Journal Pre-proof

Forecasting a clear vision: How QSAR modeling catalyzes small molecule discovery in the field of corneal endothelial regeneration

Charissa Witters , Hans De Winter , Maxim Le Compte , Martin Ondra , Debbie Le Blon , Bert Van den Bogerd , Christophe Deben , Marián Hajdúch , Carina Koppen , Hendrik Vercammen

PII: S2472-5552(26)00038-9
DOI: https://doi.org/10.1016/j.slasd.2026.100332
Reference: SLASD 100332

To appear in: *SLAS Discovery*

Received date: 6 April 2026
Revised date: 4 August 2026
Accepted date: 14 August 2026

Please cite this article as: Charissa Witters , Hans De Winter , Maxim Le Compte , Martin Ondra , Debbie Le Blon , Bert Van den Bogerd , Christophe Deben , Marián Hajdúch , Carina Koppen , Hendrik Vercammen , Forecasting a clear vision: How QSAR modeling catalyzes small molecule discovery in the field of corneal endothelial regeneration, *SLAS Discovery* (2026), doi: https://doi.org/10.1016/j.slasd.2026.100332

Highlights

- Pharmacological-based strategy to restore the corneal endothelium
- *in silico* - *in vitro* drug discovery pipeline for corneal endothelial regeneration
- QSAR model predicts compounds as lower-dose alternatives to ROCK inhibitor Y-27632
- Drug screening pipeline experimentally validates QSAR model predictions

Forecasting a clear vision: How QSAR modeling catalyzes small molecule discovery in the field of corneal endothelial regeneration

Charissa Witters[1,2,3], Hans De Winter[4], Maxim Le Compte[3,5], Martin Ondra[6], Debbie Le Blon[1,2,3], Bert Van den Bogerd[1,2,3], Christophe Deben[3,5], Marián Hajdúch[6], Carina Koppen[1,2,3], Hendrik Vercammen[1,2,3]*

Affiliations

[1]Antwerp Research Group for Ocular Science (ARGOS), Translational Neurosciences, Faculty of Medicine and Health Sciences, University of Antwerp, Wilrijk, Belgium

[2] Department of Ophthalmology, Antwerp University Hospital, Edegem, Belgium

[3] DrugVision.AI, University of Antwerp, Wilrijk, Belgium

[4] Laboratory of Medicinal Chemistry (UAMC), University of Antwerp, Wilrijk, Belgium

[5] Center for Oncological Research (CORE), University of Antwerp, Wilrijk, Belgium

[6] Laboratory of Experimental Medicine, Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University and University Hospital, Olomouc, Czech Republic

Email addresses: charissa.witters@uantwerpen.be

***Corresponding Author**

Hendrik Vercammen

Antwerp Research Group for Ocular Science (ARGOS), Translational Neurosciences, Faculty of Medicine and Health Sciences, University of Antwerp, Universiteitsplein 1, 2610 Wilrijk, Belgium.

hendrik.vercammen@uantwerpen.be

Phone no.: +32 479/ 93.61.48

Abstract

The corneal endothelium executes its crucial role in the maintenance of corneal transparency by regulating the corneal hydration levels. Dysfunction of the endothelial layer leads to corneal edema, i.e. bullous keratopathy, for which an endothelial transplantation is currently the only available treatment option. Nevertheless, the limited availability of donor tissue encourages the development of alternative treatment strategies. This study addresses a pharmacological-based alternative approach focusing on the comprehension and proof-of-concept validation of a hybrid small molecule screening approach. In this regard, a weighted voting-based quantitative structure-activity relationship (QSAR) model was developed for the *in silico* identification of small molecule building blocks promoting corneal endothelial regeneration. The established QSAR model was applied to the virtual Enamine® Hit Locator Library, composed of 460,160 compounds. Ultimately, the four most effective compounds predicted to stimulate corneal endothelial cell growth *in silico*, were selected for *in vitro* validation. All four selected compounds enhanced cell proliferation at nanomolar concentrations. Their effects were comparable to Y-27632, a ROCK inhibitor commonly used as a positive control to stimulate corneal endothelial cell proliferation and migration, supporting their potential as low-dose small molecule alternatives for corneal endothelial regeneration. These findings validate the predictive performance of our initial QSAR model and highlight the relevance for future screening of other hit locator library compounds, while feeding new biological data into the model for further refinement and accuracy improvement.

Keywords

QSAR modeling, Drug screening, Phenotypic screening, Cornea, Corneal endothelial regeneration, Live cell imaging

Abbreviations

AD, Applicability domain; AUC, Area under the curve; HCEnC-B4G12, Human corneal endothelial cell line B4G12; CADD, Computer-aided drug design; DMSO, Dimethyl sulfoxide; FECD, Fuchs' corneal endothelial dystrophy; GBR, Gradient boosting regressor; HLL, Hit Locator Library; HTS, High-throughput screening; MLR, MLP regressor model; QSAR, Quantitative structure-activity relationships; RFR, random forest regressor; RMSE, Root Mean Square Error; ROCK, Rho-associated, coiled-coil containing protein kinase; SVR, Support vector regression model

1. Introduction

The corneal endothelium is the vital hexagonal cell layer that is located at the innermost surface of our cornea [1]. This monolayer of cells ensures a crystal-clear vision by regulating corneal hydration levels via a "pump-and-leak" mechanism [2, 3]. Dysfunction of the corneal endothelial layer leads to fluid accumulation in the corneal stroma together with the formation of epithelial bullae, also known as bullous keratopathy [4]. This pathological condition can lead to impaired vision, along with pain from epithelial bullae rupture, corneal scarring, and vascularization [5]. The primary cause of corneal endothelial dysfunction is Fuchs' corneal endothelial dystrophy (FECD), a genetic disorder that affects the extracellular matrix of the corneal endothelium. The global prevalence of FECD has been estimated at 7.33% among adults and is expected to increase significantly due to the aging population [6]. Today, a diseased or damaged corneal endothelium can only be treated by means of a corneal endothelial transplantation, which requires advanced microsurgical techniques [7]. However, the availability of this treatment is limited by a global donor shortage, further complicated by insufficient eye banks and cultural barriers concerning the retrieval of human cadaveric eyes [8]. The shortage of donor tissue forces R&D into the exploration of alternative treatment solutions, including pharmacological-based strategies, which aim to restore the endothelium without the need for transplant tissue. A widely described example in the literature is treatment with Y-27632, a selective small molecule inhibitor of Rho-associated, coiled-coil containing protein kinase (ROCK). Y-27632 stimulates the proliferation and migration of corneal endothelial cells, either as a supplement for cell-based therapies or as an additive in cell culture, underscoring the promise of pharmacological approaches, more specifically the development of standalone small molecule therapies to enhance the corneal endothelium's endogenous regenerative ability [9, 10].

Traditionally, potential candidate compounds can be identified during early small molecule drug development, often achieved via a high-throughput screening (HTS) approach. Those experimental HTS offer the possibility to screen large compound libraries in an automated and robust manner [11]. However, HTS has practical limitations. On the one hand, this approach necessitates a lab equipped with an integrated platform that combines advanced machinery, enabling screening of large compound libraries [12]. The upfront investments for these robotic instruments are high, with expenses totaling up to 0.5 million euros for a single instrument [13-15]. On the other hand, it necessitates the storage and availability of synthesized HTS libraries. These libraries contain thousands of compounds, hence storage and quality control of the molecules require an adequate and advanced compound management and storage system [16, 17].

