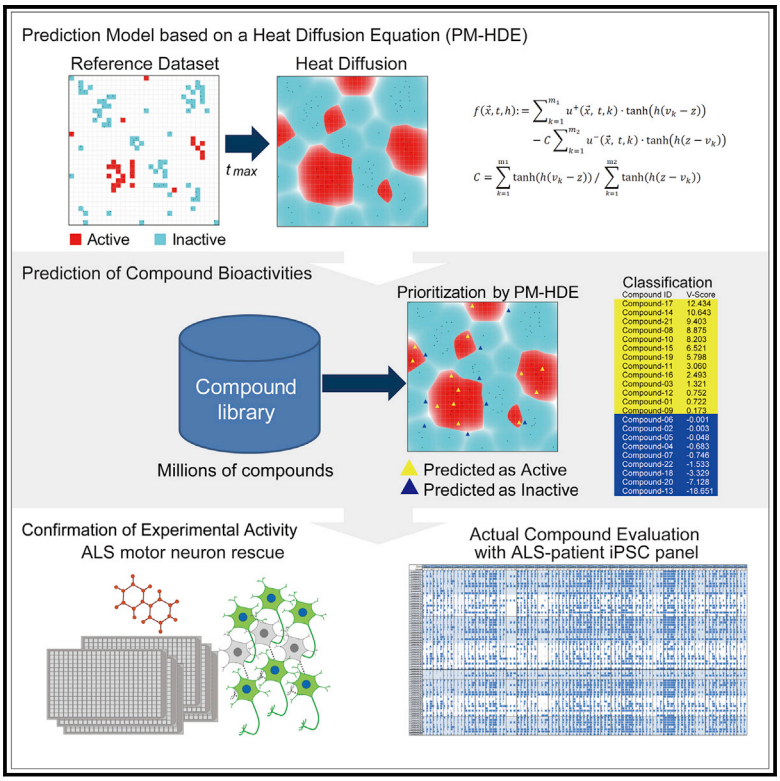
Patterns

Prediction of Compound Bioactivities Using Heat-Diffusion Equation

Graphical Abstract



Highlights

- Prediction model based on heat-diffusion equation (PM-HDE) was constructed
- PM-HDE succeeded in increasing the hit ratio and identifying potent compounds
- PM-HDE discovered new chemotypes in compound evaluation with an ALS-patient iPSC panel
- PM-HDE could represent an algorithm for future drug discovery with AI

Authors

Tadashi Hidaka, Keiko Imamura, Takeshi Hioki, ..., Takeshi Niki, Makoto Fushimi, Haruhisa Inoue

Correspondence

makoto.fushimi@takeda.com (M.F.), haruhisa@cira.kyoto-u.ac.jp (H.I.)

In Brief

Compound screening is a useful tool for discovering new candidate drugs. However, it is still a major challenge, as it is extremely time-consuming and very costly to evaluate millions of compounds. We established a prediction model based on the heat-diffusion equation (PM-HDE) to predict hits in compound screening. PM-HDE succeeded in increasing the hit ratio, identifying compounds with potent broad-spectrum efficacy, and discovering new chemotypes in actual compound evaluation in phenotypic screening.

Hidaka et al., 2020, Patterns *1*, 100140
December 11, 2020
https://doi.org/10.1016/j.patter.2020.100140

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Prediction of Compound Bioactivities Using Heat-Diffusion Equation

Tadashi Hidaka,[1,10] Keiko Imamura,[2,3,4,5,10] Takeshi Hioki,[1,3,10] Terufumi Takagi,[1] Yoshikazu Giga,[6,7] Mi-Ho Giga,[6,7] Yoshiteru Nishimura,[8] Yoshinobu Kawahara,[8,9] Satoru Hayashi,[1,3] Takeshi Niki,[2,3] Makoto Fushimi,[1,*] and Haruhisa Inoue[2,3,4,5,11,*]

[1]Research, Takeda Pharmaceutical Company Limited, Fujisawa, Japan
[2]Center for iPS Cell Research and Application (CiRA), Kyoto University, Kyoto, Japan
[3]Takeda-CiRA Joint Program (T-CiRA), Fujisawa, Japan
[4]iPSC-based Drug Discovery and Development Team, RIKEN BioResource Research Center (BRC), Kyoto, Japan
[5]Medical-risk Avoidance based on iPS Cells Team, RIKEN Center for Advanced Intelligence Project (AIP), Kyoto, Japan
[6]Graduate School of Mathematical Sciences, University of Tokyo, Tokyo, Japan
[7]Institute for Mathematics in Advanced Interdisciplinary Study, Sapporo, Japan
[8]Structured Learning Team, RIKEN Center for Advanced Intelligence Project (AIP), Fukuoka, Japan
[9]Institute of Mathematics for Industry, Kyushu University, Fukuoka, Japan
[10]These authors contributed equally
[11]Lead Contact
*Correspondence: makoto.fushimi@takeda.com (M.F.), haruhisa@cira.kyoto-u.ac.jp (H.I.)
https://doi.org/10.1016/j.patter.2020.100140

THE BIGGER PICTURE There remain many intractable diseases with no treatment available, including amyotrophic lateral sclerosis (ALS), for which the development of a cure is crucial. However, compound screening for drug development demands time, energy, and cost, and therefore artificial intelligence (AI) is expected to improve the efficiency of drug discovery. We built a novel machine-learning algorithm to predict hit compounds in compound screening using the heat-diffusion equation (HDE). This prediction model harbors the potential to solve issues that have been challenging for conventional machine learning and to exhibit accurate performance leading to the discovery of new drugs. In fact, the HDE model predicted hits with new chemotypes among millions of compounds for ALS therapeutics using a panel of large numbers of ALS patient-derived induced pluripotent stem cell models (ALS-patient iPSC panel). This algorithm could contribute to the acceleration and development of future drug discoveries using AI.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Machine learning is expected to improve low throughput and high assay cost in cell-based phenotypic screening. However, it is still a challenge to apply machine learning to achieving sufficiently complex phenotypic screening due to imbalanced datasets, non-linear prediction, and unpredictability of new chemotypes. Here, we developed a prediction model based on the heat-diffusion equation (PM-HDE) to address this issue. The algorithm was verified as feasible for virtual compound screening using biotest data of 946 assay systems registered with PubChem. PM-HDE was then applied to actual screening. Based on supervised learning of the data of about 50,000 compounds from biological phenotypic screening with motor neurons derived from ALS-patient-induced pluripotent stem cells, virtual screening of >1.6 million compounds was implemented. We confirmed that PM-HDE enriched the hit compounds and identified new chemotypes. This prediction model could overcome the inflexibility in machine learning, and our approach could provide a novel platform for drug discovery.

INTRODUCTION

During the past several years many compounds have been developed for intractable diseases, yet many diseases remain without a treatment. Compound screening, especially phenotypic screening with human induced pluripotent stem cells (iPSCs), is a useful tool for discovering new candidate drugs and disease pathways even if the disease mechanism has not

been completely clarified.[1] Nevertheless it is still a major challenge, as it is extremely time-consuming and very costly to evaluate millions of compounds. Notwithstanding these hurdles, the method for prediction of the effectiveness of compounds is expected to promote cell-based drug discovery. In the hit prediction using an inductive prediction method based on machine learning, such as quantitative structure-activity relationship (QSAR),[2,3] support vector machine (SVM),[4–6] random forest (RF),[5,7] deep learning, and clustering based on structural similarity,[8,9] compound activity can be predicted from the structure of the unassayed compound by learning the relationship between molecule descriptors and fingerprints derived from the chemical structure of the reference data and the activity value. However, due to the inherent complexity of phenotypic screening, there have been several challenges to predicting activity by conventional methods, including lack of flexibility for dealing with imbalanced datasets, non-linear prediction, and unpredictability of new chemotypes. The datasets from compound screening present an imbalanced ratio of hits/non-hits, the ratio generally being less than a few percent, with non-linear distribution, and this causes difficulty of precise prediction for conventional machine learning from compound screening. Furthermore, the prediction of new chemotypes not identified in the provided datasets is difficult for conventional machine learning, as it is difficult to determine what kind of features in chemotypes are the most important because of the lack of information related to the new chemotype.

To find further improvements regarding this issue, we established a prediction model based on the heat-diffusion equation (PM-HDE), which is a novel approach that uses a heat-diffusion equation[10] to predict hits in compound screening. The heat-diffusion equation, a partial differential equation, describes how heat conduction and material diffusion distribute in three-dimensional space over time, and this can be expanded to a multi-dimensional space for further analysis. It calculates ''active'' and ''inactive'' separately by scoring each compound and adjusts spatial integration appropriately, showing a flexibility for accurate prediction. This feature can be expected to work better than alternative conventional methods. Validation of PM-HDE was conducted using 946 PubChem datasets, after which we implemented the algorithm for actual compound screening using iPSC-based phenotypic screening for drug discovery.

