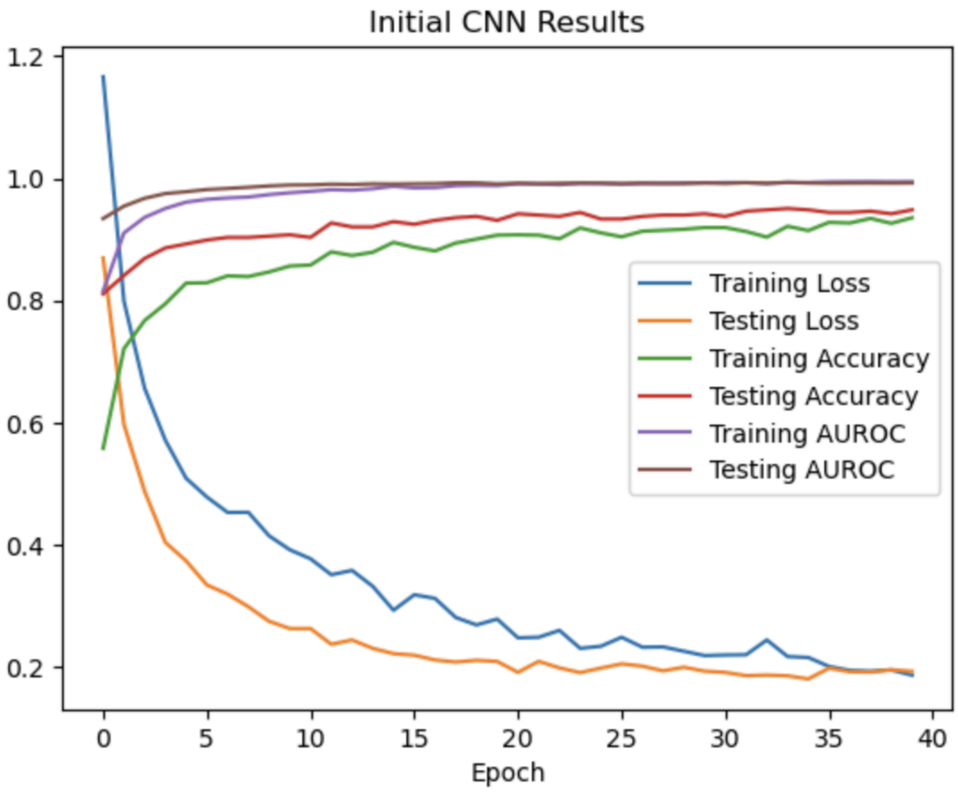
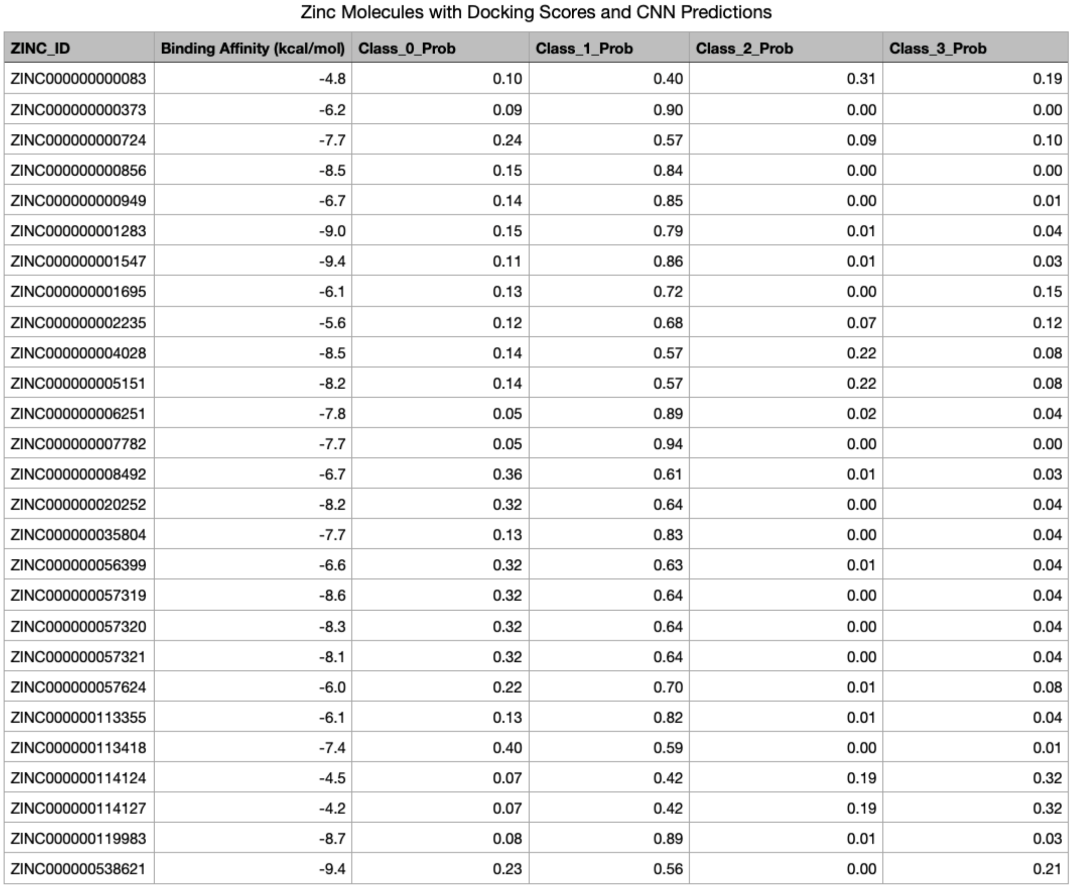


