Nuclear receptor modulators: Catching information by machine learning

Cecile Valsecchi,[1] Francesca Grisoni,[2] Viviana Consonni,[1] Davide Ballabio,[1] Roberto Todeschini[1]

[1]Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano Bicocca, Milan, Italy; [2]Department of Chemistry and Applied Biosciences, ETH Zurich, Zurich, Switzerland

Abstract

Nuclear receptors (NRs) are involved in fundamental human health processes and are a relevant target for toxicological risk assessment. To help prioritize chemicals that can mimic natural hormones and be endocrine disruptors, computational models can be a useful tool.[1,2] In this work we i) created an exhaustive collection of NR modulators and ii) applied machine learning methods to fill the data-gap and prioritize NRs modulators by building predictive models.

Introduction

The human nuclear receptor superfamily is composed by 48 key transcription factors, which constitute 16% of drug targets.[3] Although the NRs ligand binding domain is highly conserved, NRs have been shown to be promiscuous targets, as demonstrated by the endocrine interference [4]. In this framework, *in-silico* modelling based on machine learning can be used to fill the data-gap in order to reduce animal testing and protect human health. Thus, we collected bioactivity data from different sources and then applied machine learning methods, with a particular focus on the simultaneous prediction of bioactivities for 8 NRs.

Materials and Methods

We created a publicly available dataset (NURA-NUclear Receptor Activity dataset)[5] containing annotations for binding, agonism and antagonism activity for 15,206 molecules and 8 selected NRs taken from 4 public databases (*i.e.*, Tox21, ChEMBL, NR-DBIND and BindingDB). Each bioactivity type for a given receptor was considered as a task (*e.g.*, binding activity for androgen receptor), obtaining a total of 30 tasks. The structural information of each molecule was encoded in a binary vector, *i.e.*, as extended connectivity fingerprints (ECFPs).[6] We applied two multi-task neural networks (*FFNL1* and *FFNL3*), which allow to simultaneously model the 30 tasks. To assist the validation of in-silico models, molecules were randomly split into training set (11,970 molecules, 80%) and test set (2,993 molecules, 20%).

Results

NURA dataset is a useful tool to bridge the gap between toxicology-and medicinal-chemistry-related databases, as it is enriched in terms of number of molecules, structural diversity and covered atomic scaffolds compared to the single sources as shown in Figure 1 and Figure 2.

NURA dataset is intended to support decision-making in pharmacology and toxicology, as well as to contribute to data-driven applications. It can be downloaded at the following DOI: https://doi.org/10.5281/zenodo.3991561.

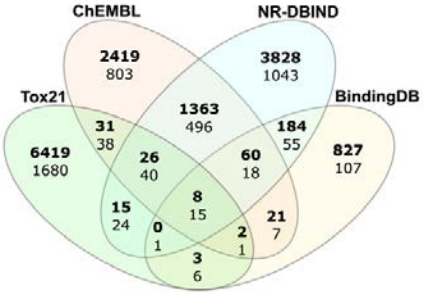

Figure 1. Venn diagram of the NURA data collected from Tox21, ChEMBL, NR-DBIND and Binding-DB and pruned. The numbers of shared and not shared molecules (in bold) and scaffolds are reported.

Correspondence: Cecile Valsecchi, Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano Bicocca, P.za della Scienza 1, 20126 Milan, Italy.
E-mail: c.valsecchi18@campus.unimib.it

Key words: Nuclear Receptors; multi-task learning; machine learning; neural networks.

Disclosures: Authors have nothing to disclose.

Conference presentation: This paper was presented at the Third Centro 3R Annual Meeting - L'era delle 3R: modelli *in silico, in vitro* e *in vivo* per promuovere la ricerca traslazionale - 30 September - 1 October 2021, Evento online organizzato dal Politecnico di Torino.

Received for publication: 9 July 2021.
Accepted for publication: 7 September 2021.

Biomedical Science and Engineering 2021; 4(s1):198
doi:10.4081/bse.2021.198

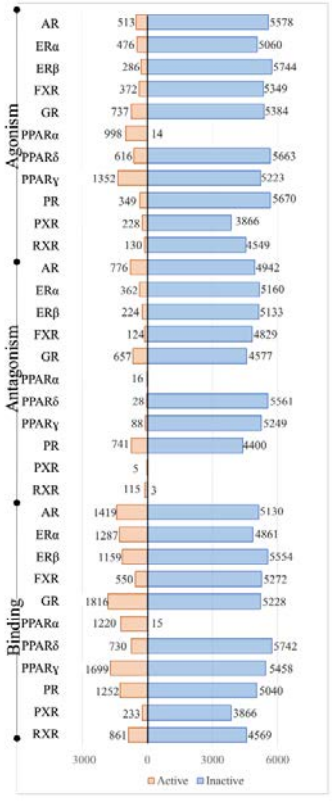

Figure 2. Distribution of molecules in the NURA dataset, divided into active (orange bar, activity lower than 10,000 nM), and inactive (blue bar, activity larger than 100,000 nM).

Table 1. Global classification measures expressed as the percentage of correctly classified active (Sn_T), inactive compounds (Sp_T) and their average (NER_T) achieved on the test set.[7]

Model	*NER_T*	*S_{pT}*	*Sn_T*
FFNL1	95.3	95.3	95.4
FFNL3	95.2	95.4	95.1

As shown in Table 1, all the *in-silico* approaches achieved good classification performances on test chemicals with a 95% of correctly classified compounds.

Discussion and Conclusions

The increased coverage of chemical and bioactivity space offered by NURA dataset result in a broader applicability domain and improved robustness of the developed computational models, which can help reducing animal testing by filling the data-gap and prioritizing chemicals. Furthermore, multi-task neural networks might offer other advantages, such as i) the possibility of leveraging information on related tasks, ii) modelling less represented tasks and iii) identifying ligands with selective or promiscuous activity.

References

1. Mansouri K, Abdelaziz A, Rybacka A, et al. CERAPP: Collaborative Estrogen Receptor Activity Prediction Project. Environ Health Perspect 2016;124:1023-33.
2. Mansouri K, Kleinstreuer N, Abdelaziz AM, et al. CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. Environ Health Perspect 2020;128:027002.
3. Santos R, Ursu O, Gaulton A, et al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov 2017;16:19-34.
4. Germain P, Staels B, Dacquet C, et al. Overview of nomenclature of nuclear receptors. Pharmacol Rev 2006;58:685-704.
5. Valsecchi C, Grisoni F, Motta S, et al. NURA: A curated dataset of nuclear receptor modulators. Toxicol Appl Pharmacol 2020;407:115244.
6. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010 May 24;50(5):742-54.
7. Valsecchi C, Collarile M, Grisoni F, et al. Predicting molecular activity on nuclear receptors by multitask neural networks. J Chemometr 2020, https://doi.org/10.1002/cem.3325