Fingerprint-based QSAR Model Generation to Identify Structural Determinants of HCV NS5B Inhibition

Berin KARAMAN MAYACK [1, 2, *] (iD), Muhammed Moyasar ALAYOUBI [3] (iD), Mikail Hakan GEZGINCI [2] (iD)

1   Department of Pharmacology, School of Medicine, University of California, Davis, Davis, CA, United States.
2   Department of Pharmaceutical Chemistry, Faculty of Pharmacy, Istanbul University, Istanbul 34116, Türkiye.
3   Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul 34956, Türkiye.

*   Corresponding Author. E-mail: bkaramanmayack@ucdavis.edu, karaman.berin@gmail.com (B.K.M.); Tel. +01-916-305 99 41.

ABSTRACT: RNA-dependent RNA polymerase, non-structural protein 5B (NS5B), is an essential enzyme of HCV for viral transcription and genome replication. Its initial validation as a promising target for the treatment of chronic hepatitis and hepatocellular carcinoma has consequently prompted different research institutes and the pharmaceutical industry to find potential inhibitors for human therapies. Among those, anthranilic acid derivatives received increasing attention because of their promising drug-like properties. In order to design promising drug candidates, the structural determinants of NS5B inhibitors were determined by a robust fingerprint-based quantitative structure-activity relationship (QSAR) model which was depicted on atomic effect contribution maps to provide visual aids for medicinal chemists. In the present work, we used a combination of computational chemistry methods including ensemble docking, binding free energy calculations, and a fingerprint-based QSAR model. We built a robust in silico protocol to accelerate the structure-based design of HCV NS5B inhibitors. The QSAR model, kpls_linear_3, constructed by KPLS fitting with linear fingerprints produced the best predictive performance (a correlation coefficient for the training set $R2 = 0.8900$, and a correlation coefficient $Q^2 = 0.9234$ and RMSE = 0.3032 for the test compounds). The atomic effect contribution map that was generated based on this model showed a good agreement between the predictions and the experimental data. To the best of our knowledge, we illustrated for the first time the use of the atomic effect contribution map as a visual aid for assessing the structural determinants of NS5B inhibitors. The computational strategy represented herein can assist pharmaceutical chemists in the rapid identification of the important features to design novel inhibitors of other protein targets as well.

KEYWORDS: NS5B; ensemble docking; binding free energy calculations; fingerprint; QSAR

1. INTRODUCTION

Hepatitis C Virus (HCV) is a blood-borne RNA virus and a member of the genus Hepacivirus in the Flaviviridae family. HCV causes both acute and chronic hepatitis leading to liver cirrhosis and hepatocellular carcinoma. It is a global health concern with the World Health Organisation (WHO) estimating that 58 million people are chronically infected and about 1.5 million new cases emerge per year [1]. HCV was first described in 1989 as a non-A non-B hepatitis [2] and shortly after in 1991, its varying genotypes and subtypes were determined [3]. There are eight known HCV genotype variants (1 to 8) with distinct geographic distributions and several subgenotypes [4, 5].

Until 2011, pegylated interferon-alpha, ribavirin combination therapy, and liver transplantation were the standard treatments for an HCV infection. However, low sustained virological response (SVR) rates, about 50% for genotype 1 and up to 80% for genotype 2 and 3, and many adverse effects including teratogenic and embryotoxic properties, depression, anemia, and low patient tolerance associated with interferon therapy, have limited its use [4, 6, 7]. In 2013, a new class of drugs called directly acting antiviral agents (DAAs) was introduced that resulted in improved patient tolerability with cure rates over 95% and a remarkable reduction in HCV-related mortality rates [4, 8].

Consequently, the non-structural protein components of HCV became the focus of many drug discovery projects. Among those HCV NS5B is a key component for viral transcription and genome replication [9]. Currently, the combination of different DAAs based on viral genotypic variations has replaced the classical line treatment options for HCV care [10]. Crystal structures of NS5B revealed a "right-hand" shaped amino-

terminal catalytic core with distinct subdomains referred to as 'fingers', 'palm', and 'thumb'. Moreover, the catalytic domain is followed by a linker sequence and a C-terminal membrane domain [11, 12]. Up to now, several scaffolds have been identified that target both the active site and the allosteric binding sites of NS5B polymerase [13]. Crystallographic fragment screening and structure-based optimization studies resulted in sub-micromolar inhibitors of gt1a and 1b replicons with improved cell permeability and good cell culture potency [6, 14-16]. Among those, anthranilic acid derivatives which target thumb site 2 demonstrated promising drug-like properties including sub-micromolar inhibitory activity against gt1a and 1b replicons, good cell permeability, and cell culture potency [14-16]. Despite a progressive improvement in potency, new structural elements that can increase the cell culture activity of these compounds with better ADMET (absorption, distribution, metabolism, excretion, and toxicity) properties and off-target profiles are still lacking.