RESULTS

The heat-diffusion equation, a partial differential equation, describes how heat conduction and material diffusion distribute in three-dimensional space over time. The state of heat diffusion over time was schematically shown from the initial state ($t = 0$) to the appropriate time points at which the unknown region is filled by heat diffusion (Figure 1A). The heat-diffusion equation puts each reference compound in a small region in the chemical space (hereafter referred to as ''mesh''), and they do not overlap with each other. Thereby, the impact of heat diffusion from each mesh to the whole space is clearly defined and their sum remains constant during the diffusion process. In addition, because the heat-diffusion equation calculates heat-diffusion processes of positive heat values from active compounds and negative heat values from inactive ones independently and do not interfere with each other, the ratio of the total quantity of positive heat values and that of negative heat values is always unchanged. Therefore, by adjusting the coefficient C as a reciprocal of the ratio, the total quantity of both positive and negative heat is always zero even in the case of an imbalanced active ratio. As a result, each point in the space can be predicted as active compound or inactive compound based on the plus and minus of the prediction score, described as Virtual Score (V-Score) (Figure 1B). The details of the mathematical methods are described in Experimental Procedures. In addition, since PM-HDE itself is a non-linear prediction model, it is possible to flexibly cope with complex relationships between descriptors and activity. Furthermore, by adjusting the diffusion time t, it is possible to predict a compound which is at a position away from the reference mesh group in the chemical space, that is, a compound that is not similar to the active compound contained in the reference data. Using these features, we applied the heat-diffusion equation to build a prediction model for phenotypic screening, PM-HDE, which can predict the hit compounds by formulating heat diffusion and applying molecular descriptors derived from the chemical structure into this formula to calculate the prediction score shown as V-Score. The V-Score is calculated as the temperature of heat diffusion, and compounds could be ranked based on the predicted score (Figure 1C).

Validation of PM-HDE Using Multiple Bioassay Datasets

First, 281 molecular descriptors were extracted from a total of 1,273 molecular descriptors to avoid imbalance of specific descriptors and delete invalid descriptor groups (Figure 2A and Table S1). To confirm the feasibility of PM-HDE, we constructed a prediction model for various assay data obtained from PubChem and performed cross-validation to confirm that PM-HDE can solve the issues of the conventional methods. Biotest data of 946 assay systems registered with PubChem were used to validate the predictive performance of PM-HDE in compound screening. The maximum area under the curve (AUC) of 946 assay systems was determined according to the procedure for parameter setting for PM-HDE (Figure 2B). To evaluate the prediction performance, we performed cross-validations on many datasets and calculated each AUC of the receiver-operating characteristics (ROC) curves (Table S2). The relationship between the proportion of active labels and AUC is shown in Figure 2C, suggesting that PM-HDE presented high prediction performance, showing high AUC levels even in datasets with a bias of the active ratio. Even large-sized data could be predicted with high accuracy (Figure 2D). The 946 assay systems we analyzed were mostly via high-throughput screening. However, these were very diverse in that they used enzymes, cells, and nucleic acids as targets, detection methods such as fluorescence, luminescence, and binding assays, focused on inhibition and promotion or agonist/antagonist in terms of the pharmacological actions, and were performed for the purpose of drug development and ADMET evaluation. We classified the prediction performance according to assay types and found that similar high-precision predictions could be obtained with both phenotypic assay and target-based assay (Figure 3A), demonstrating that the issues of conventional methods for datasets showing complex activity data such as a phenotypic assay can be solved by using PM-HDE.

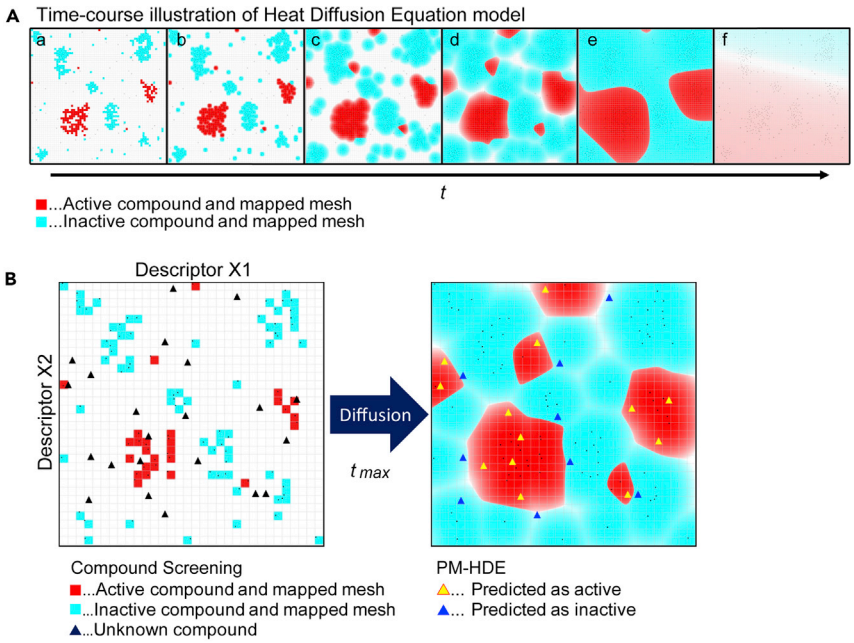


Figure 1. Schematic Representation of PM-HDE for Compound Screening

(A) Time-course illustration of PM-HDE. (a) Reference compounds with known activity are mapped in the 281-dimensional chemical space. At this point, most of the space is unoccupied by any reference meshes (white space). (b and c) As time ***t*** proceeds, plus or minus heat derived from reference meshes is gradually diffused into the space. (d) When the diffusion is stopped at an appropriate time, most of the space is filled with plus/minus heat, and each point can be predicted as active or inactive. (e and f) Beyond the appropriate diffusion time, plus heat and minus heat overlap with each other.

(B) Activity prediction of unknown compounds by PM-HDE. Left: both reference meshes and unknown compounds are mapped into the chemical space based on their calculated descriptors. Right: after the heat-diffusion process, the category of unknown compounds to which it would belong can be predicted by the V-Score.

(C) Molecular descriptors are obtained based on their chemical structures, and the V-Scores are then calculated. Plus scores mean predicted as active compounds, and the compounds in the test set can be prioritized based on the V-Score.

To investigate the relevance between V-Score and activity, the PubChem Score, and the structural similarity with the active label compound in the training set, hereafter referred to as the parent compound, were plotted for each system. It was clear that the compounds that included top-ranked compounds, within 0.0%–0.5% fraction, tended to have a much higher PubChem Score (Figure 3B). On the other hand, no apparent difference was observed in the structural similarity with the parent compound (Figure 3C). This means that the compounds ranking highest in the test set are not necessarily those with high structural similarity to the parent compound. That is, the V-Score may be expected to be high, depending on the molecular descriptors, even if structural similarity with the parent compound is not so high, and as a result PM-HDE has the potential to discover new chemotypes. This indicates that the V-Score in this method well reflects the value of the actual PubChem Score.

Performance of PM-HDE was compared with the k-nearest neighbor (kNN), RF, and SVM, as the well-known *in silico* hit predictors (Table 1). PM-HDE has a potential to show higher accuracy than kNN, especially in datasets with large entity and low active ratio. PM-HDE also presented competitive or superior accuracy compared with RF and SVM.