Therefore, *in silico* drug design, referred to as computer-aided drug design (CADD), can serve as a cost-effective and time-efficient alternative compared to the conventional HTS approach. Common ligand-based approaches include pharmacophore modeling and quantitative structure-activity relationships (QSAR) using cheminformatics-based statistical models that employ machine learning methods [18]. QSAR modeling is a computational method for predicting the biological or chemical activity of molecules based on their chemical structure. A validated QSAR model can serve as an *in silico* screening filter, enabling the identification of hits within large, virtual compound libraries. This strategy is not intended to replace *in vitro* assays, but rather as a complementary method to facilitate the drug discovery

process by reducing the number of drug candidates to be tested in an experimental setup [19]. The current machine learning methods used to build a QSAR model range from straightforward linear regressions to more advanced neural networks and are based on either unsupervised or supervised learning approaches. [20]. However, each type of machine learning model is characterized by its own strengths and weaknesses. In that respect, consensus modeling emerged as a promising strategy to overcome the limitations of a single QSAR model, by combining the predictive power of individual machine learning methods into one final predictive model [21]. Furthermore, weighted voting classifiers are considered a specific type of consensus modeling. This approach tags the individual QSAR's with a specific weight, based on its predictive performance [22].

QSAR models find their application in different stages of small molecule development, from predicting the biological outcome of robust assays in early stage hit discovery to ADME and cytotoxicity predictions during lead optimization [23-26]. Overall, the applicability domain (AD) of a specific QSAR model is inherently bounded by the (physico)chemical, structural or biological space in which the QSAR model can make reliable biological predictions. These boundaries are defined by the molecular features of the compounds in the training set. AD analysis is an important tool to safeguard the reliable application of a QSAR model [27, 28]. Different QSAR applications, involving different endpoints to be predicted, e.g. biological activity, toxicity, etc., require their own training data to extend and adjust the chemical space for relevant and adequate predictive performance. Hence, each application necessitates its own suitable QSAR model to reach the complexity of the research question they need to address [29, 30]. Predictive modeling approaches are gaining increasing interest within the field of ophthalmology. For example, QSAR models have been applied to predict relevant properties of ophthalmic anti-infective agents, supporting the development and optimization of new therapeutics for ocular infections [31]. Moreover, computational prediction approaches have also been explored beyond QSAR-based applications in small molecule development, with deep learning-based prediction models showing potential for improving ophthalmic diagnosis and disease classification, including, for instance, glaucoma stage classification and the identification and classification of eye disorders based on fundus images [32, 33]. Together, these studies illustrate the growing impact of computational prediction models across ophthalmic research and highlight their potential to address diverse biological and clinical challenges. Here, we have developed a QSAR model trained to identify small molecules that enhance corneal endothelial regeneration. We suggest a valuable strategy within early small molecule discovery that has the potential to substantially accelerate the transition from hit to lead optimization in pre-clinical research applied on corneal endothelial regeneration.

2. Materials and methods

2.1 *In silico* modeling

2.1.1 Data curation

The data set used in this study for training and validation of the QSAR model comprises a proprietary and self-assembled small molecule database of 2066 compounds, along with corresponding experimental data. This includes endpoint cell counts from an image-based proliferation assay following treatment at 50 nM concentration. Initially, the quality of the curated experimental data was validated through calculation of a robust Z-score (> 0.56), which is defined as:

$$Z' \text{ robust} = 1 - \frac{3 * \text{MAD (Negative Control)} + \text{MAD (Positive Control)}}{|\text{Median (Negative Control)} - \text{median (Positive control)}|}$$

MAD: Median absolute deviation

Salts and water molecules were removed from the structures, and compounds containing inorganic elements were excluded from the dataset. Compounds were converted into their SMILES representation and duplicate structures were subsequently discarded.

2.1.2 Calculation of molecular descriptors

To characterize the physico-chemical properties of the compounds, SMILES representations were converted into binary molecular fingerprints, representing the presence or absence of a unique structural feature or molecular fragment of the molecules. In this project, we chose MACCS keys as the molecular descriptor. These public keys comprise 166 bits implemented in the open source cheminformatics software package RDKit [34].

2.1.3 QSAR model building and validation

All model training and validation were done with SciKitLearn open source version 1.5.0. As an initial data quality assessment step supporting the definition of the QSAR model applicability domain, the biological activity space was evaluated to identify putative outliers within the dataset. Datapoints were clustered using a Kernel Density Estimation function, and compounds with readouts lower than three standard deviations below the mean were classified as "inadequate", whereas the remaining compounds were labeled as "good". A random forest classification model was subsequently generated to distinguish between "good" and "inadequate" compounds. The best classification model achieved a 5-fold cross-validated accuracy of 80.7% [35].

Next, four different regression models were developed to predict biological readouts: (1) an MLP regressor model (MLR) with a single hidden layer of 450 neurons and ReLu activation function, (2) a support vector regression model (SVR) with polynomial kernel, (3) a gradient boosting (GBR) regressor, and (4) a random forest (RFR) regressor model. The predictive performance of these four individual regression models was evaluated through internal and external validation. Specifically, the curated experimental dataset was randomly divided into a training set (70%) and an independent external test set (30%). For internal validation, the training set was subjected to ten-fold cross-validation, in which 90% of the training data was used for model fitting, i.e. training and the remaining 10% was used for validation in each iteration. Model performance was assessed based on the root mean square error (RMSE), regression coefficient (R^2), and cross-validation coefficient (Q^2). These performance metrics were calculated for each iterative training phase. In contrast, external validation was performed by evaluating model performance on the independent external test set that was not used during model training. To improve the overall predictive performance, a weighted voting consensus model was subsequently developed by combining the predictions of the individual regression models. Model-specific weights were optimized through hyperparameter optimization, resulting in final weights of 0.1, 1.0, 1.0, and 1.0 for the MLR, SVR, GBR, and RFR models, respectively. The predictive performance of the final weighted voting consensus model was subsequently evaluated using the same performance metrics.

2.1.4 Hit locator library

The QSAR consensus model was designed to predict the biological activity of the Hit Locator Library (HLL, Enamine, Kyiv, Ukraine) [36]. First, the classification model was applied to separate "good' from "inadequate" compounds, and then the consensus regression model was used to predict biological outcomes. This commercial library encompasses a chemical diversity set of 460,160 compounds. To enhance the quality of the HLL, the supplier performed sequential medchem filtering and pan assay interference compound filtering to obtain compounds with a drug-like physico-chemical profile and to reduce the number of compounds that tend to interfere with the biological assay, respectively. In this respect, the HLL library stands out for its 'lead-like' character, enhancing the synthetic feasibility of the compounds [36, 37].

