Universiteit
Leiden
The Netherlands

Proteins in harmony: Tuning selectivity in early drug discovery

Burggraaff, L.

Citation

Burggraaff, L. (2020, October 21). *Proteins in harmony: Tuning selectivity in early drug discovery*. Retrieved from https://hdl.handle.net/1887/136965

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/136965

Note: To cite this publication please use the final published version (if applicable).

Cover Page

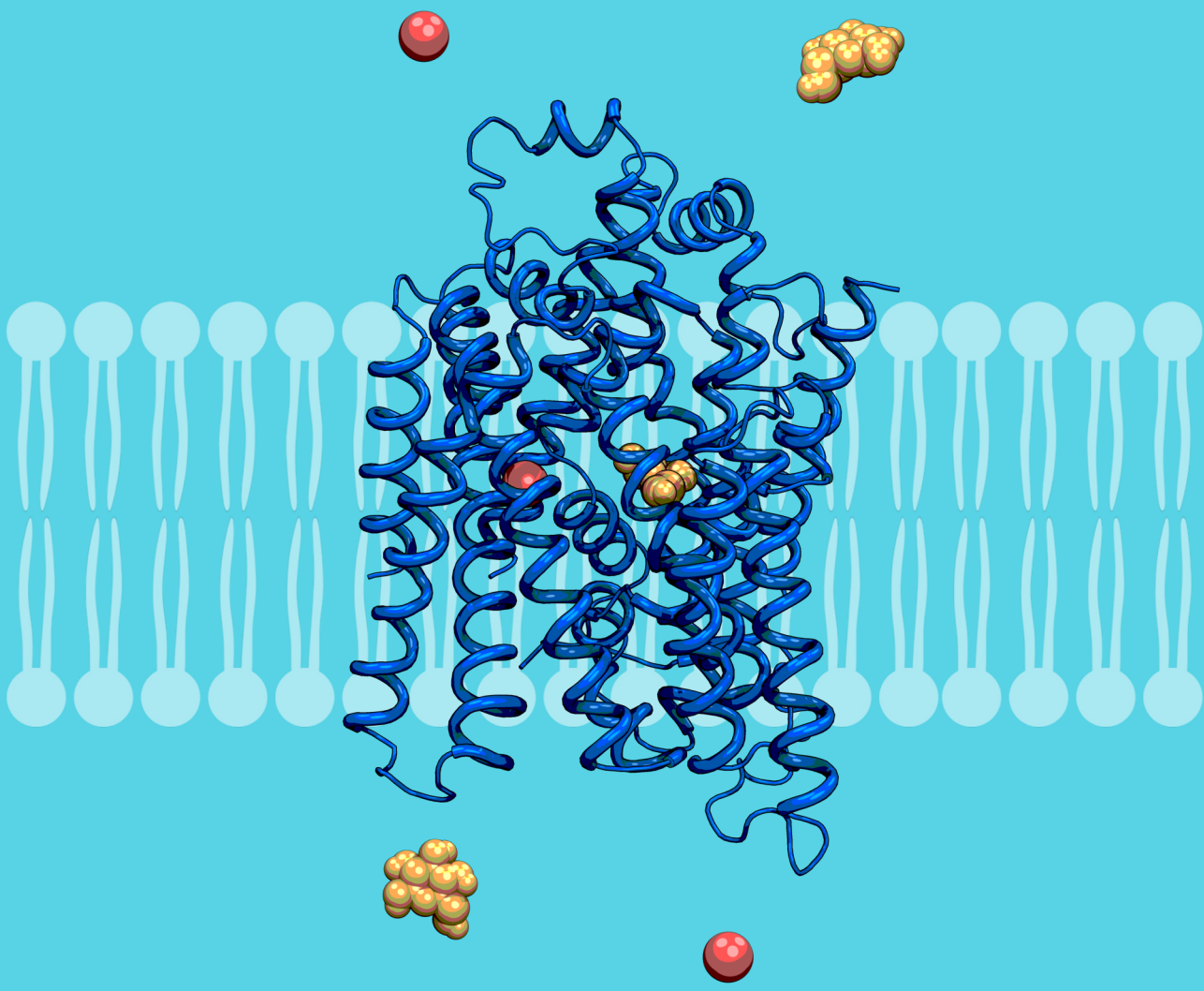
Universiteit Leiden

The handle http://hdl.handle.net/1887/136965 holds various files of this Leiden University dissertation.

Author: Burggraaff, L.
Title: Proteins in harmony: Tuning selectivity in early drug discovery
Issue date: 2020-10-21

Chapter three

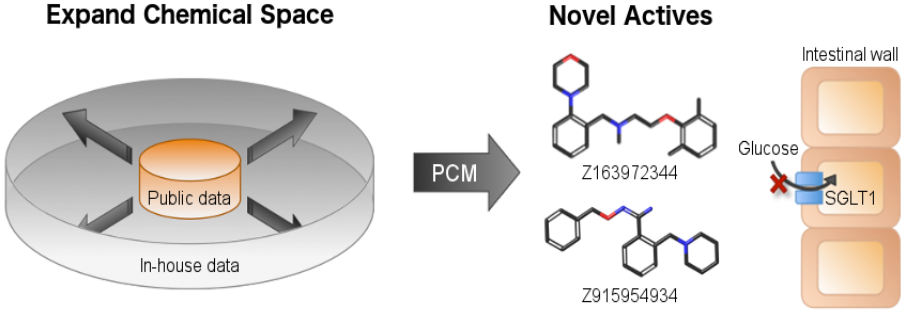
Identification of Novel Small Molecule Inhibitors for Solute Carrier SGLT1 using Proteochemometric Modeling



Lindsey Burggraaff, Paul Oranje, Robin Gouka, Pieter van der Pijl, Marian Geldof, Herman W.T. van Vlijmen, Adriaan P. IJzerman, and Gerard J.P. van Westen. Journal of Cheminformatics. 2019. doi: 10.1186/s13321-019-0337-8.

Abstract

Sodium-dependent glucose co-transporter 1 (SGLT1) is a solute carrier responsible for active glucose absorption. SGLT1 is present in both the renal tubules and small intestine. In contrast, the closely related sodium-dependent glucose co-transporter 2 (SGLT2), a protein that is targeted in the treatment of diabetes type II, is only expressed in the renal tubules. Although dual inhibitors for both SGLT1 and SGLT2 have been developed, no drugs on the market are targeted at decreasing dietary glucose uptake by SGLT1 in the gastrointestinal tract. Here we aim at identifying SGLT1 inhibitors *in silico* by applying a machine learning approach that does not require structural information, which is absent for SGLT1. We applied proteochemometrics by implementation of compound- and protein-based information into random forest models. We obtained a predictive model with a sensitivity of 0.64 ± 0.06, specificity of 0.93 ± 0.01, positive predictive value of 0.47 ± 0.07, negative predictive value of 0.96 ± 0.01, and Matthews correlation coefficient of 0.49 ± 0.05. Subsequent to model training, we applied our model in virtual screening to identify novel SGLT1 inhibitors. Of the 77 tested compounds, 30 were experimentally confirmed for SGLT1-inhibiting activity in vitro, leading to a hit rate of 39% with activities in the low micromolar range. Moreover, the hit compounds included novel molecules, which is reflected by the low similarity of these compounds with the training set (< 0.3). Conclusively, proteochemometric modeling of SGLT1 is a viable strategy for identifying active small molecules. Therefore, this method may also be applied in detection of novel small molecules for other transporter proteins.



Introduction

Sodium-dependent glucose co-transporters, or sodium-glucose linked transporters (SGLTs), are solute carriers (SLCs) that are responsible for glucose (re)absorption. SGLTs are members of the sodium-dependent transporters and are encoded by the SLC5A genes.[1] SGLTs are interesting targets in the treatment of diabetes mellitus, as their inhibition reduces the risk of hyperglycemia by decreasing glucose (re-)uptake.[2] In the human body two SGLT isoforms are involved in glucose transport: SGLT1 and SGLT2.[3] Both SGLT1 and SGLT2 are expressed in the kidney, whereas SGLT1 is also expressed in the small intestine.[4] SGLT2 is a high capacity transporter responsible for 90% of glucose reuptake in the renal tubules and multiple compounds have been developed that inhibit this solute carrier.[5,6] Furthermore, SGLT2 inhibition has been shown to decrease blood glucose levels in diabetes type 2 patients.[7] In contrast to SGLT2, SGLT1 is a low-capacity glucose transporter.[1] However, SGLT1 has a higher glucose affinity than SGLT2 and is additionally capable of transporting galactose.[1] Dual inhibitors blocking both SGLT1 and SGLT2 are currently in clinical development.[8,9] In line with previous evidence we suggest that SGLT1 inhibition in the intestine will lower blood glucose levels as well.[10,11] Compounds that do not penetrate the intestinal wall can achieve selective targeting of SGLT1 in the intestine, as they would not reach the renal tubules.[12]

The complexity and the hydrophobic nature of transporter proteins make them challenging to crystalize. Crystal structures of transporters are scarce and binding locations of small molecules to these transporters are often unknown. For human SGLTs no protein structures are available negating the use of structure-based modeling techniques. However, the publicly available compound database ChEMBL includes ligand-protein binding information for multiple SGLTs,[13-15] allowing the use of statistical modeling techniques such as quantitative structure-activity relationship analysis (QSAR) and proteochemometrics (PCM).[16] These techniques, which make use of machine learning, do not require protein structural information and can therefore more easily be applied in the context of SLCs. Although ligand-based pharmacophore modeling, QSAR, and PCM have only been applied to a few SLCs,[17,18] these techniques are well established on other drug targets including membrane proteins such as G protein-coupled receptors.[19-21]

Unfortunately, the publicly available compound interaction data for SGLTs is limited from the point of chemical diversity as the major share of ligands are glycoside-like compounds and oxopyrrolidine-carboxamides. This limited chemical space hence restricts the applicability domain of QSAR and PCM models.[22] The applicability domain of computational models can be interpreted as the theoretical ensemble of molecular structures to which a model can be applied accurately. This domain is dependent on the model input and can therefore be quantified by similarity with the training molecules.

