Supporting Information

Siamese Recurrent Neural Network with a Self-Attention Mechanism for Bioactivity Prediction

Daniel Fernández-Llaneza,* Silas Ulander, Dea Gogishvili, Eva Nittinger, Hongtao Zhao,* and Christian Tyrchan*

Department of Medicinal Chemistry, Research and Early Development, Respiratory and Immunology, Biopharmaceutical R&D, AstraZeneca, Pepparedsleden 1, SE 43183 Mölndal, Sweden

*e-mail: daniel.fernandez1@astrazeneca.com (D. Fernandez-Llaneza);
hongtao.zhao@astrazeneca.com (H. Zhao);
christian.tyrchan@astrazeneca.com (C. Tyrchan)

Mathematical Details of the Siamese Neural Network

LSTM Anatomy. In this work, LSTMs were used to extract the information encoded in the SMILES strings. LSTMs are fed with an input vector $x(t) = \{x_1, x_2,\ldots,x_t\}$ at a time step t and manage to extract the encoded sequential information by passing it through a complex of gating units that regulate the data flow. Mathematically, the system of gates is defined according to *Equation 1*.

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \boldsymbol{W} \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix} + b$$

$$c_t = f \odot c_{t-1} + i_t \odot g$$

$$h_t = o \odot \tanh(c_t) \qquad \textit{Equation 1}$$

where x_t, h_t and c_t are the input, the hidden and cell states at a given time t accordingly, h_{t-1} and c_{t-1} are the hidden and cell states from the previous timestep and i, f, o and g are the input, forget, output and new memory gates, respectively. The symbols σ and $\odot$ denote the sigmoid function and the Hadamard product and $\boldsymbol{W}$ and b are the weights and biases matrices that are backpropagated. The input gate i determines how much information from the input is allowed into the cell, the forget gate f controls the amount of data that is preserved from previous timesteps, the output gate o triages how much information is revealed via the hidden state h_t by regulating the cell state c_t update. The cell state c_t consists of information that will be passed onto the next cell and controls the influence that the previous cell state c_{t-1} and the input at the current timestep exert by operating with the forget gate f and the g gate.

Bidirectional LSTM. Standard unidirectional LSTMs can only process sequences in a temporal fashion and, as a result, they are confronted with an inability to completely distil the relevant information. In view of this limitation, bidirectional LSTMs (BiLSTM) present an elegant way of condensing all the contextual data, by exposing the training sequence both in the forward and backward direction to two independent LSTMs. Afterwards, the hidden states obtained can be processed as per *Equation 2*.

$$h_i = q\left(\vec{h}_i, \overleftarrow{h}_i\right) \qquad \textit{Equation 2}$$

where $\vec{h}_i$ is the hidden state for the forward direction, $\overleftarrow{h}_i$ is the hidden state for the backward direction and q is a function which is normally concatenation, addition or average both for $\vec{h}_i$ and $\overleftarrow{h}_i$.

Self-Attention Mechanism. The self-attention mechanism is based on concatenating the forward hidden states $(\vec{h}_0,\ldots,\vec{h}_n)$ from $\overrightarrow{\text{LSTM}}$ and the backward hidden states $(\overleftarrow{h}_0,\ldots,\overleftarrow{h}_n)$ from $\overleftarrow{\text{LSTM}}$ yielding the hidden states matrix $\boldsymbol{H} \in \mathbb{R}^{n \times 2h_s}$. The matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ containing the attention weights is then calculated

$$\boldsymbol{A} = \text{softmax}(\boldsymbol{W_2}\tanh(\boldsymbol{W_1}\boldsymbol{H}^{\boldsymbol{T}})) \qquad \textit{Equation 3}$$

where $\boldsymbol{W_1}$ and $\boldsymbol{W_2}$ are trainable weight matrices of size h_e x $2h_s$ and h_e x n, respectively. h_e is an adjustable hyperparameter which was set to 512. The attention weights $\boldsymbol{A}$ were subsequently applied to the hidden states matrix $\boldsymbol{H}$ and the output was fed through a fully connected linear layer, thereby resulting in the attentional vector $\tilde{h} \in \mathbb{R}^{2h_s}$.

$$\tilde{h} = \varphi\,(\boldsymbol{W_3}(\boldsymbol{AH})^{\mathbf{T}} + b_3) \qquad \textit{Equation 4}$$

where $\boldsymbol{W_3} \in \mathbb{R}^{n \times 1}$, $b_3 \in \mathbb{R}^n$ and $\varphi(\cdot)$ is the leaky rectified linear unit with standard hyperparameters from PyTorch.