The increasing need for the development of more potent and tolerable drugs for the treatment of HCV infection prompted us to explore the structure-activity relationship between the chemical properties of the anthranilic acid derivatives and HCV NS5B inhibition. In the present work, a combination of ensemble docking, binding free energy calculations, and QSAR analysis has been employed to establish a new in-silico protocol for predicting HCV NS5B inhibitory activity (Figure 1). Furthermore, using the newly generated QSAR model, we mapped the atomic effects of the molecular structures on a contribution map to identify the favorable and unfavorable groups or their substitution pattern. We anticipate that the presented computational strategy can be used to promote the structure-based design of novel non-nucleoside inhibitors of HCV NS5B and extend to model noncovalent inhibitors of other protein targets.
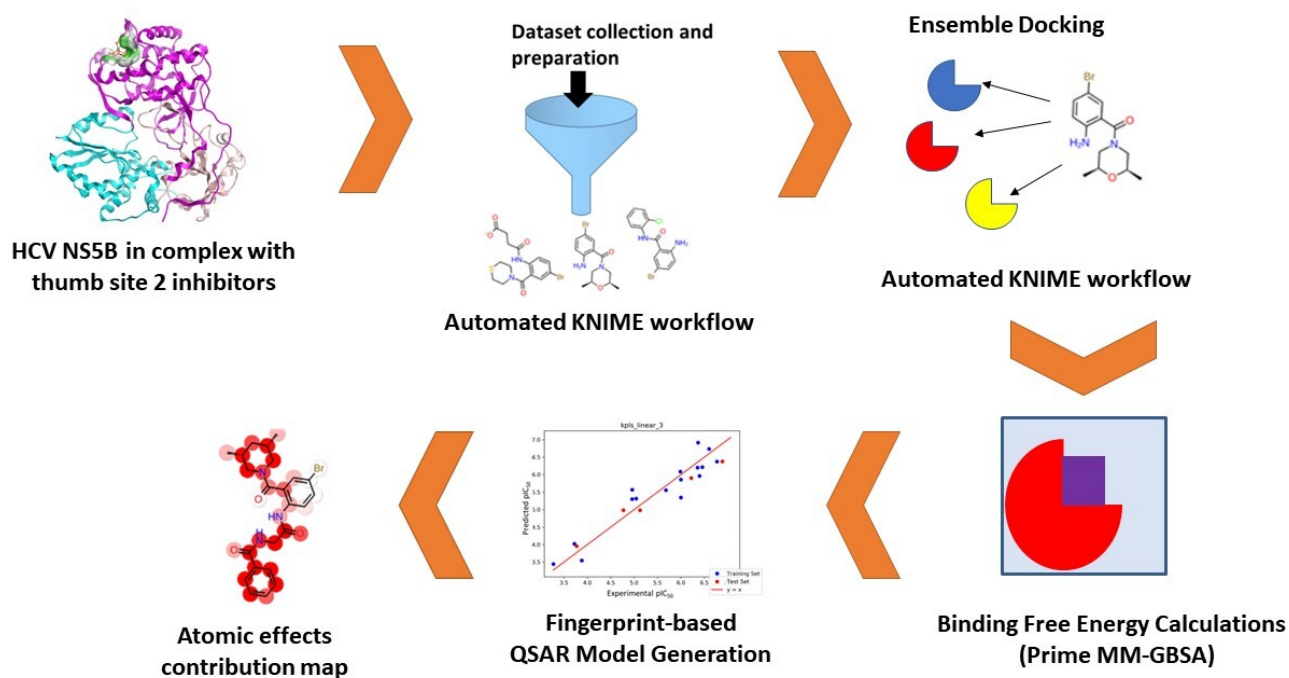
Figure 1. Overall workflow applied for fingerprint-based QSAR model generation to identify favorable and unfavorable structural determinants of HCV NS5B inhibition

2. RESULTS and DISCUSSION

2.1 Validation of Molecular Docking Setup

To assess the pose prediction power of different scoring functions available in the GLIDE program implemented within the Schrodinger Platform, a self-docking test was conducted [22-24]. A 2.0 Å RMSD between the heavy atoms of the experimental pose and docked pose has been widely accepted to define a good pose and successful docking [25, 26]. However, especially for redocking of larger ligands or for cross-docking studies, a threshold of 2.5 Å RMSD was used [27]. For redocking, around 70% of top-ranked poses were found to be within 2.0 Å RMSD of the experimental pose [26].

Karaman Mayack et al.
A short title here which is not more than 80 characters (Running head)

Journal of Research in Pharmacy
Research Article

In the present work using the first docking setup (setup 1), we obtained a docking accuracy of 74%, 78%, 75%, and 75% for XP-dockingscore, SP-dockingscore, XP-emodel score, and SP-emodel score, respectively, that corresponded with self-docking RMSD values below 2.0 Å. When we used a threshold of 2.5 Å RMSD, docking accuracy was 79%, 84%, 79%, and 84% using XP-dockingscore, SP-dockingscore, XP-emodel score, and SP-emodel score, respectively (Table S1-4). Since similar results were obtained using both docking and emodel scores, we only used docking scores for the second setup (setup 2) for ranking the compounds. In the case of the second setup, the poses within 2.0 Å of the experimental poses were given top-ranked scores in 74% and 77 % of cases for XP and SP docking scores, respectively. Similarly, the poses within 2.5 Å of the crystal pose were given the best scores in 79% and 84% of cases for XP and SP docking scores, respectively (Table S5-6). As comparable docking accuracy was obtained using two docking setups, a postdocking minimization including ten poses per ligand with SP-dockingscore was applied for further docking calculations.