In the current work we show how we expanded the chemical space of SGLT inhibitors (using an in-house dataset),[23] and with that the applicability domain of our SGLT models. We constructed PCM models based on SGLT1 and its closest family members to predict compound activity for SGLT1. We successfully identified novel SGLT1 inhibitors that display low similarity towards the training set.

Results and Discussion

SGLT chemical space

A public dataset was created based on ChEMBL version 23[13,15] which includes the target protein human SGLT1 (hSGLT1), related protein human SGLT2 (hSGLT2), and multiple other SGLTs from different species. The public dataset encompassed 2,063 data points and 1,683 unique compounds, of which 886 compounds had measured hSGLT1 activities. Additionally, this set was supplemented with an in-house dataset of 2,007 molecules previously screened for hSGLT1 and hSGLT2 inhibition.[23] This in-house dataset has as its basis the Spectrum Collection compound library[24] extended with compounds similar to primary screening hits and contained natural products and synthetic compounds. The data derived from ChEMBL was compared to the in-house dataset: the in-house dataset contained an additional 2,005 hSGLT1 activities and 140 hSGLT2 activities, which were not present in the public dataset. The difference between the public and in-house dataset is graphically represented with t-Distributed Stochastic Neighbor Embedding (t-SNE)[25] (**Figure 3.1A**). This technique was applied to decrease the high dimensionality of the datasets, making it possible to visualize them. The high dimensions are a consequence of the many descriptors that are used to describe the data, i.e. FCFP6 fingerprints. The t-SNE plot shows that the data derived for proteins similar to hSGLT1 extend the chemical space; many hSGLT2 compounds from the public domain are not tested on hSGLT1 and thus provide additional chemical information. The in-house and public datasets considerably differ from each other, with a slight overlap of only a few hSGLT1 and hSGLT2 public compounds with the in-house dataset. To further investigate the difference between the public and in-house dataset, the following physicochemical properties were considered: molecular weight, ALogP, and number of hydrogen bond donors and acceptors. The publicly available data represented mainly the drug-like space, following Lipinski's rule of five, likely resulting from the fact that hSGLT2 is a drug target investigated by pharmaceutical companies.[26] Moreover, the public data mostly includes glycoside-like compounds and oxopyrrolidine-carboxamides. In contrast, the in-house dataset encompasses more diverse molecules and captures a wider value range for the physicochemical properties mentioned above. The molecular weight and ALogP are represented in **Figure 3.1B**, where it is observed that these properties are more conserved for the public dataset than for the in-house dataset. Additionally, the number of hydrogen bond donors and acceptors is more widely captured by the in-house dataset (mean and standard deviation): public dataset hydrogen bond donor 3.6 ±1.6, hydrogen bond acceptor 6.3 ±1.8; in-house dataset hydrogen bond donor 2.0 ±2.6, hydrogen bond acceptor 5.1 ±4.1. When screening for compounds to target hSGLT1 in the intestine, it is favorable to consider compounds that do not necessarily adhere to Lipinski's rule of five, as it is preferred to minimize compound absorption from the gastrointestinal tract. Therefore, the in-house dataset contributes substantially to the applicability domain and predictive power of the statistical SGLT model.

Merging different datasets

To merge the public and in-house dataset the difference in activity units for both sets had to be resolved. The public dataset contains pChEMBL values, representing a standardized unit for affinity and potency values such as K_i, IC_{50}, EC_{50}, and K_d.[27] The potency values in the in-house dataset were available as percentage activity compared to (negative) control at a concentration of 50 μM, which could not be converted into a pChEMBL value. Hence, binary classification models were chosen over regression.

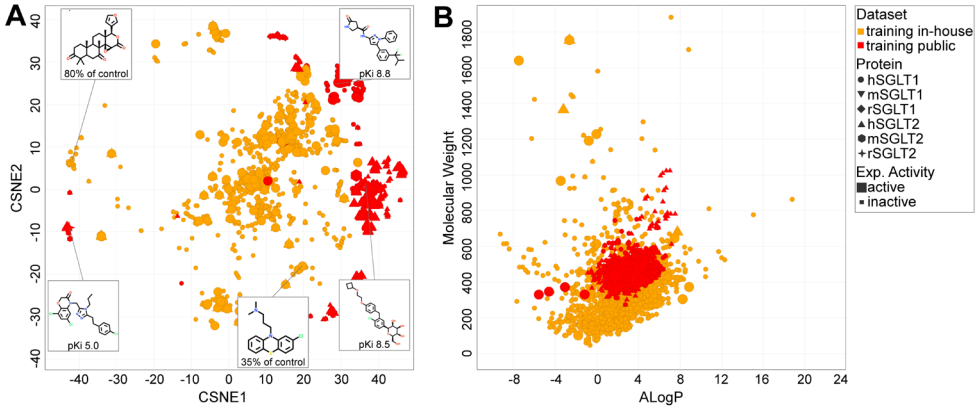


Figure 3.1. Chemical space of the public and in-house datasets. A) the t-SNE shows molecular structure and affinity (pK_i for public data and % of (negative) control for in-house data) for representative hSGLT1 compounds. B) molecular weight and ALogP distribution of compounds in the training sets.

Thresholds for compounds being 'active' were determined by grid searching cut-off values for both the public and in-house data. Activity thresholds along the grid were reviewed using hSGLT1 QSARs and external validation with a hold-out test set containing 30% of the in-house hSGLT1 data. The public domain compounds, which are mostly glycoside-like compounds and oxopyrrolidine-carboxamides, only describe a very conserved and small chemical space. However, the molecules of interest belonged to the same chemical space as the more diverse in-house compounds and therefore only compounds from the in-house set were used in validation. The activity threshold grid search showed that an activity threshold optimum for the in-house data was found at activity percentage < 70%, < 75%, and < 80% together with the threshold for public data set at pChEMBL > 8.5 (**Figure 3.2**). In further models the activity threshold was set at activity < 70% for in-house data and pChEMBL > 8.5 for public data to achieve the best performance for predicting hSGLT1 active molecules in the chemical space of the in-house compounds. Although these activity thresholds are not similar toward each other (e.g. pChEMBL > 8.5 corresponds to an in-house threshold much lower than 70%), these thresholds were determined optimal for the aim, which is the identification of novel (weak) actives that are similar in chemical space as the in-house compounds. The performance of the QSAR benchmark model using the selected thresholds was: sensitivity 0.76, specificity 0.86, positive predictive value (PPV) 0.42, negative predictive value (NPV) 0.96, and Matthews correlation coefficient (MCC) 0.48.

Proteochemometric modeling of hSGLT1

A PCM model was constructed using only public data to predict the inhibitory activity of compounds for hSGLT1. The performance of the model was tested on in-house data as these compounds represented the chemical space of interest. The model was validated using five test sets composed from in-house hSGLT1 data (5x20%). The mean performance of the public data model was very poor (mean with standard deviation): sensitivity 0.01 ±0.01, specificity 0.98 ±0.00, PPV 0.03 ±0.06, NPV 0.91 ±0.01, and MCC -0.03 ±0.03 (**Table 3.1**). Apparently, with the public data alone it was impossible to identify active compounds and the process defaulted to classification of all compounds as 'inactive'. This behavior confirms the large differences in

chemical space between the two sets as alluded to above.

Next, a PCM model was constructed based on the combined full data set consisting of all public and in-house data. To validate the performance of this model, 5-fold cross-validation was applied with the same test sets as applied in validation of performance of the public data model: rotationally 20% of the in-house hSGLT1 data was used as holdout test set; the remaining 80% was used in training. In each case the test set contained compounds not available for training. This resulted in the following performance: sensitivity 0.64 ±0.06, specificity 0.93 ±0.01, PPV 0.47 ±0.07, NPV 0.96 ±0.01, and MCC 0.49 ±0.05. Overall performance of this PCM model was regarded satisfactory for predictions of new compounds and was comparable with the QSAR benchmark model used for activity threshold determination.

Additionally the performance of models trained on in-house data only was tested to assess the effect of addition of public data. Public domain compounds contributed slightly to the predictive performance of the model in specificity, PPV, and MCC. This was observed by a minor decrease in performance upon removal of the public data from the training set: sensitivity 0.69 ±0.07, specificity 0.89 ±0.02, PPV 0.38 ±0.06, NPV 0.97 ±0.01, and MCC 0.45 ±0.05. Although the difference in performances is not significant, it is remarkable that the number of false positives decreases considerably when public data is included in training, whereas the number of true positives is only slightly negatively affected: false positives 28 ±6 vs. 43 ±6, true positives 24 ±4 vs. 26 ±4 (with and without public data, respectively). Apparently, the public data by itself is not sufficient in predicting hSGLT1 activity in the chemical space of the in-house compounds but does add value to the model when supplemented to the in-house dataset.

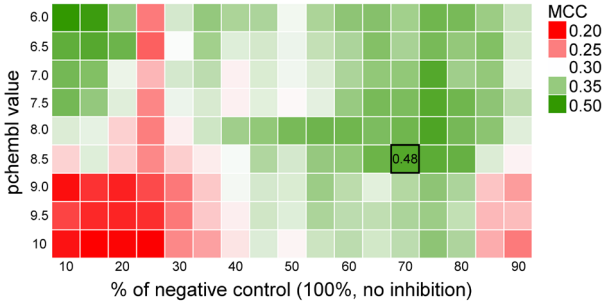


Figure 3.2. Activity threshold grid search. Searching the activity threshold grid for in-house (activity percentage compared to negative control) and public data (pChEMBL value). Model performance was measured using Matthews Correlation Coefficient (MCC), which was 0.48 for the final selected thresholds of < 70% for in-house data and pChEMBL > 8.5 for public data.

Table 3.1. Model performance depends on datasets that are used in training.