2.2 *In vitro* screening

2.2.1 Cell culture

An immortalized corneal endothelial cell line (HCEnC-B4G12) was purchased from the Deutsche Sammlung von Mikroorganismen und Zellkulturen Institute (Braunschweig, Germany) and cultured in human endothelial serum-free medium (Thermofisher Scientific, MA, USA) supplemented with 10 ng/mL basic fibroblast growth factor (bFGF, Life technologies, California, USA). Culture flasks were coated with a fibronectin-collagen mix (FNC coating mix®, Athena Enzyme Systems, Baltimore, USA) to promote cell attachment. Prior to the drug screening assays, cells were cultured for at least two weeks to establish stable growth. One day prior to drug treatment, 384 well plates (PerkinElmer cell carrier black clear, MA, USA) were coated with FNC coating mix using the Opentrons OT-2 pipetting robot (Opentrons Labworks, Inc., New York City, USA), dispensing 20 µL FNC coating/well, followed by 1 min. incubation step at 37°C followed by manual aspiration of the remaining coating. The cells, up to passage number twenty, were then seeded into the FNC-coated 384 well plates with the OT-2 pipetting robot at a density of 2.25 x 10^3 cells per well in a volume of 50 µL human endothelial serum-free medium. The 384 well plates were incubated overnight prior to drug treatment the next day.

2.2.2 QSAR hit selection

To validate the QSAR model, four of the 460,160 HLL compounds were purchased from Enamine®, containing 30 µL of 10 mM stock concentration dissolved in dimethyl sulfoxide (DMSO). To evaluate the AD of the QSAR model, the four experimentally validated compounds were projected into the high-dimensional chemical space represented by the QSAR training dataset using t-distributed stochastic neighbor embedding (t-SNE) for visualization and comparative analysis. Additionally, pairwise Tanimoto similarity scores were calculated to assess the structural similarity between the compounds.

2.2.3 Proliferation assay

An image-based phenotypic proliferation assay was performed on HCEnC-B4G12 cells. Cells were seeded into 384-well plates as described above. The Tecan D300e drug dispenser (Tecan, Männedorf, Switzerland) was used to dispense ten different concentration levels, ranging from 50 µM to 5 nM. Subsequently, the 384 well plates were imaged at day zero, i.e. immediately after drug treatment, and every other day for a measuring period of five days by the Tecan Spark Cyto plate reader (Tecan, Männedorf, Switzerland) at 37 °C / 5% CO_2 (4x objective; whole-well imaging acquisition mode) to obtain time-lapse brightfield images. The brightfield images were automatically processed using the brightfield label-free cell counting software of the Tecan Spark (Tecan, Männedorf, Switzerland). Cell counts were expressed relative to the DMSO control condition. Area under the curve (AUC) values (cells x day) were calculated, representing the growth trajectories. The experimental design included a DMSO 0.5% basal control condition, resembling the maximum DMSO solvent limit of the four HLL compounds, and staurosporine (Selleck chemicals, Houston, US), as a negative control condition. Moreover, the Y-27632 ROCK inhibitor was used as a positive control [38].

2.2.4 Statistics

All data was collected in three independent biological replicates, combined with at least a technical triplicate. Statistical analysis was performed using the Graphpad Prism v.8 software (GraphPad, San Diego, California, USA). Data was tested for normality using following statistical tests: (1) D'Agostino and Pearson test, (2) Anderson-Darling test, (3) Shapiro-Wilk test and (4) Kolmogorov-Smirnov test. Datasets that passed normality testing were statistically compared using a one-way ANOVA analysis. In contrast, data that did not met the requirements for parametric testing, were compared through a non-parametric Kruskal Wallis test. Data is represented in the figures as mean ± standard deviation or median ± 95% confidence interval. A p-value < 0.05 was considered statistically significant. The different significancy levels were displayed in the figures as *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$ and ****$p < 0.0001$.

3. Results

3.1 *In silico* HTS screening identifies hits stimulating corneal endothelial cell growth

The QSAR model was developed using a proprietary small molecule training dataset comprising structurally diverse and known compounds with lead-like properties. *In silico* screening of 460,160 HLL virtual compounds yielded predicted biological activities, specifically HCEnC-B4G12 cell counts five days post-treatment. The prediction set ranged between 5,798 and 7,626 cells, with a mean cell count of 6,728 ± 171.60 cells. Next, predictive accuracy was validated by comparing the QSAR-predicted cell counts with two experimental control conditions of the training set (Figure 1A). The DMSO solvent baseline showed a mean experimental cell count of 6,357 ± 773.30 cells, while the 2 µM staurosporine resulted in a mean experimental cell count of 577 ± 77.37 cells, representing the cytotoxic negative control condition. Notably, a smaller subset of predictions exceeded the DMSO control mean + 1SD, suggesting biological activities beyond the normal variability of the baseline condition. Hence, the model identified compounds predicted to enhance HCEnC-B4G12 cell proliferation, of which the top four were selected for further evaluation. To assess whether these predicted hits were positioned within the chemical space represented by the QSAR training dataset, the chemical diversity was visualized by projecting the small molecule training set into their high-dimensional chemical space (Figure 1B). The four top-ranked predicted hits were subsequently mapped onto this space, demonstrating their positioning relative to the broader training dataset. The t-SNE plot shows that the predicted compounds are positioned within the overall chemical space represented by the training data, although located at the boundaries of the represented chemical space supporting the need for experimental validation.