2.2 Ensemble Docking of the Dataset

Multiple docking or so-called ensemble docking is a powerful technique that takes into account conformational changes in the active pocket upon ligand binding. In this respect, already resolved crystal structures of HCV NS5B with different thumb pocket 2 inhibitors were used in our docking experiments of known inhibitors. We checked whether a correlation could be achieved between docking scores and the biological data (Table 1). However, no correlation could be obtained between the docking scores (DS) and the pIC_{50} values giving a correlation coefficient $R^2 = 0.24$ with root-means-quare error (RMSE) = 0.93. Besides a cross-validated correlation coefficient, q^2_{LOO} was also calculated using LOO (Leave One Out) cross-validation technique using the QuaSAR-model module implemented in Molecular Operating Environment (MOE) software. Ligand 8c retrieved a large Z-score and was predicted as an outlier in cross-validation studies. Nevertheless, the removal of the outlier 8c from the data resulted in only a slight improvement giving a correlation coefficient of $R^2 = 0.33$, RMSE = 0.82, and $q^2_{LOO} = 0.23$ for the 27 compounds.

Table 1. Summary of the statistics obtained for molecular docking and MM-GBSA rescoring

Statistics	DS	DS-LE1	DS-LE2	DS-LE3	BFE	BFE-LE1	BFE-LE2	BFE-LE3
R^2	0.24	0.65	0.48	0.00	0.62	0.49	0.49	0.48
RMSE	0.93	0.63	0.77	1.07	0.66	0.77	0.77	0.77
q^2_{LOO}	0.15	0.60	0.40	0.71	0.58	0.41	0.41	0.42
Outliers (Z-score)	8c: 2.62	-	-	-	8c: 3.39	-	-	8c: 3.20
R^{2*}	0.33	-	-	-	0.74	-	-	0.62
RMSE*	0.82	-	-	-	0.51	-	-	0.62
$q^2_{LOO}{}^*$	0.23	-	-	-	0.70	-	-	0.56

DS: docking score, NHA: Number of Heavy Atoms, DS-LE1: DS/NHA: DS-LE2: DS/(NHA)$^{2/3}$, DS-LE3: DS/(1 + ln(number of heavy atoms)), BFE: binding fee energy score (MM-GBSA score), BFE-LE1: BFE/NHA: BFE-LE2: BFE/(NHA)2/3, BFE-LE3: BFE/(1 + ln(number of heavy atoms)). *Statistical values obtained after the omission of molecules predicted as outliers based on their Z-score.

It is also known that scoring functions are additive in nature and that they define intermolecular energies as the sum of interactions between the ligand and the protein. Thus, the retrieved docking scores tend to be higher for larger compounds and compounds with more functional groups. Traditionally, ligand efficiency (LE) indices are calculated by scaling the binding affinity with measures such as molecular weight (MW), the number of heavy atoms (NHA), molecular or polar surface area (PSA), and partition coefficient (AlogP). We evaluated three ligand efficiency scores retrieved from Glide docking in the present work. Among them, DS-LE1 performed the best with an $R^2 = 0.62$, RMSE = 0.66, and $q^2_{LOO} = 0.58$. These results illustrate the size dependency of docking score values as reported in many other studies. Thus, relying only on the docking score may not be a sufficient criterion for ligand ranking during ligand optimization stages of these HCV NS5B polymerase inhibitors.

2.3 MM-GBSA Calculations

As previously shown, if accurate poses can be obtained through molecular docking, then a more rigorous scoring function can be used in a postprocessing step to reduce failures in binding affinity estimation [28]. We re-ranked the docking poses according to their MM-GBSA scores and checked whether there is an agreement between the relative binding free energies and the experimental data (Table 1). We also tested the performance of normalized binding free energy scores calculated similarly as for the ligand efficiency metrics in Glide docking. Interestingly, raw BFE scores performed better than any of the ligand efficiency metrics (R^2 = 0.62, RMSE = 0.66, and q^2_{LOO} = 0.58). Ligand 8c was again identified as an outlier in cross-validation studies and the removal of this ligand improved the correlation to R^2 = 0.74, RMSE = 0.51, and q^2_{LOO} = 0.70.

2.4 Generation of Fingerprint-based QSAR models

QSAR models transform the relationships between molecular descriptors and biological activity into a mathematical equation. This simplified description has been applied as one of the most effective ways to predict the biological properties of compounds accurately and guide the rational design or purchase of new chemical entities.

In this work, the AutoQSAR model implemented in Schrodinger software was used to build numeric models, using ensemble best subsets for MLR, PCR, PLS, and KPLS. Before model building, in total 495 2D descriptors, including feature counts, and molecular and topological properties were calculated. Statistical details for all top-ranked ten models are reported in Table 2. All models have a coefficient of determination (R^2) for the training set above 0.86 and the test set (Q^2) above 0.79. In general, (R^2 > 0.60 and Q^2 > 0.50) are required conditions for the predictability of a QSAR model [29, 30]. Moreover, all top-scored models were based on the KPLS model. In comparison to simpler and more traditional methods such as MLR and PLS, the KPLS method has already been shown to perform better in terms of correlation and prediction power and is a valuable QSAR tool in different drug discovery projects [31-33]. The top-ranked model, kpls_linear_3, was constructed by KPLS fitting with linear fingerprints using the 3rd split of the learning set into a training set and a test set (Table S7 and Figure S7).