Model and validation	Training	Sensitivity	Specificity	PPV	NPV	MCC
QSAR (EV)	PD + IH	0.76	0.86	0.42	0.96	0.48
Public PCM (CV)	PD	0.01 ±0.01	0.98 ±0.00	0.03 ±0.06	0.91 ±0.01	-0.03 ±0.03
In-house PCM (CV)	IH	0.69 ±0.07	0.89 ±0.02	0.38 ±0.06	0.97 ±0.01	0.45 ±0.05
Combined PCM (CV)	PD + IH	0.64 ±0.06	0.93 ±0.01	0.47 ±0.07	0.96 ±0.01	0.49 ±0.05

Abbreviations: PD - public data, IH - in-house data, EV - external validation on 30% of data, CV - 5-fold cross validation on 20% of the data per iteration.

Screening for hSGLT1 actives in a commercially available compound library

The SGLT PCM model that was trained on public and in-house data was applied to a commercially available library. This library, the Enamine high-throughput screening (HTS) library, contains over 1.8 million compounds.[28] The library covers a wide diversity regarding molecular weight and ALogP values, and encompasses a vast chemical space (**Figure 3.3**). With the PCM model, an hSGLT1 activity prediction was assigned to all 1,815,674 compounds in the library (model training time was 103 seconds; the screening speed was approximately 132 seconds for 10,000 compounds). 155,275 compounds were predicted to be in the active class based on a predicted class probability of ≥ 0.5 (score, average of the trees in agreement with the active class).

To increase confidence in the activity of compounds the screened set was pre-filtered by selecting compounds with a predicted class probability of ≥ 0.8 on a scale from 0 to 1. Here, a resulting score of 1 represents compounds predicted to be in the 'active' class, a score of 0 indicates that the compounds are predicted 'inactive'; ascending scores indicate higher certainty of compounds belonging to the 'active' class. Additionally, compounds with molecular weight ≤ 300 were removed to exclude fragment-like compounds. The final filtered set contained 672 compounds.

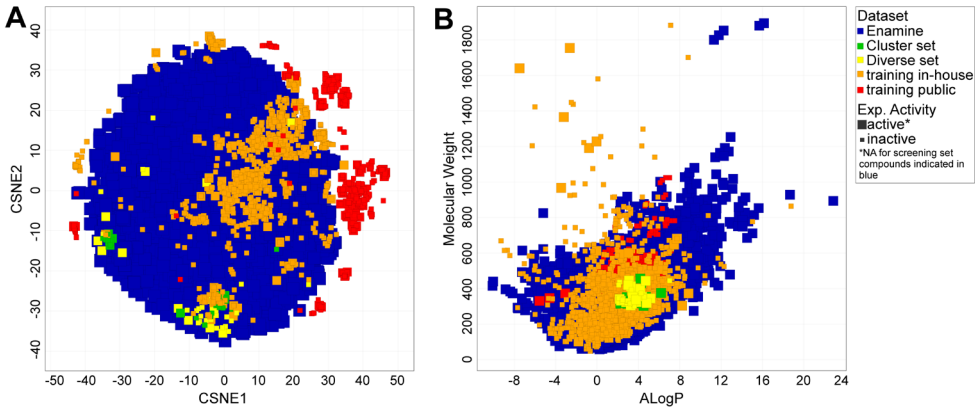


Figure 3.3. Chemical space of the selected compounds compared to the training and screening datasets. A) the Diverse set (yellow) and Cluster set (green) are displayed compared to the training (orange and red) and Enamine screening set (blue). The Enamine set is represented by a random selection of 20,000 out of the total of 1,815,674 compounds (~1%) in the screening set to limit t-SNE calculation time. B) the molecular weight and ALogP of the Diverse and Cluster set compared to the training and screening sets.

Based on the model predictions, 40 chemically diverse compounds predicted to be active were selected for experimental in vitro validation ('Diverse set'). The compounds in this set were cluster centers resulting from clustering of the remaining predicted active compounds into 40 clusters. This Diverse set was selected to increase the probability of detecting chemically novel hSGLT1 inhibitors. The selected compounds distributed widely through chemical space (**Figure 3.3** and **Supplementary Figure S3.1**), thus providing a challenging test for the SGLT PCM model. In addition to screening for novel hSGLT1 inhibitors, compounds were selected to expand the SAR around some recently identified hSGLT1 inhibitors from the in-house dataset.[23] Based on four hSGLT1 inhibitors (**Figure 3.4**) 3x10 additional compounds were selected from the pre-filtered Enamine HTS set that were predicted to be active (with top ranking scores) and that

resembled bepridil, bupivacaine, and cloperastine and 7 compounds were selected resembling trihexyphenidyl ('Cluster set'). These compounds were selected based on both model prediction (predicted class probability ≥ 0.8) and the highest similarity (Tanimoto, FCFP6) towards their known reference compound.

The total selection of 77 unique compounds was tested in vitro in cells expressing hSGLT1 in a single point measurement at a concentration of 50 μM. From the 40 diverse predicted hits that were assessed, 15 compounds were defined active as they displayed hSGLT1 inhibition in vitro with an activity reaching values below 70% compared to the negative control (100%, no inhibition) (**Supplementary Table S3.2**). From the 37 Cluster set compounds, an additional 15 compounds were confirmed to be active (**Supplementary Table S3.3**).

Cytotoxicity of hSGLT1 actives

The potential cytotoxicity of the screening compounds (Diverse set and Cluster set) was investigated by analysis of secreted adenylate kinase (AK), a marker of cell wall integrity loss. Most compounds did not show any indication of cyotoxicity, however one active from the Diverse set displayed moderate impairment of the cell wall (Z1416510792: activity 43 ±9%, cytotoxicity 25%). The cytotoxicity assay was limited by the available supernatant from the activity screen. Therefore not all compounds were measured in duplicate and cytotoxicity of one active from the Cluster set could not be determined (Z817504494: activity 45 ±3%).

Compound activity for hSGLT2

Both the Diverse set and Cluster set compounds were additionally measured for hSGLT2 inhibitory activity to assess their selectivity between the two transporters. The same cellular screening assay was performed as was used for hSGLT1 (single point measurement at a concentration of 50 μM). More actives were defined for hSGLT2 compared to hSGLT1 using the same activity threshold of 70% activity relative to negative control (100%, no inhibition): 22 actives in the Diverse set and 19 in the Cluster set. Almost all hSGLT1 actives showed activity for hSGLT2 with the possible exception of Z105569118, which only marginally surpassed the activity threshold for hSGLT2 (activity of hSGLT1 64 ±4% and hSGLT2 76 ±5%). No selective compounds were identified for hSGLT1, with 14% being the highest observed difference in inhibition (Z46160496: hSGLT1 41 ±4% and hSGLT2 55 ±2%). For hSGLT2 the biggest difference in inhibition was found for Z1318177320 that showed a difference of 39% (hSGLT1 93 ±20% and hSGLT2 54 ±0%).

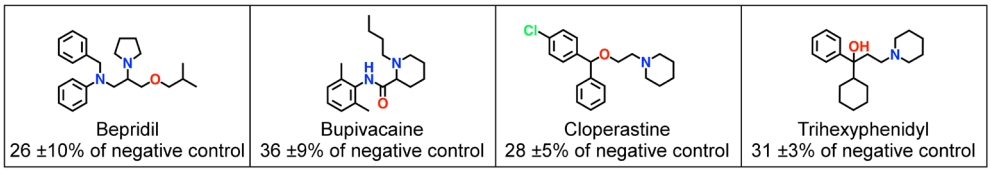


Figure 3.4. Reference hSGLT1 inhibitors for Cluster set and their inhibitory activity. Inhibitory activities (compared to negative control, where 100% is no inhibition) and chemical structures of four recently identified novel hSGLT1 inhibitors: bepridil, bupivacaine, cloperastine, and trihexyphenidyl.

Hit compound analysis

The activities of the hit compounds of the Diverse and Cluster set were analyzed. The strongest inhibitors, Z163972344 and Z915954934, were derived from the Diverse set with activities of 24 ±1% and 28 ±4% (100%: no inhibition), respectively. Z163972344 has low similarity (0.27 based on Tanimoto FCFP6) with the training set, indicating that this is a truly novel inhibitor for hSGLT1. The average similarity of actives in the Diverse set compared to training was 0.33, with Z1416510792 being the active that is most similar to the compounds in the training set with a similarity score of 0.61 (this compound showed moderate AK secretion in the cytotoxicity assay).

For the Cluster set a total of 15 actives were validated for the four different clusters. The cloperastine cluster encompassed the most actives (60% actives), whereas the trihexyphenidyl and bepridil clusters contained the least actives with 29% and 30% actives, respectively. The bupivacaine cluster had a hit rate of 40%, which is comparable with the overall hit rate of the total Cluster set (41%). The variance in hit rates between the four clusters is also reflected in the similarity of compounds toward their cluster reference: the cloperastine and bupivacaine clusters contained the most similar compounds (average similarities towards cluster reference compound were 0.43 and 0.42, respectively); the trihexyphenidyl and bepridil clusters contained less similar compounds (0.35 and 0.31, respectively).

Although the cloperastine and bupivacaine clusters contained the most similar cluster members, no conclusive SAR could be determined. The cluster members displayed variations in methyl substituents, which showed an effect for two compounds in the bupivacaine cluster (Z46224544 (45 ±10%) and Z2217101732 (74 ±8%)). This was however not observed for compounds in the cloperastine cluster: Z31367782 (36 ±4%), Z31371621 (37 ±3%), Z31367784 (43 ±7%), and Z31370217 (45 ±10%). The positions of the methyl substituents were too distinct to make solid conclusions on their relationship with compound activity.

In general, the novel active entities contain at least one aromatic ring and two hydrogen bond acceptors. Only two of the 30 actives did not adhere to Lipinski's rule of five, with an ALogP of 5.2 and 6.2 for Z1844922248 (activity 49 ±7%) and Z56906862 (activity 38 ±5%), respectively.