Siamese Neural Network. The Siamese Neural Network consists of a dual-branch network with shared weights, to guarantee that each element of the pair is passed through identical layers. Thus, the training set of a Siamese Neural Network consists of two fixed-size sequence inputs $x_1^{(i)}$ and $x_2^{(i)}$ and a binary similarity label $y^{(i)}$. Assuming that the overall mapping function into the embedding space is computed as $G_W : x_n^{(i)} \rightarrow G_W\left(x_n^{(i)}\right)$, the energy function E_W of the Siamese Neural Network is given by:

$$E_W\left(x_1^{(i)}, x_2^{(i)}\right) = cos(\theta) = \frac{G_W\left(x_1^{(i)}\right) \cdot G_W\left(x_2^{(i)}\right)}{max\left(\left\|G_W\left(x_1^{(i)}\right)\right\|_2 \cdot \left\|G_W\left(x_2^{(i)}\right)\right\|_2, \varepsilon\right)} \qquad \textit{Equation 5}$$

where $\varepsilon = 10^{-8}$, as the default hyperparameter in PyTorch. The output was clamped thereafter as defined in *Equation 6* to ensure that the estimated binary label $\tilde{y}^{(i)} \in [0,1]$:

$$\tilde{y}^{(i)} = min\left(max\left(E_W\left(x_1^{(i)}, x_2^{(i)}\right), 0\right), 1\right) \qquad \textit{Equation 6}$$

Table S1. Dataset Filtering Report

filtering step	BACE1	CCR5	DRD2	EGFR	NR1H2
Initial dataset	18 011	7 617	107 129	29 567	9 784
Standardization 1. Removal of missing biological activity 2. Filtering heavy atoms and SMILES strings	8 652	3 662	106 341	10 698	2 097
Addition of in-house inactive compounds	23 745	6 904	N/A	16 402	3 455
Removal of duplicate compounds	20 450	4 998	N/A	11 364	2 712
Curated dataset	**20 450**	**4 998**	**106 341**	**11 364**	**2 712**

Table S2. MCC Scores and False Positive Rates in *N*-shot Learning Strategies for Inference Repetition [a]

dataset	*N*	metric	$k = 1$	$k = 3$	$k = 5$
BACE1	32	MCC	0.770 (0.767 to 0.773)	**0.791** (0.787 to 0.794)	0.788 (0.787 to 0.790)
		FPR	11.610 (11.418 to 11.803)	**10.730** (10.541 to 10.918)	11.248 (11.129 to 11.368)
CCR5	16	MCC	0.744 (0.738 to 0.750)	**0.767** (0.764 to 0.770)	0.771 (0.765 to 0.776)
		FPR	19.478 (18.944 to 20.0123)	**18.570** (18.372 to 18.769)	18.289 (17.868 to 18.711)
DRD2	32	MCC	0.839 (0.834 to 0.844)	**0.848** (0.846 to 0.850)	0.845 (0.841 to 0.850)
		FPR	3.0432 (2.915 to 3.172)	**2.870** (2.810 to 2.929)	2.924 (2.835 to 3.013)
EGFR	32	MCC	0.645 (0.639 to 0.652)	**0.708** (0.705 to 0.711)	0.701 (0.698 to 0.705)
		FPR	30.906 (30.385 to 31.426)	30.308 (30.001 to 30.615)	30.331 (29.964 to 30.698)
NR1H2	8	MCC	0.699 (0.692 to 0.707)	0.705 (0.698 to 0.712)	**0.712** (0.708 to 0.717)
		FPR	16.930 (16.500 to 17.360)	17.134 (16.636 to 17.631)	16.650 (16.309 to 16.991)

[a] The numbers reported are the means and in brackets the lower and the upper bound of 95% confidence intervals (CI) from 10 runs using different support sets. The highest statistically significant score is highlighted in bold. If there is no significant difference between numbers, they were left as a regular font throughout.