Table 2. The predictive power of the top-scored ten QSAR models

Model Code	Score	SD	R^2	RMSE	Q^2
kpls_linear_3	0.8925	0.3794	0.8900	0.3032	0.9234
kpls_molprint2D_35	0.8787	0.4164	0.8748	0.2779	0.9229
kpls_radial_35	0.8689	0.4327	0.8648	0.3207	0.8973
kpls_dendritic_41	0.8684	0.4437	0.8662	0.3654	0.7887
kpls_linear_47	0.8643	0.3877	0.8864	0.3820	0.8762
kpls_molprint2D_18	0.8635	0.4088	0.8675	0.3959	0.8750
kpls_radial_8	0.8631	0.4336	0.8621	0.3884	0.8576
kpls_linear_42	0.8629	0.4286	0.8595	0.3567	0.8947
kpls_radial_2	0.8616	0.4430	0.8576	0.3495	0.8670
kpls_linear_32	0.8610	0.3791	0.8957	0.3781	0.8603

We also explored the predictive performance of the top-scored QSAR model, kpls_linear_3, on an external validation set of anthranilic acid derivatives that have not been used during the model-building process (Table S4). The correlation coefficient, R^2, was 0.58 between the observed experimental and predicted activity pIC_{50} values of external validation set compounds for the kpls_linear_3 model. Since only one model was used, the Pred SD property which shows the standard deviation in the predicted pIC_{50} values over the models used for the prediction was 0.000 for all the compounds. Whereas 'Domain Score' is based on fingerprint similarity and shows if the structure is in the applicability domain of the model used.

A domain score of 1 indicates one standard deviation from the training set average and a domain score of 0 translates to the training set average. The cutoff value for the domain score varies based on the diversity

of the molecules used in the QSAR model. When a compound gets a domain score outside the average training set domain score of ± 2.0 then that compound is flagged, and the domain alert score is set to 1 if a single predictive model is used. Therefore, compound 7b with a domain alert score of -2.0070 was identified as an outlier. As compound 7b comes from the same congeneric series as the training set, it is not an outlier. Therefore, we further repeated the binding affinity predictions using a consensus model where all the top-ten models were used to predict the HCV NS5B inhibitory activity of each external validation set of compounds and then the average of the predicted properties was used as the final predicted value (Table S5). Using a consensus model improved the correlation between the observed and predicted pIC_{50} significantly giving an R^2 = 0.86. Additionally, compound 7b retrieved a domain alert score of 0.2 which means that the structure is outside the domain of applicability only in two out of ten models. These results support the view that consensus QSAR model predictions consider different fingerprint types and therefore, are advantageous for structural characterization in comparison to single QSAR models. This allows using such consensus QSAR models for activity prediction studies of diverse ligands in drug optimization stages.

2.5 Visual Representation of the QSAR Model

The atomic effects of the molecules for the HCV NS5B inhibitory activity were investigated by generating an atomic contributions map (Figure 2) based on the top-scored KPLS model (linear, 3rd split). Such maps can be used as visual aids to assess favorable and unfavorable structural characteristics during lead optimization efforts. Based on a crystallographic fragment screen Antonysamy et. al. [14] have designed several 5-bromo anthranilic acid derivatives. However, the model suggests that bromine does not have any influence on the activity. Hence, the contribution of different halogens to the activity can be further explored.
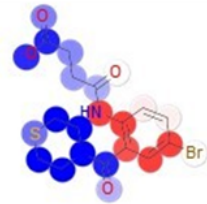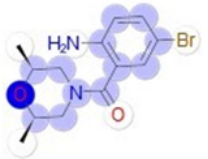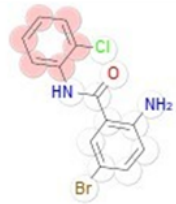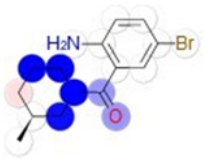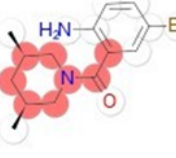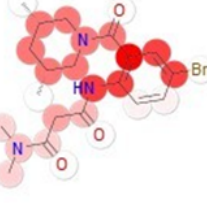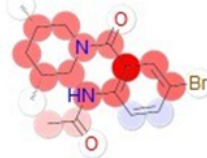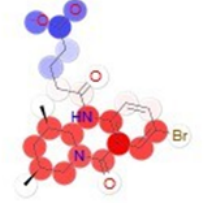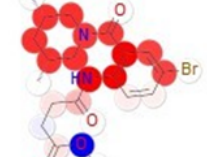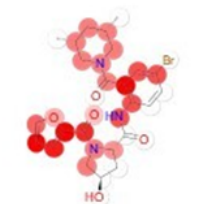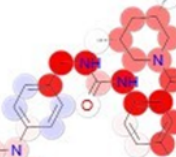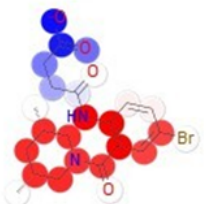
On the other hand, 3,5-dimethylpiperidine amide binding to 5-bromoaryl moiety contributed positively to the activity. However, any other analogs with ortho substitution to the aniline on this moiety including morpholine, thiomorpholine, tetrahydroisoquinoline, pyrrolidine, and chlorophenyl amide depreciated the activity or had a very small influence on the activity. This effect could be also captured with the QSAR model. In addit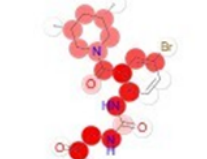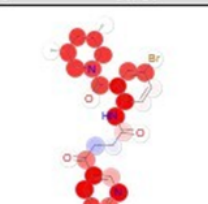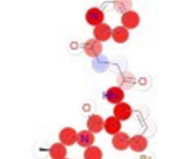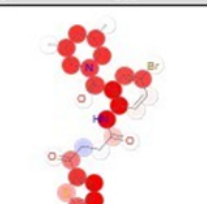ion, the relevant biological effect of including such moieties on the structure was also correctly predicted and depicted on the atomic contribution map.