Aiming for specific targeting at the gastrointestinal tract

AAs mentioned in the Introduction, hSGLT1 inhibition at the intestinal wall is desired. Based on chemical structure and physicochemical properties the identified hit compounds will most likely be absorbed. However, it is suggested that modifications can be introduced to improve specific intestinal targeting. These alterations, such as a higher molecular weight, can prevent compounds from being absorbed or transported by the intestinal wall.[29] Intestinal SGLT1 blockers are expected to display less renal damage, which is an adverse effect observed for SGLT2 inhibitors.[6] Moreover, drug action restricted to the gastrointestinal tract also limits other off-target interactions, which were observed for the marketed SGLT2 inhibitor canagliflozin.[30] An example of a compound that was optimized for specific targeting at the gastrointestinal tract is LX2761, an inhibitor aimed at intestinal SGLT1 that decreased glucose uptake in mice.[31,32] Although SGLT1 inhibition at the intestine may not comprise renal function, other adverse effects that can result from intestinal targeting need to be considered.[33,34]

Indications for alternate binding modes

Upon examination of our hSGLT1 actives, a large variety in chemical structure and physicochemical properties was observed. This indicates that different ligand types may bind to different sites on hSGLT1. It is speculated that the glycoside-like hSGLT1 inhibitors, which are represented well in the public compound domain, bind to the glucose binding site, whereas more chemically diverse hSGLT1 inhibitors are suggested to bind either there or elsewhere on the protein. The hSGLT1 actives were grouped into ten clusters. Here, the activity threshold for compounds from the public dataset was pChEMBL ≥ 6.5 to include all actives instead of only strong binders (pChEMBL > 8.5, which gave the best model performance). It was observed that the glycoside-like compounds cluster together in cluster 2 (**Figure 3.5**). Furthermore, the oxopyrrolidine-carboxamide compounds, which are also present in the public domain, are gathered in cluster 7. Cluster 4 mainly holds in-house compounds and includes the anti-histamine drug moxastine and antidepressant amitriptyline besides cloperastine. The differences in chemical structure, molecular weight, and ALogP of the clusters substantiate the possible existence of multiple binding sites. As a further example, cluster 6 differs considerably in ALogP from the other clusters. This suggests that the compounds in this cluster bind to a more hydrophilic site. The cluster centers and distribution of molecular weight, ALogP, number of hydrogen bond donors, and number of hydrogen bond acceptors for all clusters are shown in **Figure 3.6**. Additional pharmacological experiments, beyond the scope of this study, are warranted to further investigate the existence of multiple binding pockets in SGLT1. Attempts have been made to explore the binding sites of SGLT1 for substrates and inhibitor phloridzin.[35,36] Although the SGLT structure of vibrio parahaemolyticus has been used to generate hypotheses on SGLT1 binding pockets, the lack of an hSGLT1 structure hampers the detection of potential allosteric binding pockets.[37]

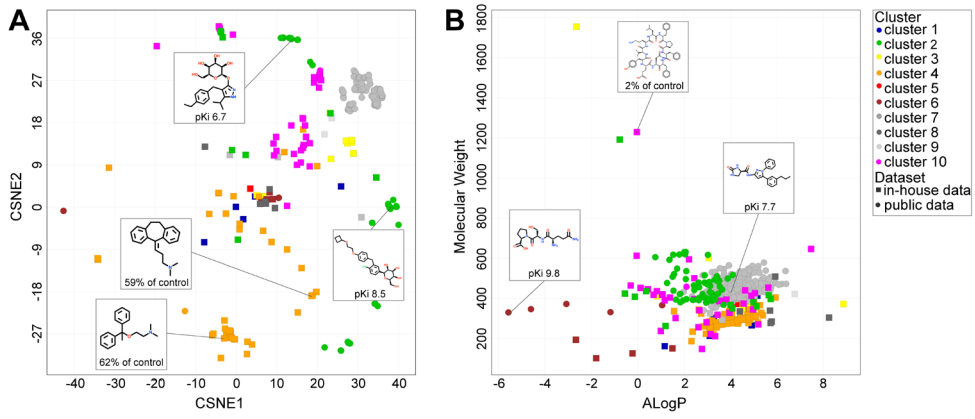


Figure 3.5. Clustering of hSGLT1 actives. Active hSGLT1 compounds in the training set clustered into ten chemical clusters (Tanimoto, FCFP6). Molecular structure and affinity (pKi for public data and % of (negative) control for in-house data) for representative cluster compounds are shown. In-house compounds with activity < 70% of (negative) control and public compounds with pChEMBL ≥ 6.5 were used in clustering. A) t-SNE plot of the chemical clusters. B) the molecular weight and ALogP distribution of compounds in the chemical clusters.

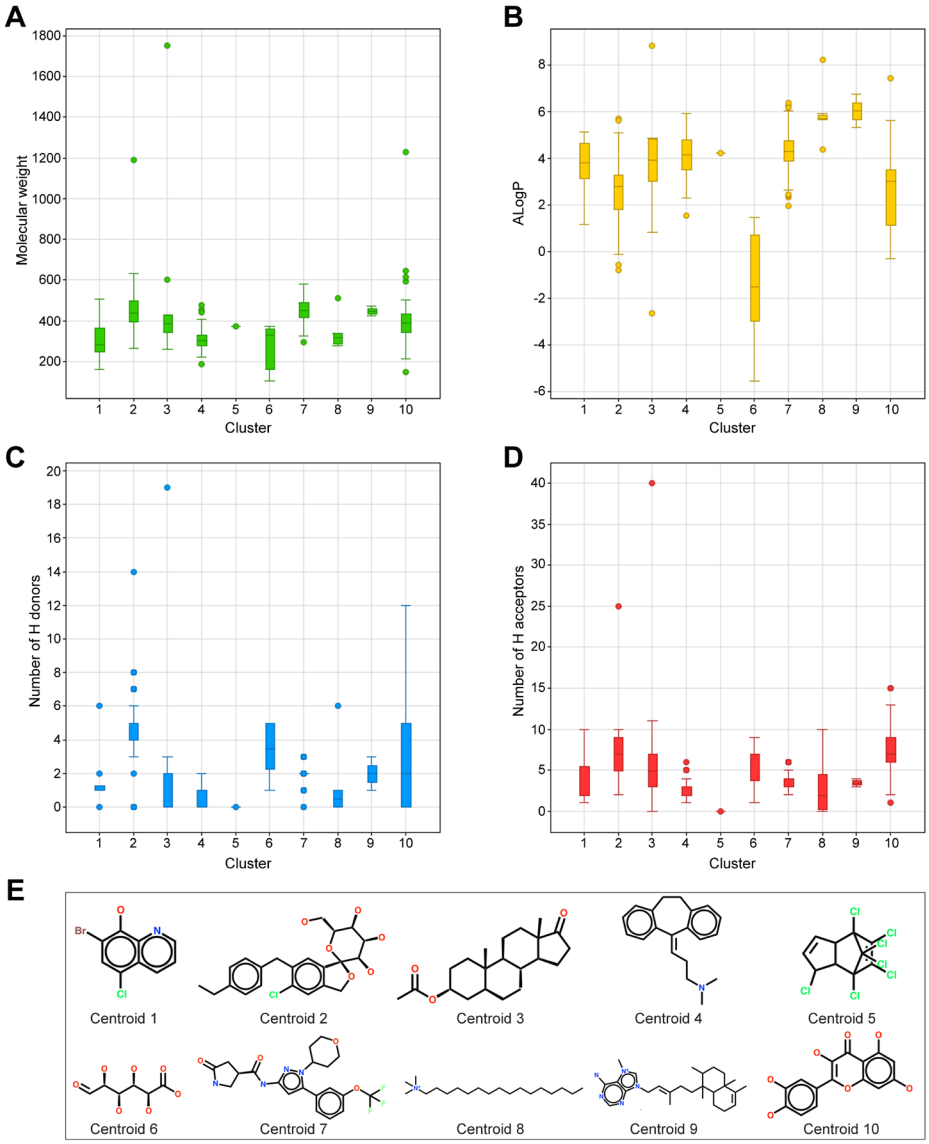


Figure 3.6. Cluster centers and distribution of physicochemical properties of hSGLT1 active compound clusters. Physicochemical properties of the ten clusters containing active hSGLT1 compounds. A) molecular weight. B) ALogP. C) number of hydrogen bond donors. D) number of hydrogen bond acceptors. E) molecular structures of the cluster centers.

Conclusion

PCM modeling appears a viable method to identify novel inhibitors for solute carrier hSGLT1 and likely any solute carrier protein. A predictive SGLT model was built with a MCC value of 0.49 ±0.05, estimated with 5-fold cross-validation. With the optimized model a hit rate of 38% was achieved when it was applied to screen for diverse molecules (Diverse set). In parallel, the model was used to boost identification of actives with a given chemotype (Cluster set). Although additional active compounds were identified, the data was too ambiguous to gain insight into the SAR of hSGLT1 inhibitors.

Diversity was found within the in-house dataset and differences were observed with the chemical space of the public dataset. Furthermore, the intrinsic variety in chemical structure of active compounds implies that there may be multiple binding sites at the transporter protein.

The novel identified inhibitors showed low similarity towards the training set and belong to the same chemical space of the in-house dataset, in contrast to the public dataset. Although the inhibitors were not optimized for specific drug delivery to the gastrointestinal tract, it is suggested that alterations (such as an increase in molecular weight and size) can make these inhibitors selective for intestinal hSGLT1.

