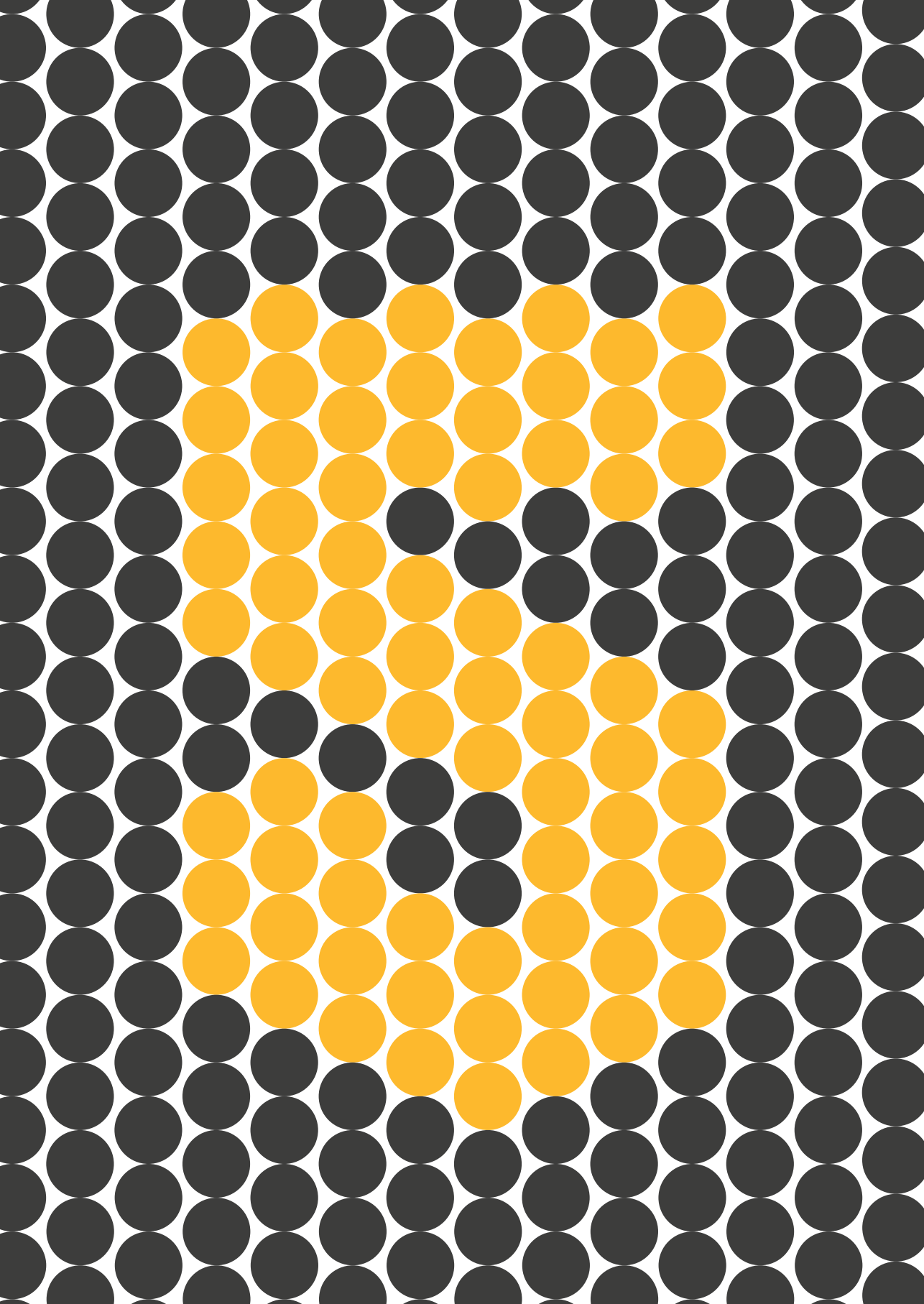
Sensing transport: label-free in vitro assays as an atTRACTive alternative for solute carrier transporter drug discovery

Sijben, H.J.



CHAPTER 5

Proteochemometric modeling identifies chemically diverse norepinephrine transporter inhibitors

Brandon J. Bongers †
Hubert J. Sijben †
Peter B.R. Hartog
Adriaan P. IJzerman
Laura H. Heitman
Gerard J.P. van Westen

Solute carriers (SLCs) are a relatively underexplored protein family compared to other major protein families such as kinases and G protein-coupled receptors (GPCRs). However, the SLC family and their role in a diverse array of diseases is known and of interest. One such SLC is the high-affinity norepinephrine transporter (NET/SLC6A2), which in contrast to most other SLCs has been relatively well studied, resulting in a large defined chemical space. Due to the low diversity of this chemical space it is challenging to identify ligands that are chemically novel. In this chapter, we aimed to find new NET inhibitors using a computational modeling screening pipeline. We applied multiple optimization steps during dataset creation, including similarity networks and stepwise feature selection, to end up with an optimal training set for our model, which was created by using proteochemometrics and stacking of several machine learning techniques. The model was applied to a large virtual database of Enamine, from which 22,000 of the 600 million predicted compounds were clustered to end up with 46 chemically diverse candidates. Of these candidates, 32 were synthesized and tested using the impedance-based TRACT assay that was developed in **Chapter 4**. We identified five hit compounds with submicromolar inhibitory potencies towards NET, which are promising for follow-up experimental research. This chapter demonstrates a comprehensive computational pipeline to predict new potential ligands for NET, which could be applied to any protein that has enough interaction data available.

Manuscript submitted
† These authors have contributed equally

5.1 – Introduction

Solute carriers (SLCs) are a divergent class of transporters and understudied compared to some of the other major receptor families, such as kinases and G protein-coupled receptors (GPCRs)[1]. Yet SLCs can play a critical role in complex diseases and as such several SLCs are interesting drug targets[2–4]. To further characterize SLCs, recently the RESOLUTE consortium was founded to develop and distribute biochemical tools and assays for *in vitro* and *in vivo* study of these transporters[5]. SLC subfamilies recognize highly divergent natural substrates and their sequence identity is low compared to other superfamilies such as kinases or GPCRs[6]. Hence, from a drug discovery perspective it is challenging to design family-wide studies to find new ligands that interact with SLCs. Instead, the focus lies on single subfamilies, or even a single SLC, to identify novel compounds.

The norepinephrine transporter (NET / SLC6A2) is involved in the rapid re-uptake of the neurotransmitter norepinephrine (NE) from the synaptic clefts of noradrenergic neurons in the peripheral and central nervous system[7]. As one of the most well characterized transporters, NET is an established drug target for depression, chronic pain and narcolepsy, with several marketed drugs available. Despite the abundance of pharmacological data on NET ligand binding, there is a need for the development of novel inhibitors with improved affinity and selectivity over other monoamine transporters[8]. Despite more structures becoming available, for NET structure-based work there is still no option for structure-based design of ligands due to the absence of a crystal or cryo-EM structure[9,10].

Computational studies such as statistical modeling and ligand docking have increased in popularity over the last decades, yet application to SLCs has been limited so far[11,12]. A 3D structure (crystal, cryo-EM or homology modeling based) of sufficient quality is required to perform structure-based drug discovery[13]. However, crystallization of SLCs is difficult given their membrane bound nature analogous to GPCRs. Hence, only limited amounts of structures are available for this family, with the promise of cryo-EM increasing that amount in the near feature. While advances in cryo-EM and machine learning, such as AlphaFold, are expected to significantly increase the available structures and alleviate some of these issues, their application in virtual screening has still to be demonstrated[14,15].

In the absence of structural information virtual screening can also be performed ligand-based using 2D chemical structures or *via* proteochemometrics (PCM), using ligand and protein information[16]. In both cases machine learning is used to identify correlation between bioactivity and structural features. Here, we will use PCM which allows us to create a comprehensive model of ligand structures of multiple proteins. This allows us to not only use the ligand space for NET, but also the most structurally related proteins, such as the dopamine and serotonin transporters. We then train these models on publically available data from ChEMBL, which contains a large amount of ligand-receptor interaction information for all ligands/proteins[17,18].

In this chapter, we applied PCM modeling to identify new chemotypes for the NET. While this transporter has been relatively well characterized compared to other SLCs, there still is a need for novel ligands that effectively, efficiently and selectively target NET[19,20]. We used

similarity networks as an approach to determine the optimal number of targets to include in our PCM model. After completion of our PCM model it was applied to the Enamine REAL database to identify novel ligands. Interestingly, the REAL database does not consist of on-the-shelf compounds but instead contains over 600 million make-on-demand molecules. These molecules can be synthesized *via* well-validated parallel synthesis protocols using a large number of building blocks. After virtual screening, the activity of our identified hits were validated experimentally with a hit rate of 5 out of 32 (16%).

5.2 – Results

5.2.1 – Dataset optimization by employing similarity networks and phylogenetic trees

The whole set derived from ChEMBL (All SLCs) was too large for model training, hence a selection of the data was made using sequence-based similarity networks (SNs). These SNs were used to highlight clusters with a pBLAST similarity to NET above a given threshold (**Figure 5.1**). Subsequently the clustered sets were used for PCM model training

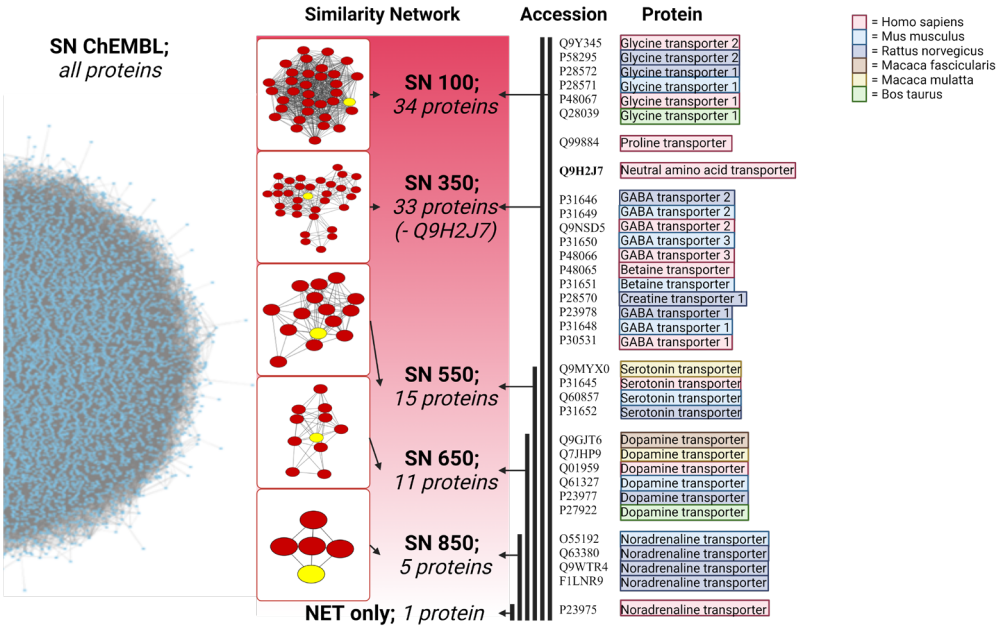


Figure 5.1 – Sequence based similarity networks obtained from SLCs in ChEMBL. Displayed are similarity networks wherein each node represents a single protein and each connection a pBLAST similarity above the chosen cutoff. Nodes in yellow denote human NET. SN25 resulted into one large cluster of almost all proteins and was discarded (left hand). From there, the following thresholds were used for the Similarity networks: Similarity network SN100 (34 proteins), including NET and related proteins from several animal species. SN350 (33 proteins), showing a smaller network with a section appearing to nearly dissociate. SN550 (15 proteins), containing the serotonin and dopamine transporters together with NET. SN650 (11 proteins) drops the serotonin transporters and the minimum viable similarity network SN850 (and all SNs above this threshold) contains solely NET from human and other species.

to determine the optimal number of related proteins for our model. A pBLAST threshold of 25 (smallest) led to a large network including all proteins (and was discarded), a threshold of 850 (largest) led to a network only including NET proteins between several species. Between these extremes several networks were obtained at intervals of 100, 350, 550, and 650, leading to a total of 6 data sets. Identification of a viable subset was also approached using phylogenetic trees calculated from protein sequence similarity.

Related proteins to NET could be identified if they were found on the same layer of the tree (**Figure 5.2**). Both the similarity networks and phylogenetic layers would then be selected for testing by modelling (up to an including layers 5, 6, and 7). In the phylogenetic approach, layer two represented the SN850 network, including three overlapped with the SN650 network, and including four represented SN550 network. Hence no separate models were trained for these groups (**Figure 5.3**).

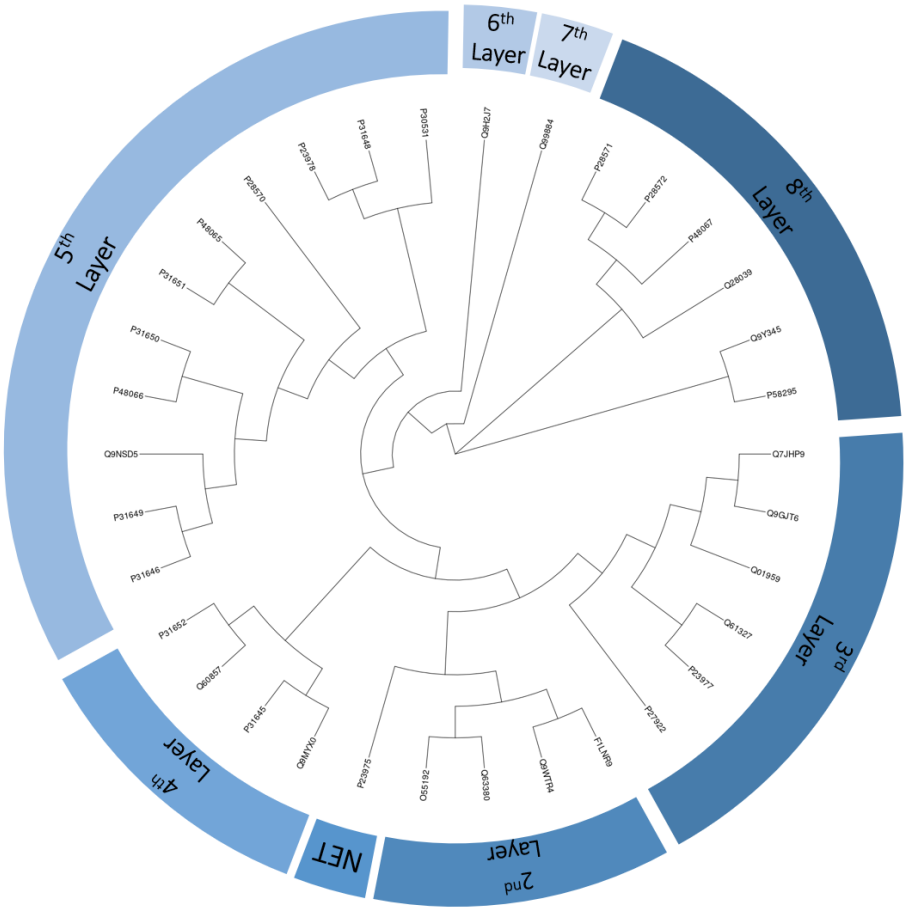


Figure 5.2 – Phylogenetic tree of maximal viable similarity network (SN100) reveals 8 individual layers. Displayed is the phylogenetic tree of the proteins (Uniprot codes) analyzed and colored with the various layers (defined as splits from the root of the tree defined by NET). This resulted in eight layers (including NET as the first layer).

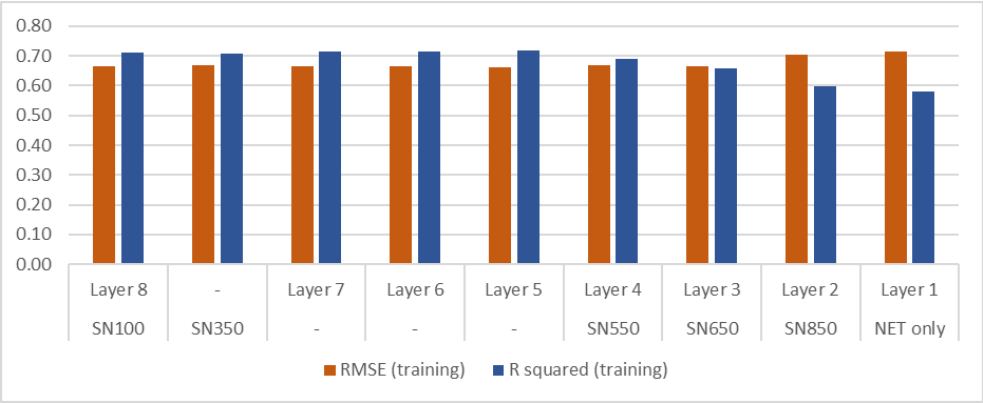


Figure 5.3 – Differences in cross validated R squared (R^2) and RMSE from models trained of the different subsets. Displayed is a plot of the R^2 and RMSE values generated during the dataset selection process. A high value for R^2 and a low value for RMSE were desired. SN100 was eventually preferred due to this due to its RMSE and R^2 values. It was preferred over layers 5, 6, 7 and SN350 as SN100 contained more data to model with.

5.2.2 – Final dataset was chosen from best scoring similarity networks

Both selection methods led to a total of nine subsets that were empirically tested to find the optimal training set. To do this, a random forest model was created and cross-validated to assess the R^2 and RMSE (**Figure 5.3**). Subsets Layer 5, 6, 7, SN350, and SN100 all scored comparatively with a R^2 of 0.71–0.72 and a RMSE of 0.66–0.67. The other sets all scored lower with R^2 0.58–0.62 and RMSE 0.66–0.75. Out of these five comparable sets, SN100 was chosen in the end as this contained the most data and produced top performing models.

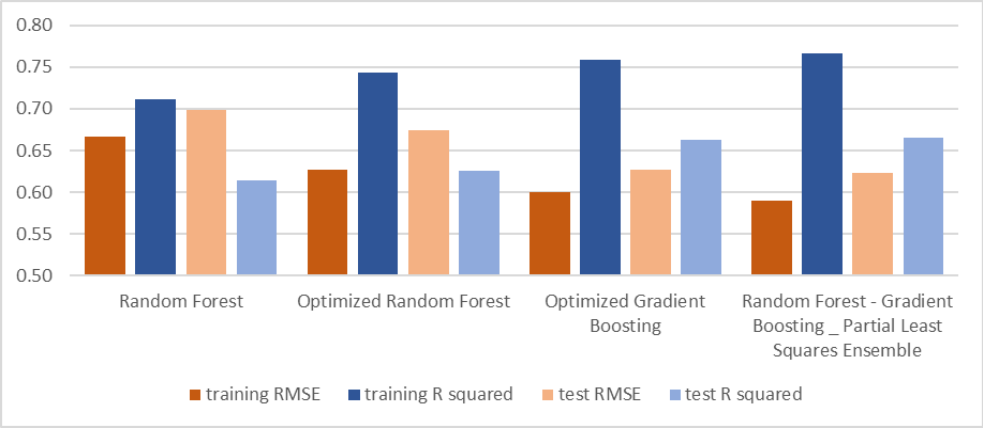


Figure 5.4 – Overview of performance of selected modeling approaches. Displayed are the internal (training, cross validation) and external (testing 30% holdout) statistics. Shown are three intermediate models, a Random Forest, an optimized Random Forest and Gradient Boosting model, with a final model consisting of a Random Forest plus Gradient Boosting ensemble with a stacked Partial Least Squares ensemble as second step. Each model improves performance over the last one, with the last being best. This optimization was applied to every combination possible, but these are not shown for brevity.

5.2.3 – Several machine learning techniques were used to create an optimized model

After selection of the optimal data subset, an optimal choice of machine learning method (ML) was determined. Three different methods were used: Random Forest (RF), Gradient Boosting (GB) and Partial Least Squares (PLS). Moreover, these methods were also further optimized and tested in an ensemble approach. Optimization was performed by both a backwards stepwise feature selection and a parameter optimization using grid search, with the best scoring model of each method continued for further analysis. Performance was determined using the R^2 and RMSE from a 30% (random based) holdout set of all NET interactions in the dataset.

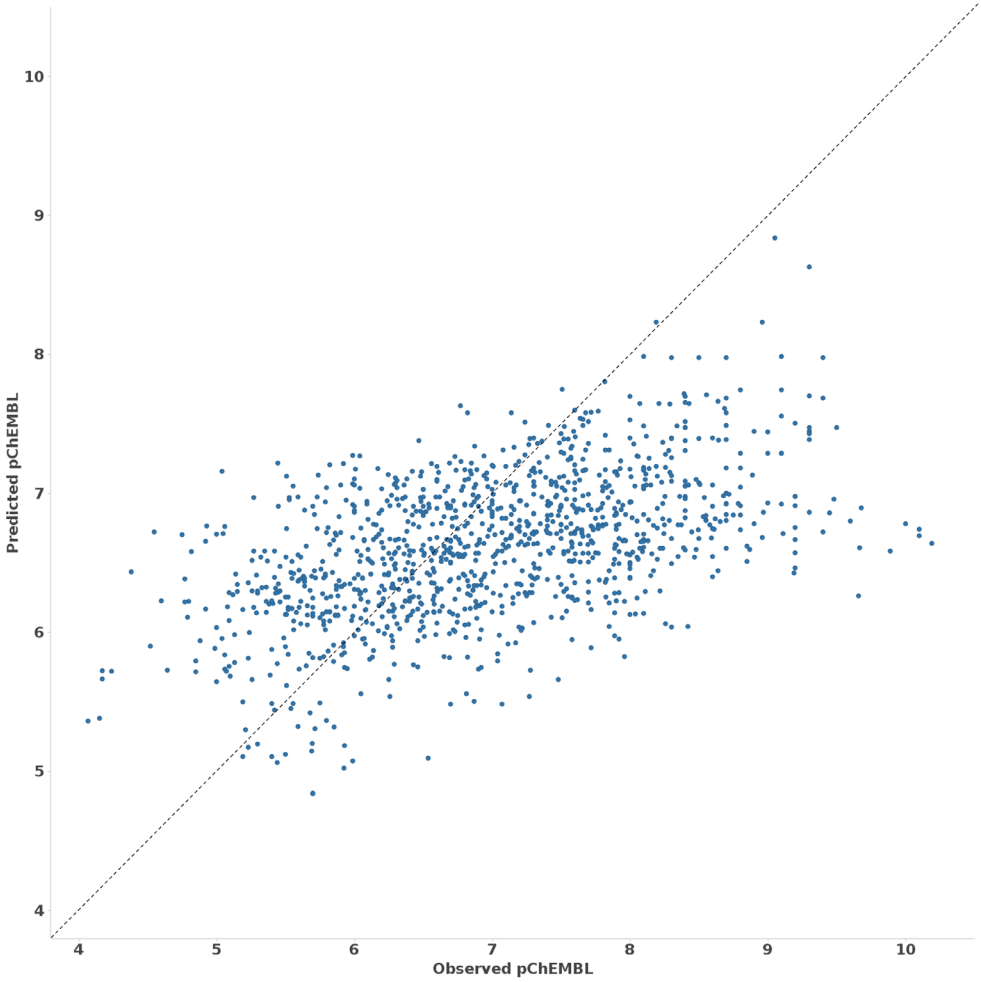


Figure 5.5 – Predicted pChEMBL values of temporal split correlated to the observed values. Temporal split prediction where all known interactions of ChEMBL of 2010 and before were used as training set (15,106 data points) and those of 2011 and later (5,083 data points) were used as the test set. R^2 was 0.24 and RMSE was 1.02.

Without optimization, PLS (R^2: 0.28; RMSE: 0.93) was underperforming compared to RF (R^2: 0.61; RMSE: 0.70) and GB (R^2: 0.65; RMSE: 0.62). Next, stepwise feature selection and parameter optimization using a grid search were performed to fine-tune the models. Optimization of both the RF and GB models showed an increase in R^2 (0.62; 0.66) and a decrease in RMSE (0.67; 0.62) (**Figure 5.4**). There was another small increase in performance when PLS was stacked as a second model after the RF and GB models. The ensemble of optimized RF and GB models, of which predictions became the descriptors for a PLS model, performed the best and will be referred to as the NET model from now on.

5.2.4 – External validation shows the robustness of the NET model

To check whether the created NET model would meet the standards of a robust model, an external validation was performed with ChEMBL data (**Figure 5.5**). This validation was a temporal split, with the training set containing data from literature published before and in 2010, and the test set 2011 and later. This resulted in a R^2 of 0.24 and a RMSE of 1.02, in line with our previous examples of a temporal split[18]. Given the challenging nature of this approach (different chemotypes that are removed from the training set) and our prior experience with expected performance of models trained on temporal split ChEMBL data, it was concluded that the NET model was robust enough to continue forward.

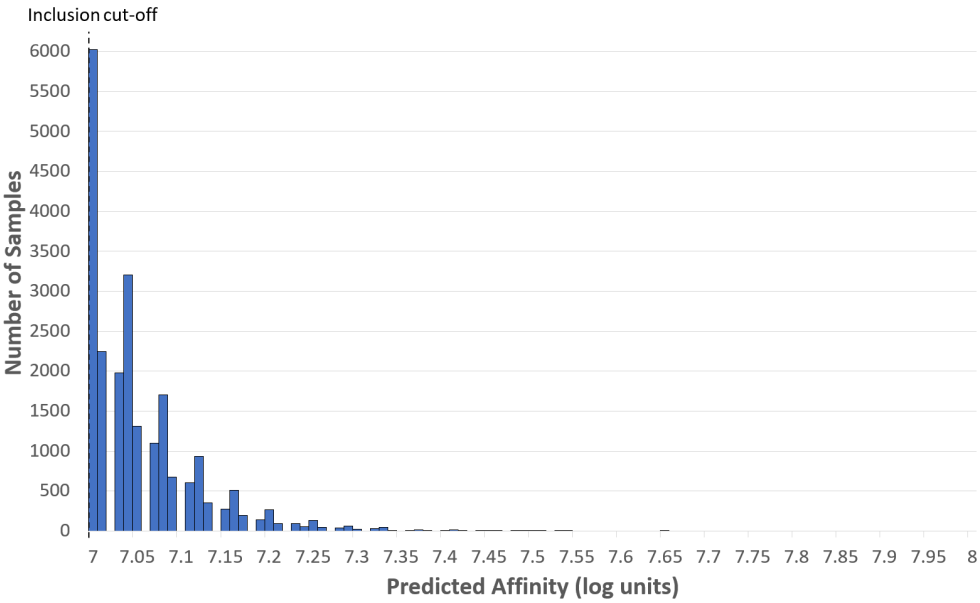


Figure 5.6 – Distribution of all predictions with affinity above 100 nM. Displayed is a histogram plot of the predicted affinities for the NET virtual screening of the Enamine compound database. Only those affinities which were predicted to have values lower than 100 nM were included (22.206 compounds).

5.2.5 – The NET model predicted 46 groups of compounds as viable candidates

The Enamine database was virtually screened with the NET model to predict the bioactivity of compounds for NET. Subsequently through several filtering steps a final selection was made as the initial database contained around 700 million compounds. In the first step only compounds with a predicted affinity towards NET better than 100 nM (7.00 log units) were considered (**Figure 5.6**). This threshold resulted in 22,206 compounds remaining, with the highest predicted affinity to reach 7.65 log units.

Subsequently compounds were clustered using HDBSCAN and visualized with t-SNE using a 1024 bit ECFP_6 fingerprint (**Figure 5.7**). HBDSCAN produced 46 clusters, with each

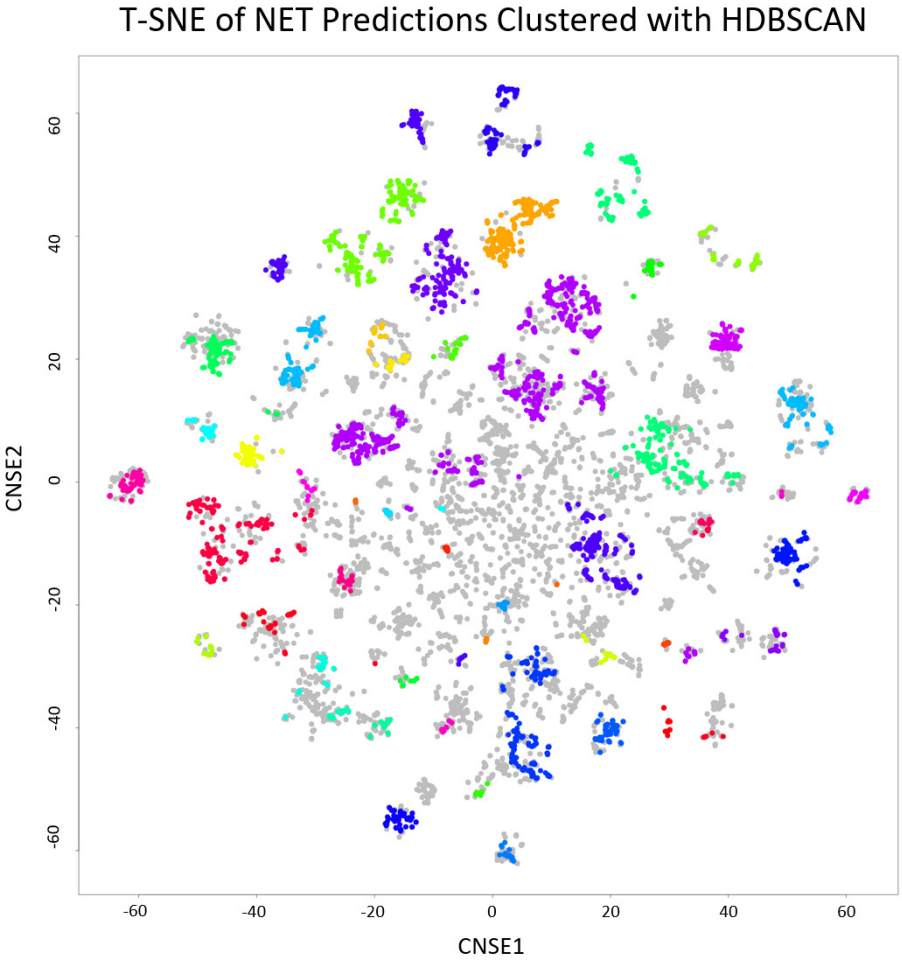


Figure 5.7 – t-SNE of the 22.206 predictions with HDBSCAN designated clusters. The t-SNE displayed was created using 1024 bits of FCFP_6. The HDBSCAN shows 46 distinct clusters with different colors. Grey points were filtered out as too similar (<90%) or too unsimilar (>50%) by HDBSCAN. The member of each cluster with the highest predicted activity was used as representative of that cluster in the prospective validation.

cluster representing structurally similar compounds. Compounds were first filtered (colored grey) by similarity to the training set, removing entries that shared either a 90% or higher similarity or a 50% or lower similarity. This was to ensure novelty and to increase the chance on NET activity in our final selection, respectively. Then, of each cluster, the compound with the highest predicted affinity was selected for a final suggested list of potential NET inhibitors. Of the 46 compounds, 32 were purchased and tested for NET activity in a label-free impedance-based assay.

5.2.6 – Experimental validation

To validate whether the predicted molecules from the NET model showed biological activity on NET, we used an impedance-based 'transport activity through receptor activation' (TRACT) assay as described in **Chapter 3 and 4**[21,22]. In this assay a HEK293 cell line with inducible expression of NET was used and the activation of endogenously expressed alpha-2 adrenergic receptors by norepinephrine (NE) was measured as a cellular response. A compound was considered a NET inhibitor if the compound was able to significantly enhance the NE-induced cellular response in a concentration-dependent manner. A single-point primary screen was performed with 10 µM test compound, using the reference NET inhibitor nisoxetine as a positive control (**Figure 5.8a**). Five of the 32 tested compounds were able to enhance the NE-induced response to a similar level as nisoxetine, which indicated that the compounds inhibited NET with a decent potency. None of the five compounds showed modulation of the NE response in cells lacking NET (**Supplementary Figure 5.S1**), confirming that the enhanced NE-induced response was specific to NET.

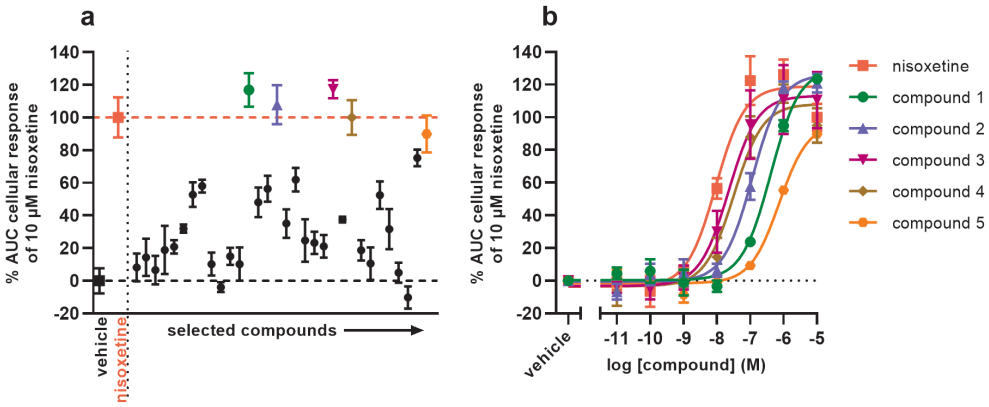
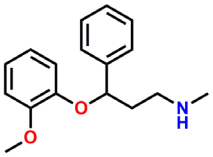
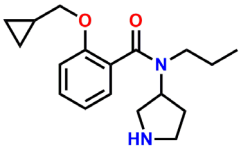
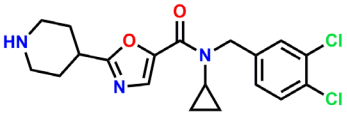
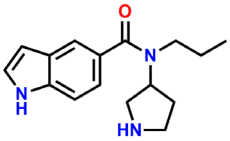
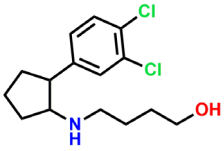
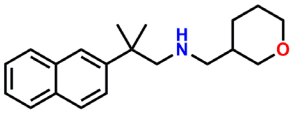


Figure 5.8 – *In vitro* functional validation of hits in a label-free impedance-based TRACT assay. (**a**) Single point screen of 32 hit compounds and (**b**) full-range concentration-inhibition curves of the top five compounds from the single point screen. Doxycycline-induced JumpIn-NET cells were pretreated for 1 h with either vehicle or (**a**) 10 µM or (**b**) increasing concentrations of nisoxetine or hit compound. Subsequently, cells were stimulated with 1 µM norepinephrine (NE) and Cell Index (CI) was measured for 30 min. Cellular responses are expressed as the net area under the curve (AUC) of the first 30 minutes after stimulation with NE. Data were normalized to the response of NE only (vehicle, 0%) and the response of NE in the presence of 10 µM nisoxetine (100%). Data are shown as the mean ± SEM of three separate experiments each performed in duplicate.

To further characterize the most potent inhibitors, full-range concentration-inhibition curves were obtained for the top five compounds and inhibitory potency (pIC_{50}) values were determined (**Figure 5.8b, Table 5.1**). The compounds on their own did not induce substantial cellular responses during pretreatment (**Supplementary Figure 5.S2**). All tested compounds showed concentration-dependent enhancement of the NE response with submicromolar inhibitory potencies (**Supplementary Figure 5.S3, Figure 5.8b**). Compounds 3 and 4 showed the highest pIC_{50} values (7.6 ± 0.1 and 7.5 ± 0.2, respectively), which were in the range of the pIC_{50} of nisoxetine (8.0 ± 0.0) (**Table 5.1**). Taken together, these results demonstrate that at least five of the 32 tested compounds were biologically active NET inhibitors in a label-free TRACT assay.

Table 5.1 – Inhibitory potency (pIC_{50}) values of tested compounds as determined in the impedance-based TRACT assay. Data are reported as the mean ± SEM of three individual experiments each performed in duplicate.

Compound	Molecular structure	pIC_{50} ± SEM
Nisoxetine		8.0 ± 0.0
Compound 1		6.4 ± 0.1
Compound 2		6.9 ± 0.1
Compound 3		7.6 ± 0.1
Compound 4		7.5 ± 0.2
Compound 5		6.1 ± 0.1

5.3 – Discussion

Major depressive disorder is one of the main causes of disability, and an increasing trend in the worldwide incidence and prevalence of depression has been observed in recent years[23,24]. Selective serotonin reuptake inhibitors (SSRIs), serotonin-norepinephrine reuptake inhibitors (SNRIs) and selective norepinephrine reuptake inhibitors (sNRIs) are established classes of prescription drugs for the first-line treatment of depression[8]. Although these drugs improve on the polypharmacological profile of tricyclic antidepressants, the current generation of reuptake inhibitors suffer from partial or non-responsiveness, relatively low remission rates, slow onset of action, and risk of adverse effects[25]. Thus, the identification of novel norepinephrine transporter (NET) inhibitors could improve on the efficacy of current antidepressants, as well as provide scaffolds for the development of (fluorescent) probes for *in vitro* imaging[26]. In this chapter, we have developed a machine learning model for the identification of novel inhibitors for human NET. After virtual screening of the Enamine database with this predictive model, we filtered out 46 compounds by clustering for experimental validation. Using the live-cell, impedance-based TRACT assay that was validated in **Chapter 4**, we identified five novel inhibitors for NET. Here, we will discuss the opportunities and limitations of this approach.

The bioactivity data that was used for training our models was obtained from ChEMBL25. However, machine learning models work best using more data and hence complimentary data from ExCAPE-DB can be included in future work[27]. Moreover, we recently released a comprehensive dataset called Papyrus that combines several datasets, that is annotated and standardized for compatibility[28]. In future applications of this pipeline, we would switch to this dataset to increase our training set size while still retaining high quality and open source data. As said, having more data improves model performance and it is known that PCM models often demonstrate better performance than single target models due to the inclusion of more data. Here we have shown that we can empirically determine an optimal set of related proteins to include in a PCM model. This is a relevant finding as prior work in the area has primarily focused on small conserved families or very large protein superfamilies[29,30]. We argue that the optimal number of included similar sequences is dependent on the (mean) similarity, the chemical variation, and the amount of data points per sequence and therefore data set dependent. Hence, good practice is to optimize this number when creating optimized models.

We used both phylogenetic trees and similarity networks to identify the optimal selection of proteins. Here, similarity networks proved to be a useful tool compared to the phylogenetic trees, as optimizing the pBLAST score threshold allowed us to vary the data set size and hence model performance. Conversely, while the trees are often used in metabolic pathway studies[31,32], here they were less useful than the networks due to the inability to tune the threshold as is the case with the networks. Similarity networks have also been used in comparative research, for example to visualize enzyme function using protein sequence, to visualize relationships between protein superfamilies, or to find similarities using gene ontology databases[33–35]. Whereas these studies mainly focused on functional similarity, we

used sequence similarity, and thus including this functional similarity used in other work to our networks could potentially create a higher quality network that could predict more accurately.

When optimizing our prediction models for R^2 and RMSE, we concluded that the ensemble-stacking model containing all three methods (Random Forest, Gradient Boosting and Partial Least Squares) performed the best. However, the values for R^2 and RMSE between different combinations of these methods were very close, including some single models. We decided to use the ensemble-stacking model, since we concluded in earlier work that these models tend to work better compared to single models[18]. Deep learning could likely improve our model even further, as was demonstrated in our earlier work, however this was deemed outside the scope of this chapter[36].

To perform our clustering we had to trim down from our initial predictions to only include compounds that had a predicted affinity of 100 nM or better (resulting in a set of 22,206 compounds). Lower thresholds resulted in a clustering that was too large and would not converge. In follow up work, by increasing the amount of computational power we should be able to include more compounds, which subsequently could reveal new interesting clusters. Next, in order to only include novel candidates, we filtered for similarity between the set of 22,206 predicted compounds compared to our initial training set. Compounds that had higher than 90% similarity were excluded, as they would be too similar to existing inhibitors. Compounds that had a similarity of 50% or lower would be discarded as well, to increase the confidence in model predictions for the compound. The thresholds were chosen arbitrarily and could be subject to another optimization finding, but this was deemed out of scope of the current work. To further limit the amount of candidates, the minimal amount of points in a cluster was set to 19; so any smaller clusters were not taken further. From each cluster the most potent compound was then selected. Finally, as not all 46 candidates could be synthesized readily we eventually obtained final set of 32 compounds that were available for experimental validation. Note, potentially exploring (analogs of) the 14 cut candidates, or centers from the smaller clusters could hence result in more hits.

After clustering, 32 compounds were initially screened for their activity on NET using the impedance-based TRACT assay that was developed in **Chapter 4**. This assay has been used previously to characterize well-known inhibitors of NET, showing a similar rank order of inhibitory potencies compared to a more traditional fluorescent substrate uptake assay. In addition, the assay was validated for screening purposes and taking into account our experience with this platform we favored the use of the TRACT assay over traditional assays. Eleven out of the 32 compounds displayed at 10 μM more than 50% enhancement of the NE-induced response, which is substantial considering that these compounds are structurally distinct from each other. This was also apparent from the five hit compounds, which all display submicromolar potencies towards NET. Although all compounds contain structural elements that are key to interacting with the sub-pockets of the norepinephrine binding site, such as a secondary amine and a substituted aromatic moiety, the scaffolds vary significantly in the substitution and size of aliphatic groups or the presence of an

amide moiety (Table 2)[37]. Thus, these scaffolds could provide a starting point for the design and synthesis of derivatives, quantitative structure-activity relationships and subsequent hit optimizations of novel NET inhibitors.

Here, we have demonstrated an approach to identify novel protein inhibitors using a combination of machine learning techniques. In contrast to prior work which focused on a single model created from only NET interaction data, the optimal set of related targets for the PCM model was determined dynamically based on data analysis and subsequent modeling, stressing the fact that multiple SLC families were investigated for model inclusion. We applied this approach to identify novel NET inhibitors, which were found by virtually screening a database containing virtual molecules that were synthesized on demand. The complete computational pipeline can be applied to other protein families with relative ease, with the same provided data, or potentially be improved on with either larger datasets or more in-depth machine learning techniques.

5.4 – Materials and methods

5.4.1 – Software

Proteochemometric modeling, data curation, feature extraction, and cluster analysis was performed in Pipeline Pilot (version 18[38]). Machine learning was performed using R (version 3.5.2) as integrated in Pipeline Pilot. Similarity network construction was done with Cytoscape (version 3.7.1[39]) in RStudio (version 3.6.0[40]). Any seeds used in randomization or model creation/prediction was set to '12345'.

5.4.2 – Interaction data

Interactions were gathered from the ChEMBL database (version 25.0[41]). Properties included were canonical SMILES for the compounds, amino acid sequence for the proteins, pChEMBL value representing the affinity (in –log M). If there was more than one pChEMBL unit assigned to a data point (combination of chemical structure and protein) the highest of the following ranked units were chosen: $K_i > IC_{50} > EC_{50} > K_d$. Any duplicate pChEMBL values left were averaged so that only a single data point for each interaction remained.

5.4.3 – Compound standardization

Pipeline Pilot was used to convert canonical SMILES to structures. Compounds were standardized as in the statistical section of Burggraaff *et al*[42]. These steps included removing salts, standardizing stereoisomers/charges, and (de)protonation based on a pH 7.0.

5.4.4 – Compound descriptors

Physicochemical properties were calculated using Pipeline Pilot built-in components. Several fingerprints were calculated Estate keys/counts, MDL fingerprints, and a selection of extended-connectivity fingerprints[43]. A full list of these compound descriptors can be found in **Supplementary Table 5.S1**, and an explanation of the letter system for the

extended-connectivity fingerprints can be found in the related article. All these descriptors were used during the feature selection process to sample which ones performed optimally.

5.4.5 – Protein descriptors

Three classes of protein descriptors were tested. The first set of protein descriptors was generated using the PROFEAT interface[44], which are alignment agonistic. Secondly, three alignment-based protein descriptors were included as used previously (Z-scales, FASGAI and BLOSUM)[45]. Finally, a third set of protein descriptors was prepared using an in-house algorithm that included a selection of protein descriptor generators and returned an autocross correlated (ACC) version[46]. An overview can be found in **Supplementary Table 5.S1**. Like the compound descriptors, these were also used in the feature selection part of the process.

5.4.6 – Similarity networks

Similarity networks were created using RStudio and package 'Rcy3' in Cytoscape, while displayed using 'yFiles'. 9,131 proteins were extracted from ChEMBL, with 5,142 proteins used in the similarity investigations as these had interactions with compounds detailed. Proteins were first analyzed using pBLAST, resulting in an all-versus-all similarity matrix. Networks were then created using a varying pBLAST threshold, a higher threshold resulting in a higher required similarity for inclusion and hence less proteins included for the network. Two networks representing the extremes: a maximum viable similarity network (required similarity >= 100) representing multiple solute carriers; and a minimum viable similarity network (required similarity >= 800) containing only NET homologs.

5.4.7 – Phylogenetic tree formation

Phylogenetic trees were created using R packages 'msa', 'seqinr' and 'ape'. Alignment was performed using the 'msa' implementation of ClustalW. Phylogenetic tree formation ended at the maximum viable similarity network (pBLAST >= 100), as it proved unfeasible to create a tree with all 5,142 proteins (pBLAST >= 25) with the available resources. Tree layers were created upwards from the minimum viable layer (NET only), with each layer above it including the previous layer. Tree creation was stopped when it reached the maximum viable similarity network. Modeling performance on the data using selected similarity networks as filter was then obtained using a 70/30 target based data split. This split was done with PCA assisted K-means, this was to ensure as much homogeneity between the datasets when validating. The R^2 and Residual Mean Squared Error (RMSE) were then calculated from a 10-fold cross validation.

5.4.8 – Feature selection

Stepwise feature selection was performed during model optimization. The maximum number of iterations were set at 25 and the number of iterations without model improvement was set to 3. Model improvement was defined as an increase in 5-fold cross validated R^2.

Table 5.2 – Grids used during the parameter optimization procedure. Models are found on the left hand side with their respective R package. Parameter grids are separated per model.

Model		Parameter Grids
Random Forest (ranger)	Number of Trees	100, 250, 500, 1000
	Number of Descriptors	Sqrt(D)*, Log2(D)*, Fraction: 10%, 50%, 90%
	Minimum Node Size	1, 5, 7
	Maximum Depth	5, 7, no max
Gradient Boosting (xgboost)	Maximum number of Trees	100, 250, 500, 1000
	Learning Rate	0.1, 0.3, 0.5
	Gamma	0, 0.3, 0.5
	Maximum Depth	5,7
	Data Fraction	0.1, 0.5, 1.0
	Descriptor Fraction	0.5, 0.7
Partial Least Squares (pls)	Number of Variables	100, 200, 300

* D represents number of descriptors.

5.4.9 – Parameter optimization

Parameter optimization was performed using a simple full grid search. Model improvement was defined as an increase in 5-fold cross validated R^2. Parameter grids are separated per model as shown in **Table 5.2**.

5.4.10 – Final validation / clustering

Clustering was used to select a diverse set of compounds for external validation. As an additional step after clustering, an identity filter was applied that removed points that had a 90% or higher identity or a 50% or lower identity with compounds found in the training data. This was to ensure that compounds were novel compared to existing compounds, but did not stray too far from the known interactions. Clustering was performed using R package 'hdbscan'. Clusters were visualized in Pipeline Pilot, including the coloring of the different clusters. Grey points were filtered out automatically as noise, and thus discarded in the final selection. Finally, these clusters were ranked based on predicted NET affinity, and the top ranked compounds were chosen for further experimental validation.

5.4.11 – Chemicals and reagents

Jump In T-REx HEK 293 cells modified for doxycycline-inducible overexpression of the wild-type human norepinephrine transporter (JumpIn-NET) were provided by CeMM (Research Center for Molecular Medicine, Medical University of Vienna, Austria) and generated as described in **Chapter 4**. Doxycycline hyclate was purchased from Sigma Aldrich (St. Louis, MO, USA). Nisoxetine hydrochloride was purchased from Santa Cruz Biotechnology (Dallas, TX, USA). The 32 selected predicted active molecules were synthesized and provided by Enamine. All other chemicals were of analytical grade and obtained from standard commercial sources.

5.4.12 – Cell culture

JumpIn-NET cells were grown as adherent cells in culture medium (high glucose Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% (v/v) fetal calf serum (FCS), 2 mM Glutamax, 100 IU/ml penicillin and 100 µg/ml streptomycin) at 37°C and 7% CO_2. Cryopreserved cells were thawed and cultured for 1–2 passages in culture medium. Cells were then cultured up to one week in culture medium supplemented with 2 mg/ml G418 and 5 µg/ml blasticidin before switching back to culture medium at least 24 h prior to an experiment. Cell cultures were split twice per week at ratios of 1:8 – 1:16 in 10 cm plates.

5.4.13 – TRACT assay

Label-free TRACT assays were performed using the xCELLigence real-time cell analysis (RTCA) platform as described in **Chapter 4**. In short, cells grown on gold-coated electrodes of 96-well E-plates impede the electric current that is generated on the electrodes. Impedance is measured at 10 kHz and is converted to the dimensionless parameter Cell Index (CI) using the following formula:

$$CI = \frac{(Z_i - Z_0)\Omega}{15\Omega}$$

where Z_i is the impedance at any given time and Z_0 is the baseline impedance measured at the start of each experiment.

Assays were performed at 37°C and 5% CO_2 in 96-well E-plates in a total volume of 100 µl. Background impedance was measured in 40 µl culture medium. JumpIn-NET cells were seeded in 50 µl at 60,000 cells/well in the presence of 1 µg/ml doxycycline (or no doxycycline for the counterscreen). The E-plate was left at room temperature for 30 min before placement in the recording station. Cells were grown for 22 hr prior to inhibitor pretreatment. All compound additions were done using a VIAFLO 96 handheld electronic 96 channel pipette (INTEGRA Biosciences, Tokyo, Japan). After 22 h, cells were pretreated for 1 h with either a single concentration (single-point primary screen, 10 µM) or increasing concentrations (full-range concentration-inhibition curves, ranging from 10 pM to 10 µM) of compound or nisoxetine (positive control). Dilutions of compounds were first made in DMSO, then in phosphate-buffered saline (PBS). Vehicle-pretreated cells received only DMSO in PBS. Final amounts of DMSO were kept at 0.1% per well. After 1 h inhibitor pretreatment, cells were stimulated with either vehicle or 1 µM norepinephrine in PBS containing 1 mM ascorbic acid (final concentration). Impedance was then measured every 15 seconds for 30 minutes.

5.4.14 – Data analysis

Raw data from TRACT assays were recorded using RTCA Software v2.0 or v2.1.1 (ACEA Biosciences). For analysis of NE-induced cellular responses CI values were normalized to the time point prior to substrate addition to obtain normalized CI (nCI) values. Data were exported from RTCA Software and analyzed in GraphPad Prism v8.1.1 (GraphPad

Software, San Diego, CA, USA). Per E-plate, nCI values of vehicle-pretreated and vehicle-stimulated cells were subtracted from all other data points to correct for any inhibitor and substrate-independent effects. NE-induced cellular responses were quantified by taking the net area under the curve (AUC) of the first 30 min after NE stimulation. Inhibitory potency (pIC_{50}) values of compounds are reported as a concentration-dependent enhancement of the NE-induced response by fitting the AUC data with non-linear regression to a sigmoidal concentration-inhibition curve with a fixed pseudo-Hill slope of 1. Data are shown as mean ± standard error of the mean (SEM) of three separate experiments each performed in duplicate.

References

1. César-Razquin, A. *et al.* (2015) A call for systematic research on solute carriers. *Cell* **162**, 478–487.
2. Rask-Andersen, M., Masuram, S., Fredriksson, R. & Schiöth, H. B. (2013) Solute carriers as drug targets: Current use, clinical trials and prospective. *Mol. Aspects Med.* **34**, 702–710.
3. Girardi, E. *et al.* (2020) A widespread role for SLC transmembrane transporters in resistance to cytotoxic drugs. *Nat. Chem. Biol.* **16**, 469–478.
4. Okabe, M. *et al.* (2008) Profiling SLCO and SLC22 genes in the NCI-60 cancer cell lines to identify drug uptake transporters. *Mol. Cancer Ther.* **7**, 3081–3091.
5. Superti-Furga, G. *et al.* (2020) The RESOLUTE consortium: unlocking SLC transporters for drug discovery. *Nat. Rev. Drug Discov.* **19**, 429–430.
6. Höglund, P. J., Nordström, K. J. V., Schiöth, H. B. & Fredriksson, R. (2011) The solute carrier families have a remarkably long evolutionary history with the majority of the human families present before divergence of Bilaterian species. *Mol. Biol. Evol.* **28**, 1531–1541.
7. Bönisch, H. & Brüss, M. (Springer Berlin Heidelberg, 2006). The norepinephrine transporter in physiology and disease. in *Neurotransmitter Transporters: Handbook of Experimental Pharmacology* (eds. Sitte, H. H. & Freissmuth, M.) vol. 175 485–524.
8. Xue, W. *et al.* (2018) Recent advances and challenges of the drugs acting on monoamine transporters. *Curr. Med. Chem.* **25**, 1–42.
9. Coleman, J. A., Green, E. M. & Gouaux, E. (2016) X-ray structures and mechanism of the human serotonin transporter. *Nature* **532**, 334–339.
10. Penmatsa, A., Wang, K. H. & Gouaux, E. (2013) X-ray structure of dopamine transporter elucidates antidepressant mechanism. *Nature* **503**, 85–90.
11. Rognan, D. (2007) Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **152**, 38–52.
12. Schlessinger, A. (Springer Berlin Heidelberg, 2014). Characterizing the structure, function, and evolution of human solute carrier (SLC) transporters using computational approaches. in *Membrane Transport Mechanism: Springer Series in Biophysics* (eds. Krämer, R. & Ziegler, C.) vol. 17 23–57.
13. Maia, E. H. B., Assis, L. C., de Oliveira, T. A., da Silva, A. M. & Taranto, A. G. (2020) Structure-based virtual screening: From classical to artificial intelligence. *Front. Chem.* **8**, 343.
14. Lee, Y. *et al.* (2019) Cryo-EM structure of the human L-type amino acid transporter 1 in complex with glycoprotein CD98hc. *Nat. Struct. Mol. Biol.* **26**, 510–517.
15. Tunyasuvunakool, K. *et al.* (2021) Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596.
16. Van Westen, G. J. P., Wegner, J. K., IJzerman, A. P., Van Vlijmen, H. W. T. & Bender, A. (2011) Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Medchemcomm* **2**, 16–30.
17. Bento, A. P. *et al.* (2014) The ChEMBL bioactivity database: An update. *Nucleic Acids Res.* **42**, 1083–1090.
18. Lenselink, E. B. *et al.* (2017) Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J. Cheminform.* **9**, 1–14.
19. Wong, E. H. F. *et al.* (2000) Reboxetine: A pharmacologically potent, selective, and specific norepinephrine reuptake inhibitor. *Biol. Psychiatry* **47**, 818–829.
20. Zhou, J. (2004) Norepinephrine transporter inhibitors and their therapeutic potential. *Drugs Future* **29**, 1235–1244.
21. Sijben, H. J., van den Berg, J. J. E., Broekhuis, J. D., IJzerman, A. P. & Heitman, L. H. (2021) A study of the dopamine transporter using the TRACT assay, a novel in vitro tool for solute carrier drug discovery. *Sci. Rep.* **11**, 1312.
22. Sijben, H. J. *et al.* (2021) Label-free high-throughput screening assay for the identification of norepinephrine transporter (NET/SLC6A2) inhibitors. *Sci. Rep.* **11**, 12290.
23. Moreno-Agostino, D. *et al.* (2021) Global trends in the prevalence and incidence of depression: a systematic review and meta-analysis. *J. Affect. Disord.* **281**, 235–243.
24. James, S. L. *et al.* (2018) Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: A systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **392**, 1789–1858.
25. Artigas, F. (2013) Future directions for serotonin and antidepressants. *ACS Chem. Neurosci.* **4**, 5–8.
26. Camacho-Hernandez, G. A. *et al.* (2021) Illuminating the norepinephrine transporter: fluorescent probes based on nisoxetine and talopram. *RSC Med. Chem.* **12**, 1174–1186.
27. Sun, J. *et al.* (2017) ExCAPE-DB: An integrated large scale dataset facilitating Big Data analysis in chemogenomics. *J. Cheminform.* **9**, 1–9.
28. Béquignon, O. J. *et al.* (2021) Papyrus – A large scale curated dataset. *ChemRxiv* **preprint**, 1–26.
29. Van Westen, G. J. P. *et al.* (2012) Identifying novel adenosine receptor ligands by simultaneous proteochemometric modeling of rat and human bioactivity data. *J. Med. Chem.* **55**, 7010–7020.

30. Lapins, M. & Wikberg, J. E. S. (2010) Kinome-wide interaction modelling using alignment-based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques. *BMC Bioinformatics* **11**, 339.

31. Oh, S. J., Joung, J. G., Chang, J. H. & Zhang, B. T. (2006) Construction of phylogenetic trees by kernel-based comparative analysis of metabolic networks. *BMC Bioinformatics* **7**, 1–12.

32. Zhou, T., Chan, K. C. C. & Wang, Z. (Springer Berlin Heidelberg, 2008). TopEVM: Using co-occurrence and topology patterns of enzymes in metabolic networks to construct phylogenetic trees. in *Pattern Recognition in Bioinformatics: Lecture Notes in Computer Science* (eds. Chetty, M., Ngom, A. & Ahmad, S.) vol. 5265 225–236.

33. Gerlt, J. A. *et al.* (2015) Enzyme function initiative-enzyme similarity tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim. Biophys. Acta - Proteins Proteomics* **1854**, 1019–1037.

34. Atkinson, H. J., Morris, J. H., Ferrin, T. E. & Babbitt, P. C. (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One* **4**, e4345.

35. Pesquita, C. *et al.* (2008) Metrics for GO based protein semantic similarity: A systematic evaluation. *BMC Bioinformatics* **9**, 1–16.

36. Koutsoukas, A., Monaghan, K. J., Li, X. & Huan, J. (2017) Deep-learning: Investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *J. Cheminform.* **9**, 1–13.

37. Pidathala, S., Mallela, A. K., Joseph, D. & Penmatsa, A. (2021) Structural basis of norepinephrine recognition and transport inhibition in neurotransmitter transporters. *Nat. Commun.* **12**, 2199.

38. http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot/ BIOVIA Pipeline Pilot | Scientific Workflow Authoring Application for Data Analysis (accessed 15-12-2015).

39. Su, G., Morris, J. H., Demchak, B. & Bader, G. D. (2014) Biological network exploration with Cytoscape 3. *Curr. Protoc. Bioinforma.* **2014**, 8.13.1-8.13.24.

40. https://support.rstudio.com/hc/en-us/articles/206212048-Citing-RStudio/ RStudio: Integrated Development for R. (accessed 31-05-2017).

41. Mendez, D. *et al.* (2019) ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940.

42. Burggraaff, L. *et al.* (2019) Identification of novel small molecule inhibitors for solute carrier SGLT1 using proteochemometric modeling. *J. Cheminform.* **11**, 15.

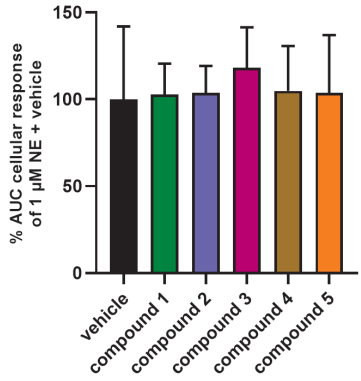
43. Rogers, D. & Hahn, M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754.

44. Chrszon, P., Dubslaff, C., Klüppelholz, S. & Baier, C. (2018) ProFeat: feature-oriented engineering for family-based probabilistic model checking. *Form. Asp. Comput.* **30**, 45–75.

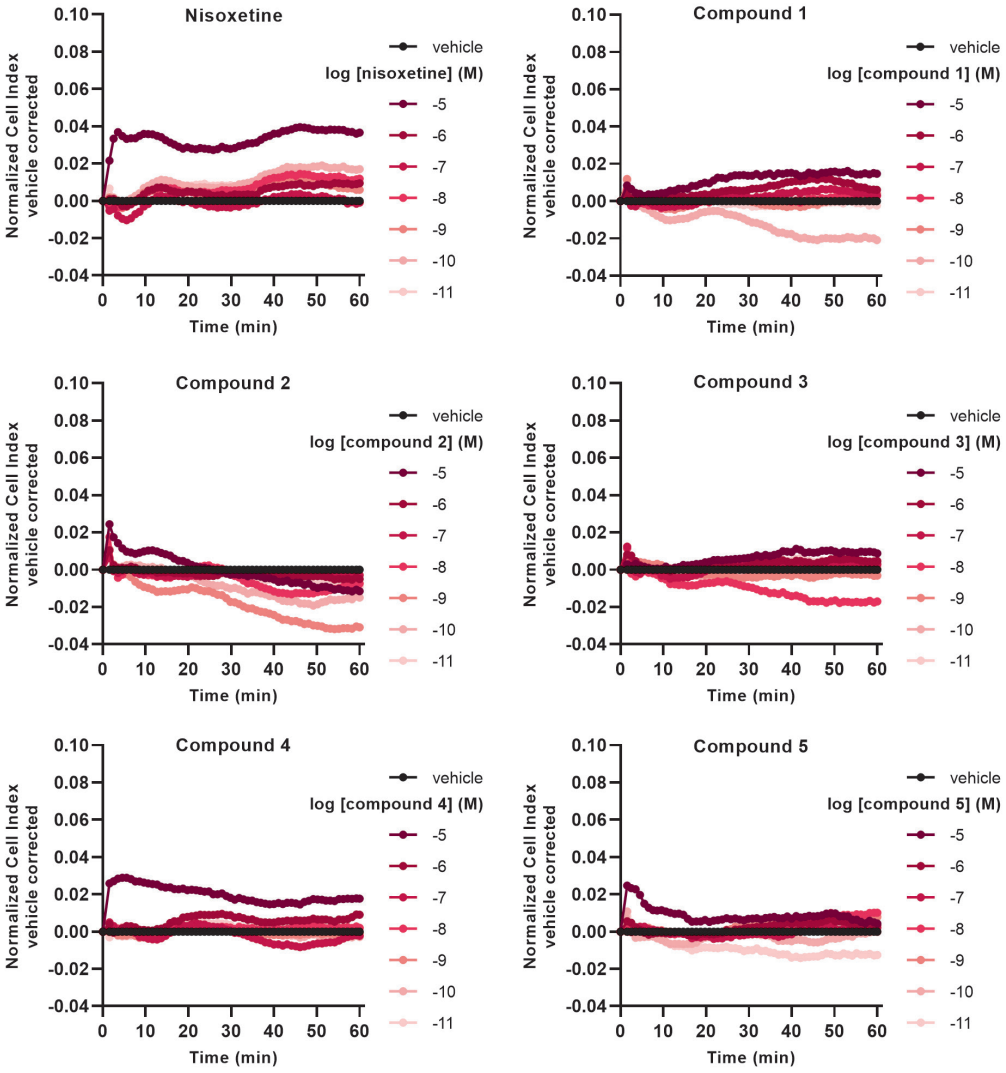
45. van Westen, G. J. *et al.* (2013) Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptor sets. *J. Cheminform.* **5**, 42.