Figure 2. A graph showing various metrics, such as loss, accuracy, and AUROC, of the convolutional neural network.



Zinc Molecules with Docking Scores and CNN Predictions

ZINC_ID	Binding Affinity (kcal/mol)	Class_0_Prob	Class_1_Prob	Class_2_Prob	Class_3_Prob
ZINC000000000083	-4.8	0.10	0.40	0.31	0.19
ZINC000000000373	-6.2	0.09	0.90	0.00	0.00
ZINC000000000724	-7.7	0.24	0.57	0.09	0.10
ZINC000000000856	-8.5	0.15	0.84	0.00	0.00
ZINC000000000949	-6.7	0.14	0.85	0.00	0.01
ZINC000000001283	-9.0	0.15	0.79	0.01	0.04
ZINC000000001547	-9.4	0.11	0.86	0.01	0.03
ZINC000000001695	-6.1	0.13	0.72	0.00	0.15
ZINC000000002235	-5.6	0.12	0.68	0.07	0.12
ZINC000000004028	-8.5	0.14	0.57	0.22	0.08
ZINC000000005151	-8.2	0.14	0.57	0.22	0.08
ZINC000000006251	-7.8	0.05	0.89	0.02	0.04
ZINC000000007782	-7.7	0.05	0.94	0.00	0.00
ZINC000000008492	-6.7	0.36	0.61	0.01	0.03
ZINC000000020252	-8.2	0.32	0.64	0.00	0.04
ZINC000000035804	-7.7	0.13	0.83	0.00	0.04
ZINC000000056399	-6.6	0.32	0.63	0.01	0.04
ZINC000000057319	-8.6	0.32	0.64	0.00	0.04
ZINC000000057320	-8.3	0.32	0.64	0.00	0.04
ZINC000000057321	-8.1	0.32	0.64	0.00	0.04
ZINC000000057624	-6.0	0.22	0.70	0.01	0.08
ZINC000000113355	-6.1	0.13	0.82	0.01	0.04
ZINC000000113418	-7.4	0.40	0.59	0.00	0.01
ZINC000000114124	-4.5	0.07	0.42	0.19	0.32
ZINC000000114127	-4.2	0.07	0.42	0.19	0.32
ZINC000000119983	-8.7	0.08	0.89	0.01	0.03
ZINC000000538621	-9.4	0.23	0.56	0.00	0.21

Figure 3. A table representing the ZINC molecules that were most likely to bind to the alpha-C helix in/DFG out conformation of BRAF V600E, their binding affinity, and their likelihoods of being assigned to a different class.

Results:

This study aimed to identify promising binders for the BRAF V600E receptor, specifically with an alpha-C helix in/DFG out conformation, using a combination of molecular docking and a convolutional neural network. The docking results provided an initial estimation of binding affinities for ZINC molecules, while the CNN further supported the classification of the molecules into four classes: non-binders, binders for conformation 1 (6p7g's conformation), binders for conformation 2 (4xv2's conformation), and binders for conformation 3 (4e26's conformation). The integration of docking scores with CNN predictions allowed us to evaluate the predictive power of the CNN in recovering molecules that fit the 6p7g conformation.

The graph in Figure 1 demonstrates the performance of the docking pipeline in distinguishing known good binders (from ChEMBL) from a large dataset of random ZINC molecules. The top 100 scoring molecules, based on docking affinity, are shown, combining the results from ChEMBL and ZINC molecules. Despite being outnumbered 20-to-1 by ZINC molecules, many ChEMBL molecules are present in the top ranks, showing the pipeline recovers a significant portion of known good binders. This suggests that the docking pipeline effectively identifies high-quality binders, as the random ZINC molecules, which have a lower likelihood of performing well, rarely outperform ChEMBL molecules. These results validate the legitimacy and robustness of the docking pipeline in differentiating strong binders from weaker or non binders.

To further validate the procedure, the graph in Figure 2 shows the training metrics of the chosen CNN. It demonstrates promising results in its ability to classify molecules into their binding conformations. The model achieves high testing accuracy and AUROC, with both metrics stabilizing after approximately 20 epochs. The testing AUROC, which measures the model's ability to distinguish between classes, is consistently above 0.9, highlighting its strong predictive capability. However, the training and testing loss metric indicates areas for improvement. While the losses decrease rapidly during the initial epochs, the loss plateaus around 0.2. This suggests that the model may not be performing as well as the accuracy reflects, but tuning to bring the training loss down results in overfitting.

Once the procedure had been validated, random ZINC molecules were docked. The docking procedure identified a range of binding affinities among the random ZINC molecules, with values ranging from weak binding (> -5 kcal/mol) to strong binding (< -8 kcal/mol). The CNN assigned probabilities for each molecule across the four classes, enabling a comparison between the docking-derived affinities and the CNN's predictions. Molecules like ZINC000000000856 and ZINC000000000949 were assigned a high probability of belonging to Class 1 (0.84 and 0.85, respectively), aligning with their strong docking scores (-8.5 kcal/mol and -7.7 kcal/mol, respectively) and indicating a good fit for the alpha-C helix in/DFG out conformation.

In fact, a significant overlap was observed between high-affinity docked molecules and those classified as Class 1 by the CNN. This suggests that the CNN can reliably identify molecules likely to bind to the alpha-C helix in/DFG out conformation. Molecules that were classified as Class 1 by the CNN but had poor binding scores often had high probabilities of being assigned to a different binding class, in addition to the 1st class (but not the non-binding class). This can be seen in Figure 3, where molecules like ZINC000000114124 and ZINC000000114127 had poor binding scores and were classified as most-likely to bind to Class 1, but also had relatively high probabilities of being assigned to Class 3. The alignment between docking and CNN predictions demonstrates the utility of combining these approaches to prioritize molecules for further experimental validation.

Conclusion:

Overall, this study aimed to identify potential inhibitors for the BRAF V600E alpha-C helix in/DFG out conformation by integrating molecular docking, clustering, and machine learning techniques. The docking simulations provided binding affinities for ZINC molecules, while a convolutional neural network classified these molecules into four categories, predicting their suitability for binding to distinct receptor conformations. The results indicate that the combination of docking and machine learning is a powerful tool for prioritizing molecules for experimental validation.