Implementation of PM-HDE in Compound Screening Using Human Motor Neurons

To make this model practical, we applied PM-HDE to perform hit prediction of iPSC-based phenotypic screening for amyotrophic lateral sclerosis (ALS). ALS is a lethal neurological disease in which motor neuron death causes muscle weakness and atrophy.[11,12] Since there have been no radical treatments, the development of a medicine for ALS is an urgent medical need. Motor neurons were derived from sporadic ALS-patient iPSCs (Figure S1; Table S3), and a phenotypic screening system to evaluate compounds that inhibit motor neuron death was constructed as previously reported.[13] This system produces motor neurons from iPSCs in 7 days, followed by motor neuron death in the next 7 days. After culturing motor neurons in 384-well plates with compounds for 7 days, the number of surviving motor neurons was automatically measured by high content analysis (Figure 4A). Figures of motor neurons on day 6 and day 14, which were treated with inactive compound or active compound from day 6 to day 14, are shown in Figure 4B. The stability and quality of the screening system were proved by the evidence of a low coefficient of variation (Figure 4C) and a high Z' factor (Figure 4D). The hit criterion was defined as a compound that inhibited 60% or more of motor neuron death. Using this screening system, 48,415 compounds

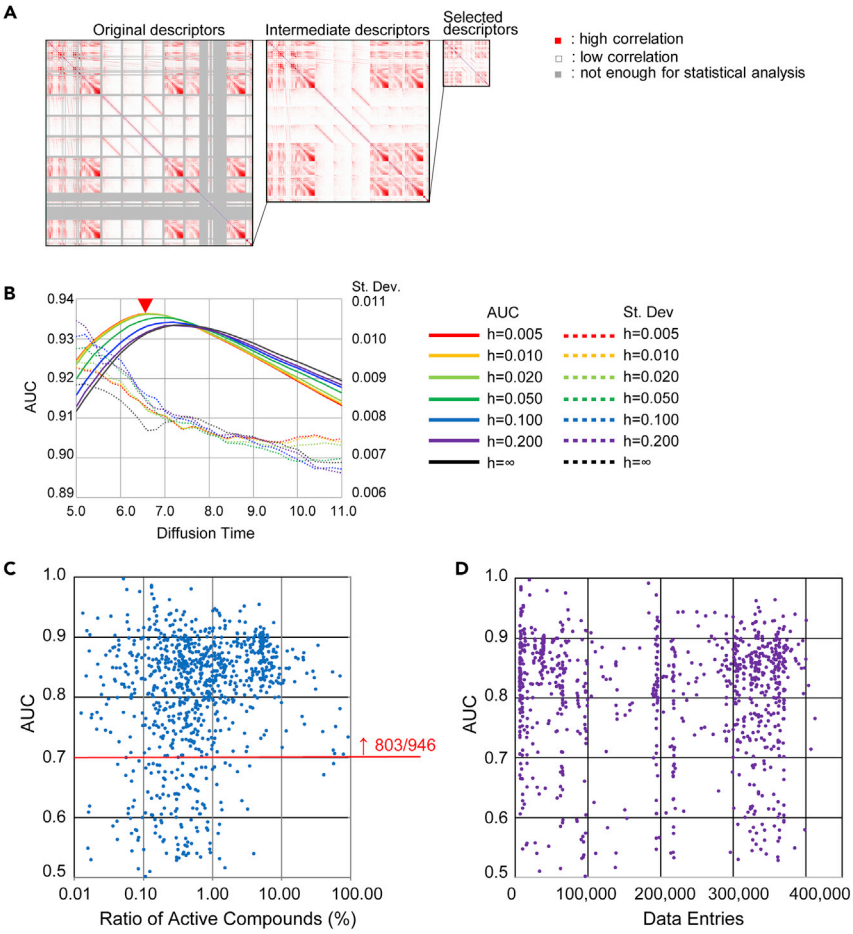


Figure 2. Virtual Screening Using Public Datasets

(A) Setting of PM-HDE parameters. Left: correlation (depth of red) map among 1,273 descriptors generated by CORINA Symphony. Descriptor: some descriptors shown in gray color are not applicable for training (e.g., values could not be calculated for certain kinds of compounds). Middle: correlation map among 999 descriptors without invalid ones. Right: correlation map among selected 281 descriptors.

(B) Determination of appropriate prediction parameters. Example of iteration to search for the best diffusion time ***t*** and weighting parameter ***h***. According to the point where the maximum AUC value was observed (red arrowhead), the best parameter set was determined.

(C) PM-HDE prediction performance. In total, 946 BioAssay data were evaluated by PM-HDE. Each plot corresponds to 1 of 946 datasets obtained from the PubChem Score. The x axis indicates the ratio of active compounds included in that assay dataset and the y axis indicates AUC of the ROC curve of 10-fold cross-validation. The mean performance value of PubChem bioassay was 0.81. The dataset ratio of AUC > 0.7 in each group with active ratio was as follows: 93.90% in 10%–100% actives, 94.90% in 1.0%–10% actives, 80.10% in 0.1%–1.0% actives, and 77.50% in 0.0%–0.1% actives. The predicted performance is independent of the actives ratio.

(D) PM-HDE prediction performance and size of datasets. Each plot corresponds to one of 946 datasets obtained from PubChem Score. The x axis indicates the number of compounds in that assay dataset and the y axis indicates AUC of the ROC curve of 10-fold cross-validation. The predicted performance is independent of the size of the dataset.

selected from the corporate library were used for first screening, and 174 hits were obtained (Figure 4E). The hit ratio was 0.36%.

Although phenotypic screening has the advantage of finding an active compound that directly reflects the disease phenotypes, evaluation of millions of compounds presents a major hurdle in terms of time and cost. Therefore, we applied PM-HDE to predict active compounds (Figure 5A) following determination of the best parameter sets (Figures 5B and 5C). Simulation of >1.6 million compounds was conducted based on first screening data, and 5,875 compounds were prioritized for evaluation of their suppressive effects against motor neuron death based on the V-Scores. The predicted active compounds were evaluated using ALS motor neurons by *in vitro* experiments. Details of the screening procedure are described in Figure 5D. In total, 5,875 compounds were added to ALS motor neurons derived from patient iPSCs to evaluate their effects on motor neuron death, and 252 hits were obtained (Figure 5E). The hit rate was 4.3%, a 10-fold increase compared with the first screening. Furthermore, when the structures of the 252 hits were examined, 16.7% of the compounds presented a new chemotype showing 50% or less similarity with any parent compounds (Figure 6A). These results suggested that PM-HDE could be useful for increasing hit efficiency and finding new chemotypes.

Next, we selected compounds without effects of cell proliferation, antioxidant, or kinase inhibition for further evaluation from the compounds presenting mainly 50% or less similarity with any parent compounds. The PM-HDE strategy successfully discovered five new chemical series that had not been identified by other screening approaches, showing little similarity between respective compounds as well as with riluzole and edaravone, which have been approved by the Food and Drug Administration for the treatment of ALS (Figure 6B).

Verification of Output of PM-HDE Using Biological Assays

Finally, the compounds identified by PM-HDE followed by the further selection described above were verified using ALS iPSC panels consisting of motor neurons derived from multiple ALS-patient iPSCs. Since sporadic ALS is known to harbor genetic heterogeneity, evaluation of drug responsiveness using iPSCs from multiple ALS patients is required. To evaluate the robustness of compound efficacy, we conducted ALS iPSC panel trials composed of motor neurons from 29 iPSC clones derived from multiple sporadic ALS patients as previously reported[13] (Figures 7A and S1), and the efficacy of five compounds identified by PM-HDE was evaluated (Figure 7B). In addition, the efficacy of each compound was compared with riluzole and edaravone (Figure 7C). The compounds identified by PM-HDE demonstrated broad and potent effectiveness against motor neuron death in clones derived from various sporadic ALS patients.

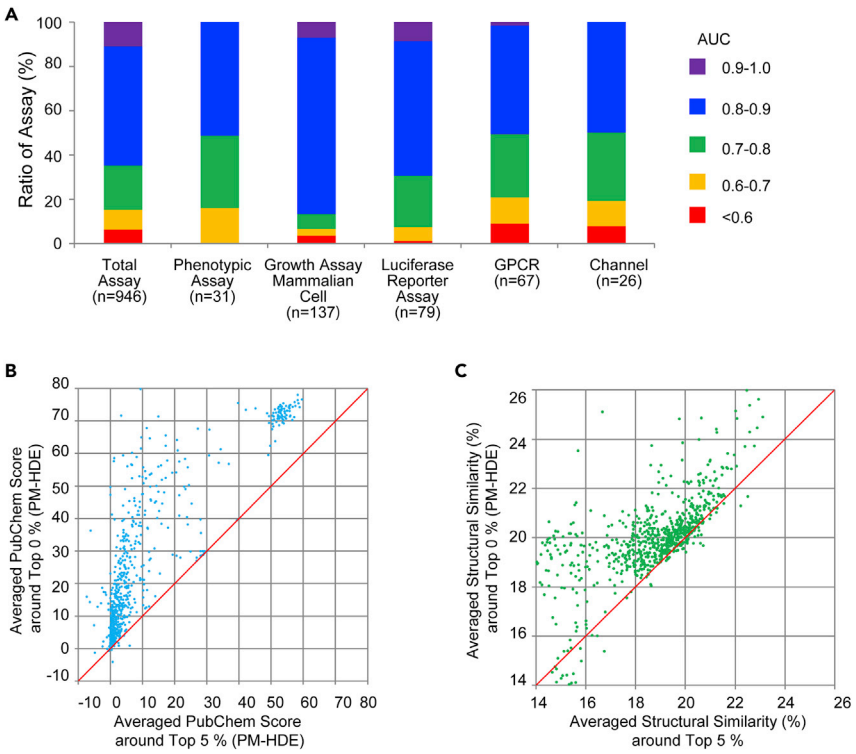


Figure 3. Validation of PM-HDE Using Public Datasets

(A) PM-HDE prediction performance and target types. The graph shows the population of prediction performance (AUC) for each target type.

(B) V-Score reflection and actual potency. Each plot corresponds to one of 946 datasets obtained from PubChem Score. The x and y axes indicate averaged PubChem Scores of the top 5.0%–5.5% ranked and top 0.0%–0.5% ranked compounds, respectively. High prediction scores correlate with the potency of the actual activity.

(C) PM-HDE and structural similarity to the parent compounds. Each plot corresponds to one of 946 datasets obtained from PubChem Score. The x and y axes indicate the average of the structural similarities of the top 5.0%–5.5% ranked and top 0.0%–0.5% ranked compounds to the parent compounds, respectively. The prediction score is independent of the structural similarity to the parent compound, an active compound included in the reference data for prediction.

DISCUSSION

PM-HDE was used in million-scale compound screening to identify promising lead compounds and new chemotypes. The robustness of PM-HDE was verified using PubChem data and was then applied to ALS iPSC-based phenotypic screening, resulting in the identification of potent compounds with a broad spectrum covering multiple iPSC lines from ALS patients. With regard to the heat-diffusion equation, it is a differential equation that requires integration constants in order to have a unique solution describing the distribution of heat in a particular body over time.[10] The equation has other important applications in mathematics, statistical mechanics, probability theory, and financial mathematics. PM-HDE is the first application of the heat-diffusion equation to drug discovery.

In past years, although molecular target-based drug screening has been the main drug discovery paradigm, there appears to be recent renewed interest in phenotypic screens for drug discovery.[14–17] Phenotypic screening by pathologically relevant cellular models using patient-derived iPSCs has the potential to find new candidate drugs and disease pathways even for the disease whose etiology has not yet been completely clarified, and it is possible to discover the potential efficacy of compounds beyond the limitations of target-based screening.

In the heat-diffusion equation, the ratio of active compounds to inactive compounds is defined as a volume ratio to the entire space by considering the mesh (described as *C* in Experimental Procedures). In PM-HDE, the reference data are calculated by mapping them to a mesh in chemical space. Since meshes have a definite size, we can give each one a heat value and define its ratio to the total chemical space. In this case, *C* adjusts the sum of the heat value of the active and inactive data groups so that the total heat value is always zero at any diffusion time. Wherever the test data are located in the chemical space, the sign of the prediction score (V-Score) can be used to determine whether the data are in the active or inactive region at diffusion time *t*. If one does not adjust for *C*, any region in the chemical space may be marked as active, depending on the ratio of actives in the reference data. Thus, PM-HDE can be used for imbalanced datasets since it can set the sum of the entire space to zero by taking the inverse of *C*. Thus, PM-HDE provides flexibility by adjusting *h*, even if it is a separation problem or a quantitative prediction that reflects the intensity of compound activity.

In phenotypic assays, as the factors affecting activity are diverse and complex, the relationship between molecular descriptors and activity also becomes complex and is not linear.[18] Therefore, in QSAR and SVM, which are prediction models of linear regression, it is necessary to make an adjustment so as to enable linear separation using a kernel function, but there is a risk of overfitting by this adjustment, and there is a concern that extrapolation may be reduced. Furthermore, while prediction based on structural similarity can find active compounds with a high probability, it is challenging to find new chemotypes because only compounds similar to the parent compounds in the reference data can be selected. PM-HDE can be expected to maintain prediction performance even with datasets with large bias in the ratio of active compounds in the reference data. In addition, since PM-HDE itself is a non-linear equation, it is possible to flexibly cope with complex relationships between descriptors and activity. By adjusting the diffusion time *t*, it is possible to predict a compound which is at a position away from the reference mesh group in the chemical space, that is, a compound that is not similar to the active compound contained in the reference data. Furthermore, the PM-HDE algorithm is suitable for parallelization, and it can learn at high speed even when dealing with myriad reference data and many descriptors.

Table 1. Comparison of PM-HDE with *In Silico* Hit Predictors

AID	Assay Description	Entries in Dataset	Actives	Active Ratio	PM-HDE AUC	kNN AUC	RF AUC	SVM AUC
0000097	NCI human tumor cell line growth inhibition assay. Data for the HS 578T breast cell line	25,421	1,378	5.42%	0.86	0.85	0.77	0.85
0000361	Pyruvate kinase	50,585	602	1.19%	0.86	0.85	0.83	0.86
0000881	qHTS assay for inhibitors of 15-hLO-2 (15-human lipoxygenase 2)	104,068	574	0.55%	0.91	0.86	0.86	0.90
0000885	qHTS assay for activators of cytochrome P450 3A4	12,561	159	1.27%	0.92	0.92	0.89	0.93
0001457	qHTS assay for identifying the cell-membrane permeable IMPase inhibitors: potentiation with lithium	202,356	722	0.36%	0.83	0.79	0.8	0.82
0002842	High-throughput screen of a putative kinase compound library to identify inhibitors of *Mycobacterium tuberculosis* H37Rv	23,462	1,248	5.32%	0.91	0.91	0.85	0.89
0624173	qHTS of *Trypanosoma brucei* inhibitors	400,566	483	0.12%	0.93	0.83	0.87	0.91
1159515	qHTS assay to identify small-molecule agonists of the nuclear factor κB signaling pathway-cell viability counter screen	6,228	190	3.05%	0.84	0.84	0.85	0.84

qHTS, quantitative high-throughput screening.

It is easy to interpret the influence of each piece of activity information contained in reference data and each descriptor used for training for the prediction result.

In this study, although we have compared PM-HDE with other machine-learning methods using several benchmark datasets, the number of comparisons was limited and the only performance comparison in actual screening was against random screening. Further verifications are desired for the widespread application of PM-HDE. Recently, conventional machine leaning has evolved into new methods to improve the performance.[19] Furthermore, deep-learning methods have also been developed and applied to drug discovery with flexibility.[20] These methods are expected to contribute to drug development, and the choice of each method may depend on experimental uncertainty regarding the data and dataset size.[20]

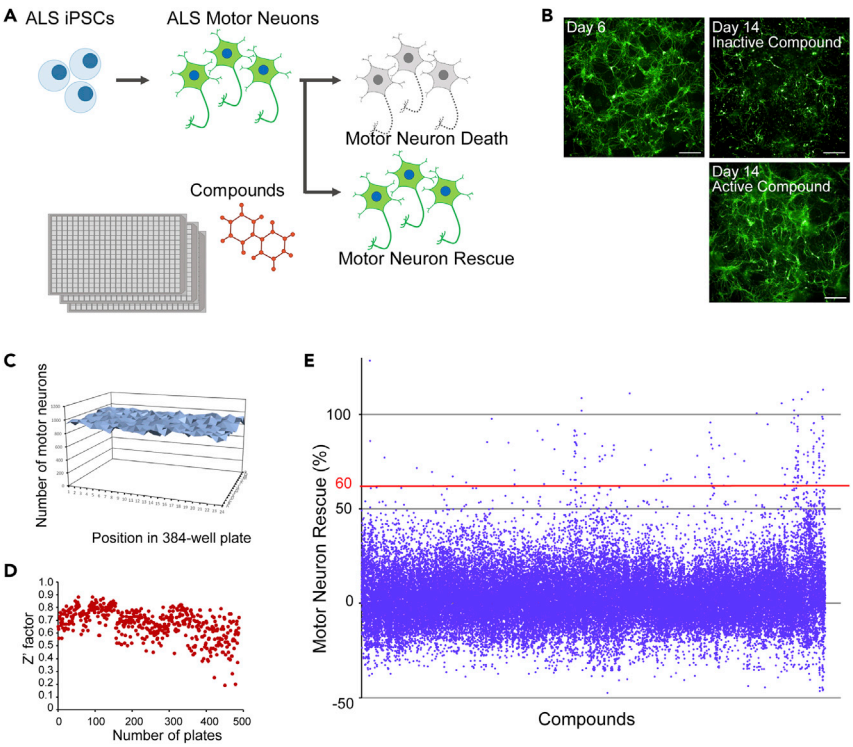


Figure 4. iPSC-Based Phenotypic Screening and Constructing Datasets for PM-HDE

(A) Screening system for the evaluation of ALS motor neuron survival.

(B) ALS motor neurons derived from patient iPSCs. Cells were stained with TUBB3, a marker for neurons.

(C) Flat test of the screening.