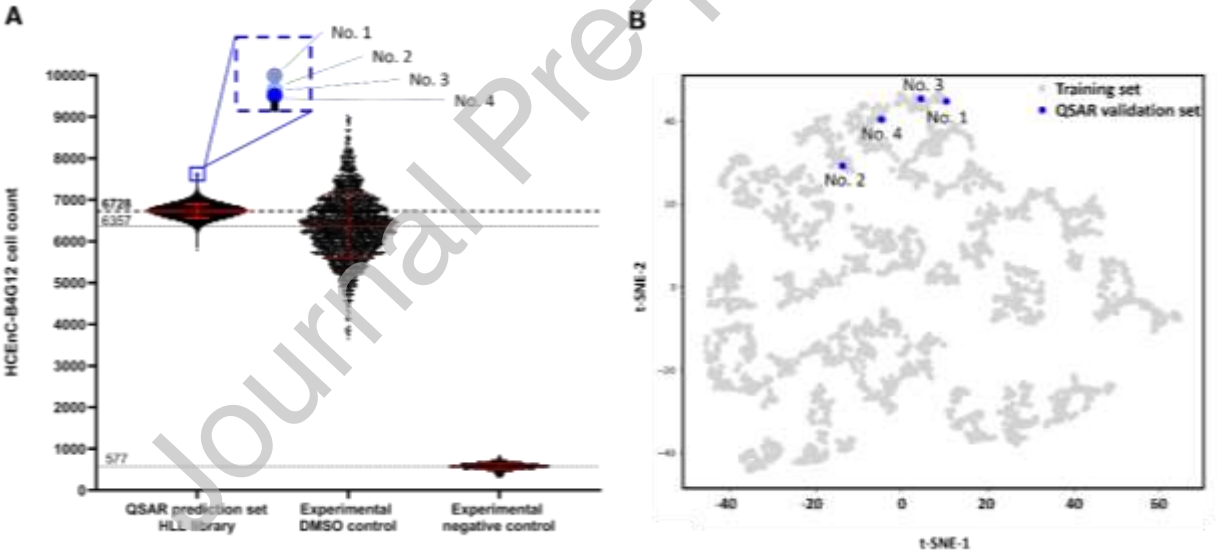
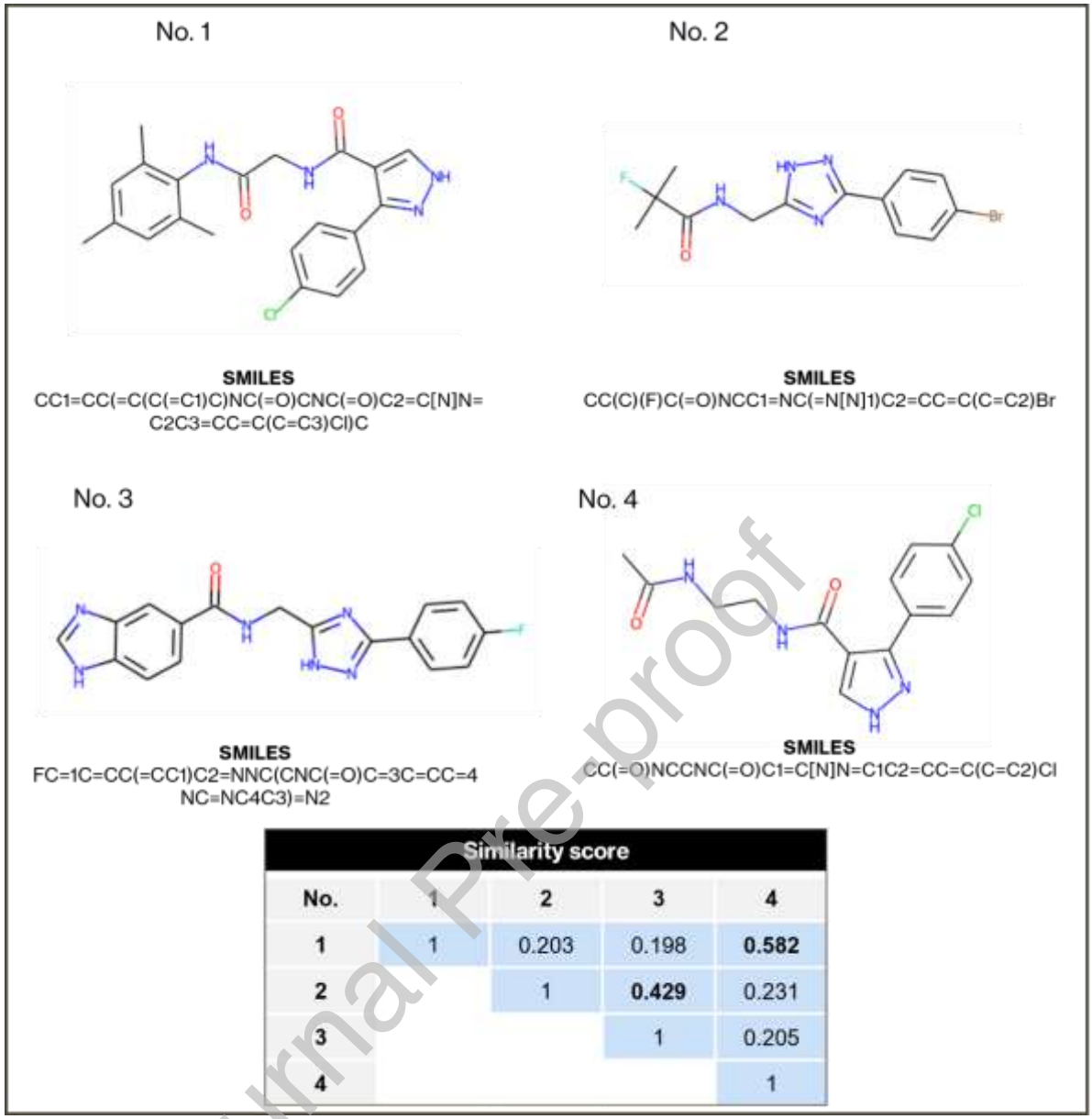


Fig. 1. *In silico* screening of the virtual hit locator library. (A) Distribution of the predicted biological activities for each of the HLL compounds plotted against an in-house set of experimental control data, i.e. a basal DMSO control and a negative control treated with staurosporine. Mean ± SD is indicated in red. The blue zoom box indicates the top four selected compounds for *in vitro* validation of the QSAR model. **(B)** t-SNE plot visualizing the experimental validation set (blue-colored data points) in the high-dimensional chemical space of the QSAR training points (grey-colored data points).



Similarity score

No.	1	2	3	4
1	1	0.203	0.198	**0.582**
2		1	**0.429**	0.231
3			1	0.205
4				1

Fig. 2: **2D chemical structures and similarity analysis of the selected experimental compounds**. Tanimoto similarity is calculated for each pair. The Tanimoto similarity score ranges from 0 to 1, where 1 indicates identical chemical fingerprints and 0 indicates no similarity. The highest similarity among different compounds was observed between compound 1 and 4 (score = 0.582). *Chemical structures are drawn with the RDKit chem.draw package and are accompanied with their SMILES notation.*

The four top-ranked compounds define two distinct types of compound groups, i.e. pyrazole-chlorobenzene urea derivatives (compound No. 1 and No. 4) and aryl-triazole hybrids (compound No. 2 and No. 3), each characterized by halogenated aromatic systems flanking a central urea or amide moiety, yielding amphiphilic architectures (Table 1). Compound No. 1 and No. 4 differ markedly at the distal terminus: a bulky mesityl group in compound No. 1 versus a minimal acetyl group in compound No. 4. Remarkably, both retain nM potency, indicating that the pyrazole-urea-aryl scaffold constitutes the core pharmacophore and that steric modulation at the periphery is well tolerated. The urea-linked framework further supports bidentate hydrogen bonding and π-π interactions, mirroring features derived from the trainings dataset of known kinase inhibitors (Figure 2). Compound No. 2 and No. 3 share a triazole-aryl-amide core. Compound No. 2 is the smallest and most flexible, while compound No. 3 features a fused benzimidazole and greater planarity. Furthermore, to assess chemical similarity among the top four predicted compounds, pairwise Tanimoto similarity scores were calculated based on their molecular fingerprints. The Tanimoto coefficient ranges from 0 (no similarity) to 1 (identical structures).

The highest degree of similarity was observed between compound No. 1 and compound No. 4 (Tanimoto score = 0.582), suggesting moderate structural overlap. In contrast, compound pairs such as compound No. 1 and compound No. 3 displayed low similarity (score = 0.198), reflecting distinct chemical architectures (Figure 2). These Tanimoto scores shows the model's ability to identify structurally varied compounds with similar phenotypic bioactivity.

Table 1. QSAR validation compounds and their aromaticity.