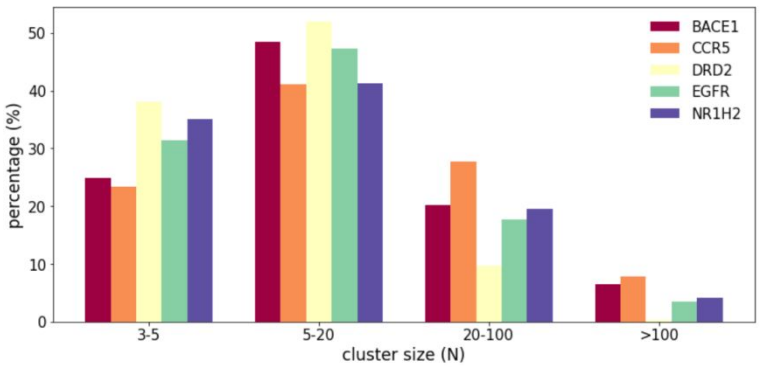

Figure S1. Distribution of cluster sizes across datasets.

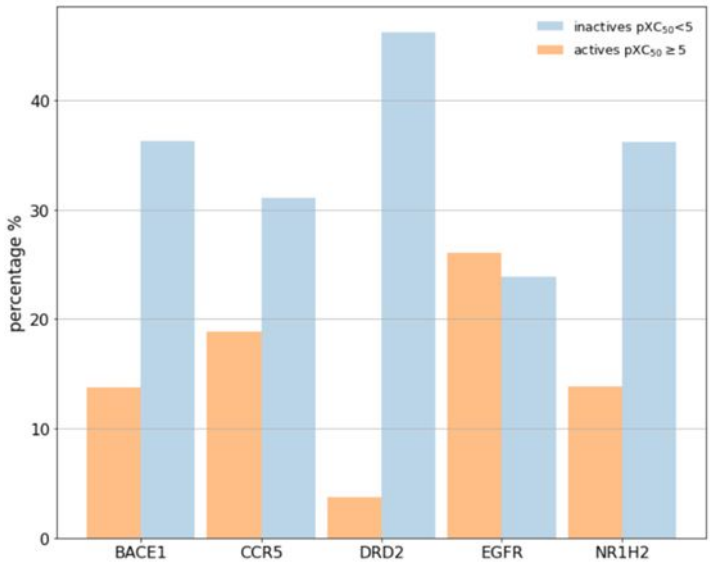

Figure S2. Distribution of active and inactive compounds by dataset.

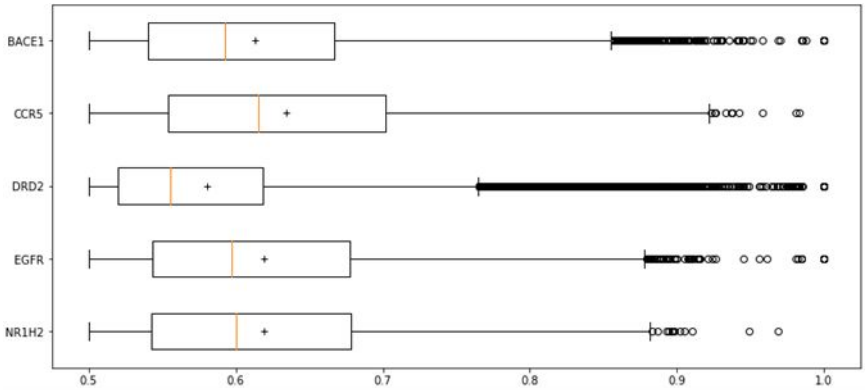

Figure S3. Boxplots showing Tanimoto coefficients between the centroid and the rest of the compounds within the same cluster across datasets.