46. Wold, S., Jonsson, J., Sjöström, M., Sandberg, M. & Rännar, S. (1993) DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. *Anal. Chim. Acta* **277**, 239–253.

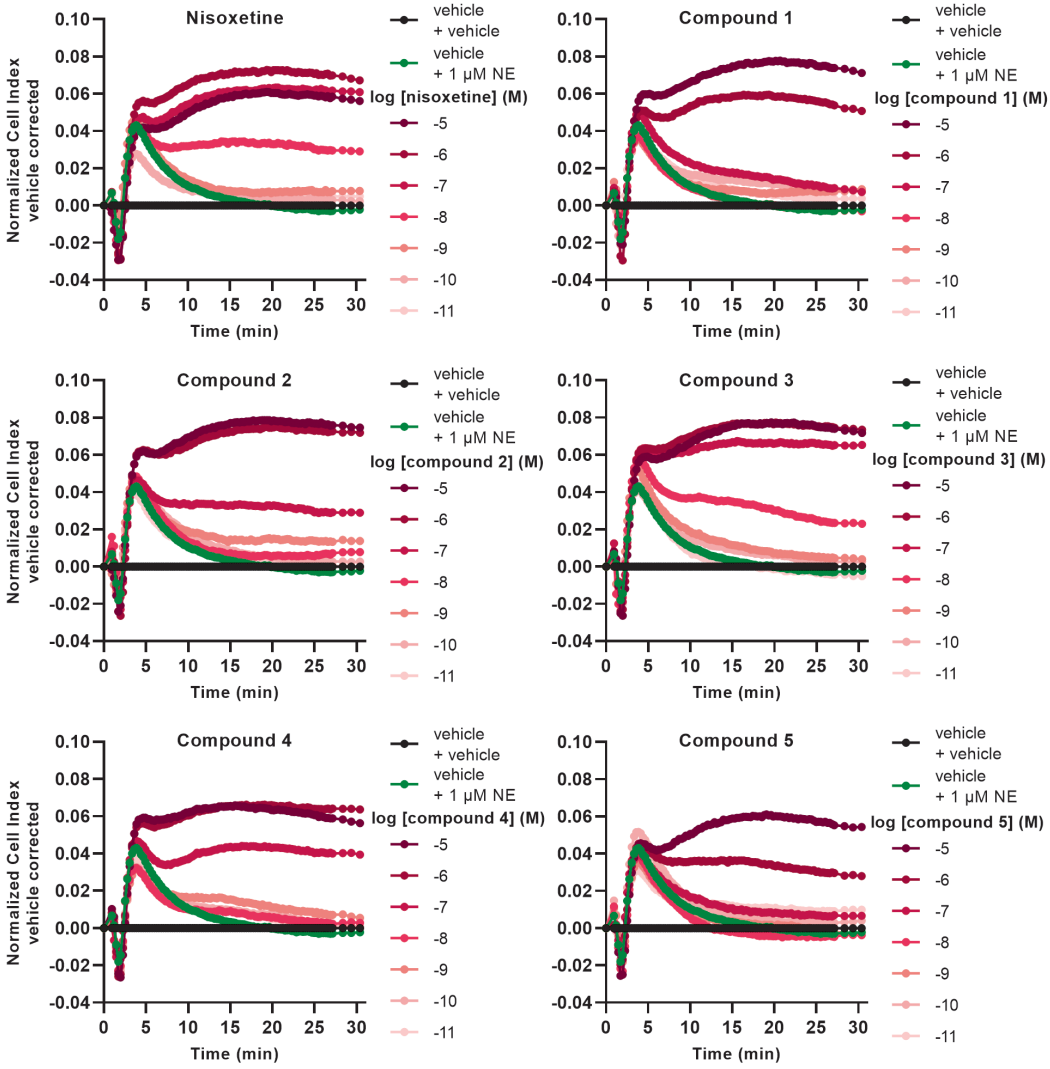
Supplementary Information



Supplementary Figure 5.S1 – Counterscreen of the five hit compounds in a label-free impedance-based TRACT assay. JumpIn-NET were not induced with doxycycline and as such did not express NET. Cells were pretreated for 1 h with either vehicle or 10 µM of the hit compound. Subsequently, cells were stimulated with 1 µM norepinephrine (NE) and Cell Index (CI) was measured for 30 min. Cellular responses are expressed as the net area under the curve (AUC) of the first 30 minutes after stimulation with NE. Data were normalized to the response of NE only (vehicle, 100%). Data are shown as the mean ± SD of two separate experiments each performed in duplicate.



Supplementary Figure 5.S2 – Representative xCELLigence traces of JumpIn-NET cells during inhibitor pretreatment in a label-free impedance-based TRACT assay. Cells were pretreated for 1 h with either vehicle or increasing concentrations of nisoxetine or hit compound. Cell Index was normalized to the time point prior to inhibitor addition (t = 0 min). Data are shown as the mean of a representative experiment.



Supplementary Figure 5.S3 – Representative xCELLigence traces of JumpIn-NET cells during norepinephrine (NE) stimulation in a label-free impedance-based TRACT assay. Cells were pretreated for 1 h with either vehicle or increasing concentrations of nisoxetine or hit compound, and subsequently stimulated with vehicle or 1 µM NE. Cell Index was normalized to the time point prior to NE addition (t = 0 min). Data are shown as the mean of a representative experiment.

Supplementary Table 5.S1 – Descriptors used throughout the model building process.

Molecular Descriptors		Protein Descriptors	
ALogP	SCFC 6	Amino Acid Composition	ACC c scales
Molecular Weight	FPFC 6	Dipeptide Composition	ACC DPPS
Number of Hydrogen Donors	EPFC 6	Auto Correlation Descriptors	ACC E scales
Number of Hydrogen Acceptors	LPFC 6	Composition Transition Distribution	ACC G scales
Number of Rotatable Bonds	SPFC 6	Quasi Sequence Order Descriptors	ACC HESH
Number of Bridge Bonds	FEFC 6	Pseudo Amino Acid Composition	ACC HSEHPCSV
Number of Atoms	EEFC 6	Amphiphilic Pseudo Amino Acid Composition	ACC Norinder
Number of Rings	LEFC 6	Total Amino Acid Properties	ACC Kidera
Number of Aromatic Rings	SEFC 6	Aligned Z scales Sandberg	ACC P scales
Number of Fragments	FHFC 6	Aligned FASGAI	ACC QCP
N Plus O Count	EHFC 6	Aligned BLOSUM	ACC Sneath
Molecular Solubility	LHFC 6	ACC Z scales Hellberg	ACC SVEEVA
Molecular Surface Area	SHFC 6	ACC Z scales Jonsson	ACC SVHEHS
Molecular Polar Surface Area	FCFP 6	ACC Z scales Sandberg	ACC SVRG
Molecular Polar Solvent-Accessible Surface Area (SASA)	ECFP 6	ACC Z scales binary	ACC SVWG
Estate Keys	LCFP 6	ACC T scales	ACC V scales
Estate Counts	SCFP 6	ACC ST scales	ACC VSGETAWAY
MDLPublicKeys	FPFP 6	ACC VHSE	ACC VSTPV
MDL2DKeys960	EPFP 6	ACC ISA ECI	ACC VSW
MDL2DKeys166	LPFP 6	ACC GRID t-score	ACC VTSA
PHFP 2-4	SPFP 6	ACC VSTV	ACC SVGER
PHRFP 2-4	FEFP 6	ACC MSWHIM	ACC PSM
PHPFP 2-4	EEFP 6	ACC_FASGAI	ACC SSIA AM1
PHFC 2-4	LEFP 6	ACC_BLOSUM	ACC SSIA PM3
PHPFC 2-4	SEFP 6	ACC_VARIMAX	ACC SSIA HF
PHRFC 2-4	FHFP 6	ACC Protein fingerprint numerical	ACC SSIA DFT
FCFC 6	EHFP 6	ACC Protein fingerprint hash	
ECFC 6	LHFP 6		
LCFC 6	SHFP 6		