Interestingly, removing one (7c) or two methyl groups (7g), which did not have an exact IC_{50} value and was not considered during QSAR model generation) of 3,5-dimethylpiperidine reduced the activity significantly. The model also suggested that removing one methyl group is unfavorable and turns also the associated atoms from red to blue on the contribution map. This happens upon the removal of atoms from the fixed regions and this loss of "good bits" changes the net effect also for the fragments that comprise the fixed backbone atoms.

Moreover, the incorporation of succinic acid (8c) or glutaric acid (8k) to the aniline of 5-bromo anthranilic acid derivatives or replacing the aniline with an acetamide (8f) or a sulfonamide (8m) was not favorable and these effects could be also captured well by our contribution map. The contribution map also could predict that the extension of aniline with a 1,4-dicarbonyl linker (either extended or incorporated in a ring system) favor the presence of a substituted amide group in it.

Also, these terminal aromatic or heteroaromatic rings were predicted to have a positive contribution to the activity in most cases. This was also justified as the thumb site 2 of the HCV NS5B protein has residues such as Ser476, His475, and Tyr477 which can interact productively with such structural arrangements. Besides, interactions with these residues take place in a solvent-exposed region, and therefore, different substituents could be accommodated.

In the case of succinate derivatives 11d and 11e which have a proline ring between the carbonyl moieties on the linker were predicted to have unfavorable or no effects based on the position of the atom. This finding was also in agreement with the experimental findings that analogs with restricted rotation between the carbonyl moieties decrease the inhibitory activity against NS5B. Based on the biochemical assay results, compound 11d (IC_{50} = 0.35) was more potent than compound 11e (IC_{50} = 1.03).

Figure 2. Visual representation of atomic effects for the top-scored KPLS model (kpls_linear_3) built from linear fingerprints. Observed pIC$_{50}$ values are reported as PIC50. Atoms that have positive contributions to the predicted activity are colored red, whereas atoms that have neutral or negative contributions are colored white and blue, respectively. Color intensity shows the strength of the effect.
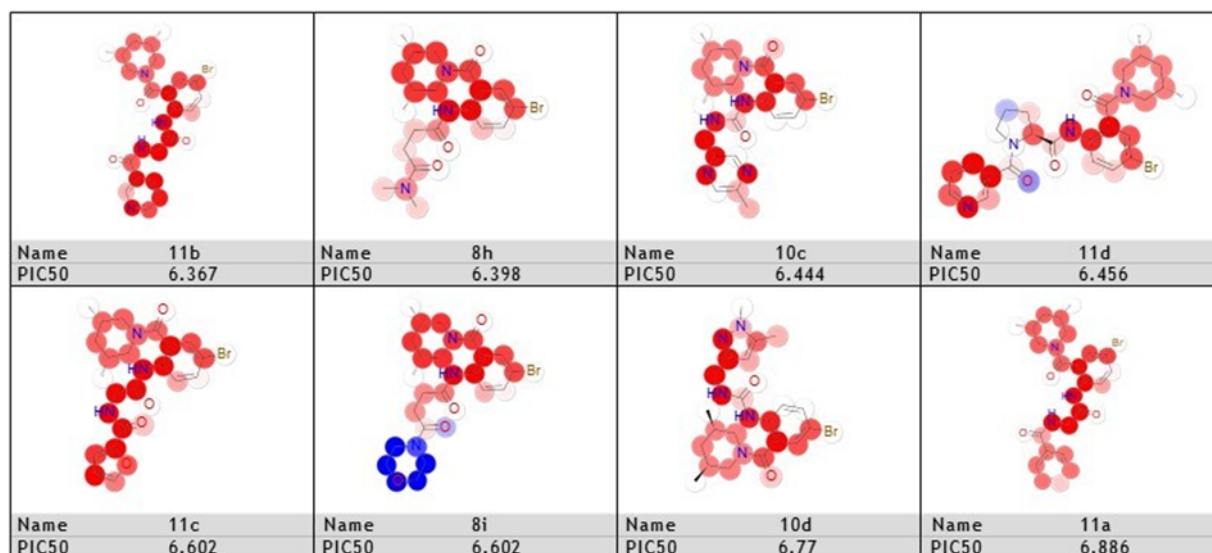
Name	11b	Name	8h	Name	10c	Name	11d
PIC50	6.367	PIC50	6.398	PIC50	6.444	PIC50	6.456
Name	11c	Name	8i	Name	10d	Name	11a
PIC50	6.602	PIC50	6.602	PIC50	6.77	PIC50	6.886

Figure 2. Continued.

Nonetheless, the predicted atomic effects on our contribution map were showing a more unfavorable substitution pattern for compound 11d in comparison to 11e. To our surprise, the replicon activity assay as well showed that 11e was a more active (EC_{50} = 3.7 μM) proline analog than 11d (EC_{50} = 12.9 μM). These findings also support the prediction reliability of our newly generated QSAR model as well as the useful insights that medicinal chemists can retrieve in characterizing the important features to design novel inhibitors by depicting the atomic effects on a contribution map.

3. CONCLUSION

In the present work, we aimed to develop a cost- and time-effective in silico strategy that can guide the optimization of congeneric series of NS5B inhibitors before synthesis and experimental testing. We observed a significant correlation between the calculated relative binding free energies and the experimental pIC50 values for the congeneric series of HCV NS5B inhibitors studied. These results suggest that a combination of an ensemble docking and rescoring of the poses with a more sophisticated binding free energy model such as MM-GBSA can be used to account for the protein flexibility and the prediction of binding affinities of congeneric NS5B inhibitors. Next, we investigated several fingerprint-based QSAR models for binding affinity prediction. The QSAR model, kpls_linear_3, constructed by KPLS fitting with linear fingerprints produced the best predictive performance and was used to build an atomic effect contributions map. This map could assess the favorable and unfavorable structural characteristics of NS5B inhibitors. In conclusion, obtained results demonstrated that this in silico protocol can be used in the refinement of HCV NS5B inhibitors and represent a useful tool to guide the rational design of new chemical entities for the treatment of chronic hepatitis and hepatocellular carcinoma. Moreover, the KNIME workflows herein disclosed represent highly valuable tools to perform rapid redocking analysis of a large number of crystal structures and ensemble docking of large datasets that can be easily modified for various virtual screening purposes of different therapeutic targets.

4. MATERIALS AND METHODS

4.1. Dataset Preparation

A validated dataset of anthranilic acid derivatives with HCV NS5B polymerase inhibitory activities on genotype 1b (gt1b) was collected from the literature [14]. Only the ligands with the exact IC50 values were considered. The structure of the molecules was drawn using the 2D Sketcher module implemented in the Maestro interface of the Schrodinger platform. In total twenty-eight compounds were prepared using LigPrep (Schrödinger Release 2019: LigPrep, Schrödinger, LLC, New York, NY, 2019) within the Schrodinger program. Molecular structures of all compounds, in vitro biological activities, as well as calculated pIC50 values of the dataset, are listed in Figure S1 in the supplementary material.

4.2. Preparation of protein-inhibitor complexes

We first collected crystal structures of HCV NS5B complexed with ligands bound to thumb Site 2. In total 47 protein complexes were deposited in the Protein Data Bank (PDB) [17] by April 2020. These protein-ligand complexes were retrieved from PDB and curated for further *in silico* studies. An automated workflow (Figure S2) was generated for this purpose using the KNIME [18] and Schrödinger platforms.

4.3. Molecular Docking

All docking calculations were performed using the Glide docking tool [19] implemented in Schrödinger. We generated a workflow (Figure S3-5) in KNIME to redock the co-crystallized ligands to the respective protein structure. This workflow was used for automated grid generation, redocking of native ligands, and RMSD calculation.

4.4. Ensemble Docking of Dataset

We generated a workflow (Figure S6) in KNIME to dock ligands into the thumb site 2 pocket of multiple HCV NS5B protein structures. We saved two different output files for further analysis: (1) the top-scored pose per grid file was saved for each ligand to be used further in binding free energy calculations, and (2) the top-scored pose out of all grid solutions was saved for each ligand as the best docking solution. To avoid data heterogeneity in QSAR studies, only the top-scored enantiomer from each pair was considered.

4.5. Rescoring with MM-GBSA Calculations

The top-ranked docking solution for each ligand retrieved for each grid file with the corresponding protein structure was used for binding free energy calculations. Prime MM-GBSA module and VSGB 2.0 implicit solvent model implemented in the Schrödinger platform were used for binding free energy calculations using default options.

4.6. Fingerprint-based Quantitative Structure-Activity Relationship (QSAR) Model Generation

In the present work, we used the AutoQSAR task panel implemented in Schrödinger to generate the 2D-QSAR models. For the studied dataset, physicochemical descriptors, topological descriptors, and QikProp properties were calculated automatically for model building. Moreover, we included the MM-GBSA score as a descriptor prior to the model generation step. Descriptors were removed before the model-building stage if more than 90% of the ligands in the dataset had the same value for that particular property.

In AutoQSAR, the generation of numeric QSAR models was done using four different techniques: multiple linear regression (MLR) [20], partial least-squares regression (PLS) [21], kernel-based partial least-squares regression (KPLS), and principal components regression (PCR). In addition, four hashed type 2D fingerprints including linear, radial, dendritic, and molprint2D were generated for KPLS models and ten thousand most informative bits for each fingerprint type were set to be retained by default. The maximum allowed correlation between any pair of descriptors was set to 0.80 by default. The data set was randomly divided into a learning set (75%, 21 compounds) and an external validation set (25%, 7 compounds). During model building, the learning set was randomly split into a training set (75%, 17 compounds) and a test set (25%, 6 compounds). Division of compounds into training and test set is an important step of any QSAR model generation. To capture all the features of the dataset, this process was repeated 50 times to generate 50 models per supervised learning techniques (MLR, PCR, PLS, and KPLS) used in model generation. Only the top ten models were saved for further evaluation. In addition, we used the contribution map analysis option in the AutoQSAR module to explore the atomic effects of different substituents on compounds with an anthranilic acid core in the inhibitory activity against HCV NS5B.