Compound No.	Formula	SMILES	Five-membered rings	Six-membered rings	Bicyclic rings
1	$C_{21}H_{21}ClN_4O_2$	CC=1C=C(C)C(NC(=O)CNC(=O)C2=CNN=C2C=3C=CC(Cl)=CC3)=C(C)C1	Pyrazole	Chlorobenzene, Tri methyl benzene	N/A
2	$C_{13}H_{14}BrFN_4O$	CC(C)(F)C(=O)NCC1=NC(=NN1)C=2C=CC(Br)=CC2	Triazole	Bromobenzene	N/A
3	$C_{17}H_{13}FN_6O$	FC=1C=CC(=CC1)C2=NNC(CNC(=O)C=3C=CC=4NC=NC4C3)=N2	Triazole, imidazole	Fluorobenzene	Benzimidazole
4	$C_{14}H_{15}ClN_4O_2$	CC(=O)NCCNC(=O)C1=CNN=C1C=2C=CC(Cl)=CC2	Pyrazole	Chlorobenzene	N/A

3.2 Experimental QSAR validation by a phenotypic proliferation assay

3.2.1 QSAR hits enhance HCEnC-B4G12 cell growth

To experimentally validate the performance of QSAR model, an image-based proliferation assay was performed by using HCEnC-B4G12 cells treated with the top four selected compounds from the HLL library. Three control conditions were included: 2 µM staurosporine (cytotoxic negative control), 0.5% DMSO (basal control), and 25 µM Y-27632 (ROCK inhibitor, positive control), which significantly increased proliferation relative to DMSO by 57%. The heatmap displays the endpoint cell counts normalized to baseline DMSO growth conditions (Figure 3A). A positive hit was defined as a >20% increase in proliferation compared to DMSO baseline conditions, indicated by the orange gradient. All four compounds demonstrated significant stimulation of HCEnC-B4G12 cell growth at nanomolar concentrations. Specifically, compound No. 3 showed a significant increase of 47% in HCEnC-B4G12 cells at 10 nM. Given that the QSAR model was trained on 50 nM treatment data, kinetic analysis over five days showed that all four compounds significantly increased HCEnC-B4G12 cell growth at 50 nM compared to baseline growth condition (Figure 3B). The area under the curve metric showed significantly higher growth relative to DMSO. Furthermore, all four compounds achieved biological effects comparable to those of the positive control, despite being used at lower concentrations. Brightfield imaging revealed that HCEnC-B4G12 cells adopted an elongated morphology 24 hours post-treatment, representing a transient migratory and/or proliferative response (Figure 3C). By day five, cells had re-established a dense monolayer with a characteristic hexagonal shape, consistent with the restoration of their barrier-like phenotype.

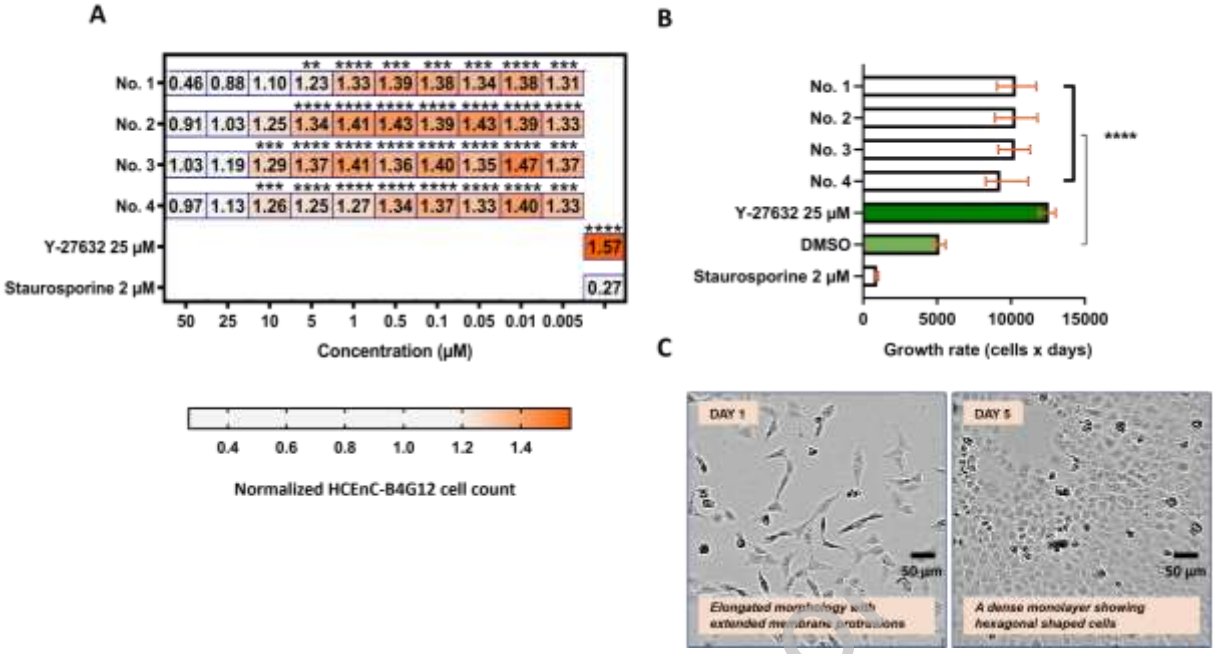


Fig. 3: Image-based phenotypic proliferation assay of HCEnC-B4G12 cells. (A) End point analysis of HCEnC-B4G12 cells growth. Normalized cell counts of HCEnC-B4G12 cells treated with lead compounds to DMSO-treated (vehicle control) cells five days post-treatment. The heatmap displays ten different compound concentrations, ranging from 50 µM to 5 nM. The values highlighted in orange represent compound conditions that exceed the DMSO control by a minimal 20% growth increase. P-values were calculated using a non-parametric Kruskal-Wallis test, **p < 0.01, ***p < 0.001 and ****p < 0.0001. **(B)** Proliferation measured as AUC of cell growth over time post 50 nM treatment of selected compounds (1–4). The AUC values (cells x day) encapsulate the five-day growth profile and are represented by the median value within a 95% confidence interval. P-values were calculated using a non-parametric Kruskal-Wallis test, ****p < 0.0001. **(C)** Brightfield images show an example of the HCEnC-B4G12 phenotype at low (day 1 – 24 hours post-treatment) and high (day 5 – endpoint) cell confluency, respectively. Scale bar: 50 µm.

3.2.2 Correlation analysis of the QSAR predicted and experimental obtained results

To evaluate the predictive accuracy of the QSAR model, predicted cell proliferation values were compared with label-free experimental cell counts five days post treatment at 50 nM (Figure 4). The comparison revealed a consistent overestimation by the QSAR model, with differences ranging from 508 to 1036 cells. As previously visualized by the t-SNE chemical space analysis, the validated compounds were positioned within the chemical space represented by the training dataset, although located towards less densely represented regions of this space. This positioning may contribute to increased prediction uncertainty and highlights the importance of further prospective experimental validation.