Methods

Compounds and assay materials

DMEM-F12 (Biowest, Cat. No. L0092-500), DMEM (Lonza, BE12-604F/U1), Heat Inactivated Foetal Bovine Serum (HI-FBS, Biowest, Cat. No. S181H-500) and HBSS without Ca and Mg (HyClone, Cat. No. SH30588.01), DPBS (HyClone, Cat. No. SH30028.02), isopropanol (20842.312), clear-bottom black 96 well plates (Greiner, Cat. No. 655090) and polypropylene 96-well plates (Nunc, Cat. No. 151193) were all obtained from VWR (Amsterdam, the Netherlands). TrypLE Express (Gibco, Cat. No. 12605010), geneticin (Gibco, Cat. No. 10131027), D-glucose free DMEM (Gibco, Cat. No. 11966025), water soluble probenecid (Invitrogen, Cat. No. P36400), 5000 U/mL penicillin-streptomycin (Gibco, Cat. No. 15070063) were all ordered from Thermo Fisher Scientific (Breda, the Netherlands). 1-NBD-Glucose was custom synthesized by Mercachem (Nijmegen, the Netherlands). Bovine serum albumin (Cat. No. A8806), poly-L-lysine hydrobromide mol. wt. 30,000-70,000 (Cat. No. P2636), cell culture grade DMSO (Cat. No. D2650) were all acquired from Sigma-Aldrich Chemie (Zwijndrecht, the Netherlands). The hSGLT1 cDNA cloned in the pCMV6-neo vector was purchased from Origene Technologies (Rockville, USA, Cat. No. SC119918). The hSGLT2 cDNA was custom synthesized and cloned into the pcDNA3.1 vector by Thermo Fisher Scientific (Breda, the Netherlands). The experimentally tested Enamine screening compounds were acquired from Enamine (Kyiv, Ukraine).

Assay procedure

Two days in advance, CHO-hSGLT1 or CHO-hSGLT2 cells were seeded in maintenance medium (DMEM-F12 supplemented with 10% HI-FBS and 400 µg/mL geneticin) at 60.000 cells/well in clear-bottom black 96 well plates, pre-coated with 100 µg/mL poly-lysine. Cells were washed with 240 µL/well D-glucose free DMEM. Dilutions of test compounds and controls prepared

in D-glucose free DMEM with 350 μM 1-NBD-Glucose, 0.3% BSA, and 2 mM probenecid were added at 90 μL/well and placed in a humidified incubator at 37 °C with 5% CO_2 for 30 min. Subsequently cells were washed once with ice-cold DMEM-F12 and once with ice-cold HBSS, both at 240 μL/well. Finally, 1-NBD-Glucose was extracted from the cells with 100 μL/well isopropanol for 10 min at 600 rpm on an orbital shaker. Fluorescence was measured on a Flexstation 3 (Molecular Devices, San Jose, USA) with excitation at 445 nm, emission at 525 nm and cut off 515 nm. The uptake of 1-NBD-Glucose was normalized to the dynamic range between minimal inhibition (0.2% DMSO vehicle control) and maximal inhibition (100 μM phloridzin, >100x SGLT1/2 IC_{50}). Phloridzin is a strong inhibitor of SGLT1 and SGLT2 and was used as 0% reference, with 100% being no inhibition. A concentration of 100 μM phloridzin was used to ensure full SGLT1/2 inhibition. The Z-factor for the controls was determined and only data with Z > 0.4 (average Z SGLT1 assays: 0.8±0.1, average Z SGLT2 assays: 0.6±0.1) was used.[38]

Cytotoxicity assay

The cytotoxicity of compounds was tested with the ToxiLight bioassay kit (Lonza, obtained from VWR, Amsterdam, The Netherlands) according to the supplier's instructions. This non-destructive assay measures leakage of the enzyme AK from damaged cells into the CHO-hSGLT1/2 inhibition assay media, i.e. the degree of cytolysis. AK converts ADP into ATP and the enzyme luciferase subsequently catalyzes the formation of light from ATP and luciferin. Briefly, 20 mL of CHO-SGLT1/2 inhibition assay medium was added to 100 mL reconstituted AK detection reagent in white 96 wells Cellstar plates (Greiner bio-one, obtained from VWR, Amsterdam, The Netherlands) and incubated for 5 min at room temperature. Next, bioluminescence was measured on a FlexStation 3 Multi-Mode Microplate Reader (Molecular Devices, San Jose, USA) by 1 second integrated reading. Cytotoxicity was expressed as the percentage of bioluminescence of the 0.5% DMSO vehicle control which was set at 0%. The average cytotoxicity was calculated from biological replicates as indicated and average values > 20% were considered toxic (arbitrary threshold).

Dataset

Publicly available data from ChEMBL (version 23) was extracted for human SGLT1 (accession: P13866), human SGLT2 (P31639), and related proteins human SGLT3 (Q9NY91), rat SGLT1 (P53790), rat SGLT2 (P53792), mouse SGLT1 (Q9QXI6), mouse SGLT2 (Q923I7), and mouse SGLT3 (Q8R479). The retrieved compounds were standardized by removing salts, keeping the largest fragment, standardizing stereoisomers, standardizing charges, deprotonating bases, protonating acids, and optimizing the 2D structure by correcting bond lengths and angles. Activity values with confidence score 7 and 9 were kept and duplicate activity values were discarded based on activity standard unit ranking: K_i>IC_{50}>EC_{50}>K_d. For duplicate compounds with similar activity standard units (e.g. a compound with two Ki values), the average pChEMBL value was calculated.

An additional in-house dataset was provided by Unilever, Vlaardingen.[23] This dataset was based on the Spectrum Collection compound library (MicroSource Discovery Systems) extended with additional compounds that were similar to primary bioassay screening hits. This dataset consisted of compound activity data for hSGLT1 and hSGLT2. The activity was expressed as percentage

1-NBD-Glucose uptake compared to control at 50 μM, with control being the absence of inhibitor (=100%). Molecular structures were standardized in the same manner as the public data. The final dataset (public and in-house datasets combined, no duplicates) encompassed 3,686 unique compounds with 4,208 derived activities, of which 2,888 for hSGLT1.

Compound descriptors

Compounds were described using 512 FCFP6 fingerprint bits and the following physicochemical properties: molecular weight, ALogP, number of hydrogen bond acceptors, number of hydrogen bond donors, number of rotatable bonds, number of bridge bonds, and number of aromatic rings. Fingerprints and physicochemical descriptors were calculated in Pipeline Pilot (version 16.1.0).[39]

Protein descriptors

Protein sequences were aligned using whole sequence alignment in Clustal Omega (version 1.2.2).[40] Subsequently the sequences were converted to protein descriptors using Z-scales.[41] The first three Z-scales were implemented as protein descriptor as these were shown to perform well in previous work.[42] These three Z-scales include information on residue lipophilicity, size, and polarity.

Machine learning

Models were trained using the Random Forest R component in Pipeline Pilot (version 16.1.0). The number of trees was 500 and number of variables tried at each split was 38 (square root of the number of descriptors). Remaining settings were kept default.

T-Distributed Stochastic Neighbor Embedding

T-SNE was calculated on FCFP6 fingerprint descriptors that were converted to 2024 bits. The t-SNE component in Pipeline Pilot (version 18.1.0) was used to perform tSNE. The derived t-SNE values are represented by two components: CSNE1 and CSNE2.

Clustering of hSGLT1 actives to explore binding modes

hSGLT1 active compounds in the training set were clustered into ten clusters using the cluster molecules component in Pipeline Pilot (version 16.1.0). Compounds from the in-house set were included as 'active' when percentage of (negative) control was < 70%. Compounds from the public data set were termed 'active' when pChEMBL value ≥ 6.5.

Computational Hardware

Experiments were performed on a server running CentOS 6.9 equipped with a dual Xeon E-5 2630 v2 processor and 128 GB of RAM.