(D) Z' factor of the screening.

(E) Activity distribution of 48,415 compounds in the first screening. The hit criterion was defined as a compound that inhibited 60% or more of motor neuron death (shown as red line), and 174 hits were obtained.

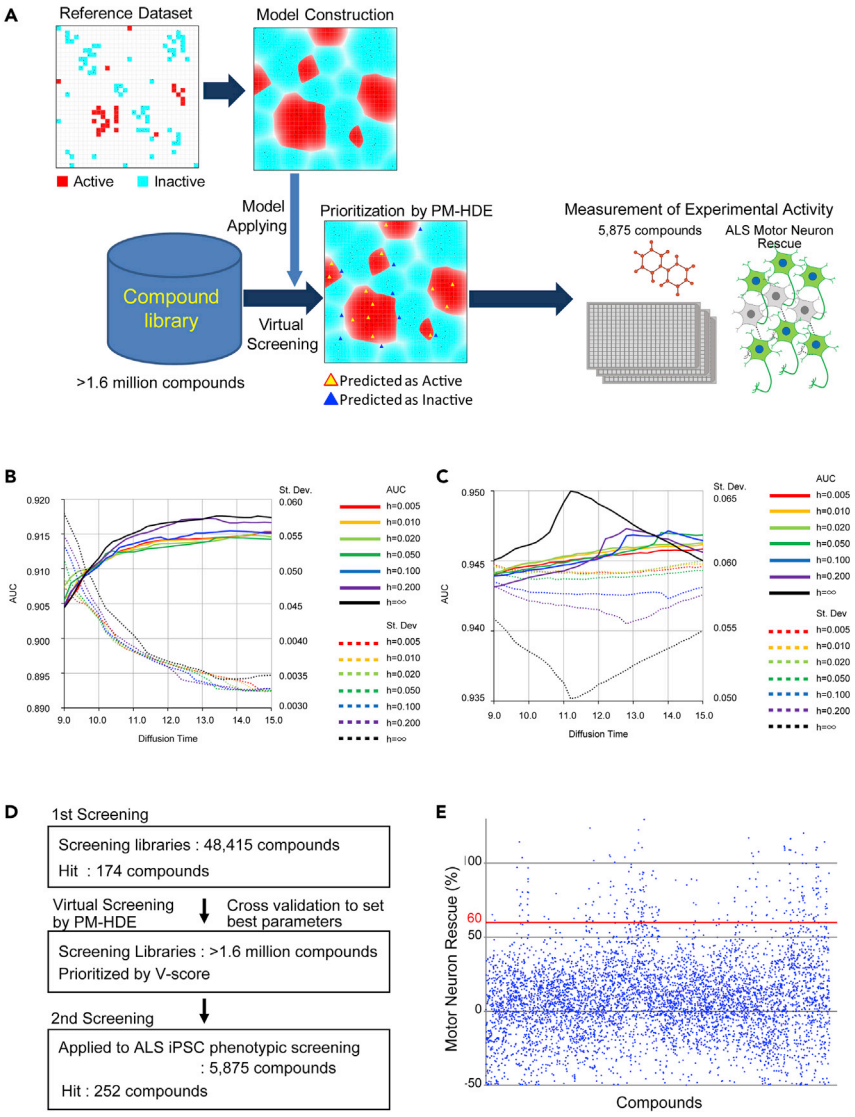


Figure 5. Virtual Compound Screening Using PM-HDE and Its Verification by ALS iPSC-Based Phenotypic Assay

(A) Workflow of simulation by PM-HDE.

(B and C) Determination of best parameter sets for PM-HDE. Two patterns for PM-HDE were built using different hit criteria: one criterion was more than 60% at 3 μM (B) and the other was more than 80% at 3 μM (C).

(D) The flow of virtual screening and *in vitro* screening.

(E) Activity distribution of 5,875 compounds selected by virtual screening. Virtual screening by PM-HDE selected 5,875 compounds from >1.6 million compounds. The 5,875 compounds were evaluated in ALS iPSC-based phenotypic assay. The hit criterion was defined as a compound that inhibited 60% or more of motor neuron death (shown as red line), and 252 hits were obtained.

PM-HDE could be a powerful tool to discover promising lead compounds for drug development. It may be possible to screen unlimited numbers of compounds, and widening its application to drug discovery for various other diseases is also expected.

EXPERIMENTAL PROCEDURES

Resource Availability

Lead Contact

Haruhisa Inoue, haruhisa@cira.kyoto-u.ac.jp.

Materials Availability

The study did not generate new unique reagents.

Data and Code Availability

The BioActivity data, descriptions, and SD-file can be downloaded from PubChem's FTP site (BioActivity data: ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay/Concise/CSV/Data/; descriptions: ftp://ftp.ncbi.nlm.nih.gov/pubchem/Bioassay/Concise/CSV/Description/; SD-files: ftp://ftp.ncbi.nlm.nih.gov/pubchem/Compound/CURRENT-Full/SDF/). The code supporting the current study have not been deposited in a public repository because the authors are currently in the process of patenting the code, but are available from the Lead Contact on request.

Study Design

The objective of our study was to develop a prediction model for utilization in drug discovery and development, including ALS iPSC-based phenotypic screening. PM-HDE was verified using PubChem data obtained from the published database and was then applied to the hit prediction of ALS iPSC-based phenotypic screening with a readout of motor neuron survival. Generation of human iPSCs was approved by the Ethics Committees of the respective departments, including Kyoto University. The use of human iPSCs for the experiments was approved by the Ethics Committees of the respective departments, including Kyoto University and the Research Ethics Review Committee of Takeda. All methods were performed in accordance with approved guidelines. Formal informed consent was obtained from all subjects.

Mathematical Methods for Heat-Diffusion Equation

The heat-diffusion equation was derived by the following mathematical theory.

(1) Given data

Reference data are calculated separately for the active and inactive groups.
z: (the value of) criterion.
[active reference data] $\vec{P}^{+}(k)$:descriptor (n–dim.) vector,
v_k^+ : score of the kth data (e.g., value of biological activity), $>z$ for $k = 1, \cdots, m_1$.
[inactive reference data] $\vec{P}^{-}(k)$:descriptor (n–dim.) vector,

A

		Similarity to Parent Compound						
		<50%	50-60%	60-70%	70-80%	80-90%	90-100%	Total
Motor neuron Rescue	60-70%	20	4	14	14	22	17	91
	70-80%	12	3	4	12	9	14	54
	80-90%	7	1	4	8	19	15	54
	90-100%	3	3	7	2	20	18	53
	Total	42	11	29	36	70	64	252

B

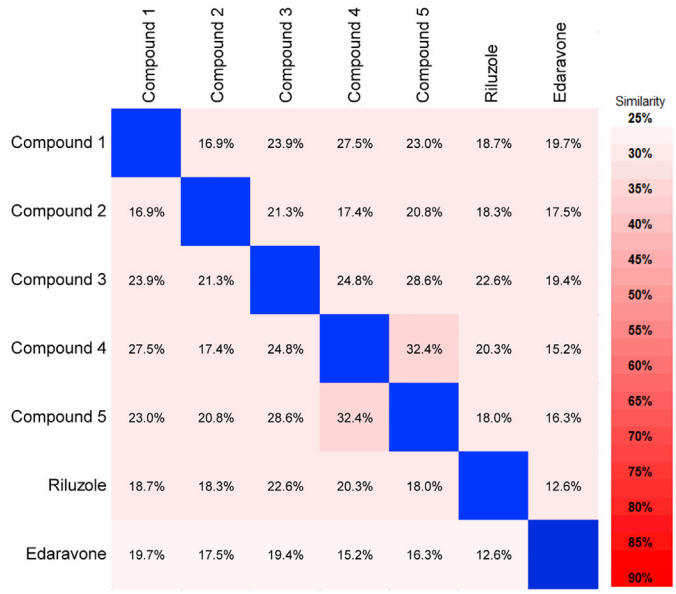


Figure 6. Validation of Hit Compounds

(A) Distributions of structural similarities to the parent compounds for 252 hit compounds identified by the second screening. 16.7% of the compounds presented a new chemotype showing 50% or less similarity with any of the parent compounds.

(B) Similarity of chemical structures for five selected compounds and two approved drugs, riluzole and edaravone.

v_k^-: score of the kth data (e.g., value of biological activity), <$z$for $k = 1, \cdots, m_2$.

Ansatz (A1): Given data are preconditioned so that there are no two reference data with the same descriptor vectors.

(2) Mesh for the space of descriptor vectors

Normalization is performed for each descriptor, then binned and allocated to one of the ranges. By allowing all descriptors to do this, one reference datum is assigned to a mesh somewhere in the chemical space.

Fix mesh size $\eta>0$. Set cube $Q_J := \prod_{i=1}^n [(j_i-1)\eta, j_i\eta)$ for $J = (j_1, \cdots j_n)$.

Ansatz (A2): Given reference data are preconditioned so that each cube Q_J contains at most one reference datum. If Q_J contains $\vec{P}^+(k)$ (resp. $\vec{P}^-(k)$), J is denoted by $J^+(k) = (j_1^+(k), \cdots, j_n^+(k))$ (resp. $J^-(k) = (j_1^-(k), \cdots, j_n^-(k))$).

When compounds are assigned to a mesh, a mesh with a composition of both active and inactive compounds should be excluded from the reference data. If a mesh is assigned to multiple active or inactive members, the activity value of the mesh is represented by the average of the members.

(3) Heat-diffusion equation

The contribution from one reference datum to the test data is calculated for each mesh using the formula below.

$u^\pm(\vec{x}, t, k) := \frac{1}{(4\pi t)^{n/2}} \prod_{i=1}^n \int_{(j_i^\pm(k)-1)\eta}^{j_i^\pm(k)\eta} e^{\frac{-(x_i-y_i)^2}{4t}} dy_i$ for Method A,

$u^\pm(\vec{x}, t, k) := \frac{1}{(4\pi t)^{n/2}} \prod_{i=1}^n e^{\frac{-(x_i-b_i^\pm(k))^2}{4t}}$ with $b_i^\pm(k) := (j_i^\pm(k) - 1/2)\eta$ for Method B

for $\vec{x} = (x_1, \cdots, x_n)$ and $t>0$ (decoding the same order).

In actuality, the coefficient of $1/(4\pi t)^{n/2}$ does not need to be calculated.

(4) V-Score $f(\vec{x}, t, h)$:

$$= \sum_{k=1}^{m_1} u^+(\vec{x}, t, k) \cdot \tanh(h(v_k - z)) - C \sum_{k=1}^{m_2} u^-(\vec{x}, t, k) \cdot \tanh(h(z - v_k))$$

for $\vec{x}$ and $t>0$ with $C := \sum_{k=1}^{m_1} \tanh(h(v_k - z)) / \sum_{k=1}^{m_2} \tanh(h(z - v_k))$ and some $h>0$ or $h = +\infty$, where $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ for $x, -\infty < x < +\infty$. Here we interpret $\tanh(hx) = \text{sign}(x)$ when $h = +\infty$.

(5) Our criteria: we expect that for suitable small $\hat{t}>0$, $\vec{x}$ would be active (resp. inactive) if $f(\vec{x}, \hat{t}, h)>0$ (resp. $f(\vec{x}, \hat{t}, h)<0$).

Remark: It is enough to check the sign of $(4\pi\hat{t})^{n/2} f(\vec{x}, \hat{t}, h)$ instead of $f(\vec{x}, \hat{t}, h)$ itself.

(6) f satisfies the heat-diffusion equation:

$$\begin{cases} \frac{\partial f}{\partial t} = \frac{\partial^2 f}{\partial x_1^2} + \frac{\partial^2 f}{\partial x_2^2} + \cdots + \frac{\partial^2 f}{\partial x_n^2} & t>0, \\ \text{initial condition } f(\vec{x}, 0) = M(\vec{x}), \end{cases} \quad \text{with}$$

$$M(\vec{x}) := \sum_{k=1}^{m_1} u_0^+(\vec{x}, k) \cdot \tanh(h(v_k - z)) - C \sum_{k=1}^{m_2} u_0^-(\vec{x}, k) \cdot \tanh(h(z - v_k)),$$

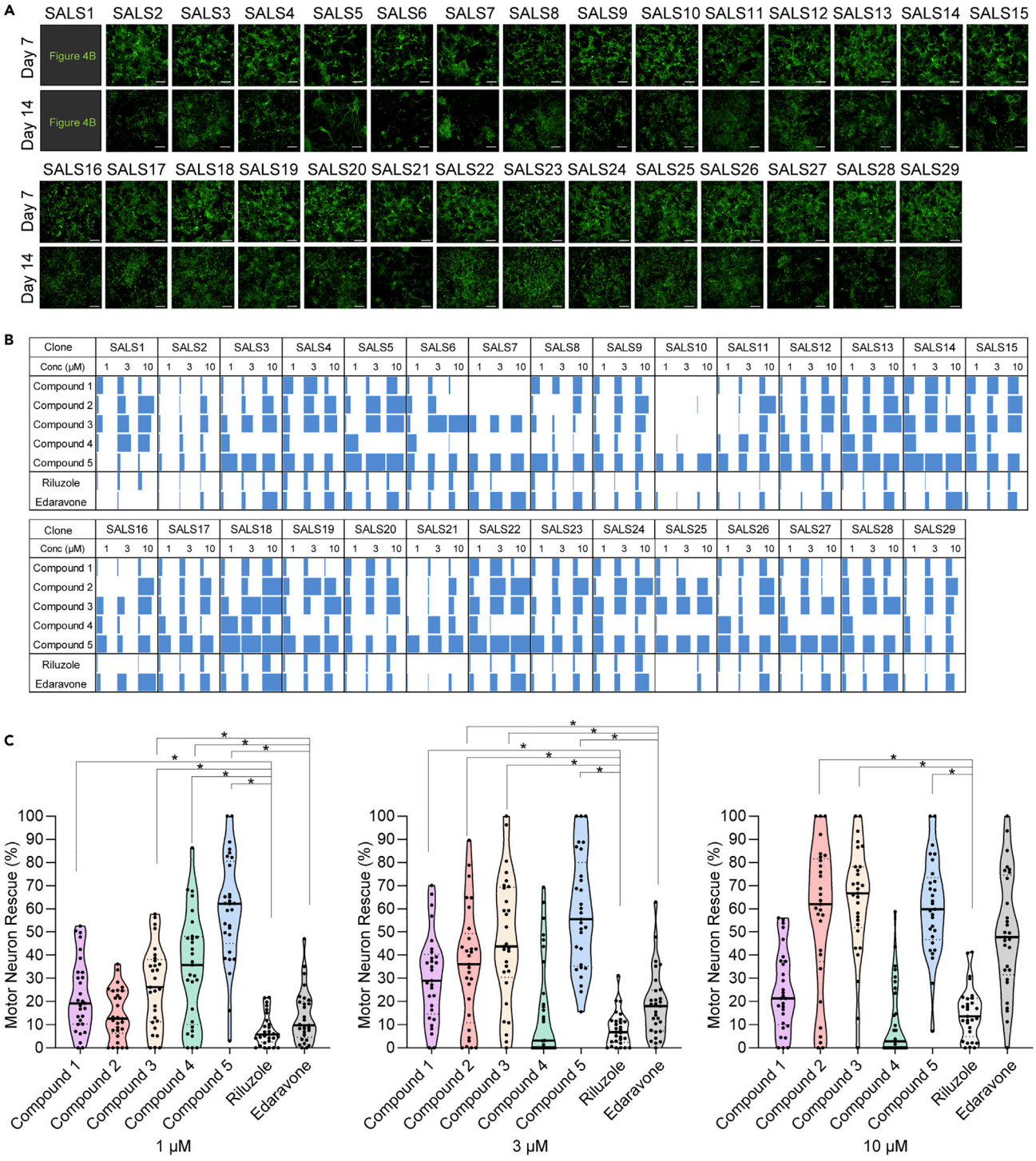


Figure 7. ALS iPSC Panel Trials for Compound Evaluation

(A) Representative figures of motor neurons listed in an ALS iPSC panel consisting of motor neurons derived from 29 clones from sporadic ALS patients. Motor neurons in the ALS iPSC panel recapitulated ALS motor neuron death.

(B) Evaluation of the active compounds extracted from PM-HDE using the ALS iPSC panel. 1 μM, 3 μM, and 10 μM concentrations of five compounds were evaluated using 29 ALS clones (n = 3 in each clone). Effects of riluzole and edaravone were also evaluated.

(C) Efficacy of the compounds was compared with those of riluzole and edaravone showing potent efficacy on ALS motor neuron rescue (n = 29 clones, one-way ANOVA; *p < 0.05).

where $u_0^{\pm}(\vec{x}, k) := \begin{cases} 1, & \vec{x} \in Q_{J^{\pm}(k)}, \\ 0, & \text{otherwise}, \end{cases}$ for Method A,

$u_0^{\pm}(\vec{x}, k) := \prod_{i=1}^{n} \delta(x_i - b_i^{\pm}(k))$ for Method B,

and $\delta$is Dirac's δ-function (decoding the same order).

