# D3R GC2016 Challenge: Dataset Description and Instructions

## Overview

The challenge involves a protein-ligand dataset for the farnesoid x receptor (FXR) target, generously donated by Roche Pharmaceuticals. The dataset comprises:

- IC50 data for 102 compounds in total, 96 in four chemical series (benzimidazoles, isoxazoles, spiros and sulfonamides) and 6 miscellaneous compounds.
- Potency range of 0.000335 – 62.37 μM for 92 compounds, and 10 having potency > 100 μM.
- 36 co-crystal structures and one apo, with representatives from each of the four chemical series. Resolutions range from 1.8 – 2.6 Å.

As with the GC2015 exercise, this challenge will involve two stages. **Stage 1** of the Challenge is to predict the ligand poses of the available crystal structures and also to predict or rank the potencies of all ligands, including those for which crystal structures are not available. After Stage 1 has closed, all available co-crystal structures will be made public. The **Stage 2** Challenge is to repeat the affinity predictions or rankings, this time using the additional disclosed ligand-pose information.

Please note that the compounds whose IDs are listed in Appendix A (at the end of these instructions) were prepared and tested as 50:50 racemic mixtures. However, the following citation[1], available co-crystal structures and discussions with Roche scientists indicate that the S isomer is much more active than the R. The provided SMILES are the S form but you're free to consider the R. You'll be evaluated against the uncorrected experimental IC50s.

Two subsets have been selected, 15 sulfonamides and 18 spiros, which contain chemically similar compounds and thus lend themselves to the calculation of relative binding affinities by alchemical methods, such as free energy perturbation. The compound IDs for these sets are provided in Appendix B and C, at the end of these instructions. The free energy (FE) challenge can be valid for not only **Stage 1** but also **Stage 2**.

We are providing the apo protein structure, compounds in the form of canonical SMILES strings and SDfiles, and information on the experimental conditions for the crystallography and binding measurements. The ligand IDs are from **FXR_1** to **FXR_102** and the IDs for ligands with available co-crystal structures are from **FXR_1** to **FXR_36**. These 36 are therefore for the pose prediction part of the Challenge.

**Factors that make this Challenge interesting:**

- There are two helices adjacent to the ligand binding site that can adopt varied conformations. The conformations observed in the blinded dataset are well exemplified in publicly available co-crystal structures of FXR in the Protein Data Bank.

- There are PDB entries in the Protein Data Bank with ligands in the benzimidazole[1,2] and isoxazole[3] chemical series, but co-crystal structures with the spiros or sulfonamides are not publicly available, as of today.

- Water mediated protein interactions are important for the binding of some but not all ligands.

- Some ligands have rings with nontrivial puckering options.

NOTE: Our own test calculations docking the Challenge ligands into structures in the public domain showed no clear disadvantage to using these structures, relative to the structures provided by Roche, except for the self-docking cases.

**The information packet provided**

The packet for this Challenge includes the apoprotein crystal structure provided by Roche. *This is not necessarily meant for use in your predictions: you are free to use any structure(s) of your choice from the PDB*. However, when you upload your pose predictions, your structures must be translationally and rotationally aligned to the apo structure. We request that these alignments be based on the coordinates of the α-carbon backbone atoms and emphasized secondary structure elements.

The compound IDs and SMILES strings for all ligands are provided as a csv file named **ALL_FXR_compounds_D3R_GC2016.csv,** and the compounds are also provided in the form of an SDfile named **ALL_FXR_compounds_D3R_GC2016.sdf**. Further, the compound IDs for the two FEP sets are provided in Appendices B and C, at the end of these instructions.

No attempt was made to set appropriate starting conformations or optimal protonation or tautomer states for the ligands, or to generate alternative tautomer states. It is up to you to choose and set these states for your calculations.

**Binding assay and crystallization conditions**

The FXR binding affinities were carried out using the Scintillation Proximity Assay (SPA)[4], a radioligand displacement assay. The assay buffer contained 50 mM HEPES (pH 7.4), 10 mM NaCl, 5 mM MgCl2 and 0.01% CHAPS. The reactions were incubated for 30 min in the presence of [$^{3}$H]2,*N*-dicyclohexyl-2-[2-(2,4-dimethoxy-phenyl)-benzoimidazol-1-yl]acetamide, the test compound and the buffer. The amount of radioligand that remained bound was determined; dose response curves were then generated and the IC50s calculated.