References

(1) Wood, I. S.; Trayhurn, P. Glucose Transporters (GLUT and SGLT): Expanded Families of Sugar Transport Proteins. 2003. https://doi.org/10.1079/BJN2002763.
(2) Tsujihara, K.; Hongu, M.; Saito, K.; Kawanishi, H.; Kuriyama, K.; Matsumoto, M.; Oku, A.; Ueta, K.; Tsuda, M.; Saito, A. Na+-Glucose Cotransporter (SGLT) Inhibitors as Antidiabetic Agents. 4. Synthesis and Pharmacological Properties of 4'-Dehydroxyphlorizin Derivatives Substituted on the B Ring. J. Med. Chem. 1999, 42, 5311–5324. https://doi.org/10.1021/jm990175n.
(3) Wright, E. M.; Loo, D. D.; Hirayama, B. A. Biology of Human Sodium Glucose Transporters. Physiol. Rev. 2011, 91, 733–794.
(4) Gorboulev, V.; Schürmann, A.; Vallon, V.; Kipp, H.; Jaschke, A.; Klessen, D.; Friedrich, A.; Scherneck, S.; Rieg, T.; Cunard, R.; et al. Na(+)-D-Glucose Cotransporter SGLT1 Is Pivotal for Intestinal Glucose Absorption and Glucose-Dependent Incretin Secretion. Diabetes 2012, 61, 187–196. https://doi.org/10.2337/db11-1029.
(5) Clar, C.; Gill, J. A.; Court, R.; Waugh, N. Systematic Review of SGLT2 Receptor Inhibitors in Dual or Triple Therapy in Type 2 Diabetes. BMJ Open 2012, 2, e001-007. https://doi.org/10.1136/bmjopen-2012-001007.
(6) Rosenstock, J.; Seman, L. J.; Jelaska, A.; Hantel, S.; Pinnetti, S.; Hach, T.; Woerle, H. J. Efficacy and Safety of Empagliflozin, a Sodium Glucose Cotransporter 2 (SGLT2) Inhibitor, as Add-on to Metformin in Type 2 Diabetes with Mild Hyperglycaemia. Diabetes, Obes. Metab. 2013, 15, 1154–1160. https://doi.org/10.1111/dom.12185.
(7) Komoroski, B.; Vachharajani, N.; Feng, Y.; Li, L.; Kornhauser, D.; Pfister, M. Dapagliflozin, a Novel, Selective SGLT2 Inhibitor, Improved Glycemic Control over 2 Weeks in Patients with Type 2 Diabetes Mellitus. Clin. Pharmacol. Ther. 2009, 85, 513–519. https://doi.org/10.1038/clpt.2008.250.
(8) Sands, A. T.; Zambrowicz, B. P.; Rosenstock, J.; Lapuerta, P.; Bode, B. W.; Garg, S. K.; Buse, J. B.; Banks, P.; Heptulla, R.; Rendell, M.; Cefalu, W. T.; Strumph, P. Sotagliflozin, a Dual SGLT1 and SGLT2 Inhibitor, as Adjunct Therapy to Insulin in Type 1 Diabetes. Diabetes Care 2015, 38, 1181–1188. https://doi.org/10.2337/dc14-2806.
(9) Rendell, M. S. Efficacy and Safety of Sotagliflozin in Treating Diabetes Type 1. Expert Opin. Pharmacother. 2018, 19, 307–315. https://doi.org/10.1080/14656566.2017.1414801.
(10) Masumoto, S.; Akimoto, Y.; Oike, H.; Kobori, M. Dietary Phloridzin Reduces Blood Glucose Levels and Reverses SGLT1 Expression in the Small Intestine in Streptozotocin-Induced Diabetic Mice. J. Agric. Food Chem. 2009, 57, 4651–4656. https://doi.org/10.1021/jf9008197.
(11) Rieg, T.; Vallon, V. Development of SGLT1 and SGLT2 Inhibitors. Diabetologia 2018, 61, 2079–2086. https://doi.org/10.1007/s00125-018-4654-7.
(12) Spatola, L.; Finazzi, S.; Angelini, C.; Dauriz, M.; Badalamenti, S. SGLT1 and SGLT1 Inhibitors: A Role to Be Assessed in the Current Clinical Practice. Diabetes Ther. 2018, 9, 427–430. https://doi.org/10.1007/s13300-017-0342-8.
(13) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. Nucleic Acids Res. 2017, 45, D945–D954. https://doi.org/10.1093/nar/gkw1074.
(14) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL Bioactivity Database: An Update. Nucleic Acids Res. 2014, 42, D1083-90. https://doi.org/10.1093/nar/gkt1031.
(15) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. Nucleic Acids Res. 2012, 40, D1100–D1107. https://doi.org/10.1093/nar/gkr777.
(16) van Westen, G. J. P.; Wegner, J. K.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Proteochemometric Modeling as a Tool to Design Selective Compounds and for Extrapolating to Novel Targets. Med. Chem. Commun. 2011, 2, 16–30. https://doi.org/10.1039/C0MD00165A.
(17) Lin, L.; Yee, S. W.; Kim, R. B.; Giacomini, K. M. SLC Transporters as Therapeutic Targets: Emerging Opportunities. Nat. Rev. Drug Discov. 2015, 14, 543–560. https://doi.org/10.1038/nrd4626.
(18) De Bruyn, T.; van Westen, G. J. P.; IJzerman, A. P.; Stieger, B.; de Witte, P.; Augustijns, P. F.; Annaert, P. P. Structure-Based Identification of OATP1B1/3 Inhibitors. Mol. Pharmacol. 2013, 83, 1257–1267.
(19) van Westen, G. J. P.; van den Hoven, O. O.; van der Pijl, R.; Mulder-Krieger, T.; de Vries, H.; Wegner, J. K.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Identifying Novel Adenosine Receptor Ligands by Simultaneous Proteochemometric Modeling of Rat and Human Bioactivity Data. J. Med. Chem. 2012, 55, 7010–7020. https://doi.org/10.1021/jm3003069.
(20) Tresadern, G.; Trabanco, A. A.; Pérez-Benito, L.; Overington, J. P.; van Vlijmen, H. W. T.; van Westen, G. J. P. Identification of Allosteric Modulators of Metabotropic Glutamate 7 Receptor Using Proteochemometric Modeling. J. Chem. Inf. Model. 2017, 57, 2976–2985. https://doi.org/10.1021/acs.jcim.7b00338.

(21) van Westen, G. J. P.; Wegner, J. K.; Geluykens, P.; Kwanten, L.; Vereycken, I.; Peeters, A.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Which Compound to Select in Lead Optimization? Prospectively Validated Proteochemometric Models Guide Preclinical Development. PLoS One 2011, 6, e27518. https://doi.org/10.1371/journal.pone.0027518.
(22) Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR Applicabilty Domain Estimation by Projection of the Training Set Descriptor Space: A Review. Altern. Lab. Anim. 2005, 33, 445–459.
(23) Oranje, P.; Gouka, R.; Burggraaff, L.; Vermeer, M.; Chalet, C.; Duchateau, G.; van der Pijl, P.; Geldof, M.; de Roo, N.; Clauwaert, F.; Vanpaeschen, T.; Nicolaï, J.; de Bruyn, T.; Annaert, P.; IJzerman, A. P.; van Westen, G. J. P. Novel Natural and Synthetic Inhibitors of Solute Carriers SGLT1 and SGLT2. Pharmacol. Res. Perspect. 2019, 7, e00504. https://doi.org/10.1002/prp2.504.
(24) MicroSource Discovery Systems. Spectrum Collection http://www.msdiscovery.com/spectrum.html (accessed Nov 25, 2015).
(25) Van Der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. J. Mach. Learn. Res. 2008, 9, 2579–2605.
(26) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. Adv. Drug Deliv. Rev. 2001, 46, 3–26.
(27) Papadatos, G.; Gaulton, A.; Hersey, A.; Overington, J. P. Activity, Assay and Target Data Curation and Quality in the ChEMBL Database. J. Comput. Aided. Mol. Des. 2015, 29, 885–896. https://doi.org/10.1007/s10822-015-9860-5.
(28) Enamine. Enamine HTS Collection https://enamine.net (accessed Oct 1, 2017).
(29) Charmot, D. Non-Systemic Drugs: A Critical Review. Curr. Pharm. Des. 2012, 18, 1434–1445. https://doi.org/10.2174/138161212799504858.
(30) Secker, P. F.; Beneke, S.; Schlichenmaier, N.; Delp, J.; Gutbier, S.; Leist, M.; Dietrich, D. R. Canagliflozin Mediated Dual Inhibition of Mitochondrial Glutamate Dehydrogenase and Complex I: An off-Target Adverse Effect. Cell Death Dis. 2018, 9, 226. https://doi.org/10.1038/s41419-018-0273-y.
(31) Goodwin, N. C.; Ding, Z.-M.; Harrison, B. A.; Strobel, E. D.; Harris, A. L.; Smith, M.; Thompson, A. Y.; Xiong, W.; Mseeh, F.; Bruce, D. J.; et al. Discovery of LX2761, a Sodium-Dependent Glucose Cotransporter 1 (SGLT1) Inhibitor Restricted to the Intestinal Lumen, for the Treatment of Diabetes. J. Med. Chem. 2017, 60, 710–721. https://doi.org/10.1021/acs.jmedchem.6b01541.
(32) Powell, D. R.; Smith, M. G.; Doree, D. D.; Harris, A. L.; Greer, J.; DaCosta, C. M.; Thompson, A.; Jeter-Jones, S.; Xiong, W.; Carson, K. G.; et al. LX2761, a Sodium/Glucose Cotransporter 1 Inhibitor Restricted to the Intestine, Improves Glycemic Control in Mice. J. Pharmacol. Exp. Ther. 2017, 362, 85–97.
(33) Lehmann, A.; Hornby, P. J. Intestinal SGLT1 in Metabolic Health and Disease. Am. J. Physiol. Liver Physiol. 2016, 310, G887–G898. https://doi.org/10.1152/ajpgi.00068.2016.
(34) Poulsen, S. B.; Fenton, R. A.; Rieg, T. Sodium-Glucose Cotransport. Curr. Opin. Nephrol. Hypertens. 2015, 24, 463–469. https://doi.org/10.1097/MNH.0000000000000152.
(35) Lostao, M. P.; Hirayama, B. A.; Loo, D. D. F.; Wright, E. M. Phenylglucosides and the Na+/Glucose Cotransporter (SGLT1): Analysis of Interactions. J. Membr. Biol. 1994, 142, 161–170. https://doi.org/10.1007/BF00234938.
(36) Bisignano, P.; Ghezzi, C.; Jo, H.; Polizzi, N. F.; Althoff, T.; Kalyanaraman, C.; Friemann, R.; Jacobson, M. P.; Wright, E. M.; Grabe, M. Inhibitor Binding Mode and Allosteric Regulation of Na+-Glucose Symporters. Nat. Commun. 2018, 9, 5245. https://doi.org/10.1038/s41467-018-07700-1.
(37) Watanabe, A.; Choe, S.; Chaptal, V.; Rosenberg, J. M.; Wright, E. M.; Grabe, M.; Abramson, J. The Mechanism of Sodium and Substrate Release from the Binding Pocket of VSGLT. Nature 2010, 468, 988–991.
(38) Zhang, J.-H.; Chung, T. D. Y.; Oldenburg, K. R. A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. J. Biomol. Screen. 1999, 4, 67–73. https://doi.org/10.1177/108705719900400206.
(39) Dassault Systèmes BIOVIA. BIOVIA Pipeline Pilot. Dassault Systèmes: San Diego 2016.
(40) Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; Thompson, J. D.; Higgins, D. G. Fast, Scalable Generation of High-quality Protein Multiple Sequence Alignments Using Clustal Omega. Mol. Syst. Biol. 2011, 7, 539.
(41) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. J. Med. Chem. 1998, 41, 2481–2491. https://doi.org/10.1021/jm9700575.
(42) van Westen, G. J. P.; Swier, R. F.; Cortes-Ciriano, I.; Wegner, J. K.; Overington, J. P.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Benchmarking of Protein Descriptor Sets in Proteochemometric Modeling (Part 2): Modeling Performance of 13 Amino Acid Descriptor Sets. J. Cheminform. 2013, 5, 42. https://doi.org/10.1186/1758-2946-5-42.

.