Using the equation in (3), we integrate (4) for all meshes to obtain a predicted score for the test data. For each reference data value, the difference between the activity value (v_k) and the criteria (z) is taken, the difference is multiplied by seven different h in hyperbolic tangent tanh(), and the weight is given according to the activity value. In other words, we can reflect the intensity of the activity as the heat value, because the weight of the activity differs between ones that greatly exceed the criteria and ones that are close to the criteria. In addition, by using tanh, the correspondence between the activity value and the weighting can vary smoothly from near-linear to logarithmic to determine the appropriate area. On the other hand, by using $h = \infty$, it is possible to provide only active/inactive information instead of activity values.

In this study, prioritization of unknown compounds by their V-Score was performed according to "Method B" in the definitions as follows.

1. Contribution from each mesh:

$$u^{\pm}\left(\vec{x}, t, k\right) = \prod_{i=1}^{n}\left(e^{\frac{-\left(x_i - b_i^{\pm}(k)\right)^2}{4t}}\right) = e^{\frac{-\sum_{i=1}^{n}\left(x_i - b_i^{\pm}(k)\right)^2}{4t}},$$

$n = 281$ (data dimension)

2. V-Score for an unknown compound:

$$f(\vec{x}, t, h) := \sum_{k=1}^{m_1} u^{+}(\vec{x}, t, k) \cdot \tanh(h(v_k - z)) - C\sum_{k=1}^{m_2} u^{-}(\vec{x}, t, k) \cdot \tanh(h(z - v_k)),$$

$$C = \sum_{k=1}^{m1} \tanh(h(v_k - z)) \Big/ \sum_{k=1}^{m2} \tanh(h(z - v_k)),$$

where $\vec{x}$: description vector of unknown compound, t: diffusion time, k: reference data ID, $b_i^{\pm}(k)$: the ith descriptor's value of the center point of the mesh in which the kth reference data (+, active; −, inactive) are contained, v_k: activity score of the kth reference data,

z : criterion of the assay, $m1$: number of active meshes, $m2$: number of inactive meshes,

h : weighting parameter 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, ∞.

In the prediction, the parameter set t, h should be determined beforehand for maximum AUC based on cross-validation.

Multiple reference compounds with the same label in a mesh are also calculated as a single reference compound. On the other hand, if reference compounds with different labels correspond to a mesh, the mesh will be considered as noise data and will be considered invalid and not used in model generation. By including this kind of preprocessing, the noise in the reference compound can be greatly reduced.

Acquisition of Biological Activity Data and Chemical Structures from Public Data

The BioActivity Data (concise version), descriptions, and SD-file were collectively downloaded from PubChem's FTP site on May 30, 2016. Based on these files, we first created a biological activity data file for each Assay ID (AID) and an assay name list. In the biological activity data file, only those records in which ACTIVITY_OUTCOME was labeled active or inactive were included. Furthermore, conditions for selecting a system to be analyzed were defined as the number of compounds of 8,000 or more, excluding duplications based on Compound ID (CID), the number of active labels of 10 or more, the ACTIVITY_SCORE range of 2 or more, after which CID, ACTIVITY_OUTCOME, and ACTIVITY_SCORE (hereafter abbreviated as "PubChem Score") were output. The criterion of each assay system was defined as the midpoint of the ACTIVITY_SCORE boundary values of the Active label and the Inactive label. Biological activity data files were prepared for the 946 assay systems selected above. A part of the list is shown in Table S1.

Next, structure information described in the Connection Table corresponding to all the CIDs was extracted from the SD-files and desalted into a single molecule; then those containing isotopes, fraudulent valence, inappropriate atomic species, radicals, and super macromolecules were excluded. To eliminate mistakes of structural formulas, only those records in which the Connection Table in the SD-file and isomeric SMILES contained in the Data field including stereoscopic information were consistent were considered as valid structural data.

Generation and Selection of Molecular Descriptors

The SD-file was converted into isomeric SMILES using mol2smi ver.4.95 (Daylight Chemical Information Systems, Aliso Viejo, CA), then the most stable 3D-conformer was generated using Omega ver.2.5.1.4 (OpenEye Scientific, Santa Fe, NM). Using the SD-file with a three-dimensional coordinate as input, we calculated all 1,273 molecular descriptors that can be generated using CORINA Symphony Descriptor ver.1.0 (Molecular Networks, Nürnberg, Germany). Because the original 1,273 molecular descriptors contain many vectorial data, analyzing all of them is not practical due to the overfitting of specific descriptors. Therefore, the proportions of records with invalid values, correlation coefficient, skewness, and kurtosis were calculated for all molecular descriptors, and 281 descriptors considered necessary and sufficient for the subsequent analyses were selected (Table S2). These 281 descriptors were fixed throughout this study. We merged these data with the biological activity data using CID as a key and created an input file for analysis for each assay system.

Normalization and Data Cleansing

Irrelevant data contained in the input file were cleansed by following these procedures.

(1) Mean and standard deviation were determined for each of the 281 molecular descriptors and normalized for all values.
(2) The values of (1) were binned every 0.1 and assigned to 281-dimensional mesh space using them as grid points.
(3) When multiple data corresponded to the same mesh, all data in the mesh were invalidated and excluded from the analysis if there were pairs with different labels as Active/Inactive.

These normalization/cleansing processes were also applied to the preparation of the training set in cross-validation, as described later.

Coding and Cross-Validation of the Heat-Diffusion Equation Program

Coding was performed using the 64-bit version of Visual C++ 2013 (Microsoft, Redmond, DC) according to the algorithms shown in Figure 1B. The prediction score for one record in a test set is referred to as the V-Score in Table S1. It is expressed by $f(x, t)$, and corresponds to the temperature in the original definition.

After cleansing the data groups of each assay system, we divided the dataset into a training set (test set = 9:1) and performed cross-validation. For each test set, by giving the diffusion time t = 6.0 to 12.0 in 0.2 increments in 31 levels and the hyperbolic tangent coefficient h in 7 levels, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, and ∞, AUC of the ROC curve was calculated for each of the 31 × 7 = 217 patterns of parameters as calculation parameters, and the mean and standard deviation of 217 values for ten test sets were calculated to determine the best-suited parameters. These series of calculations were performed in a batchwise manner for 946 assay systems.

Analysis of the Relationship between V-Score and Structural Similarity

The means of active label capture rate, PubChem Score, and structural similarity between training set and test set were examined for the 0.0%–0.5% fraction and the 5.0%–5.5% fraction from the top of the V-Score in each of the 713 systems that gave a local maximum of AUC and showed a standard deviation

of ≤0.12 with a maximum AUC of ≧0.65 among the 946 assay systems. Structural similarity was obtained by calculation of the Tanimoto coefficient following the generation of 2,048 bit fingerprints using Daylight's Toolkit ver. 4.95 based on isomeric SMILES.

The k-Nearest Neighbor, Support Vector Machine, and Random Forest

The Python library "scikit-learn" was employed for the development of kNN (sklearn.neighbors.KNeighborsClassifier), SVM (sklearn.svm.SVC), and RF (sklearn.ensemble.RandomForestClassifier) classification models. All methods employed used 10-fold cross-validation to identify their hyper-parameters, i.e., the number of neighbors to use k for kNN, penalty parameter C, and nuclear parameter γ for SVM, and maximum depth, the minimum number of samples required to split an internal node, and the minimum number of samples required to be at a leaf node for RF. A 10-fold cross-validation was implemented by first dividing the training set into ten equal groups, nine of which were used for model construction and the tenth for validation. Series of hyper-parameter values were respectively assigned to construct the models, and by determining the best AUC values the optimal ones were identified.

Generation of iPSCs

iPSCs were generated from skin fibroblasts or peripheral blood mononuclear cells of sporadic ALS patients using episomal vectors (Sox2, Klf4, Oct3/4, L-Myc, Lin28, and p53-short hairpin RNA) as reported previously.[21,22] Established iPSCs were cultured on an SNL feeder layer with human iPSC medium (primate embryonic stem cell medium; ReproCELL, Yokohama, Japan) supplemented with 4 ng/mL basic fibroblast growth factor (Wako Chemicals, Osaka, Japan) and penicillin/streptomycin or cultured under feeder-free conditions on laminin-511-E8 (Nippi, Tokyo, Japan)-coated plates with StemFit AK01 (Ajinomoto, Tokyo, Japan). The information on iPSC clones is listed in Table S3. Karyotyping was performed by Nihon Gene Research Laboratories (Sendai, Japan).