There are multiple improvements that can be made to this process. For the CNN model, improving its performance is vital. Using more data, or using oversampling techniques to enrich the data, could also improve the model’s robustness. However, adjusting regularization techniques will be necessary to mitigate overfitting. Reducing the number of epochs can also prevent overfitting. Additionally, this study assumed that molecular fingerprints were the optimal structure to spatially represent molecules as an input that could go into a CNN. Exploring other methods of spatially-representing a molecule in a vectorized form could enhance the CNN’s performance. Currently, the CNN has been evaluated for its ability to classify binders for the alpha-C helix in/DFG out conformation. Comparing the CNN’s predictions across the docking results for all three conformations will provide a more comprehensive understanding of its capability to classify. Also, further narrowing down the potential inhibitors is necessary – this should incorporate known chemical information of the receptor, such as necessary H-bonds or salt bridges. Finally, a VAE could be implemented to generate new molecular fingerprints for unexplored chemical spaces. Using a VAE in tandem with the CNN and docking pipeline could improve clustering and classification by augmenting the dataset with structurally diverse samples.

References:

A. Šali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234, 779-815, 1993.

Böcker A;Derksen S;Schmidt E;Teckentrup A;Schneider G; (2003, May 3). *A hierarchical clustering approach for large compound libraries*. Journal of chemical information and modeling. https://pubmed.ncbi.nlm.nih.gov/16045274/

Cotto-Rios, X.M., Agianian, B., Gitego, N. *et al.* The co-crystal structure of BRAF(V600E) with PHI1 (2020) https://doi.org/10.2210/pdb6p7g/pdb

Cotto-Rios, X.M., Agianian, B., Gitego, N. *et al.* Inhibitors of BRAF dimers using an allosteric site. *Nat Commun* 11, 4370 (2020). https://doi.org/10.1038/s41467-020-18123-2

Degirmenci, Ufuk & Wang, Mei & Hu, Jiancheng. (2020). Targeting Aberrant RAS/RAF/MEK/ERK Signaling for Cancer Therapy. Cells. 9. 10.3390/cells9010198.

Hanrahan, A.J., Chen, Z., Rosen, N. *et al.* BRAF — a tumour-agnostic drug target with lineage-specific dependencies. *Nat Rev Clin Oncol* 21, 224–247 (2024). https://doi.org/10.1038/s41571-023-00852-0

Irwin, Sterling, Mysinger, Bolstad and Coleman, *J. Chem. Inf. Model. 2012* DOI: 10.1021/ci3001277.

Kowalewski, A., Durślewicz, J., Zdrenka, M., Grzanka, D., & Szylberg, Ł. (2020, August). *Clinical relevance of BRAF V600E mutation status in brain tumors with a focus on a novel management algorithm*. Targeted oncology. https://pmc.ncbi.nlm.nih.gov/articles/PMC7434793/

Leonardi, Giulia & Candido, Saverio & Carbone, Maurizio & Raiti, Fabio & Colaianni, Valeria & Garozzo, Sebastiano & Cinà, Diana & McCubrey, James & Libra, Massimo. (2012). BRAF mutations in papillary thyroid carcinoma and emerging targeted therapies (Review). Molecular medicine reports. 6. 687-94. 10.3892/mmr.2012.1016.

Qin, J., Xie, P., Ventocilla, C. *et al*. BRAF in complex with an organic inhibitor 7898734 (2012) https://doi.org/10.2210/pdb4e26/pdb

Śmiech, M., Leszczyński, P., Kono, H., Wardell, C., & Taniguchi, H. (2020, November 12). *Emerging BRAF mutations in cancer progression and their possible effects on transcriptional networks*. Genes. https://pmc.ncbi.nlm.nih.gov/articles/PMC7697059/

Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, de Veij M, Ioannidis H, Lopez DM, Mosquera JF, Magarinos MP, Bosc N, Arcila R, Kizilören T, Gaulton A, Bento AP, Adasme MF, Monecke P, Landrum GA, Leach AR. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res. 2024 Jan 5;52(D1):D1180-D1192.doi: 10.1093/nar/gkad1004. PMID: 37933841; PMCID: PMC10767899.

Zhang, C., Spevak, W., Zhang, Y. *et al*. B-Raf Kinase V600E oncogenic mutant in complex with Dabrafenib (2015) https://doi.org/10.2210/pdb4xv2/pdb