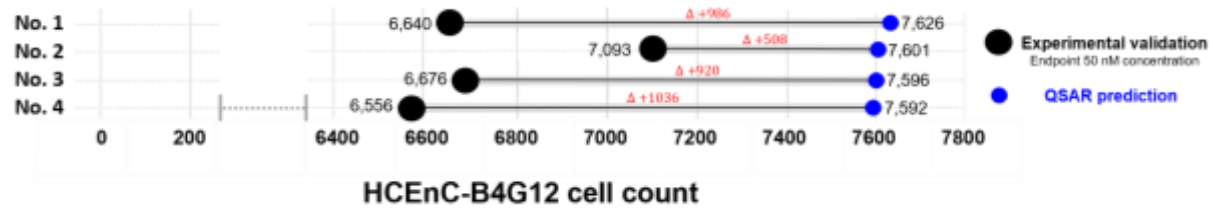


Fig. 4: **Relationship between the QSAR-predicted and experimentally obtained biological activities.** Comparison between the predicted and experimental cell counts. The cell count difference is indicated in red. Size differences between the representative circles are related to the respective datasets,

i.e. median experimental values in contrast to single prediction values. *Figure is created with Datawrapper*.

4. Discussion

This study presents a hybrid *in silico - in vitro* drug discovery pipeline that effectively integrates a weighted voting-based QSAR model with phenotypic screening to identify novel small molecules promoting corneal endothelial regeneration [39]. This drug discovery approach addresses the need for seeking non-surgical treatment alternatives to corneal endothelial transplantations in diseases such as Fuchs' corneal endothelial dystrophy [6]. An ensemble QSAR model was developed combining four machine learning regressors (MLR, SVR, GBR, RFR), optimized through a weighted voting scheme [40]. This consensus approach improves predictive performance compared to individual models, demonstrating its robustness in navigating through a high-dimensional chemical space [41, 42]. Furthermore, a self-made experimental trainings dataset of over 2000 lead-like compounds was used to build and train the QSAR model instead of relying on public available datasets to ensure translational and biological relevance. Internal and external model validations ensured reliable activity predictions across the Enamine® Hit Locator Library (460,160 compounds), from which the top four compounds were selected [43-45]. Importantly, the AD of the QSAR model is inherently limited by the chemical and biological space represented in its training dataset. Consequently, new chemical structures must be reasonably similar to compounds within the training dataset for reliable predictions to be expected. To maximize the AD of the developed model, the initial training dataset of this study was therefore designed to encompass a diverse range of chemical scaffolds and compound classes, including regeneration-related chemotypes such as ROCK inhibitors [46].

All four selected hits enhanced proliferation of HCEnC-B4G12 corneal endothelial cells with nanomolar potency (< 10 nM) five days after treatment, matching the benchmark ROCK inhibitor Y-27632 which is active in micromolar range [10]. Two distinct types of compounds were identified: pyrazole-chlorobenzene urea derivatives (compound No. 1 and No. 4) and aryl-triazole hybrids (compound No. 2 and No. 3). Importantly, all four compounds promoted a corneal endothelial-like morphology at their active concentrations, characterized by the formation of a dense monolayer and hexagonal-shaped HCEnC-B4G12 cells. However, as the current study employed a phenotypic screening strategy, the observed biological effects were evaluated without identification of the underlying molecular targets. Therefore, the relationship between the identified chemical structures and their biological activity, including potential binding specificity, remains to be further elucidated. Future studies will focus on mechanistic validation using target-specific assays, including biochemical kinase profiling and multi-omics approaches (e.g., transcriptomics), to determine whether these compounds exert their effects through ROCK-dependent mechanisms, modulation of upstream and/or downstream ROCK signaling pathways, or alternative signaling cascades that promote corneal endothelial regeneration. Nevertheless, the composition of the Enamine® Hit Locator Library should be considered when interpreting the identified hits. As this library contains a substantial number of kinase-focused molecules, the enrichment of this subset of molecules may have influenced the chemical space explored during virtual screening. Furthermore, the four identified hits possess structural features resembling kinase hinge binders, suggesting potential similarity with kinase-binding chemotypes [47].

The pipeline offers a drug discovery strategy to overcome the limitations of corneal donor shortages and invasive surgeries, thereby seeking towards a pharmacological donor independent treatment [1]. Pharmacological regeneration via topical or intracameral application of such compounds could serve as a minimally invasive therapy, with potential synergy in combination with ROCK inhibitors or cell-based approaches [10]. Given that the QSAR model was trained on 50 nM treatment data, kinetic analysis over five days showed that all four compounds significantly increased HCEnC-B4G12 cell growth at 50 nM compared to baseline growth condition, which confirms the performance of the model. The integration of QSAR-driven screening with high-throughput, image-based phenotypic assays represents a scalable discovery process to identify bioactive small molecules [48]. In this regard, this model makes it possible to virtually screen large compound libraries for computationally guided hit identification and drastically reduces experimental costs and risk of failures compared to traditional HTS methodologies.

However, several limitations need to be addressed in the future to further strengthen the applicability of this QSAR-based hit identification approach. First, the current work was designed as a low-cost proof-of-concept validation of the proposed hit discovery strategy, which included only four experimentally validated compounds. Therefore, extensive prospective validation will be critical to provide stronger evidence supporting the model's predictive performance. Such validation can be addressed by experimentally testing additional HLL compounds, spanning the full prediction range, including compounds with intermediate and low predicted activity scores, as well as structurally related analogues predicted to have reduced activity. Hence it will further assess the model's discriminatory capacity and robustness. Second, the use of a single immortalized corneal endothelial cell line (HCEnC-B4G12) represents another limitation of the current experimental validation. On the one hand, validating the identified hits in additional corneal endothelial cell models, such as the immortalized cell line HCEnC-21(T), would strengthen the biological validation by providing orthogonal screening systems and reducing the likelihood of model-specific false-positive or false-negative findings [49]. On the other hand, although immortalized cell lines provide robust, reproducible, and scalable models that are well suited for hit identification, the HCEnC-B4G12 cell line has inherent limitations as a model for corneal endothelial cell regeneration. This includes its pleomorphic phenotype during routine culture, its immortalized nature, and the absence of the native corneal tissue architecture. Therefore, future hit confirmation in primary human corneal endothelial cells and *ex vivo* human corneal models will be essential to confirm the biological relevance and translational applicability of the identified compounds [50]. Moreover, these additional biological datasets can be incorporated into future iterative QSAR model refinement, further improving the predictive robustness and generalizability of the developed prediction model.

Following biological and mechanistic validation of the identified hit compounds, subsequent optimization efforts may focus on structure-activity relationship (SAR) studies in which phenotypic morphological readouts are systematically integrated with target-related biological data to guide iterative compound optimization. This SAR-driven approach can be complemented by scaffold decoration and other medicinal chemistry strategies to generate novel analogues and chemically distinct derivatives with improved potency, selectivity, and drug-like properties. In addition, advanced physiologically relevant models, such as cornea-on-a-chip platforms, may provide a more dynamic and biologically complex testing environment for evaluating candidate performance, thereby supporting hit-to-lead optimization and increasing the translational relevance of selected hit compounds [51-53].

Finally, future QSAR model refinements will consist of the expansion of the training set, adoption of graph neural networks in complex unstructured data, and integration of transcriptomic and/or proteomic response profiles to enable mechanism-informed predictions [18, 54].