Author contributions: Concept – B.K.M.; Design – B.K.M.; Data Collection and/or Processing – B.K.M., M.M.A.; Analysis and/or Interpretation – B.K.M., M.M.A.; Literature Search – B.K.M., M.M.A., M.H.G.; Writing – B.K.M.; Critical Reviews – B.K.M., M.M.A.

Conflict of interest statement: The authors declare no conflict of interest.

REFERENCES

[1]   World Health Organization. Hepatitis C, 27 July 2021. Available from: https://www.who.int/news-room/fact-sheets/detail/hepatitis-c

[2]   Choo QL, Kuo G, Weiner AJ, Overby LR, Bradley DW, Houghton M. Isolation of a cDNA clone derived from a blood-borne non-A, non-B viral hepatitis genome. Science. 1989;244(4902):359-362. https://doi.org/10.1126/science.2523562.

[3]   Choo QL, Richman KH, Han JH, Berger K, Lee C, Dong C, Gallegos C, Coit D, Medina-Selby R, Barr PJ, Weiner AJ, Bradley DW, Kuo G, Houghton M. Genetic organization and diversity of the hepatitis C virus. Proc Natl Acad Sci U S A. 1991;88(6):2451-2455. https://doi.org/10.1073/pnas.88.6.2451.

[4]   Dennis BB, Naji L, Jajarmi Y, Ahmed A, Kim D. New hope for hepatitis C virus: Summary of global epidemiologic changes and novel innovations over 20 years. World J Gastroenterol. 2021;27(29):4818-4830. https://doi.org/10.3748/wjg.v27.i29.4818.

[5]   Pisano MB, Giadans CG, Flichman DM, Ré VE, Preciado MV, Valva P. Viral hepatitis update: Progress and perspectives. World J Gastroenterol. 2021;27(26):4018-4044. https://doi.org/10.3748/wjg.v27.i26.4018.

[6]   Nittoli T, Curran K, Insaf S, DiGrandi M, Orlowski M, Chopra R, Agarwal A, Howe AY, Prashad A, Floyd MB, Johnson B, Sutherland A, Wheless K, Feld B, O'Connell J, Mansour TS, Bloom J. Identification of anthranilic acid derivatives as a novel class of allosteric inhibitors of hepatitis C NS5B polymerase. J Med Chem. 2007;50(9):2108-2116. https://doi.org/10.1021/jm061428x.

[7]   Worachartcheewan A, Prachayasittikul V, Toropova AP, Toropov AA, Nantasenamat C. Large-scale structure-activity relationship study of hepatitis C virus NS5B polymerase inhibition using SMILES-based descriptors. Mol Divers. 2015;19(4):955-964. https://doi.org/10.1007/s11030-015-9614-2.

[8]   Higuera-de la Tijera F, Servín-Caamaño A, Servín-Abad L. Progress and challenges in the comprehensive management of chronic viral hepatitis: Key ways to achieve the elimination. World J Gastroenterol. 2021;27(26):4004-4017. https://doi.org/10.3748/wjg.v27.i26.4004.

[9]   Li HC, Yang CH, Lo SY. Cellular factors involved in the hepatitis C virus life cycle. World J Gastroenterol. 2021;27(28):4555-4581. https://doi.org/10.3748/wjg.v27.i28.4555.

[10]   Wang LS, D'Souza LS, Jacobson IM. Hepatitis C-A clinical review. J Med Virol. 2016;88(11):1844-1855. https://doi.org/10.1002/jmv.24554.

[11]   Götte M, Feld JJ. Direct-acting antiviral agents for hepatitis C: structural and mechanistic insights. Nat Rev Gastroenterol Hepatol. 2016;13(6):338-351. https://doi.org/10.1038/nrgastro.2016.60.

[12]   Lohmann V. Hepatitis C virus RNA replication. Curr Top Microbiol Immunol. 2013;369:167-198. https://doi.org/10.1007/978-3-642-27340-7_7.

[13]   Powdrill MH, Bernatchez JA, Götte M. Inhibitors of the Hepatitis C Virus RNA-Dependent RNA Polymerase NS5B. Viruses. 2010;2(10):2169-2195. https://doi.org/10.3390/v2102169.

[14]   Antonysamy SS, Aubol B, Blaney J, Browner MF, Giannetti AM, Harris SF, Hébert N, Hendle J, Hopkins S, Jefferson E, Kissinger C, Leveque V, Marciano D, McGee E, Nájera I, Nolan B, Tomimoto M, Torres E, Wright T. Fragment-based discovery of hepatitis C virus NS5b RNA polymerase inhibitors. Bioorg Med Chem Lett. 2008;18(9):2990-2995. https://doi.org/10.1016/j.bmcl.2008.03.056.