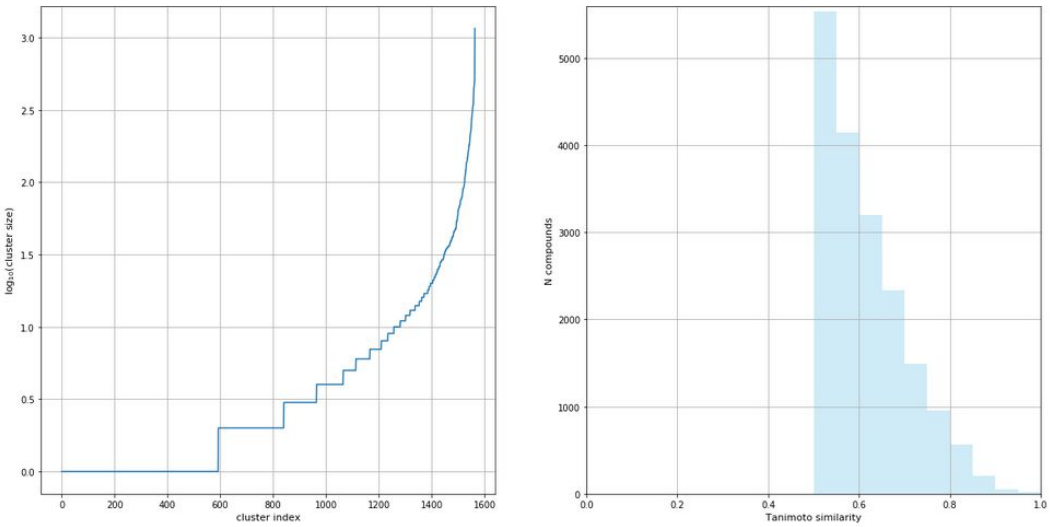

Figure S4. Cluster size distribution (*left*) and Tanimoto similarity distribution (*right*) for BACE1

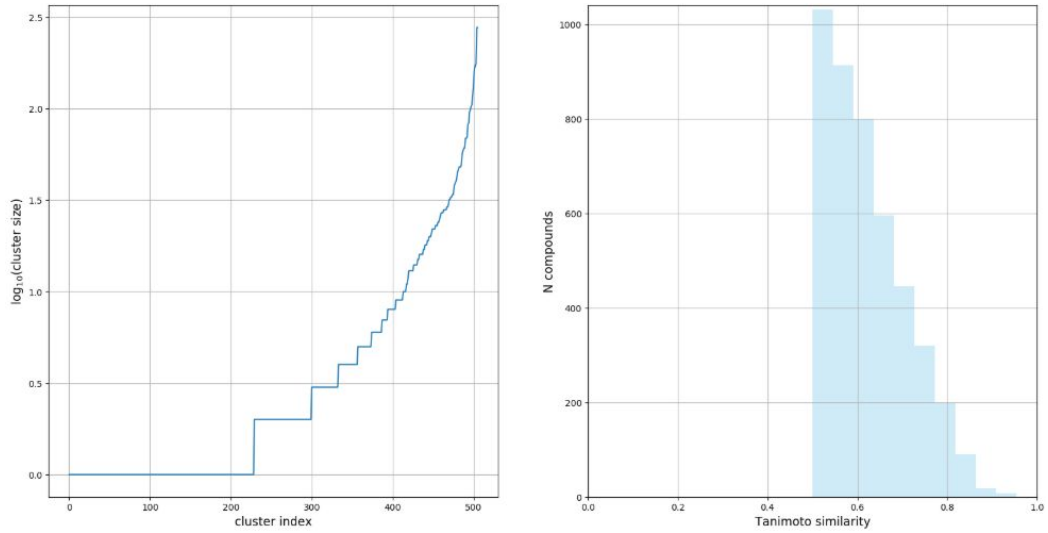

Figure S5. Cluster size distribution (*left*) and Tanimoto similarity distribution (*right*) for CCR5

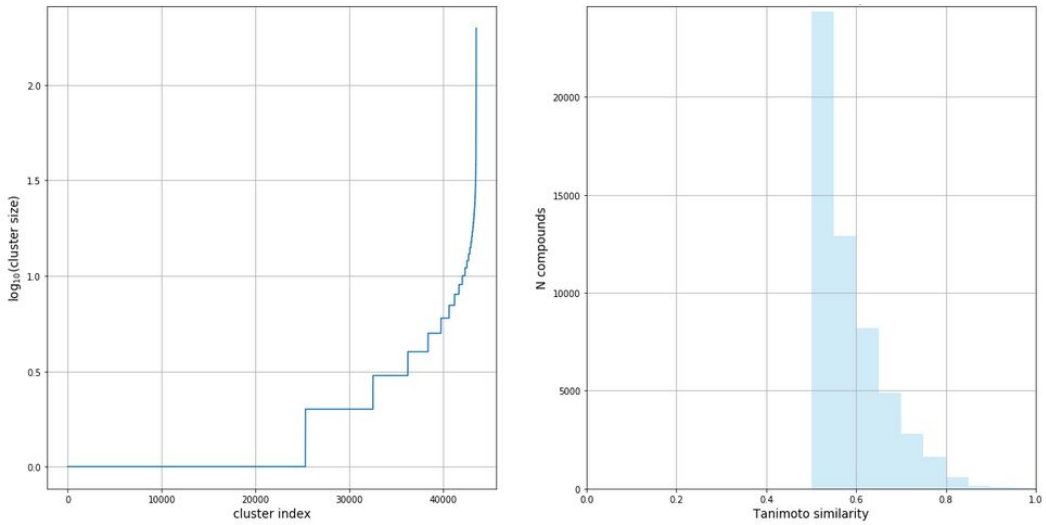

Figure S6. Cluster size distribution (*left*) and Tanimoto similarity distribution (*right*) for DRD2.