5. Conclusion

This study presents a robust methodology that bridges computational modeling and experimental validation to accelerate small molecule discovery for corneal endothelial regeneration applications. By combining a weighted voting-based QSAR model with phenotypic screening, we identified four chemically distinct small molecules that effectively stimulated corneal endothelial cell proliferation at nanomolar concentrations. Their performance matched the performance of Y-27632 at micromolar range, a widely investigated ROCK inhibitor candidate for corneal endothelial regeneration. These primary hits promoted HCEnC-B4G12 cell growth five days after treatment and induced the desired morphological phenotype characterized by hexagonal barrier-like structures. While the validated QSAR model initially overestimated the cell count, it will be refined through iterative testing. Nonetheless, experimental validation confirmed the compounds' bioactivity, underscoring the value of consensus-based modeling in early-stage drug discovery. Beyond ophthalmology, this pipeline highlights a scalable, adaptable approach to target tissue regeneration.

Declaration of Conflicting Interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be considered as a potential conflict of interest. All authors have read the journal's policy on conflicts of interest and authorship agreement.

Data availability

Data will be made available on request.

Funding

This study was supported by the Funds Scientific Research Flanders (FWO) via a strategic basic research grant (project No. 1SH3T24N) that CW received and a fundamental research project (project No. G073623N). Furthermore, this research project was also supported by the Institute of Molecular and Translational Medicine (IMTM) and by the Ministry of Education, Youth and Sports of the Czech Republic (European Infrastructure for Translational Medicine — EATRIS-CZ, Project No. LM2023053) and the National Institute for Cancer Research — Program EXCELES, ID Project No. LX22NPO5102, Funded by the European Union — Next Generation EU. The publication costs are supported by the SLAS2025 International conference and exhibition student poster award.

ORCID iD

Charissa Witters https://orcid.org/0009-0000-1499-2790

Acknowledgements

The authors thank the Institute of Molecular and Translational Medicine (IMTM) for the strong and sustainable collaboration. The authors also thank the SLAS Society for Laboratory Automation and Screening.

References

1. Price, M.O., et al., *Corneal endothelial dysfunction: Evolving understanding and treatment options.* Prog Retin Eye Res, 2021. **82**: p. 100904.
2. Li, S., E. Kim, and J.A. Bonanno, *Fluid transport by the cornea endothelium is dependent on buffering lactic acid efflux.* Am J Physiol Cell Physiol, 2016. **311**(1): p. C116–26.
3. Bonanno, J.A., *Molecular mechanisms underlying the corneal endothelial pump.* Exp Eye Res, 2012. **95**(1): p. 2–7.
4. Kaur, A., et al., *Indications and types of keratoplasties: Trends in the past 10 years (2011-2020) in Eastern India.* Indian J Ophthalmol, 2023. **71**(9): p. 3166–3170.
5. Feizi, S., *Corneal endothelial cell dysfunction: etiologies and management.* Ther Adv Ophthalmol, 2018. **10**: p. 2515841418815802.
6. Aiello, F., et al., *Global Prevalence of Fuchs Endothelial Corneal Dystrophy (FECD) in Adult Population: A Systematic Review and Meta-Analysis.* J Ophthalmol, 2022. **2022**: p. 3091695.
7. Price, M.O., et al., *EK (DLEK, DSEK, DMEK): New Frontier in Cornea Surgery.* Annu Rev Vis Sci, 2017. **3**: p. 69–90.
8. Gain, P., et al., *Global Survey of Corneal Transplantation and Eye Banking.* JAMA Ophthalmol, 2016. **134**(2): p. 167–73.
9. Peh, G.S.L., et al., *Effects of Rho-Associated Kinase (Rock) Inhibitors (Alternative to Y-27632) on Primary Human Corneal Endothelial Cells.* Cells, 2023. **12**(9).
10. Kinoshita, S., et al., *Injection of Cultured Cells with a ROCK Inhibitor for Bullous Keratopathy.* N Engl J Med, 2018. **378**(11): p. 995–1003.
11. Hughes, J.P., et al., *Principles of early drug discovery.* Br J Pharmacol, 2011. **162**(6): p. 1239–49.