[15]   Beaulieu PL, Coulombe R, Duan J, Fazal G, Godbout C, Hucke O, Jakalian A, Joly MA, Lepage O, Llinàs-Brunet M, Naud J, Poirier M, Rioux N, Thavonekham B, Kukolj G, Stammers TA. Structure-based design of novel HCV NS5B thumb pocket 2 allosteric inhibitors with submicromolar gt1 replicon potency: discovery of a quinazolinone chemotype. Bioorg Med Chem Lett. 2013;23(14):4132-4140. https://doi.org/10.1016/j.bmcl.2013.05.037.

[16]   Stammers TA, Coulombe R, Duplessis M, Fazal G, Gagnon A, Garneau M, Goulet S, Jakalian A, LaPlante S, Rancourt J, Thavonekham B, Wernic D, Kukolj G, Beaulieu PL. Anthranilic acid-based Thumb Pocket 2 HCV NS5B polymerase inhibitors with sub-micromolar potency in the cell-based replicon assay. Bioorg Med Chem Lett. 2013;23(24):6879-6885. https://doi.org/10.1016/j.bmcl.2013.09.102.

[17]   Burley SK, Berman HM, Kleywegt GJ, Markley JL, Nakamura H, Velankar S. Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. Methods Mol Biol. 2017;1607:627-641. https://doi.org/10.1007/978-1-4939-7000-1_26.

[18]   Roughley SD. Five Years of the KNIME Vernalis Cheminformatics Community Contribution. Curr Med Chem. 2020;27(38):6495-6522. https://doi.org/10.2174/0929867325666180904113616.

[19] Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, Sanschagrin PC, Mainz DT. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. J Med Chem. 2006;49(21):6177-6196. https://doi.org/10.1021/jm051256o.

[20] Peter SC, Dhanjal JK, Malik V, Radhakrishnan N, Jayakanthan M, Sundar D. Quantitative Structure-Activity Relationship (QSAR): Modeling Approaches to Biological Applications. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, editors. Encyclopedia of Bioinformatics and Computational Biology. Oxford: Academic Press; 2019. p. 661-76.

[21] Höskuldsson A. PLS regression methods. Journal of Chemometrics. 1988;2(3):211-228. https://doi.org/10.1002/cem.1180020306.

[22] Bolcato G, Cuzzolin A, Bissaro M, Moro S, Sturlese M. Can We Still Trust Docking Results? An Extension of the Applicability of DockBench on PDBbind Database. Int J Mol Sci. 2019; 20(14):3558. https://doi.org/10.3390/ijms20143558.

[23] Cheng T, Li X, Li Y, Liu Z, Wang R. Comparative assessment of scoring functions on a diverse test set. J Chem Inf Model. 2009;49(4):1079-1093. https://doi.org/10.1021/ci9000053.

[24] Plewczynski D, Łaźniewski M, Augustyniak R, Ginalski K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. J Comput Chem. 2011;32(4):742-755. https://doi.org/10.1002/jcc.21643.

[25] Erickson JA, Jalaie M, Robertson DH, Lewis RA, Vieth M. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. J Med Chem. 2004;47(1):45-55. https://doi.org/10.1021/jm030209y.

[26] Leach AR, Shoichet BK, Peishoff CE. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. J Med Chem. 2006;49(20):5851-5855. https://doi.org/10.1021/jm060999m.

[27] Sutherland JJ, Nandigam RK, Erickson JA, Vieth M. Lessons in Molecular Recognition. 2. Assessing and Improving Cross-Docking Accuracy. J Chem Inf Model. 2007;47(6):2293-2302. https://doi.org/10.1021/ci700253h.

[28] Wichapong K, Lawson M, Pianwanit S, Kokpol S, Sippl W. Postprocessing of protein-ligand docking poses using linear response MM-PB/SA: Application to Wee1 kinase inhibitors. J Chem Inf Model. 2010;50(9):1574-1588. https://doi.org/10.1021/ci1002153.

[29] Golbraikh A, Tropsha A. Beware of q2! J Mol Graph Model. 2002;20(4):269-276. https://doi.org/10.1016/s1093-3263(01)00123-1.

[30] Tropsha A, Gramatica P, Gombar VK. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. QSAR & Combinatorial Sci. 2003;22(1):69-77. https://doi.org/10.1002/qsar.200390007.

[31] Falchi F, Bertozzi SM, Ottonello G, Ruda GF, Colombano G, Fiorelli C, Martucci C, Bortorelli R, Scarpelli R, Cavalli A, Bandiera T, Armirotti A. Kernel-Based, Partial Least Squares Quantitative Structure-Retention Relationship Model for UPLC Retention Time Prediction: A Useful Tool for Metabolite Identification. Anal Chem. 2016;88(19):9510-9517. https://doi.org/10.1021/acs.analchem.6b02075.

[32] Deokar H, Deokar M, Wang W, Zhang R, Buolamwini JK. QSAR Studies of New Pyrido[3,4-b]indole Derivatives as Inhibitors of Colon and Pancreatic Cancer Cell Proliferation. Med Chem Res. 2018;27(11-12):2466-2481. https://doi.org/10.1007/s00044-018-2250-5.

[33] Byadi S, Eddine MH, Sadik K, Podlipnik Č, Aboulmouhajir A. Fingerprint-based 2D-QSAR Models for Predicting Bcl-2 Inhibitors Affinity. Lett Drug Des Discov. 2020;17(10):1206-1215. http://dx.doi.org/10.2174/1570180817999200414155403.