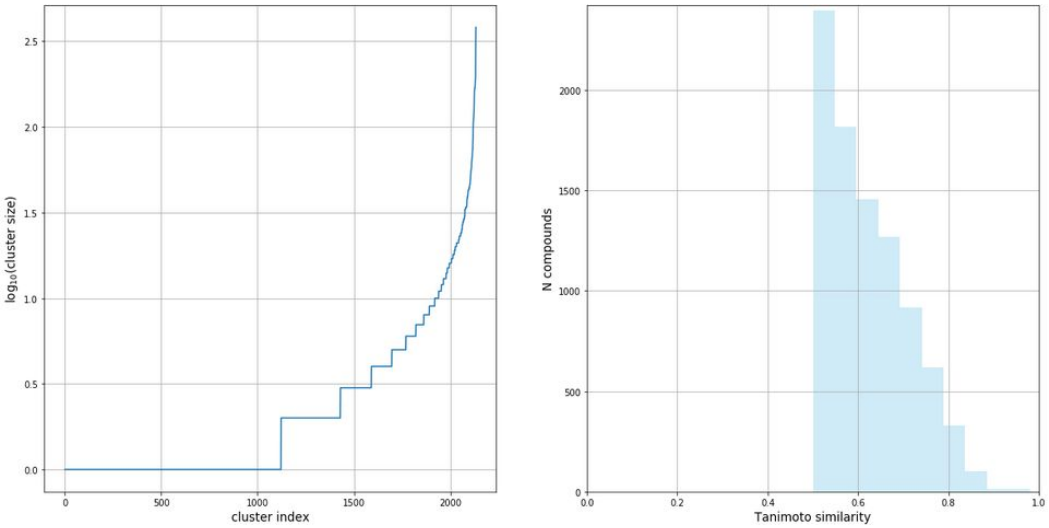

Figure S7. Cluster size distribution (*left*) and Tanimoto similarity distribution (*right*) for EGFR.

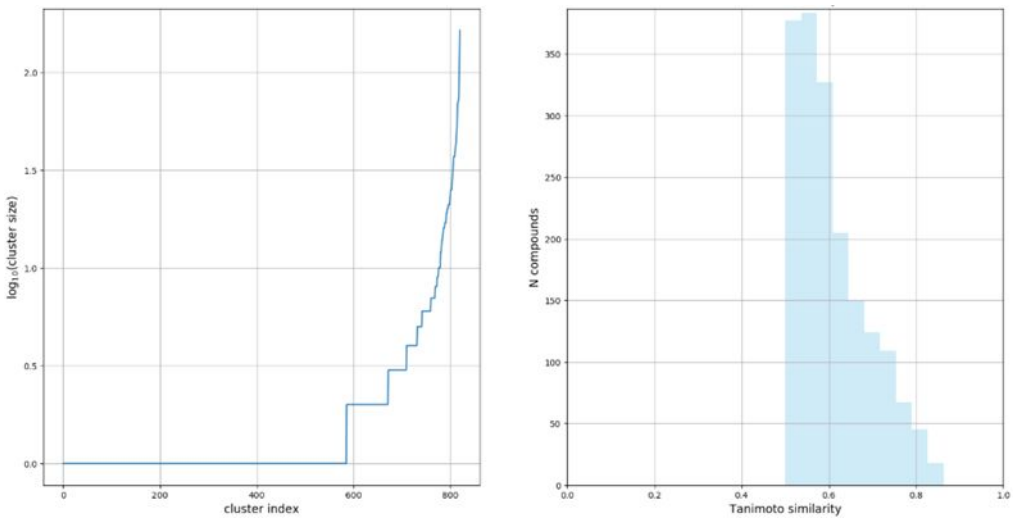

Figure S8. Cluster size distribution (*left*) and Tanimoto similarity distribution (*right*) for NR1H2.