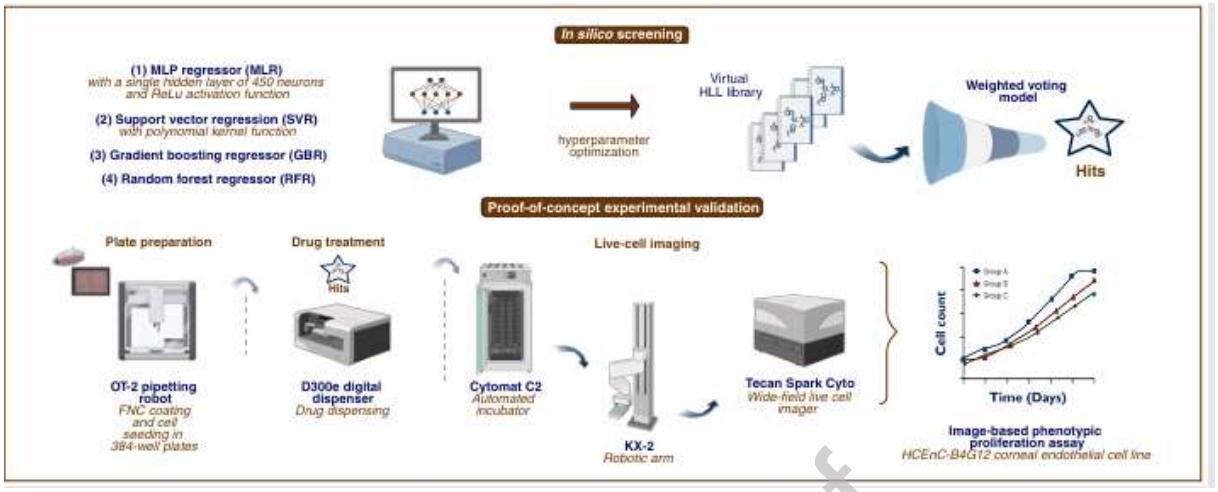
12. Michael, S., et al., *A robotic platform for quantitative high-throughput screening.* Assay Drug Dev Technol, 2008. **6**(5): p. 637–57.
13. Blay, V., et al., *High-Throughput Screening: today's biochemical and cell-based approaches.* Drug Discov Today, 2020. **25**(10): p. 1807–1821.
14. Albukhari, F.F.B.a.A., *Design and Implementation of High Throughput Screening Assays for Drug Discoveries*, in *High-Throughput Screening for Drug Discovery*, S.K. Saxena, Editor. 2021.
15. William Downey, D.H.N.a.D.L.S. *Needs and challenges in High Throughput Screening*. 2007 [cited 2024; Available from: https://www.ddw-online.com/needs-and-challenges-in-high-throughput-screening-554-200712/#:~:text=High%20throughput%20screening%20(HTS)%20groups,operate%20under%20tight%20spending%20budgets.
16. Baillargeon, P., et al., *Monitoring of HTS compound library quality via a high-resolution image acquisition and processing instrument.* J Lab Autom, 2011. **16**(3): p. 197–203.
17. Yasgar, A., et al., *Compound Management for Quantitative High-Throughput Screening.* JALA Charlottesv Va, 2008. **13**(2): p. 79–89.
18. Chang, Y., et al., *A Guide to In Silico Drug Design.* Pharmaceutics, 2022. **15**(1).
19. Wang, Y.L., et al., *Web-Based Quantitative Structure-Activity Relationship Resources Facilitate Effective Drug Discovery.* Top Curr Chem (Cham), 2021. **379**(6): p. 37.
20. Niazi, S.K. and Z. Mariam, *Recent Advances in Machine-Learning-Based Chemoinformatics: A Comprehensive Review.* Int J Mol Sci, 2023. **24**(14).
21. Valsecchi, C., et al., *Consensus versus Individual QSARs in Classification: Comparison on a Large-Scale Case Study.* J Chem Inf Model, 2020. **60**(3): p. 1215–1223.
22. Rodríguez, L.I.K.J.J., *A weighted voting framework for classifiers ensembles.* Knowledge and Information Systems, 2012. **38**: p. 259–275.
23. Siramshetty, V., et al., *Validating ADME QSAR Models Using Marketed Drugs.* SLAS Discov, 2021. **26**(10): p. 1326–1336.
24. Khan, M.T. and I. Sylte, *Predictive QSAR modeling for the successful predictions of the ADMET properties of candidate drug molecules.* Curr Drug Discov Technol, 2007. **4**(3): p. 141–9.
25. Halder, A.K., A.H.S. Delgado, and M. Cordeiro, *First multi-target QSAR model for predicting the cytotoxicity of acrylic acid-based dental monomers.* Dent Mater, 2022. **38**(2): p. 333–346.
26. Nagai, J., et al., *QSAR Prediction Model to Search for Compounds with Selective Cytotoxicity Against Oral Cell Cancer.* Medicines (Basel), 2019. **6**(2).
27. Sahigara, F., et al., *Comparison of different approaches to define the applicability domain of QSAR models.* Molecules, 2012. **17**(5): p. 4791–810.
28. Muratov, E.N., et al., *QSAR without borders.* Chem Soc Rev, 2020. **49**(11): p. 3525–3564.
29. Jaworska, J., N. Nikolova-Jeliazkova, and T. Aldenberg, *QSAR applicabilty domain estimation by projection of the training set descriptor space: a review.* Altern Lab Anim, 2005. **33**(5): p. 445–59.
30. Mora, J.R., et al., *Rethinking the applicability domain analysis in QSAR models.* J Comput Aided Mol Des, 2024. **38**(1): p. 9.
31. Rasheed, M.W., A. Mahboob, and I. Hanif, *On QSAR modeling with novel degree-based indices and thermodynamics properties of eye infection therapeutics.* Front Chem, 2024. **12**: p. 1383206.
32. Uyar, K., M. Yurdakul, and Ş. Taşdemir, *Abc-based weighted voting deep ensemble learning model for multiple eye disease detection.* Biomedical Signal Processing and Control, 2024: p. p. 106617.
33. Yurdakul, M., K. Uyar, and S. Tasdemir, *MaxGlaViT: A Novel Lightweight Vision Transformer-Based Approach for Early Diagnosis of Glaucoma Stages From Fundus Images.* International Journal of Imaging Systems and Technology, 2025. **35**(4).
34. *RDKit: Open-Source Cheminformatics Software* Available from: https://www.rdkit.org/.
35. al., Pedregosa e., *Scikit-learn: Machine Learning in Python.* JMLR 2011. **12**: p. pp. 2825–2830.

36. enamine. *Hit Locator Library - 460*. Available from: https://enamine.net/compound-libraries/diversity-libraries/hit-locator-library-460.
37. (PBR), P.b.r. *Hit Locator Library*

Enamine. 2024; Available from: https://www.pharmaceutical-business-review.com/products/hit-locator-library/.
38. Vercammen, H., et al., *"Keep on ROCKIn": Repurposed ROCK inhibitors to boost corneal endothelial regeneration.* Biomed Pharmacother, 2024. **174**: p. 116435.
39. Niazi, S.K. and Z. Mariam, *Recent Advances in Machine-Learning-Based Chemoinformatics: A Comprehensive Review.* International Journal of Molecular Sciences, 2023. **24**(14): p. 11488.
40. Chen, S. and W. Zheng, *RRMSE-enhanced weighted voting regressor for improved ensemble regression.* PLoS One, 2025. **20**(3): p. e0319515.
41. desiani, A., et al., *Weighted Voting Ensemble Learning of CNN Architectures for Diabetic Retinopathy Classification.* JURNAL INFOTEL, 2024. **16**.
42. Uyar, K., M. Yurdakul, and Ş. Taşdemir, *Abc-based weighted voting deep ensemble learning model for multiple eye disease detection.* Biomedical Signal Processing and Control, 2024. **96**: p. 106617.
43. Zhong, S., et al., *Enlarging Applicability Domain of Quantitative Structure–Activity Relationship Models through Uncertainty-Based Active Learning.* ACS ES&T Engineering, 2022. **2**(7): p. 1211–1220.
44. Ruiz, I.L. and M. Gómez-Nieto, *Study of the Applicability Domain of the QSAR Classification Models by Means of the Rivality and Modelability Indexes.* Molecules, 2018. **23**(11).
45. Hao, Y., et al., *Developing a Semi-Supervised Approach Using a PU-Learning-Based Data Augmentation Strategy for Multitarget Drug Discovery.* Int J Mol Sci, 2024. **25**(15).
46. Trinh, C., et al., *On the Development of Descriptor-Based Machine Learning Models for Thermodynamic Properties: Part 2-Applicability Domain and Outliers.* Algorithms, 2023. **16**(12).
47. Liu, S.D., et al., *Charting the evolving landscape of kinase hinge-binding chemotypes beyond flatland: a systematic analysis and opportunities for sp*

-rich binders. European Journal of Medicinal Chemistry, 2026. **306**.
48. Hughes, J., et al., *Principles of early drug discovery.* British Journal of Pharmacology, 2011. **162**(6): p. 1239–1249.
49. Schmedt, T., et al., *Telomerase immortalization of human corneal endothelial cells yields functional hexagonal monolayers.* PLoS One, 2012. **7**(12): p. e51427.
50. Tsai, M.C., A. Kureshi, and J.T. Daniels, *Establishment of an Ex Vivo Human Corneal Endothelium Wound Model.* Transl Vis Sci Technol, 2025. **14**(1): p. 24.
51. Van Meenen, J., et al., *An Overview of Advanced In Vitro Corneal Models: Implications for Pharmacological Testing.* Tissue Engineering Part B: Reviews, 2021. **28**(3): p. 506–516.
52. Macarron, R. and R.P. Hertzberg, *Design and implementation of high throughput screening assays.* Methods Mol Biol, 2002. **190**: p. 1–29.
53. Van Hoorick, J., et al., *Designer Descemet Membranes Containing PDLLA and Functionalized Gelatins as Corneal Endothelial Scaffold.* Adv Healthc Mater, 2020. **9**(16): p. e2000760.
54. Wang, C., G.A. Kumar, and J.C. Rajapakse, *Drug discovery and mechanism prediction with explainable graph neural networks.* Sci Rep, 2025. **15**(1): p. 179.

Graphical abstract



Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: