Molecular Diversity (2023) 27:1661–1674
https://doi.org/10.1007/s11030-022-10519-0

ORIGINAL ARTICLE

Homology modeling of Forkhead box protein C2: identification of potential inhibitors using ligand and structure-based virtual screening

Mayar Tarek Ibrahim[1] · **Jiyong Lee**[2] · **Peng Tao**[1]

Received: 22 March 2022 / Accepted: 19 August 2022 / Published online: 1 September 2022

Abstract

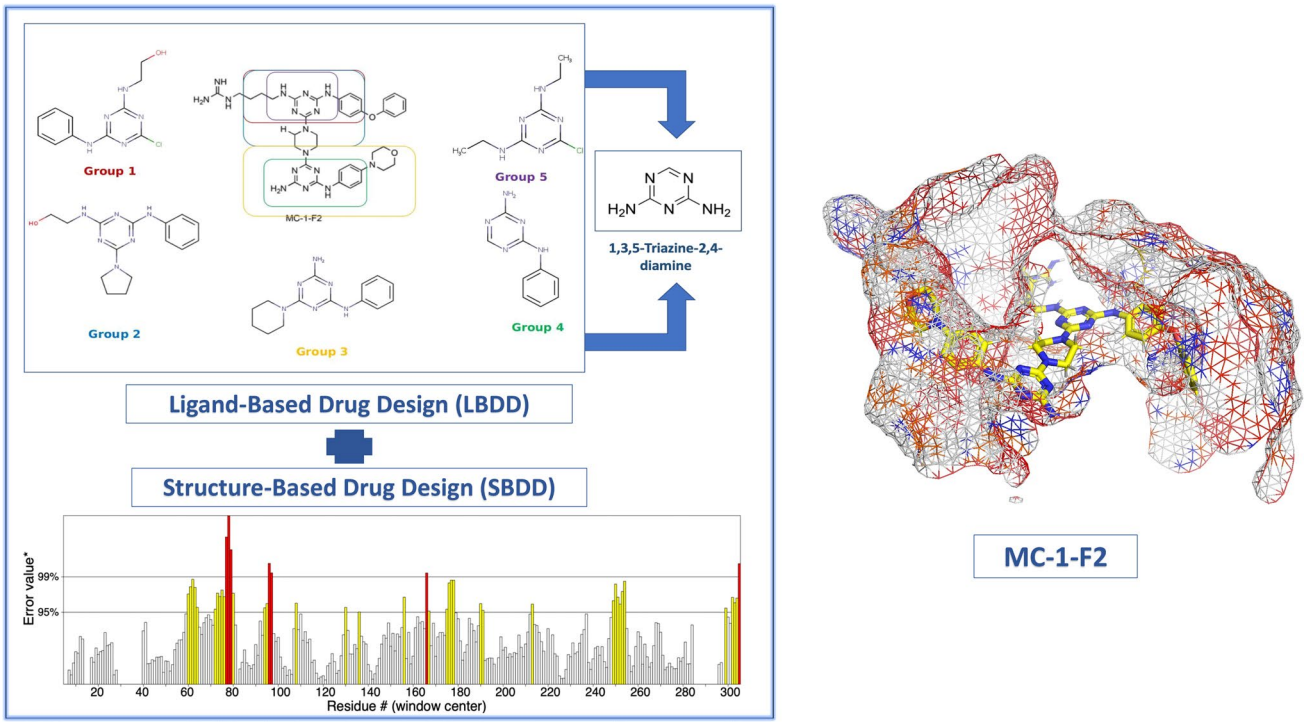
Overexpression of Forkhead box protein C2 (FOXC2) has been associated with different types of carcinomas. FOXC2 plays an important role in the initiation and maintenance of the epithelial–mesenchymal transition (EMT) process, which is essential for the development of higher-grade tumors with an enhanced ability for metastasis. Thus, FOXC2 has become a therapeutic target for the development of anticancer drugs. MC-1-F2, the only identified experimental inhibitor of FOXC2, interacts with the full length of FOXC2. However, only the DNA-binding domain (DBD) of FOXC2 has resolved crystal structure. In this work, a three-dimensional (3D) structure of the full-length FOXC2 using homology modeling was developed and used for structure-based drug design (SBDD). The quality of this 3D model of the full-length FOXC2 was evaluated using MolProbity, ERRAT, and ProSA modules. Molecular dynamics (MD) simulation was also carried out to verify its stability. Ligand-based drug design (LBDD) was carried out to identify similar analogues for MC-1-F2 against 15 million compounds from ChEMBL and ZINC databases. 792 molecules were retrieved from this similarity search. De novo SBDD was performed against the full-length 3D structure of FOXC2 through homology modeling to identify novel inhibitors. The combination of LBDD and SBDD helped in gaining a better insight into the binding of MC-1-F2 and its analogues against the full length of the FOXC2. The binding free energy of the top hits was further investigated using MD simulations and MM/GBSA calculations to result in eight promising hits as lead compounds targeting FOXC2.

✉ Peng Tao
ptao@smu.edu

[1] Department of Chemistry, Center for Research Computing, Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, TX, USA

[2] Department of Chemistry and Biochemistry, The University of Texas at Tyler, Tyler, TX, USA

Graphical abstract



Keywords Similarity search · Homology modeling · De novo structure-based design · MC-1-F2 · Forkhead box protein C2 · MM/GBSA

Introduction

The epithelial–mesenchymal transition (EMT) is a biological process in which the epithelial cells undergo biochemical changes to become mesenchymal cells [1, 2]. EMT is essential for cancer metastasis through the generation of more invasive cells with higher migration capacity and increased resistance to apoptosis with stem cell-like features [2, 3]. The generated cancer stem cells (CSCs) from the EMT process have the ability of self-renewal leading to the formation of secondary tumors [2, 4, 5]. Several transcription factors are activated to initiate the EMT process, including zinc finger E-box-binding homeobox 1 (ZEB1), Snail, Slug, Twist, Goosecoid, and Forkhead Box Protein C2 (FOXC2) [6–9].

FOXC2 is a member of the FOX family of transcription factors that are characterized by the presence of conserved winged-helix (forkhead) DNA-binding domain (DBD). This conserved DBD consists of three α-helices, three β-sheets, and two less conserved winged loops. The main region for DNA recognition among these secondary structures is located near the third helix H3 [10–15]. FOX family of transcription factors is divided into 19 subgroups, including the FOXC subgroup [16]. FOXC subfamily consists of two proteins: FOXC1 and FOXC2. Both proteins are involved in different developmental processes [17]. Mutations in the FOXC2 gene were found to be associated with different genetic diseases, such as Lymphoedema distichiasis syndrome (LDS) [18]. Overexpression of FOXC2 was found to be associated with breast and hepatocellular carcinoma through initiating and maintaining EMT [19–21].

Indirect inhibition of FOXC2 resulted in attenuating the cancer metastasis by reversing the effects of EMT, making FOXC2 an appropriate drug target for the treatment of metastatic cancer [22–24]. This drives the work of Castaneda et al. to develop the first small molecule inhibiting FOXC2 [25]. One-bead-one-compound (OBOC) library of α-helix mimetics was used to find a proper inhibitor for the full-length structure of FOXC2. OBOC library was used because of the presence of the H3 domain, in which an α helix is essential in the functional role of FOXC2. Screening of nearly 20,400 compounds of OBOC library was carried out followed by elimination of false positives to find the top 16 hits and further filtration of false positives [25].

The most probable hit identified was MC-1-F2 (Fig. 1). The MC-1-F2 is the first inhibitor of FOXC2 in vitro.

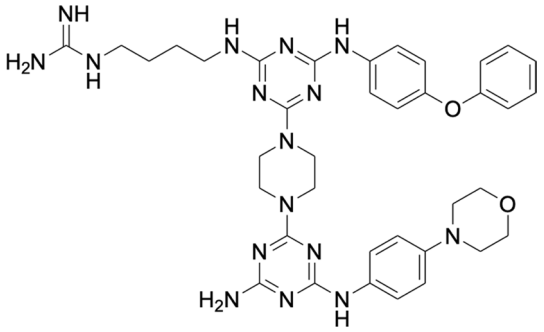


Fig. 1 The structure of MC-1-F2 (the only identified experimental inhibitor of FOXC2 protein)

MC-1-F2 was found to induce apoptosis and inhibit colony formation capacity in the cell lines with high expression of FOXC2. Importantly, the inhibition of FOXC2 by MC-1-F2 leads to a reversal of EMT and inhibition of cancer metastasis in vitro [25].

The dissociation constant value (K_d) for the binding of MC-1-F2 to the DBD of FOXC2 was higher than 100 μM, this indicates that the binding was weak and insufficient for the inhibitor to bind to the protein. This insinuates that other domains of the protein are involved in the effective binding of the inhibitor [25]. In other words, MC-1-F2 was found to bind to the full-length structure of FOXC2 not only to the DBD. The binding between MC-1-F2 and FOXC2 involved both the main scaffold of MC-1-F2 as well as the side chains [25].

MC-1-F2 violates the four rules of Lipinski's rule of five, which decreases its probability to be an active drug because of its poor physicochemical properties. Therefore in this work, we applied both ligand-based drug design and structure-based drug design (SBDD) computational methods to develop novel compounds with improved physicochemical properties as potent inhibitors for FOXC2.

Methodology

Ligand-based drug design (LBDD)

Multi-fingerprint browser based on city-block distance using daylight-type substructure fingerprint (sFP) was utilized to search for similar analogues to MC-1-F2 in more than 13 million structures of commercially available compounds in the ZINC database and more than 2 million compounds in the ChEMBL database [26–30]. sFP is a binary fingerprint that represents the molecule in the form of 0 and 1 to indicate the absence and the presence of certain substructures, respectively [28, 29].

sFP is used to calculate the city-block distance (CBD) through the following equation to measure the similarity between molecules A and B, represented as two vectors in K-dimensional space.

$$\mathrm{CBD}_{A,B} = \sum_{j=1}^{K} |A_j - B_j|. \quad (1)$$

This search resulted in retrieving nearly 1000 compounds from the ZINC database and nearly 10,000 compounds from the ChEMBL database.

Lipinski's rule of five was applied to the retrieved similar analogues, followed by the removal of compounds containing toxicity alerts as well as the removal of pan-assay interference compounds (PAINS) using KNIME [31] to get a total of 5414 compounds. This filtration was carried out to ensure that the molecules have desirable physicochemical properties as drug molecules.

A molecular substructure miner, MoSS, using pair-wise maximum common substructure similarity (MCSS) metric, implemented in KNIME was used for reliable similarity search [32]. MoSS focuses on finding the frequent substructures and the discriminative fragments in a database of molecules. Similarity search based on city-block distance implemented in the multi-fingerprint browser is fast, thus suitable to screen large databases. MoSS MCSS molecule similarity was applied as an additional similarity metric on the 5414 compounds to ensure the proximity of the retrieved compounds to MC-1-F2. This results in a reduction of the 5414 compounds to 792 compounds. Furthermore, these 792 compounds were clustered into five different groups based on K-Medioids using the values generated by MoSS MCSS molecule similarity [31]. Clustering of the molecules was carried out to identify common functional groups within the retrieved molecules similar to MC-1-F2.

Homology modeling of the full-length structure of FOXC2

The available crystal structures of FOXC2 in the Protein Data Bank (PDB) only contain the DBD but not the full length of the protein. MC-1-F2, the only identified experimental inhibitor for FOXC2, interacts with other domains of FOXC2 in addition to the DBD. Thus, it is necessary to develop a three-dimensional (3D) structure of the full-length FOXC2 using homology modeling. The target 3D structure of FOXC2 consists of 501 residues, this represented the full length of the FOXC2 protein including the N-terminal, C-terminal, and DBD.

I-TASSER server, implementing Threading ASSEmbly Refinement, was used to build 3D structures of full-length FOXC2 (UniprotKB Q99958) [33–35]. I-TASSER depends on enhanced Profile–Profile threading Alignment (PPA) method, in which the target sequence is aligned with the protein structures in the PDB, whose pair-wise sequence

identity with the target sequence is more than 70%. This alignment is carried out to search for similar protein folds depending on the sequence using the PPA method. The identified folds are excised from the threading aligned regions. On the other hand, the unaligned parts of the sequence, representing the loops, are modeled using ab initio modeling [34]. The final structure is an average of the coordinates of all the clustered structures generated by replica-exchange Monte Carlo simulations [35], which is carried out to explore the conformational space.

The developed 3D model for the full-length FOXC2 protein was associated with a *Z*-score of 3.93. *Z*-score measures the difference between the native raw fold of a protein and the average score of misfolded structures divided by the standard deviation. The normalized *Z*-score assesses the quality of the model as well as the alignment [36]. 1VTN is a forkhead protein used as a template sequence for the generation of the homology model of FOXC2. The normalized *Z*-score of 3.93, larger than 1, indicates a good alignment of the FOXC2 protein sequence with the sequence of 1VTN protein.

I-TASSER uses the confidence score (C-score) to evaluate the quality of the model.

$$C\text{-score} = \ln\left(\frac{M}{M_{\text{tot}}} \cdot \frac{1}{\langle \text{RMSD} \rangle} \cdot \frac{\prod_{i=1}^{4} Z(i)}{\prod_{i=1}^{4} Z_0(i)}\right), \tag{2}$$

where M is the multiplicity of the clustered structures; M_{tot} is the total number of the I-TASSER structure decoys; $\langle \text{RMSD} \rangle$ is the average root mean square deviation (RMSD); $Z(i)$ is the highest *Z*-score of the templates used in the PPA threading, and $Z_0(i)$ is a *Z*-score cutoff distinguishing between good and bad templates [33–35].

The C-score, with a range between – 5 and 2, was 1.1 for the final model, suggesting the high quality of the homology model. Furthermore, Molprobity, ERRAT, and ProSA modules [37–40] were employed to validate the quality of the model.

Molecular dynamics simulations for the homology model of FOXC2

Molecular dynamics (MD) simulations were carried out for the homology model. The protein was prepared for the MD simulation using CHARMM program package version 41b1 [41]. The protein was solvated in a cubic box using the TIP3P water model [42]. Counter ions Cl^- and Na^+ were added to neutralize the charges and maintain the ionic strength of the system. The model was minimized followed by heating from 0 to 300 K. The prepared system was equilibrated for 10 ns as an isothermal–isobaric ensemble (NPT) using OpenMM [43] running on GPU. Production of 100-ns simulations was carried out as a canonical ensemble (NVT). Electrostatic interactions were computed using the particle mesh Ewald (PME) [44] method. Hydrogen covalent bonds were constrained during the simulation using the SHAKE method [44].

RMSD was calculated using the MdTraj package [45], where the first frame of the simulation was the reference structure. RMSD was used to evaluate the stability of the generated model and the convergence of the simulation and calculated as

$$\text{RMSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Ur_i - r_i^{\text{ref}})^2}, \tag{3}$$

where N is the number of atoms, r_i is the coordinate of atom i, r_i^{ref} is the coordinate of atom i in the reference structure, and U is the best-fit rotational matrix to align a given structure onto the reference structure.

Identification of the binding site of the full-length homology model of FOXC2

The binding site of the full-length homology model of FOXC2 was determined using the MakeReceptor module implemented in the OpenEye scientific package [46–48]. MakeReceptor module identifies the active site by determining the target mask and by generating a negative image for the active site. The target mask is the subset of the protein suitable for docking, while a negative image is the shape of the active site where ligands interact with the active site without clashes [46–48]. MC-1-F2 was docked against the three binding sites detected by the MakeReceptor module. The binding site associated with the highest score upon docking of MC-1-F2 was chosen to be the active site. The docking score of MC-1-F2 against the identified active site was based on four factors: shape complementarity of − 13 kcal/mol, hydrogen bond component of − 2 kcal/mol, protein desolvation of 6 kcal/mol, and ligand desolvation of 8 kcal/mol. Besides the shape complementarity component, the ligand desolvation attributes immensely to the final docking score due to the large size of MC-1-F2.

Furthermore, the binding of MC-1-F2 against both the full-length model generated by I-TASSER [33, 49, 50] and the crystal structure of DBD of FOXC2 (PDB ID: 6O3T) was investigated using the BindScope server, a convolutional neural network to predict the binding probability of compounds [51–53]. This was performed to validate the generated model by I-TASSER [33–35]. The binding probability of MC-1-F2 to the 3D structure including the full length of the protein (the generated homology model) is expected to be higher than the PDB structure of DBD only.

Screening through molecular docking

The 792 compounds obtained from the similarity search were screened using molecular docking against the identified active site of the full-length FOXC2 3D structure. A conformational search was carried out for these 792 compounds using the OMEGA2 package [54] to determine the most probable conformation for each compound. OMEGA is a knowledge-based method efficient in producing bioactive conformations of the molecular structures with high quality. OMEGA method takes the effect of ionization, tautomerism, and isomerism into account by generating different conformers for the same molecule based on energetic and geometric criteria. This is achieved by combining a library of fragments along the sigma bond [54].

FRED docking package provided by OpenEye Scientific Software Inc. was used to dock these compounds against the target active site [46, 47]. In the FRED docking method, a systematic examination is carried out for the possible docking poses. The top hits are determined using the Chemgauss4 scoring function, which depends on shape complementarity [46, 47]. The top 15 hits based on the docking score were further validated by calculating the binding probability using the BindScope server [51–53]. BindScope server was employed as a further filtration step in the virtual screening process, as it outperforms the classical virtual screening methods that depend on simple docking scoring functions [55].

De novo structure-based drug design (SBDD) on the full-length model

Complimentary to the above virtual screening studies to find potent inhibitors from the existing databases, de novo SBDD using the LIGANN server [56–58] was performed based on the active site of the 3D structure of the full-length FOXC2 identified by the MakeReceptor module. LIGANN is a generative neural network (GNN) that produces ligands' shapes complementary to the protein shape, represented as a generative adversarial network. The ligand shape is decoded into SMILES strings representing the ligands generated from the design (SBDD). Using GNN in the de novo structure-based design bypasses the challenges faced by the conventional structure-based compound design. This is achieved by including the 3D information of the compounds' binding poses during the design process. This was found to top the fragment-based drug design (FBDD) by 200% when the size of the linker was limited to five atoms [56–58].

A total of 79 new compounds were generated and screened through molecular docking using FRED [47, 48] against the identified active site of the full-length FOXC2. Similarly, the top 15 hits concerning the docking score were further validated by calculating the binding probability using the BindScope server [51–53].

MM/GBSA calculation

The stability of the binding of the top ten hits to the developed full-length model of FOXC2 was assessed using MD simulations. Amber ff14SB forcefield implemented in Amber package was used for the protein [59–61]. The ligands were parameterized using General AMBER force field 2 (GAFF2) implemented in the AnteChamber module [62, 63].

The complexes of the top ten hits were solvated in a cubic water box using the TIP3P water model, followed by neutralizing the system using Na^+ and Cl^- counter ions. The system was then minimized by the steepest descent algorithm and equilibrated for 50 picoseconds (ps) as an isothermal–isobaric ensemble (NPT) at 300 K. 50-ns simulations for the top ten hits were carried out in the form of a canonical ensemble (NVT). Electrostatics interactions were accounted for using the particle mesh Ewald method. The hydrogen bonds were constrained using the SHAKE algorithm [44]. The binding free energy of the protein–ligand complexes was calculated using Molecular Mechanics/Generalized Born Surface Area (MM/GBSA) algorithm implemented in AmberTools20 [64].

Results

Ligand-based drug design: similarity search of MC-1-F2

Similarity search for analogues to MC-1-F2 compound was performed against 15 million commercially available compounds in ChEMBL and ZINC databases using a multi-fingerprint browser [26–30]. This search resulted in nearly 11,000 compounds. An additional filtration step was employed utilizing MoSS MCSS molecule similarity implemented in KNIME to ensure the proximity of the similar analogues retrieved from the initial similarity search [31]. The output molecules from this similarity search were further filtered by removing the PAINS and the toxicity alerts and applying Lipinski's rule of five. This extensive filtration resulted in 792 compounds.

K-Medoids clustering based on MoSS MCSS molecule similarity was applied to the 792 compounds identified from the similarity searching resulting in five groups. Group 1 contains 209 molecules, group 2 contains 88 molecules, group 3 contains 75 molecules, group 4 contains 308 molecules, and group 5 contains 112 molecules. Each group is represented by a representative molecule illustrated in Fig. 2. These five representative molecules were selected

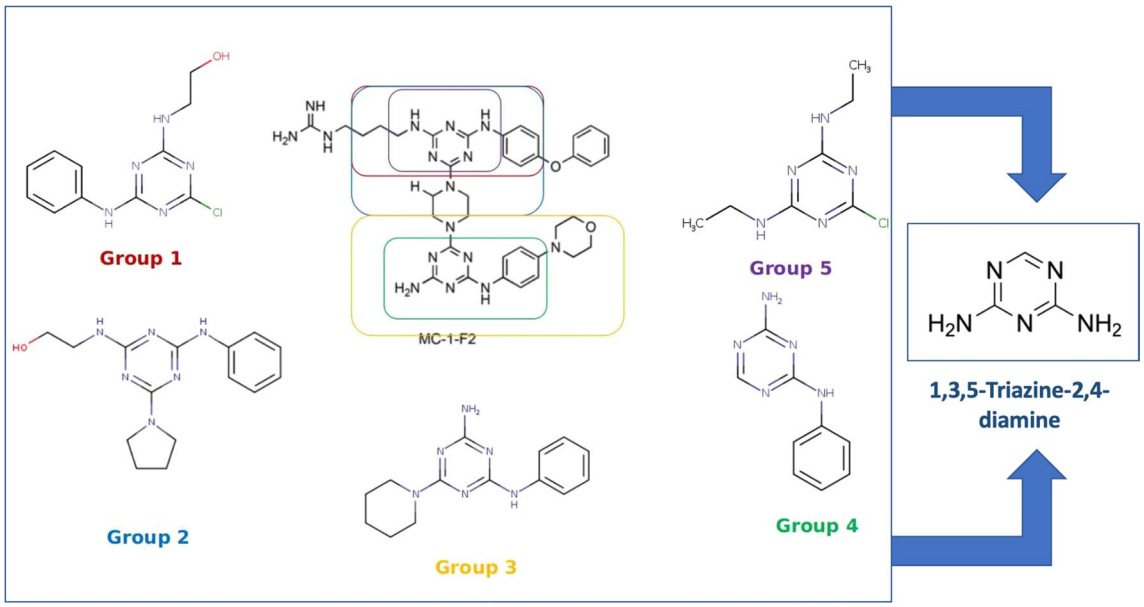


Fig. 2 Five groups were generated from clustering the 792 compounds obtained from the similarity search of MC-1-F2 against 15 million commercially available compounds in ChEMBL and ZINC databases [26–30]. The representative structure of each group is shown in comparison to the structure of MC-1-F2 (the only identified experimental inhibitor of FOXC2). 1,3,5-Triazine-2,4-diamine was found to be the common functional group in the representative structures of all five groups

as the cluster centers of above five groups from K-Medoids clustering analysis. Group 1 is the largest among all five groups and is the most similar to MC-1-F2 based on MoSS MCSS molecule similarity. 1,3,5-Triazine-2,4-diamine was found to be the common functional group in the representative structures of all five groups. The role of this functional group in the binding of MC-1-F2 and the similar analogues to FOXC2 protein should be further investigated.

Quality of the generated 3D structure of the full-length FOXC2

The developed full-length model of FOXC2, generated by I-TASSER [33–35], shows 70% sequence identity to the 1VTN target sequence. In addition, the *Z*-score of the developed model is 3.93, reflecting the good quality of the model. The quality of the developed model was also assessed using MolProbity [33]. 85.4% of the residues lie within a favorable area in the Ramachandran plot. Only 3.6% of the residues, 17 residues out of the total of 501 residues, were identified to be outliers in the Ramachandran plot (Fig. 3). Moreover, no residues were forming unfavorable bonds or violating the acceptable tetrahedral geometry or the chiral volume of the protein structure. This supports that the developed structure of the full-length FOXC2 protein has good quality and should be suitable for the consequent SBDD studies.

The homology model of the full-length protein was also validated using the ERRAT server [34] to check the quality of the model by calculating the non-bonded interactions

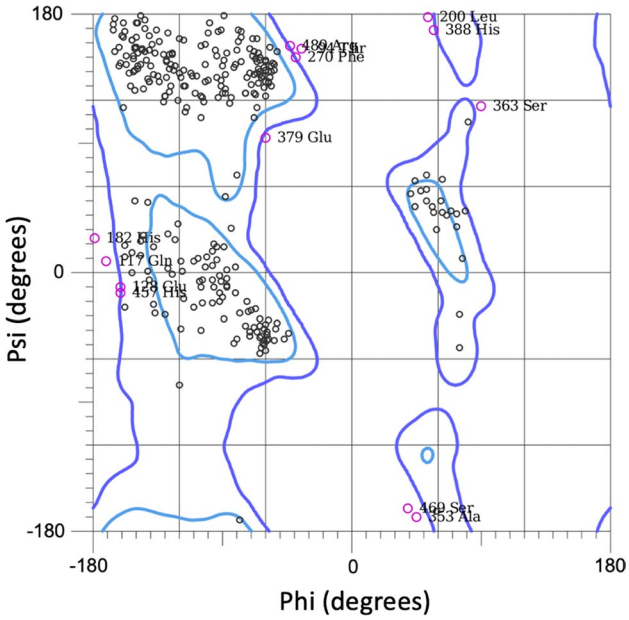


Fig. 3 Ramachandran plot, produced by MolProbity module [33], to validate the quality of the 3D structure of the full-length FOXC2 protein. 85.4% of the residues, 428 residues, lie within favorable regions of the Ramachandran plot. Only 3.6% of the residues are outliers

between different atom types. The model is considered acceptable if the overall quality is higher than 50%. In addition, the higher the overall quality score, the more reliable the homology model. The overall quality of the

first 300 residues (0–305) of the model is 81.8% (Fig. 4a). The overall quality of the last 200 residues (306–501) of the protein is 78.9% (Fig. 4b). With the weighted average of the model's quality surpassing 81.2%, the developed model of the full-length FOXC2 protein is considered as a good model. Moreover, less than 3% of the residues exceed the 99% error value for the non-bonded interaction. This further supports and validates the quality of the generated model.

ProSA module [39, 40] was employed to evaluate the quality of the model by calculating the energy of the residues as well as the *Z*-score of the overall structure. ProSA *Z*-score indicates the statistical significance of the model by comparing its score to the knowledge-based scores of proteins of similar size with the same folds [39, 40]. The quality of the protein folds was evaluated based on the energy of the amino acid residues. The lower the energy of the residues the better the model. The energy of the residues of the developed model was associated with energy ranging from − 1 to 1 kcal/mol (Fig. 5a). This indicates the low energy of the protein folds comparing to random conformers of similar size. *Z*-score measures the deviation of the total energy of the model for 3D structures generated by X-ray and NMR with similar sizes. The model was associated with a *Z*-score of – 4.29 (Fig. 5b). This reflects that the generated model of the full-length FOXC2 is stable with low energy and has quality well comparable with proteins with similar sizes from PDB.

Homology modeling and molecular dynamics simulations

RMSD was calculated over the time course of 100-ns MD simulation to check the stability of the generated model as well as the convergence of the simulation. The simulation remains stable after the first 10 ns. A high rise in the RMSD value from 0 to 1.15 nm was observed during the first 10 ns of simulations. However, the RMSD remains within the range of 0.05 nm for the rest of the simulations (Fig. 6), indicating the stability of the generated 3D model of the full-length FOXC2.

Root mean square fluctuation (RMSF) was calculated for the full-length model of FOXC2 after the first 10 ns. The first frame was used as the reference structure. The fluctuations of the RMSF values are less than 0.1 nm for all the residues, except for the first ten residues with high fluctuations of nearly 0.25 nm to 0.5 nm. This indicates the high flexibility of the N-terminal. RMSF shows that model is stable as indicated by the limited fluctuation profile of the residue throughout the simulation (Fig. 7).

Identification of the generated model active site

The interaction of MC-1-F2 was investigated with the three detected sites from the MakeReceptor module [46–48]. The binding site associated with the highest score upon docking of MC-1-F2 was chosen as the active site. The interaction

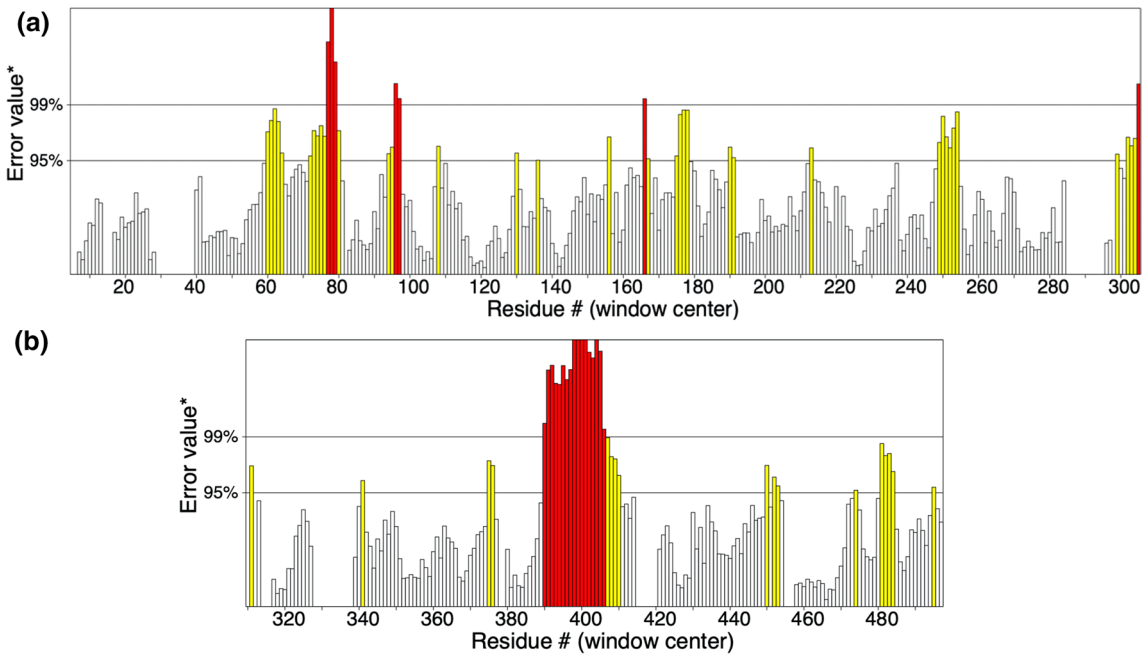


Fig. 4 ERRAT plot [34] for the 3D structure of FOXC2. Misfolded regions are represented by the red bars. Regions with low error rate are represented by the white bars. Regions with an error rate between 95 and 99% are represented by the yellow bars. **a** ERRAT plot for residues 1–305; **b** ERRAT plot for residues 306–501

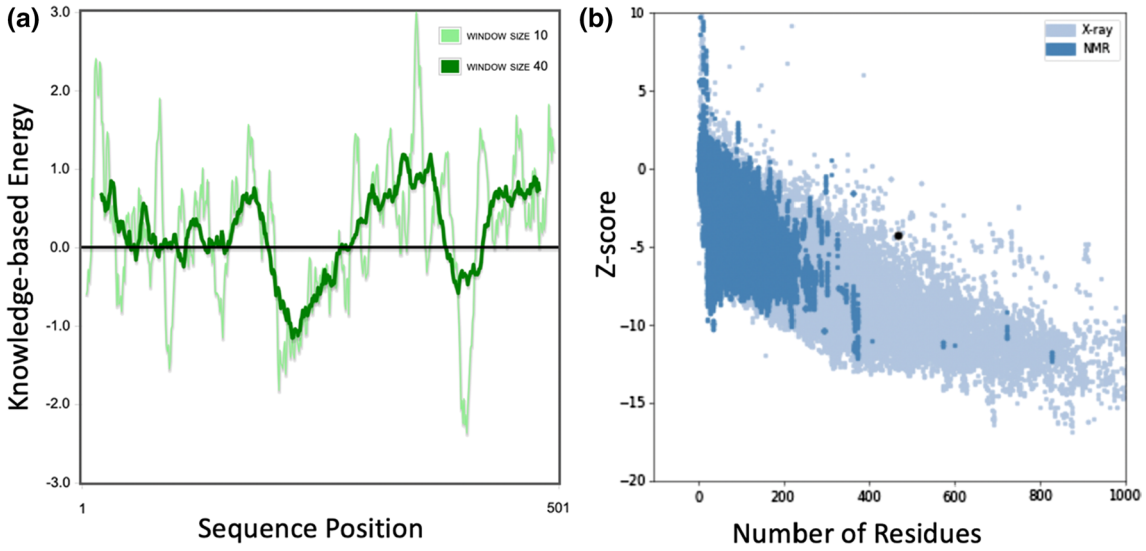


Fig. 5 ProSA plot for the 3D structure of full-length FOXC2. **a** Energy plot for the 501 residues of the full-length FOXC2 lie between 1 kcal/mol and − 1 kcal/mol. This plot represents the local model quality of the model. **b** The model is associated with a *Z*-score of − 4.29, represented by the black dot. This plot represents the overall quality of the model quality. *Z*-score was calculated for the proteins of the same size in PDB, including structures obtained through X-ray crystallography (light blue) and NMR spectroscopy (dark blue)

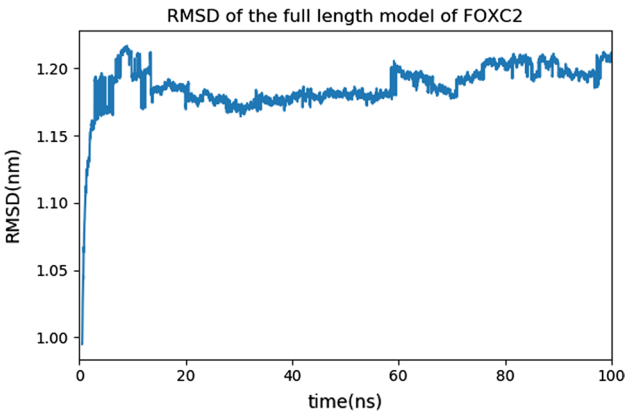


Fig. 6 RMSD of the 100-ns molecular dynamics simulation of full-length FOXC2 homology model 3D structure

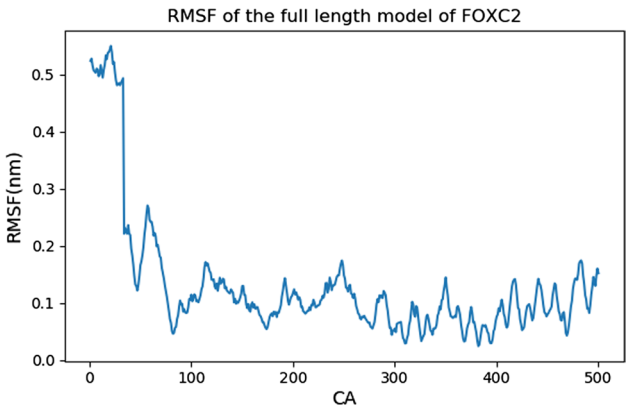


Fig. 7 RMSF of the full-length FOXC2 homology model 3D structure

of MC-1-F2 with FOXC2 involved not only the DBD of the protein but also part of the C-terminal. This agrees with the experimental data that MC-1-F2 inhibitor binds to the full length of the protein [25]. The binding mode of MC-1-F2 upon docking involves important residues ASN 449, LEU 459, SER 476, TYR 481, THR 484, and TYR 488. This may be because the C-terminal of the protein contains extra secondary structures in the full-length FOXC2 comparing to DBD, providing a stable binding site (Fig. 8). Using the BindScope server, the binding probability of MC-1-F2 to the full-length protein was compared to that of the DBD. This was implied to validate and ensure the better binding of MC-1-F2 to the identified active site comparing to the DBD. It was found that the binding probability of MC-1-F2 against the full-length model was higher than the DBD (Table 1). This agrees with the experimental data that MC-1-F2 interacts with not only the DBD but the full length of FOXC2 [25].

Screening using the docking method against the determined active site (LBDD)

The 792 compounds obtained from the similarity search for the analogues to MC-1-F2 were docked against the determined active sites of the full-length FOXC2 using FRED [46, 47]. The docking scores of the top 15 molecules range

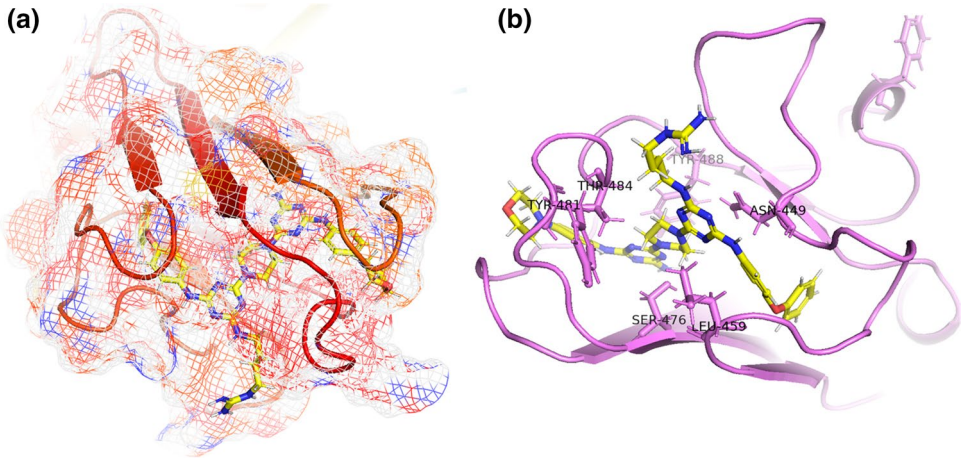


Fig. 8 The detected active site, associated with the best binding of MC-1-F2 upon docking, of the full-length model of FOXC2 using the MakeReceptor module in the OpenScientific package [46–48]. **a** Shape complementarity against MC-1-F2. MC-1-F2 is illustrated in the form of sticks. The detected active sites are illustrated in the form of the mesh showing the C-terminal helices, the domain to which MC-1-F2 binds. **b** The binding mode of MC-1-F2 upon docking. ASN 449, LEU 459, SER 476, TYR 481, THR 484, and TYR 488 are identified as important residues for binding with MC-1-F2

Table 1 The binding probability of MC-1-F2 against the developed model of the full-length FOXC2 using I-TASSER and the DBD crystal structure of FOXC2

Model used	Binding probability
I-TASSER	0.9757
DBD Crystal structure (PDB ID:6O3T)	0.8483

from – 13 to – 6 kcal/mol (Table S1). The docking scores of similar analogues are higher than MC-1-F2. This may be attributed to the smaller size of the analogues compared to MC-1-F2.

The top 15 molecules based on FRED docking scores contain the main scaffold of MC-1-F2 inhibitor, 1,3,5-triazine-2,4-diamine functional group in particular. The top hits were further evaluated using the BindScope server [51–53] to calculate the binding probability and to identify the most probable hits. The compounds with binding probability equal to or higher than 0.85 were selected. This additional filtration using binding probability resulted in choosing only 3 compounds from the top 15 hits. The selected 3 compounds are illustrated in Fig. 9.

De novo design of inhibitors against full-length FOXC2 (SBDD)

Docking of the 79 molecules, obtained from de novo structure-based design using LIGANN [56–58], against the detected active site of the full-length FOXC2 was carried

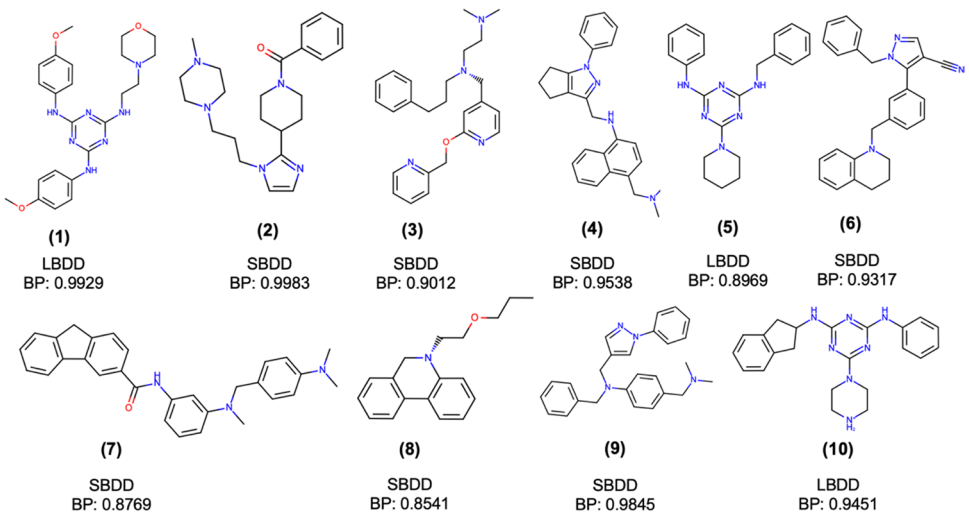


Fig. 9 The top 10 hits identified by augmenting both LBDD and SBDD by docking the compounds obtained from similarity search and de novo structure-based design against the full-length FOXC2. The molecules are labeled for the method used to generate the molecule either LBDD (similarity search) or SBDD (de novo structure-based design). In addition, the binding probability (BP) of the molecule against the full-length FOXC2 obtained from the BindScope server is also listed [52–54]

out using FRED [46, 47]. Compounds generated from SBDD have higher docking scores than the compounds generated from LBDD. The docking score of the top 15 hits generated from SBDD ranges from – 17 to – 11 kcal/mol (Table S2). The higher docking scores could be attributed to the smaller and the simpler compounds obtained from de novo SBDD comparing to the compounds obtained from LBDD. The compounds with binding probability equal to or higher than 0.85 as evaluated using BindScope server [51–53] were selected. This additional filtration resulted in choosing 7 compounds from the top 15 hits. The selected 7 compounds, illustrated in Fig. 9, do not have the 1,3,5-triazine-2,4-diamine moiety which was an essential functional group in the hits from the similarity search.

In summary, the top 10 hits obtained from this filtration included 3 compounds from the similarity search (LBDD) and 7 compounds from de novo structure-based design (SBDD). Some of the compounds, especially those obtained from LBDD, were structurally similar to MC-1-F2 (Fig. 1).

Since the main problem with the experimentally identified inhibitor MC-1-F2 was the unfavorable physicochemical properties which results in difficulty purifying the compound using HPLC, the physicochemical properties of the top ten hits were inspected. The physicochemical properties of the compounds generated from the similarity search were found to have a higher polar surface area (PSA) comparing to those generated from the de novo structure-based design (Table 2). The top ten hits, obtained from augmenting both LBDD and SBDD, have a larger number of hydrogen bond acceptor moieties than hydrogen bond donor moieties. This is also a feature in the experimentally identified inhibitor MC-1-F2. In addition, the top ten hits follow both the Veber's rule [65] and Lipinski's rule of five, where the PSA for all the top ten hits is less than 140 $Å^2$ unlike MC-1-F2 (Table 2). The physicochemical properties of the top 10 hits are favorable comparing to MC-1-F2. This will provide better bioavailability and a higher chance of being an efficient drug, as well as an easier synthesis process.

The binding modes of the top ten hits were investigated in comparison to MC-1-F2. All the top ten hits and MC-1-F2 share interactions with four specific residues, ASN449, LEU459, SER476, and THR484, in the identified active site (Fig. 10).

The stability of the identified binding mode was further evaluated using MD simulations of the top ten hits. The low deviation of the RMSD values reflects the convergence of the simulations (Fig. 11). The fluctuation of the RMSD values for each trajectory is less than 0.25 nm. This indicates the stability of the binding between the top ten hits and the identified active site of the developed full-length of FOXC2 (Fig. 11).

The binding free energies of the top ten hits targeting FOXC2 were calculated using MM/GBSA method. The binding affinities of eight identified ligands are higher than the experimental inhibitor (MC-1-F2), except for compound 5 with a slightly lower binding affinity and compounds 4 and 10 with positive binding free energy (Fig. 12). This implies the effectiveness of augmenting LBDD and SBDD in identifying new inhibitors for FOXC2 with better physicochemical properties.

Discussion and conclusion

MC-1-F2 is the first identified inhibitor of FOXC2 protein, but has unfavorable physiochemical properties. This resulted in difficulty purifying the drug during its synthesis. The two main computational drug design approaches: ligand-based drug design (LBDD) and SBDD were applied to find new hits with desirable physicochemical properties. The top eight hits identified from both LBDD and SBDD have favorable physicochemical properties following both Lipinski's and Veber's rules [65]. We anticipate that the favorable physicochemical properties will play important roles in our future efforts to examine the efficacy of small

Table 2 The physiochemical properties calculated by FRED [46, 47] of the top ten hits obtained from both LBDD and SBDD in comparison to the experimentally identified inhibitor MC-1-F2

Compound no.	Generated from	Molecular weight	No. of ON	No. of OHNH	PSA
1	Similarity search (LBDD)	451.5	3	10	105.7
2	LIGANN (SBDD)	395.5	0	6	44.6
3	LIGANN (SBDD)	404.5	0	5	41.5
4	LIGANN (SBDD)	396.5	1	4	33.1
5	Similarity search (LBDD)	391.5	2	6	92.4
6	LIGANN (SBDD)	404.5	0	4	44.9
7	LIGANN (SBDD)	447.6	1	4	35.6
8	LIGANN (SBDD)	267.4	0	2	12.5
9	LIGANN (SBDD)	396.5	0	4	24.3
10	Similarity search (LBDD)	388.5	3	7	82.6
MC-1-F2	Experimental Inhibitor	746.9	9	18	229.5

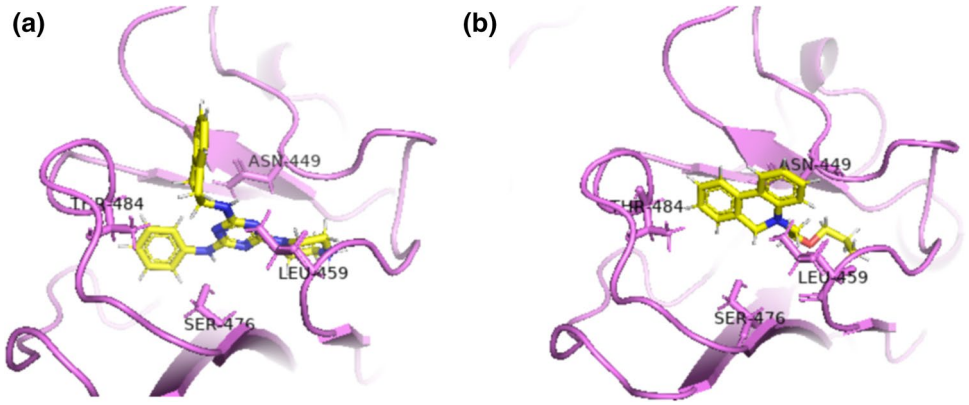


Fig. 10 The binding modes of the top ten hits obtained from both LBDD and SBDD show common interactions with four residues in the active site (ASN 449, LEU 459, SER476, and THR484). Only one molecule from each strategy is shown for simplification. **a** The interaction of compound 10 as representation for the interaction of the compounds obtained from LBDD; **b** The interaction of compound 8 as representation for the interaction of the compounds obtained from SBDD

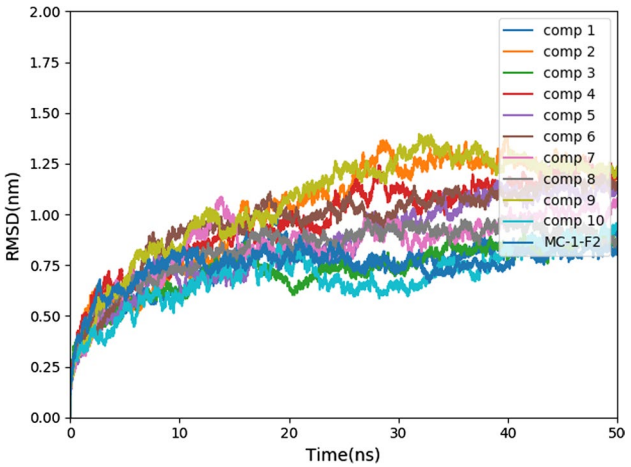


Fig. 11 RMSD plot of the top ten hits and MC-1-F2 calculated over the 50-ns simulations

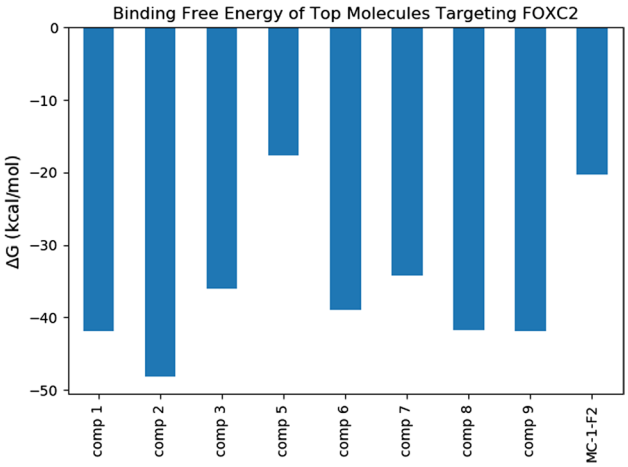


Fig. 12 The binding free energies of the top ligands targeting the developed full-length FOXC2 protein. Compounds 4 and 10 are not plotted due to their positive binding free energy

molecule inhibitors of FOXC2 in vivo as well as to perform clinical and translational research.

The top eight hits were found to share a common binding mode, including four residues: ASN449, LEU459, SER476, and THR484. That is why more studies should be carried out to investigate the role of these four residues and the C-terminal of FOXC2 protein in its role in the EMT process. The binding modes of MC-1-F2 and the top eight hits suggest that these could be allosteric compounds. In a previous study, it has been shown that MC-1-F2 retains FOXC2 in the cytoplasm resulting in the inhibition of FOXC2 nuclear localization, and the cytoplasm-retained FOXC2 becomes subjected to 26S-proteasome-mediated degradation [25]. Examining the roles that these four residues and the C terminus of FOXC2 play in the FOXC2 protein–protein interactions associated with FOXC2 nuclear localization could lead to a mechanistic understanding of FOXC2 inhibitors.

The developed model for the full length of FOXC2 with I-TASSER was found to be stable as indicated by the RMSD value ranging between 1.15 and 1.20 nm after the first 10 ns and throughout the remainder of the 100-ns simulations. This reflects the stability of the model and supports the use of the model as the basis for the SBDD.

FOXC2 is a new interesting therapeutic target for the development of anticancer drugs by inhibiting its role in initiating and maintaining the EMT process. Augmenting both LBDD by carrying out a similarity search for MC-1-F2 against both ZINC and ChEMBL databases and SBDD by carrying out homology modeling, MD, de novo structure-based design, and MM/GBSA calculations helped in identifying new possible hits. Moreover, applying several

filtration steps by carrying out docking using FRED and using BindScope, a convolutional neural network approach, and binding free energy calculations ensured that the top eight hits obtained had the desirable physicochemical properties unlike the experimental inhibitor identified MC-1-F2.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11030-022-10519-0.

Acknowledgements Research reported in this paper was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award no. R15GM122013. Computational time was generously provided by Southern Methodist University's Center for Research Computing. The authors would like to thank OpenEye Scientific Software Inc. for providing an academic license for their computer-aided drug design programs.

Author contributions MTI contributed to conceptualization, methodology, formal analysis and investigation, and writing and preparation of the original draft. JL contributed to writing, reviewing, and editing of the manuscript and supervision. PT contributed to conceptualization and writing, reviewing, and editing of the manuscript, funding acquisition, and supervision.

Declarations

Conflict of interest The authors declare no competing financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

Data availability Experimental data: The crystal structure of DBD of Forkhead Box Protein C2 (PDB ID: 6O3T)) can be obtained free of charge from the Protein Data Bank (https://www.rcsb.org/). The sequence of Forkhead Box Protein C2 (UniprotKB Q99958) can be obtained free of charge from Uniprot (https://www.uniprot.org). Packages: The multi-fingerprint browser can be used free of charge from http://dcb-reymond23.unibe.ch:8080/MCSS/. KNIME package can be downloaded free of charge from (https://www.knime.com/). OpenMM. The package can be downloaded free of charge from (https://openmm.org). I-TASSER server can be used free of charge from (https://zhanggroup.org/I-TASSER/). BindScope server can be used free of charge from (https://playmolecule.com/BindScope/). LIGANN server can be used free of charge from (https://playmolecule.com/LiGANN/). MakeReceptor module, Omega, and FRED can be obtained from OpenEye Scientific using (https://www.eyesopen.com). PyMOL 2.4.0 (can be obtained free of charge from (https://pymol.org/2/). Molprobity can be used free of charge from (http://molprobity.biochem.duke.edu). ProSA module can be used free of charge from https://prosa.services.came.sbg.ac.at/prosa.php.

References

1. Kalluri R, Neilson EG et al (2003) Epithelial–mesenchymal transition and its implications for fibrosis. J Clin Invest 112:1776–1784. https://doi.org/10.1172/JCI20530
2. Kalluri R, Weinberg RA et al (2009) The basics of epithelial–mesenchymal transition. J Clin Invest 119:1420–1428. https://doi.org/10.1172/JCI39104
3. Tsai JH, Yang J (2013) Epithelial–mesenchymal plasticity in carcinoma metastasis. Genes Dev 27:2192–2206. https://doi.org/10.1101/gad.225334.113
4. Mitra A, Mishra L, Li S (2015) EMT, CTCs and CSCs in tumor relapse and drug-resistance. Oncotarget 6:10697–10711. https://doi.org/10.18632/oncotarget.4037
5. Sato R, Semba T, Saya H, Arima Y (2016) Concise review: stem cells and epithelial–mesenchymal transition in cancer: biological implications and therapeutic targets. Stem Cells 34:1997–2007. https://doi.org/10.1002/stem.2406
6. Shi Y, Massagué J (2003) Mechanisms of TGF-β signaling from cell membrane to the nucleus. Cell 113:685–700. https://doi.org/10.1016/S0092-8674(03)00432-X
7. Medici D, Hay ED, Olsen BR (2008) Snail and Slug promote epithelial-mesenchymal transition through β-catenin–T-cell factor-4-dependent expression of transforming growth factor-β3. Mol Biol Cell 19:4875–4887. https://doi.org/10.1074/jbc.M111.276311
8. Niessen K, Fu Y, Chang L, Hoodless PA, McFadden D, Karsan A (2008) Slug is a direct Notch target required for initiation of cardiac cushion cellularization. J Cell Biol 182:315–325. https://doi.org/10.1083/jcb.200710067
9. Mani SA, Yang J, Brooks M, Weinberg RA (2007) Mesenchyme Forkhead 1 (FOXC2) plays a key role in metastasis and is associated with aggressive basal-like breast cancers. Proc Natl Acad Sci USA 104:10069–10074. https://doi.org/10.1073/pnas.0703900104
10. Chen X et al (2019) Structural basis for DNA recognition by FOXC2. Nucleic Acids Res 47:3752–3764. https://doi.org/10.1093/nar/gkz077
11. Golson ML, Kaestner KH (2016) Fox transcription factors: from development to disease. Development 143:4558–4570. https://doi.org/10.1242/dev.112672
12. Clark KL, Halay ED, Lai E, Burley SK (1993) Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5. Nature 364:412–420. https://doi.org/10.1038/364412a0
13. Brent MM, Anand R, Marmorstein R (2008) Structural basis for DNA recognition by FoxO1 and its regulation by posttranslational modification. Structure 16:1407–1416. https://doi.org/10.1016/j.bbamcr.2010.11.025
14. Boura E, Rezabkova L, Brynda J, Obsilova V, Obsil T (2010) Structure of the human FOXO4-DBD–DNA complex at 1.9 Å resolution reveals new details of FOXO binding to the DNA. Acta Crystallogr D Biol Crystallogr 66:1351–1357. https://doi.org/10.1107/s0907444910042228
15. Littler DR, Alvarez-Fernández M, Stein A, Hibbert RG, Heidebrecht T, Aloy P, Medema RH, Perrakis A (2010) Structure of the FoxM1 DNA-recognition domain bound to a promoter sequence. Nucleic Acids Res 38:4527–4538. https://doi.org/10.1093/nar/gkq194
16. Singh P, Han EH, Endrizzi JA, O'Brien RM, Chi Y-I (2017) Crystal structures reveal a new and novel FoxO1 binding site within the human glucose-6-phosphatase catalytic subunit 1 gene promoter. J Struct Biol 198:54–64. https://doi.org/10.1016/j.jsb.2017.02.006
17. Fang J, Dagenais SL, Erickson RP, Arlt MF, Glynn MW, Gorski JL, Seaver LH, Glover TW (2000) Mutations in FOXC2 (MFH-1), a forkhead family transcription factor, are responsible for the hereditary lymphedema-distichiasis syndrome. Am J Hum Genet 67:1382–1388. https://doi.org/10.1086/316915
18. Bell R, Brice G, Child A, Murday V, Mansour S, Sandy C, Collin J, Brady A, Callen D, Burnand K, Mortimer P, Jeffery S (2001) Analysis of lymphoedema-distichiasis families for FOXC2 mutations reveals small insertions and deletions throughout the gene. Hum Genet 108:546–551. https://doi.org/10.1007/s004390100528
19. Yu M et al (2013) Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. Science 339:580–584. https://doi.org/10.1126/science.1228522