Supporting Information

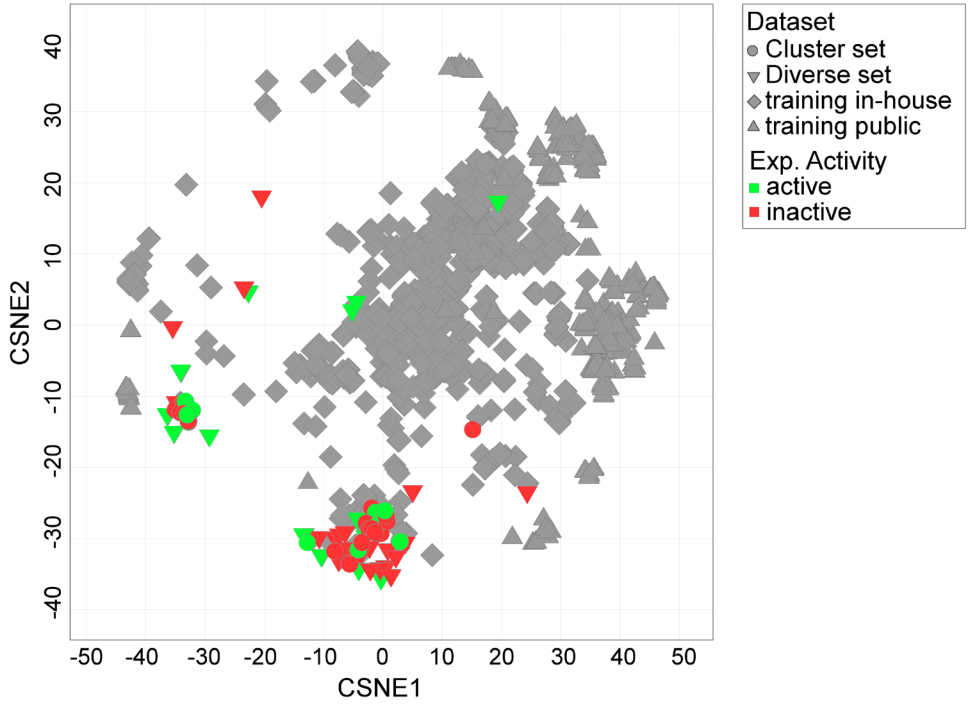


Figure S3.1. T-SNE representation of the chemical space of actives and inactives of selected compounds compared to the training set. Experimental activities of compounds in the Diverse set and Cluster set compared to the training set (grey).

Table S3.2. Bioactivities, cytotoxicities and Tanimoto similarities of the Diverse set.

Structure	Compound ID	SGLT1 activity (% of control)		Activity class	SGLT2 activity (% of control)		Similarity towards training (Tanimoto)	Toxicity (AK secretion)				
		Average	St.dev		Average	St.dev		Average	St.dev	n	n1	n2
CN(CCOc1c(C)cccc1C)Cc2ccccc2N3CCOCC3	Z163972344	24	1.1	active	29	5.5	0.271	-8.80	2.10	2	-10.3	-7.3
N\C(=N\OCc1ccccc1)\c2ccccc2CN3CCCCC3	Z915954934	28	3.7	active	42	1.7	0.431	2.24	10.38	2	9.6	-5.1
CN(C)CCCN1\C(=N/c2ccccc2)\SC=C1c3ccccc3	Z56792856	35	4.1	active	29	3.6	0.328	7.10	23.30	2	23.6	-9.4
CC(N(C)S(=O)(=O)N1CCCCCC1)c2cccc(c2)[N+](=O)[O-]	Z237884404	39	3.0	active	35	5.0	0.295	-14.70		1		
CN(C)CCN(Cc1occc1)c2ccc(Cl)c(n2)C#N	Z1416510792	43	9.2	active	45	15.6	0.614	25.30		1		
CC1CCCN(CCC(C#N)(c2ccccc2)c3ccccc3)C1	Z137703320	47	6.9	active	28	10.9	0.321	4.71	15.42	5		
Fc1ccc(CN(CCc2ccccn2)CC#Cc3ccccc3)cc1	Z1844922248	49	7.0	active	48	12.1	0.366	-9.20		1		
OC1(CC2CCC(C1)N2CCOC3CCCCO3)c4ccccc4	Z2677914007	51	3.1	active	50	6.1	0.329	8.10	14.50	2	18.4	-2.1
O=C(CN1CCC2CCCCC2C1)Nc3cccc(c3)S(=O)(=O)N4CCCCC4	Z46498483	56	3.8	active	27	5.8	0.269	8.70	18.60	2	21.9	-4.5
CCOC(=O)C1CCCCN1S(=O)(=O)c2ccc(cc2)C(C)(C)C	Z45484979	57	0.7	active	42	2.9	0.321	-0.02	9.05	5		
CC1=CSC(=O)N1CCC(=O)N2CCN(Cc3ccccc3)C4(CCCCC4)C2	Z1268385265	61	12.7	active	69	15.0	0.265	-3.49	5.79	2	0.6	-7.6
Fc1ccc(cc1Cl)C2CN(CCS(=O)(=O)c3ccccc3)CCO2	Z1070584838	62	6.1	active	61	4.7	0.343	3.59	10.50	4		
CN(Cc1ccccc1NC(=O)N2CCCC(C2)N3CCCC3)C4CCCCC4	Z594462480	63	5.6	active	54	5.2	0.312	2.39	10.75	4		
CN(CC(=O)N1CC(c2ccccc2)c3ccccc3C1)S(=O)(=O)c4ccc(cc4)C#N	Z289555820	63	18.6	active	60	2.1	0.257	-3.35	12.25	2	-12.0	5.3
Cc1c(C)n(C2CCCCC2)c(NC(=O)CN3CCCC3C(C#N)c4ccccc4)c1C#N	Z1301318288	65	8.6	active	56	3.2	0.261	10.55	2.19	2	12.1	9.0
CN(CCOC(c1ccccc1)c2ccccc2)S(=O)(=O)C	Z1642825375	72	4.4	inactive	67	15.7	0.743	14.10	14.70	2	24.5	3.7
O=S(=O)(N1CCC1c2ccccc2)c3cccc4N=S=Nc34	Z741018970	73	6.6	inactive	86	4.7	0.232	7.00	5.60	4		

Structure	Compound ID	SGLT1 activity (% of control)		Activity class	SGLT2 activity (% of control)		Similarity towards training (Tanimoto)	Toxicity (AK secretion)				
		Average	St.dev		Average	St.dev		Average	St.dev	n	n1	n2
CCS(=O)(=O)c1ccccc1C(=O)N2CCCN(CC3CCCCC3)CC2	Z783393074	76	13.0	inactive	77	6.6	0.297	10.80	12.12	5		
CN(CCC#N)Cc1c(C)nn(Cc2ccccc2Cl)c1C	Z103956902	76	4.6	inactive	74	7.1	0.258	-6.47	15.72	5		
CC1CC(CCN1Cc2ccccc2)N(C)Cc3cnns3	Z2161555222	76	3.0	inactive	80	4.7	0.316	-1.30	9.99	5		
CCC1CN2CCCC2CN1Cc3ccc(OC(F)F)cc3	Z1318121286	78	7.6	inactive	69	10.2	0.348	7.14	11.83	5		
CN(C1CCOC1)S(=O)(=O)c2ccccc2C(=O)c3ccccc3	Z1725346307	82	2.7	inactive	80	13.4	0.295	7.86	18.23	4		
CCC(CN1CCC(CC1)OCC2CCOC2)c3ccccc3	Z1661722958	83	5.2	inactive	78	13.2	0.322	0.30	12.95	5		
N#Cc1ccccc1c2ccc(CN3CCC4(CC3)OCCO4)cc2	Z368852630	83	5.4	inactive	67	15.4	0.345	-4.79	4.50	5		
CCN(CC)CCN1CN(Cc2ccccc2)c3nc4ccccc4n3C1	Z104342888	86	5.6	inactive	66	1.7	0.263	-3.10	12.27	5		
CCCCN(CC#Cc1ccc(cc1)C(F)(F)F)C2CCS(=O)(=O)C2	Z1332354136	86	7.6	inactive	67	7.3	0.286	-9.27	9.75	5		
CCCCC(N1CCCS1(=O)=O)C(=O)Nc2cc(C)c(F)cc2Cl	Z2151290710	88	2.9	inactive	93	8.9	0.397	2.00	13.77	4		
C(C1CCCN(Cc2cccc(c2)n3cccn3)C1)N4CCOCC4	Z1568339019	89	10.2	inactive	80	4.8	0.279	6.21	12.79	5		
FC(F)(F)C(N1CCN(CC2OCCO2)CC1)c3ccccc3	Z1612085883	89	1.9	inactive	84	11.7	0.259	7.06	21.91	4		
C(C1CCOCC1)N(Cc2ccccc2)Cc3cn4cccnc4n3	Z1176622491	90	7.3	inactive	93	10.1	0.250	-0.12	9.90	4		
N#CC1CCC(C1)N2CCC(CN3CCc4ccccc34)CC2	Z1912611670	90	11.4	inactive	64	10.5	0.286	-9.32	18.88	5		
C\C(=C\c1ccccc1)\C=N\N2CCN(Cc3ccccc3Cl)CC2	Z56796910	91	4.5	inactive	59	12.8	0.276	8.70		1		
N#Cc1c(CN2CCC(=CC2)c3ccccc3)cn4ccccc14	Z220381090	94	14.7	inactive	84	16.2	0.250	-0.96	10.32	5		
CC1(C)CC(N(Cc2ccccc2)C(=O)CCCS(=O)(=O)N3CCCC3)C(C)(C)O1	Z1609106419	95	1.0	inactive	89	12.8	0.271	8.73	10.20	4		
Cn1cc(CN2CCCN(CC2)C3Cc4ccccc4C3)cn1	Z1176394041	96	6.3	inactive	79	10.2	0.254	4.62	10.59	4		
CN1CCCC1CCSc2nnc(C)n2Cc3ccccc3	Z1526821488	97	10.5	inactive	82	3.8	0.338	9.66	21.34	4		

Structure	Compound ID	SGLT1 activity (% of control)		Activity class	SGLT2 activity (% of control)		Similarity towards training (Tanimoto)	Toxicity (AK secretion)				
		Average	St.dev		Average	St.dev		Average	St.dev	n	n1	n2
C(\C=C\c1ccccc1)N2CCN(CC2)C(c3ccccc3)c4ccc(cc4)c5ccccc5	Z1172208353	103	7.2	inactive	84	10.9	0.667	12.26	18.33	5		
O=C(CN1CCN(CC1)C(=O)\C(=C\c2ccccc2)\c3ccccc3)N4CCOCC4	Z73457041	114	5.6	inactive	88	15.2	0.273	-0.63	20.90	5		
S=C1N=C(NN1CN2CCCCC2)\C=C\c3ccccc3	Z238423340	115	10.5	inactive	107	9.5	0.288	-33.53	10.44	5		
Clc1nc2sccn2c1\C=N\N3CCN(CC3)C(c4ccccc4)c5ccccc5	Z49982582	122	6.6	inactive	109	13.6	0.286	15.01	18.31	5		

Table S3.3. Bioactivities, cytotoxicities and Tanimoto similarities of the Cluster set.