Generation of Motor Neurons and ALS iPSC-Based Phenotypic Screening

Motor neurons were generated from iPSCs as previously described.[13] A polycistronic vector containing mouse *Lhx3*, mouse *Ngn2*, and mouse *Isl1* under control of the tetracycline operator, which was generated from KW110_PB_TA_ERN (Ef1a_rtTA_*neo*) vector backbone with rtTA and neomycin resistance gene, was co-transfected along with a pHL-EF1α-hcP-Bace-A encoding transposase into iPSCs using lipofectamine LTX (Thermo Fisher Scientific, Waltham, MA). After clone selection using neomycin, iPSCs carrying the tetracycline-inducible motor neuron differentiation cassette were established.

These iPSCs were dissociated to single cells using Accutase (Innovative Cell Technologies, San Diego, CA) and plated onto Matrigel-coated 384-well plates (PerkinElmer, Waltham, MA) with Neuronal Medium DMEM/F12 (Thermo Fisher Scientific), N2 (Thermo Fisher Scientific) containing 1 μM retinoic acid (Sigma), 1 μM Smoothened agonist, 10 ng/mL brain-derived neurotrophic factor (R&D Systems, Minneapolis, MN), 10 ng/mL glial-cell-line-derived neurotrophic factor (R&D Systems), and 10 ng/mL NT-3 (R&D Systems) with 1 μg/mL doxycycline (TaKaRa, Kusatsu, Japan), and cultured for 7 days. The compounds were added on day 6 or day 7, and cells were fixed in 4% paraformaldehyde and stained with anti-βIII-tubulin (TUBB3) antibody conjugated with Alexa 488 (Sigma) on day 14. Images of motor neurons were acquired by Opera Phenix (PerkinElmer) followed by quantification of the number of surviving motor neurons stained with neuronal marker TUBB3 using the analysis software Harmony (PerkinElmer). The effect of compounds was described as "motor neuron rescue." Motor neuron rescue (%) = $([X - C]/[T - C]) \times 100$, where X: number of motor neurons treated with compounds on day 14, C: number of motor neurons treated with dimethyl sulfoxide group on day 14, T: Number of motor neurons in pretreatment on day 6 or day 7.

Statistical Analysis

Results were analyzed using one-way ANOVA followed by Dunnett's post hoc test to determine statistical significance of the data. Differences were considered significant at $p < 0.05$. Analyses were performed using SPSS software (IBM).

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.patter.2020.100140.

ACKNOWLEDGMENTS

We would like to express our sincere gratitude to all our coworkers and collaborators; to Misako Takemoto, Eri Ejiri, Ayami Onodera, Yuka Hirabayashi, Yumiko Nakagaito, Shigeru Kondo, Hitoshi Nakamura, Hiromitsu Fuse, Keisuke Imamura, Masashi Toyofuku, Yumi Imai, and Sachiko Itono for their experimental support; and to Shinya Yamanaka, Seigo Izumo, Haruhide Kimura, Takanobu Kuroita, Toshimasa Tanaka, Atsushi Nakanishi, Hidetoshi Shimodaira, and Yuichiro Yada for their scientific discussions. This work was funded in part by a grant from the Research Center Network for Realization of Regenerative Medicine from AMED (H.I.), Research Project for Practical Applications of Regenerative Medicine from AMED (H.I.), and grant-in-aid for scientific research (18K18452 to H.I.). Takeda Pharmaceutical Company Limited was the sponsor of this work. Takeda Pharmaceutical Company Limited was paying the salary of T.N. in relation to this work.

AUTHOR CONTRIBUTIONS

H.I. and M.F. conceived the project. T. Hidaka, K.I., T. Hioki, T.T., and H.I. designed the research and wrote the manuscript. T. Hidaka, K.I., T. Hioki, Y.N., Y.K., S.H., and T.N. performed the experiments. Y.G. and M.G. provided expertise and feedback regarding the heat-diffusion equation.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 17, 2020
Revised: September 7, 2020
Accepted: October 14, 2020
Published: November 11, 2020

REFERENCES

1. Farkhondeh, A., Li, R., Gorshkov, K., Chen, K.G., Might, M., Rodems, S., Lo, D.C., and Zheng, W. (2019). Induced pluripotent stem cells for neural drug discovery. Drug Discov. Today *24*, 992–999.
2. Neves, B.J., Braga, R.C., Melo-Filho, C.C., Moreira-Filho, J.T., Muratov, E.N., and Andrade, C.H. (2018). QSAR-based virtual screening: advances and applications in drug discovery. Front. Pharmacol. *9*, 1275.
3. Cherkasov, A., Muratov, E.N., Fourches, D., Varnek, A., Baskin, I.I., Cronin, M., Dearden, J., Gramatica, P., Martin, Y.C., Todeschini, R., et al. (2014). QSAR modeling: where have you been? Where are you going to? J. Med. Chem. *57*, 4977–5010.
4. Li, Y., Wang, L., Liu, Z., Li, C., Xu, J., and Gu, Q. (2015). Predicting selective liver X receptor beta agonists using multiple machine learning methods. Mol. Biosyst. *11*, 1241–1250.
5. Deshmukh, A.L., Chandra, S., Singh, D.K., Siddiqi, M.I., and Banerjee, D. (2017). Identification of human flap endonuclease 1 (FEN1) inhibitors using a machine learning based consensus virtual screening. Mol. Biosyst. *13*, 1630–1639.
6. Chandra, S., Pandey, J., Tamrakar, A.K., and Siddiqi, M.I. (2017). Multiple machine learning based descriptive and predictive workflow for the identification of potential PTP1B inhibitors. J. Mol. Graph. Model. *71*, 242–256.
7. Lee, K., Lee, M., and Kim, D. (2017). Utilizing random Forest QSAR models with optimized parameters for target identification and its application to target-fishing server. BMC Bioinformatics *18*, 567.

8. Swamidass, S.J., and Baldi, P. (2007). Mathematical correction for fingerprint similarity measures to improve chemical retrieval. J. Chem. Inf. Model. *47*, 952–964.
9. Racz, A., Bajusz, D., and Heberger, K. (2018). Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints. J. Cheminform. *10*, 48.
10. Giga, M., Giga, Y., Ohtsuka, T., and Noriaki, U. (2013). On behavior of signs for the heat equation and a diffusion method for data separation. Commun. Pure Appl. Anal. *12*, 2277–2296.
11. Ravits, J., Appel, S., Baloh, R.H., Barohn, R., Brooks, B.R., Elman, L., Floeter, M.K., Henderson, C., Lomen-Hoerth, C., Macklis, J.D., et al. (2013). Deciphering amyotrophic lateral sclerosis: what phenotype, neuropathology and genetics are telling us about pathogenesis. Amyotroph. Lateral Scler. Frontotemporal Degener. *14* (*Suppl 1*), 5–18.
12. Ling, S.C., Polymenidou, M., and Cleveland, D.W. (2013). Converging mechanisms in ALS and FTD: disrupted RNA and protein homeostasis. Neuron *79*, 416–438.
13. Imamura, K., Izumi, Y., Watanabe, A., Tsukita, K., Woltjen, K., Yamamoto, T., Hotta, A., Kondo, T., Kitaoka, S., Ohta, A., et al. (2017). The Src/c-Abl pathway is a potential therapeutic target in amyotrophic lateral sclerosis. Sci. Transl Med. *9*, eaaf3962.
14. Zheng, W., Thorne, N., and McKew, J.C. (2013). Phenotypic screens as a renewed approach for drug discovery. Drug Discov. Today *18*, 1067–1073.
15. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. (2019). Applications of machine learning in drug discovery and development. Nat. Rev. Drug Discov. *18*, 463–477.
16. Carpenter, K.A., and Huang, X. (2018). Machine learning-based virtual screening and its applications to Alzheimer's drug discovery: a Review. Curr. Pharm. Des. *24*, 3347–3358.
17. Rifaioglu, A.S., Atas, H., Martin, M.J., Cetin-Atalay, R., Atalay, V., and Dogan, T. (2019). Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. Brief. Bioinform. *20*, 1878–1912.
18. Prosperi, M.C., and De Luca, A. (2012). Computational models for prediction of response to antiretroviral therapies. AIDS Rev. *14*, 145–153.
19. Zhang, L., Tan, J., Han, D., and Zhu, H. (2017). From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discov. Today *22*, 1680–1685.
20. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The rise of deep learning in drug discovery. Drug Discov. Today *23*, 1241–1250.
21. Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861–872.
22. Okita, K., Yamakawa, T., Matsumura, Y., Sato, Y., Amano, N., Watanabe, A., Goshima, N., and Yamanaka, S. (2013). An efficient nonviral method to generate integration-free human-induced pluripotent stem cells from cord blood and peripheral blood cells. Stem Cells *31*, 458–466.