Different crystallization conditions had to be used for the diverse chemotypes; the crystallization conditions for the compound IDs for the pose prediction challenge are provided as a csv file named **Data_set_fxr_crystallization_conditions.csv.**

Note that, per Roche, the crystallization solutions for FXR_10 and FXR_26 were unbuffered, and their pHs are not known.

**Predicted Aggregators**

All compounds have been subjected to Open Eye FILTER (http://www.eyesopen.com/filter) to identify known and predicted aggregators. None of the Challenge compounds are known aggregators, but 82 out of the 102 compounds are "predicted" as aggregators using the QSAR model within FILTER. Whether they are true aggregators or not has not been tested.

NOTE: Open Eye states this QSAR model is very aggressive in its predictions.

**Due dates**

Your predictions (poses, ligand affinities/rankings and/or FE set predictions) for **Stage 1** must be uploaded to the D3R website by 5pm (PST), 21$^{st}$ November 2016. The experimental ligand-protein poses will be released immediately after Stage 1 closes.

Your **Stage 2** predictions (ligand affinities/ rankings) are due by 5pm (PST), 1st February 2017. The experimental ligand-protein IC50s will be released immediately after Stage 2 closes.

**Computational methods allowed**
You may use any method(s) you like to generate your pose and affinity ranking predictions; e.g., docking and scoring, MM-PB(GB)/SA, FEP, quantum-based methods, etc. However, for the FE sets, we emphasize these are designed for methods that can be categorized as alchemical free energy methods.

**Anonymous versus public participation**
When you sign up for the challenge, you are given the option of participating anonymously. Anonymous participation means that we may report on your predictions and methods, but your identity will not be disclosed. Public participation means we may also disclose who you are. Please note that, although we are committed to protecting the identity of anonymous participants, we cannot make any guarantees.

For each stage, you may use the D3R website to change your anonymous/public status until the stage has closed. However, after the stage has closed, you may not change your anonymous/public status.

**Submitting your predictions**
Two separate files, in different file formats, must represent each ligand pose prediction you submit: a PDB file for the protein, and an MDL mol file for the ligand. Any ligand coordinates provided in PDB format or included in the protein PDB files will be ignored. Both the protein and the ligand coordinates must be aligned with the coordinate frame of the apoprotein structure provided in this packet. You may submit up to five predicted poses for each ligand. If you submit more than one pose for a given ligand, then a docking score or energy should be provided for each pose.

For compound affinities, we anticipate accepting one or more of the following for each challenge set:

- Predicted affinities, in units of nM, for each compound.
- Relative affinities for each compound, where the ligand with the lowest ID number is arbitrarily set to a potency of 1.
- An ordinal ranking of ligands, where 1 indicates maximum affinity (e.g., lowest IC50).

A template file will be provided for submitting affinity or ranking predictions and detailed instructions for uploading your predictions for each stage will be provided in the coming weeks. Please note that there will be some adjustments to the template files and the submission procedures relative to prior challenges.

**Evaluation of predictions**
Pose predictions will be evaluated based on, at minimum, symmetry-corrected RMSD to crystallographic conformations. Additional criteria may be based on ligand-protein contacts and/or overlap of predicted and experimental electron densities. We also might evaluate predicted conformational changes of the protein binding site. Affinity predictions/rankings will be evaluated based on, at minimum, accuracy of ranking.

**Pending items, error reports, questions**
We will email you during the Challenge regarding templates and instructions for uploading your predictions. We will also email you if necessary to share additional information or changes to the

Challenge. Please feel free to contact us if you notice any errors in the information provided or have questions about D3R GC2016 Challenge: drugdesigndata@gmail.com.

**D3R Webinar, March 2017**

Participants are invited to share and discuss their results, at a webinar hosted by D3R, which is scheduled for March, 2017. The exact dates will be communicated later.

**References**

(1) Richter, H. G.; Benson, G. M.; Bleicher, K. H.; Blum, D.; Chaput, E.; Clemann, N.; Feng, S.; Gardes, C.; Grether, U.; Hartman, P.; Kuhn, B.; Martin, R. E.; Plancher, J. M.; Rudolph, M. G.; Schuler, F.; Taylor, S. *Bioorganic & medicinal chemistry letters* **2011**, *21*, 1134.

(2) Richter, H. G.; Benson, G. M.; Blum, D.; Chaput, E.; Feng, S.; Gardes, C.; Grether, U.; Hartman, P.; Kuhn, B.; Martin, R. E.; Plancher, J. M.; Rudolph, M. G.; Schuler, F.; Taylor, S.; Bleicher, K. H. *Bioorganic & medicinal chemistry letters* **2011**, *21*, 191.

(3) Feng, S.; Yang, M.; Zhang, Z.; Wang, Z.; Hong, D.; Richter, H.; Benson, G. M.; Bleicher, K.; Grether, U.; Martin, R. E.; Plancher, J. M.; Kuhn, B.; Rudolph, M. G.; Chen, L. *Bioorganic & medicinal chemistry letters* **2009**, *19*, 2595.

(4) Gardes, C.; Blum, D.; Bleicher, K.; Chaput, E.; Ebeling, M.; Hartman, P.; Handschin, C.; Richter, H.; Benson, G. M. *Journal of lipid research* **2011**, *52*, 1188.

**Appendix A. Compounds experimentally tested as 50:50 racemic mixtures**

FXR_6, FXR_7, FXR_8, FXR_9, FXR_13, FXR_18, FXR_20, FXR_22, FXR_25, FXR_31, FXR_35, FXR_36, FXR_37, FXR_39, FXR_40, FXR_42, FXR_50, FXR_51, FXR_52, FXR_53, FXR_54, FXR_55, FXR_56, FXR_57, FXR_58, FXR_59, FXR_60, FXR_61, FXR_62, FXR_63, FXR_64, FXR_66, FXR_67, FXR_68, FXR_69, FXR_70, FXR_71, FXR_72.

**Appendix B. FE Set 1**

| | |
|---|---|
| FXR_17 | CCOC(=O)c1ccc(NC(=O)c2c3CN(CCc3nn2c4ccccc4)S(=O)(=O)c5cccs5)cc1 |
| FXR_45 | CCOC(=O)c1ccc(NC(=O)c2c3CN(CCc3nn2c4ccc(OC(F)(F)F)cc4)S(=O)(=O)c5cccs5)cc1 |
| FXR_46 | NC(=O)c1ccc(NC(=O)c2c3CN(CCc3nn2c4ccccc4)S(=O)(=O)c5cccs5)cc1 |
| FXR_47 | CCOC(=O)c1cccc(NC(=O)c2c3CN(CCc3nn2c4ccccc4)S(=O)(=O)c5cccs5)c1 |
| FXR_48 | CCOC(=O)Cc1ccc(NC(=O)c2c3CN(CCc3nn2c4ccccc4)S(=O)(=O)c5cccs5)cc1 |
| FXR_49 | CC(=O)c1ccc(NC(=O)c2c3CN(CCc3nn2c4ccccc4)S(=O)(=O)c5cccs5)cc1 |
| FXR_91 | O=C(Nc1ccccc1)c2c3CN(CCc3nn2c4ccccc4)S(=O)(=O)c5cccs5 |
| FXR_93 | O=C(Nc1ccccc1)n2c3CN(CCc3cc2c4ccccc4)S(=O)(=O)c5cccs5 |
| FXR_95 | CC(=O)Nc1ccc(NC(=O)c2c3CN(CCc3nn2c4ccccc4)S(=O)(=O)c5cccs5)cc1 |
| FXR_96 | CN(C)C(=O)c1ccc(NC(=O)c2c3CN(CCc3nn2c4ccccc4)S(=O)(=O)c5cccs5)cc1 |
| FXR_98 | CNC(=O)c1ccc(NC(=O)c2c3CN(CCc3nn2c4ccccc4)S(=O)(=O)c5cccs5)cc1 |
| FXR_99 | COc1ccc(NC(=O)c2c3CN(CCc3nn2c4ccccc4)S(=O)(=O)c5cccs5)cc1 |
| FXR_100 | NS(=O)(=O)c1ccc(NC(=O)c2c3CN(CCc3nn2c4ccccc4)S(=O)(=O)c5cccs5)cc1 |
| FXR_101 | OC(=O)c1ccc(NC(=O)c2c3CN(CCc3nn2c4ccccc4)S(=O)(=O)c5cccs5)cc1 |
| FXR_102 | O=C(Nc1ccc(cc1)C(=O)N2CCOCC2)c3c4CN(CCc4nn3c5ccccc5)S(=O)(=O)c6cccs6 |

**Appendix C. FE Set 2**

| | |
|---|---|
| FXR_10 | OC(=O)c1ccc(CN2C(=O)C3(CCN(CC3)S(=O)(=O)c4cccs4)c5cc(Br)ccc25)cc1 |
| FXR_12 | OC(=O)c1ccc(CN2C(=O)C3(CCN(CC3)S(=O)(=O)c4ccccc4Cl)c5cc(Br)ccc25)cc1 |
| FXR_38 | COC(=O)c1ccc(CN2C(=O)C3(CCN(CC3)S(=O)(=O)c4cccs4)c5cc(Br)ccc25)cc1 |
| FXR_41 | COC(=O)c1ccc(CN2C(=O)C3(CCN(CC3)S(=O)(=O)c4ccccc4Cl)c5cc(Br)ccc25)cc1 |
| FXR_73 | Oc1ccc(CN2C(=O)C3(CCN(CC3)S(=O)(=O)c4cccs4)c5cc(Br)ccc25)cc1 |
| FXR_74 | OC(=O)c1ccc(CN2C(=O)C3(CCN(CC3)S(=O)(=O)c4ccccc4Br)c5cc(Br)ccc25)cc1 |
| FXR_75 | Brc1ccc2N(Cc3ccncc3)C(=O)C4(CCN(CC4)S(=O)(=O)c5cccs5)c2c1 |
| FXR_76 | OC(=O)c1ccc(CN2C(=O)C3(CCN(CC3)S(=O)(=O)c4ccccc4)c5cc(Br)ccc25)cc1 |
| FXR_77 | OC(=O)c1ccc(CN2C(=O)C3(CCN(CC3)S(=O)(=O)c4cccc(Cl)c4Cl)c5cc(Br)ccc25)cc1 |
| FXR_78 | OC(=O)c1ccc(CN2C(=O)C3(CCN(CC3)S(=O)(=O)c4c(Cl)cccc4Cl)c5cc(Br)ccc25)cc1 |
| FXR_79 | OC(=O)c1cccc(CN2C(=O)C3(CCN(CC3)S(=O)(=O)c4cccs4)c5cc(Br)ccc25)c1 |
| FXR_81 | Cc1c(Cl)cccc1S(=O)(=O)N2CCC3(CC2)C(=O)N(Cc4ccc(cc4)C(=O)O)c5ccc(Br)cc35 |
| FXR_82 | OC(=O)c1ccc(CN2C(=O)C3(CCN(CC3)S(=O)(=O)c4cccc(Cl)c4F)c5cc(Br)ccc25)cc1 |
| FXR_83 | OC(=O)c1ccc(CN2C(=O)C3(CCN(CC3)S(=O)(=O)c4cc(Cl)ccc4Cl)c5cc(Br)ccc25)cc1 |
| FXR_84 | OC(=O)c1ccc(CN2C(=O)C3(CCN(CC3)S(=O)(=O)c4ccccc4F)c5cc(Br)ccc25)cc1 |
| FXR_85 | Cc1ccccc1S(=O)(=O)N2CCC3(CC2)C(=O)N(Cc4ccc(cc4)C(=O)O)c5ccc(Br)cc35 |
| FXR_88 | OC(=O)c1ccc(CN2C(=O)C3(CCN(CC3)S(=O)(=O)c4ccccc4C(F)(F)F)c5cc(Br)ccc25)cc1 |
| FXR_89 | OC(=O)c1ccc(CN2C(=O)C3(CCN(CC3)S(=O)(=O)c4ccc(Cl)cc4)c5cc(Br)ccc25)cc1 |