Structure	Compound ID	Reference compound	SGLT1 activity (% of control); n=2				Activity class	SGLT2 activity (% of control), n=2				Similarity towards training (Tanimoto)	Toxicity (AK secretion)		
			Average	St.dev	n1	n2		Average	St.dev	n1	n2		Average	St.dev	n
Clc1ccc(cc1)C(N2CCN(CCOCC(=O)N3CCCCC3)CC2)c4ccccc4	Z827380946	Cloperastine	31	11.0	22.8	38.9	active	30	5.5	25.9	33.6	0.370	-0.27	8.13	3
Cc1ccc(cc1)C(OCC(O)CN2CCCCC2)c3ccccc3	Z31367782	Cloperastine	36	4.4	33.3	39.6	active	35	0.5	34.4	35.1	0.492	-10.06	14.06	4
CC1CCCN(CC(O)COC(c2ccccc2)c3ccccc3)C1	Z31371621	Cloperastine	37	3.3	34.8	39.4	active	27	2.9	29.3	25.3	0.419	-5.72	8.22	3
CC(C)N(CC#CCC(O)(C1CCCC1)c2ccccc2)Cc3ccccc3	Z56906862	Trihexyphenidyl	38	5.4	41.6	34.0	active	25	5.0	21.5	28.5	0.441	-10.40		1
CC(CN1CCCCC1)OC(=O)C(c2ccccc2)c3ccccc3	Z204157928	Trihexyphenidyl	41	7.3	36.3	46.6	active	27	4.1	24.2	30.1	0.322	-18.49	10.29	4
CC(C)c1cccc(C(C)C)c1NC(=O)CN2CCCCCC2	Z46160496	Bupivacaine	41	4.3	43.7	37.7	active	55	2.1	56.6	53.6	0.414	-8.60		1
Cc1ccccc1C(OCC(O)CN2CCCCC2)c3ccccc3	Z31367784	Cloperastine	43	7.3	37.4	47.8	active	38	0.6	37.6	38.4	0.459	-14.87	17.28	3
CN(CC(O)COC(c1ccccc1)c2ccccc2)C3CCCCC3	Z31370217	Cloperastine	45	10.0	38.0	52.9	active	24	7.0	19.2	29.0	0.365	15.30	16.73	4
CC1CCCCN1CC(=O)Nc2ccccc2C(=O)c3ccccc3	Z46224544	Bupivacaine	45	9.6	38.2	51.8	active	20	0.1	20.1	20.3	0.403	-1.30		1
CC(C)CCSCC(=O)N1CC2CCCN2Cc3ccccc13	Z817504494	Bepridil	45	3.3	47.4	42.7	active	27	11.9	35.0	18.1	0.305			0
CN(CCC1CCCCC1)CC(=O)Nc2c(C)cccc2C	Z1003034804	Bupivacaine	55	1.0	54.5	55.9	active	41	1.3	41.8	40.1	0.426	1.67	8.02	4
CC(N1CCCCCCC1)C(=O)Nc2ccccc2C(F)(F)F	Z124096378	Bupivacaine	57	6.9	61.7	52.0	active	45	2.7	46.9	43.0	0.413	0.78	4.98	4

Structure	Compound ID	Reference compound	SGLT1 activity (% of control); n=2				Activity class	SGLT2 activity (% of control), n=2				Similarity towards training (Tanimoto)	Toxicity (AK secretion)		
			Average	St.dev	n1	n2		Average	St.dev	n1	n2		Average	St.dev	n
CC1CN(CCOC(c2ccccc2)c3ccccc3)CCC1=O	Z1751527676	Cloperastine	62	8.6	55.5	67.6	active	56	10.6	48.4	63.4	0.518	-15.40		1
CCOC(=O)C1CCCCN1CC(=O)N2CCCc3ccccc23	Z105569118	Bepridil	64	4.0	60.8	66.4	active	76	4.5	79.4	73.0	0.286	15.96	30.89	3
O=C(CN1CCCCCC1)N(Cc2ccccc2)c3ccccc3	Z199542950	Bepridil	65	0.4	65.7	65.1	active	53	4.2	49.6	55.6	0.394	-0.80		1
COCC(C)N1CCN(CCCOCc2ccccc2)CC1	Z1665888845	Bepridil	71	6.8	66.2	75.7	inactive	74	1.6	74.8	72.6	0.324	5.75	10.97	3
NC(=O)C1CCCCN1CC(=O)N(Cc2ccccc2)Cc3ccccc3	Z46442988	Bepridil	71	18.0	83.4	57.8	inactive	78	14.4	88.6	68.2	0.284	4.60		1
C(CN1CCC(CC1)OCc2ccccc2)Oc3ccccc3	Z196036526	Bepridil	72	1.9	70.7	73.4	inactive	47	0.7	46.1	47.1	0.301	4.82	14.47	3
CC(N1CCCCC1)C(=O)Nc2cc(C)ccc2C(=O)c3ccccc3	Z2217101732	Bupivacaine	74	8.1	67.8	79.3	inactive	43	7.2	47.9	37.7	0.403	-4.30		1
CC1CCCCN1CC(=O)Nc2ccccc2C(F)(F)F	Z46224455	Bupivacaine	75	2.4	77.1	73.8	inactive	47	11.5	55.3	39.0	0.422	-2.08	6.50	4
Cc1ccccc1NC(=O)CCN2CCCCC2C3OCCO3	Z1298879676	Bupivacaine	77	14.0	87.0	67.4	inactive	74	18.0	86.8	61.3	0.426	7.10		1
CCN(C)CC1CCN(C1)C(=O)CSC(c2ccccc2)c3ccc(Cl)cc3	Z1276034798	Cloperastine	78	3.1	80.0	75.7	inactive	55	5.9	59.4	51.0	0.435			0
CC(CCN1CC(C)C(C1)N2CCOCC2)c3ccccc3	Z1897146788	Bepridil	82	2.2	83.9	80.8	inactive	77	7.9	71.8	83.0	0.280	-6.70		1
CC1CCCCN1CC2CCCN(C2)C(=O)Nc3c(C)cccc3C	Z1230260964	Bupivacaine	83	1.7	81.3	83.7	inactive	82	12.7	91.3	73.3	0.449			0
Fc1ccccc1CCN(CCC#N)CCN2CCOCC2	Z1704430244	Cloperastine	84	8.4	90.0	78.1	inactive	87	10.3	94.4	79.8	0.343	0.50		1

Structure	Compound ID	Reference compound	SGLT1 activity (% of control); n=2				Activity class	SGLT2 activity (% of control), n=2				Similarity towards training (Tanimoto)	Toxicity (AK secretion)		
			Average	St.dev	n1	n2		Average	St.dev	n1	n2		Average	St.dev	n
CCCCC(N1CCC(CC1)OCc2ccccc2)C(=O)N	Z1084070528	Bepridil	85	18.0	97.8	72.3	inactive	89	3.6	91.4	86.3	0.284			0
Cc1cccc(C)c1NC(=O)CCCN2CCN(CC#C)CC2	Z1535223420	Bupivacaine	85	13.0	93.7	75.8	inactive	92	0.0	92.2	92.2	0.433			0
C[Si](C)(CN1CCCC(CCC(=O)N)C1)c2ccccc2	Z1912600813	Trihexyphenidyl	86	3.3	88.8	84.1	inactive	79	9.5	85.4	72.0	0.328	8.57	6.24	4
C(CN1CCC(CN2CCCCCC2)CC1)Cc3ccccc3	Z436106106	Trihexyphenidyl	88	0.8	87.5	88.6	inactive	82	8.3	75.7	87.4	0.357	-11.90		1
CCN(Cc1ccccc1)c2ccc(\C=N\N3CCN(C)CC3)cc2	Z49564244	Bepridil	91	3.0	88.6	92.9	inactive	93	3.9	89.8	95.3	0.361	-22.37	24.25	4
Fc1ccc(cc1)C(N2CCC(CN3CCOCC3)CC2)c4ccccc4	Z1318177320	Cloperastine	93	20.0	106.7	79.1	inactive	54	0.1	54.6	54.4	0.422			0
O=C(N1CCN(CCOC(c2ccccc2)c3ccccc3)CC1)c4ccccc4	Z195623196	Cloperastine	95	9.7	101.6	87.9	inactive	72	1.9	73.0	70.4	0.474	-14.70		1
NC(=O)C(N1CCC(CN2CCCC2)CC1)c3ccccc3	Z1103127942	Trihexyphenidyl	98	14.0	87.9	107.7	inactive	89	10.0	82.2	96.3	0.322	-13.70		1
NC(C(=O)N1CCCC(CN2CCCCCC2)C1)c3ccccc3	Z1459737806	Trihexyphenidyl	98	16.0	86.9	109.3	inactive	101	7.5	106.7	96.1	0.333	-7.90		1
OC(CN1CCC(CN2CCCCCC2)CC1)c3ccccc3Cl	Z416806614	Trihexyphenidyl	98	5.6	93.8	101.6	inactive	79	1.8	78.1	80.6	0.317	-11.30		1
CC1CN2CCCCC2CN1CC(=O)Nc3ccc(C)cc3	Z1223130644	Bupivacaine	104	2.3	102.3	105.5	inactive	95	15.1	84.7	106.1	0.415	-5.70		1
CCOCC1CCN(C1)C(=O)C(CN2CCCC2)c3ccccc3	Z2229127630	Bepridil	112	6.6	107.2	116.5	inactive	92	3.9	89.0	94.5	0.278	8.28	13.99	4