20. Wang T, Zheng L, Wang Q, Hu Y-W (2018) Emerging roles and mechanisms of FOXC2 in cancer. Clin Chim Acta 479:84–93. https://doi.org/10.1186/s12935-020-01265-0
21. Xia L et al (2013) Overexpression of forkhead box C1 promotes tumor metastasis and indicates poor prognosis in hepatocellular carcinoma. Hepatology 57:610–624. https://doi.org/10.1002/hep.26029
22. Hollier BG et al (2013) FOXC2 expression links epithelial-mesenchymal transition and stem cell properties in breast cancer. Cancer Res 73:1981–1992. https://doi.org/10.1158/0008-5472.CAN-12-2962
23. Werden SJ et al (2016) Phosphorylation of serine 367 of FOXC2 by p38 regulates ZEB1 and breast cancer metastasis, without impacting primary tumor growth. Oncogene 35:5977–5988. https://doi.org/10.1038/onc.2016.203
24. Yu YH et al (2013) MiR-520h-mediated FOXC2 regulation is critical for inhibition of lung cancer progression by resveratrol. Oncogene 32:431–443. https://doi.org/10.1038/onc.2012.74
25. Castaneda M, Chen L, Pradhan L, Li S, Zein R (2018) A Forkhead box protein C2 inhibitor: targeting epithelial–mesenchymal transition and cancer metastasis. ChemBioChem 19:1359–1364. https://doi.org/10.1002/cbic.201800022
26. Awale M, Reymond J-L (2014) A multi-fingerprint browser for the ZINC database. Nucleic Acids Res 42:W234–W239. https://doi.org/10.1093/nar/gku379
27. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG (2012) ZINC: a free tool to discover chemistry for biology. J Chem Inf Model 52:1757–1768. https://doi.org/10.1021/ci3001277
28. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742–754. https://doi.org/10.1021/ci100050t
29. Hagadone TR (1992) Molecular substructure similarity searching: efficient retrieval in two-dimensional structure databases. J Chem Inf Comput Sci 32:515–521. https://doi.org/10.1021/ci00009a019
30. Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F, Bellis L, Overington JP (2015) ChEMBL web services: streamlining access to drug discovery data and utilities. Nucleic Acids Res 43:W612–W620. https://doi.org/10.1093/nar/gkv352
31. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Thiel K, Wiswedel B (2009) KNIME—the Konstanz information miner: version 2.0 and beyond. ACM SIGKDD Explor Newsl 11:26–31. https://doi.org/10.1145/1656274.1656280
32. Borgelt C, Meinl T (2009) Full perfect extension pruning for frequent subgraph mining. In: Zighed DA, Tsumoto S, Ras ZW, Hacid H (eds) Mining complex data, vol 165. Studies in computational intelligence. Springer, Berlin, pp 189–205. https://doi.org/10.1007/978-3-540-88067-7_11
33. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y (2015) The I-TASSER Suite: protein structure and function prediction. Nat Methods 12:7–8. https://doi.org/10.1038/nmeth.3213
34. Zhang Y, Kolinski A, Skolnick J (2003) TOUCHSTONE II: a new approach to ab initio protein structure prediction. Biophys J 85:1145–1164. https://doi.org/10.1016/S0006-3495(03)74551-2
35. Zhang Y, Kihara D, Skolnick J (2002) Local energy landscape flattening: parallel hyperbolic Monte Carlo sampling of protein folding. Proteins Struct Funct Bioinform 48:192–201. https://doi.org/10.1002/prot.10141
36. Benkert P, Biasini M, Schwede T (2011) Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics 27:343–350. https://doi.org/10.1093/bioinformatics/btq662
37. Williams CJ et al (2018) MolProbity: more and better reference data for improved all-atom structure validation. Protein Sci 27:293–315. https://doi.org/10.1002/pro.3330
38. Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci 2:1511–1519. https://doi.org/10.1002/pro.5560020916
39. Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. Proteins Struct Funct Bioinform 17:355–362. https://doi.org/10.1002/prot.340170404
40. Wiederstein M, Sippl MJ (2007) ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35:W407–W410. https://doi.org/10.1093/nar/gkm290
41. Brooks BR et al (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30:1545–1614. https://doi.org/10.1002/jcc.21287
42. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935. https://doi.org/10.1063/1.445869
43. Eastman P, Pande V (2010) OpenMM: A hardware-independent framework for molecular simulations. Comput Sci Eng 12:34–39. https://doi.org/10.1109/MCSE.2010.27
44. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG (1995) A smooth particle mesh Ewald method. J Chem Phys 103:8577–8593. https://doi.org/10.1063/1.470117
45. McGibbon RT et al (2015) MDTraj: a modern open library for the analysis of molecular dynamics trajectories. Biophys J 109:1528–1532. https://doi.org/10.1016/j.bpj.2015.08.015
46. McGann M (2012) FRED pose prediction and virtual screening accuracy. J Chem Inf Model 51:578–596. https://doi.org/10.1021/ci100436p
47. McGann M (2012) FRED and HYBRID docking performance on standardized datasets. J Comput Aided Mol Des 26:897–906. https://doi.org/10.1007/s10822-012-9584-8
48. Kelley BP, Brown SP, Warren GL, Muchmore SW (2015) POSIT: flexible shape-guided docking for pose prediction. J Chem Inf Model 55:1771–1780. https://doi.org/10.1021/acs.jcim.5b00142
49. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. BMC Bioinform 9:1–8. https://doi.org/10.1186/1471-2105-9-40
50. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5:725–738. https://doi.org/10.1038/nprot.2010.5
51. Koes DR, Baumgartner MP, Camacho CJ (2013) Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. J Chem Inf Model 53:1893–1904. https://doi.org/10.1021/ci300604z
52. Jiménez J, Skalic M, Martinez-Rosell G, de Fabritiis G (2018) K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. J Chem Inf Model 58:287–296. https://doi.org/10.1021/acs.jcim.7b00650
53. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK (2012) Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. J Med Chem 55:6582–6594. https://doi.org/10.1021/jm300687e
54. Hawkins PCD, Skillman AG, Warren GL, Ellingson BA, Stahl MT (2010) Conformer generation with OMEGA: algorithm and validation using high quality structures from the Protein Databank and Cambridge Structural Database. J Chem Inf Model 50:572–584. https://doi.org/10.1021/ci100031x
55. Skalic M, Martinez-Rosell G, Jiménez J, de Fabritiis G (2019) PlayMolecule BindScope: large scale CNN-based virtual screening on the web. Bioinformatics 35:1237–1238. https://doi.org/10.1093/bioinformatics/bty758
56. Aggarwala A, Mittalb M, Battinenic G (2021) Generative adversarial network: an overview of theory and applications. Int J Inf Manag Data Insights 1:100004. https://doi.org/10.1016/j.jjimei.2020.100004

57. Skalic M, Jiménez J, Sabbadin D, de Fabritiis G (2019) Shape-based generative modeling for de novo drug design. J Chem Inf Model 59:1205–1214. https://doi.org/10.1021/acs.jcim.8b00706
58. Skalic M, Sabbadin D, Sattarov B, Sciabola S, Fabritiis GD (2019) From target to drug: generative modeling for the multimodal structure-based ligand design. Mol Pharm 16:4282–4291. https://doi.org/10.1021/acs.molpharmaceut.9b00634
59. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C (2015) ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. J Chem Theory Comput 11:3696–3713. https://doi.org/10.1021/acs.jctc.5b00255
60. Salomon-Ferrer R, Case DA, Walker RC (2013) An overview of the Amber biomolecular simulation package. WIREs Comput Mol Sci 3:198–210. https://doi.org/10.1002/wcms.1121
61. Case DA et al (2005) The Amber biomolecular simulation programs. J Comput Chem 26:1668–1688. https://doi.org/10.1002/jcc.20290
62. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. J Comput Chem 25:1157–1174. https://doi.org/10.1002/jcc.20035
63. Wang J, Wang W, Kollman PA, Case DA (2006) Automatic atom type and bond type perception in molecular mechanical calculations. J Mol Graph Model 25:247–260. https://doi.org/10.1016/j.jmgm.2005.12.005
64. Miller BR III, McGee TD Jr, Swails JM, Homeyer N, Gohlke H, Roitberg AE (2012) MMPBSA. py: an efficient program for end-state free energy calculations. J Chem Theory Comput 8:3314–3321. https://doi.org/10.1021/ct300418h
65. Veber DF, Johnson SR, Cheng H-Y, Smith BR, Ward KW, Kopple KD (2002) Molecular properties that influence the oral bioavailability of drug candidates. J Med Chem 45:2615–2623. https://doi.org/10.1021/jm020017n