Hyperparameters used for RFR, SVMR and NN

Randomforest package [1] as implemented in R version 3.4.2 [2-3] was used to build all models using RFR algorithm. The number of trees was varied from 20 to 100, with the number of cross validations ranging from 2 to 10. Optimal prediction accuracy was obtained for combinations involving 40 decision trees with 5 cross validations. The split point used was 5 with maximum depth of 20. Bootstrap aggregation with replacement was used for sampling the data and the decision tree classifier used was 1 (default). Another package known as e1071 [4] was used to build models using SVMR algorithm in the same version of R. A standard grid search was followed to find optimized values for the two tunable parameters such as the cost (500) and epsilon parameters (0.05). The package Neuralnet [5] as implemented in R that uses a feed-forward neural network with 2 hidden layers with 100 repetitions was applied for building the models with NN algorithm.

Table SM-1: Descriptions of all machine learning models developed in this work.

Model Development	
Type	**Dataset Classification and Choice of Descriptors**
Descriptors Analysis	All 88 APIs and all chemically intuitive descriptors
AO-initial	78 APIs and all physicochemical properties descriptors
AO-initial	78 APIs, all physicochemical properties descriptors and all MOE-2D descriptors
AO1	78 APIs, all physicochemical properties descriptors and 10 most important MOE-2D descriptors.
AO2	Best model from above rebuilt with 66 APIs, as categorized by excluding APIs with average crystallization propensity strictly 0 and 1.
AO3	Best AO1 and AO2 models with datasets classified under various experimental factors as mentioned below.
AO3(a)	59 APIs after excluding single experimental factor such as: 'no crystals found'
AO3(b)	64 APIs after excluding single experimental factor such as: '>95% crystallized'.
AO3(c)	48 APIs after excluding single experimental factor such as: 'High propensity to crystallize due to multiple anhydrous, hydrate/solvated forms >5 forms/conversion from salt to free form'.
AO3(d)	63 APIs after excluding single experimental factor such as: 'conversion from salt to free form'.
AO3(e)	56 APIs after excluding single experimental factor such as: 'very high sol in most solvents >50 mg/ml'.
AO3(f)	57 APIs after excluding single experimental factor such as: 'color change'.
AO3(g)	61 APIs after excluding single experimental factor such as: 'amorphous/ oil SM'.
AO3(h)	63 APIs after excluding single experimental factor such as: 'chemical degradation'.
AO3(i)	60 APIs after excluding single experimental factor such as: 'purity of SM is low'.
AO3(j)	54 APIs after excluding both experimental factors such as: 'chemical degradation' and 'amorphous/ oil SM'.
AO3(k)	54 APIs after excluding both experimental factors such as: 'High propensity to crystallize due to multiple anhydrous, hydrate/solvated forms >5 forms/conversion from salt to free form' and 'conversion from salt to free form'.

AO3(l)	54 APIs after excluding both experimental factors such as: ‘chemical degradation’ and ‘conversion from salt to free form’.
AS1 and AS2	Best AO family model rebuilt using solvent identity as a descriptor. Model AS1 has 58 API entries with >=10 experimental outcomes, while model AS2 has 212 API entries with >=5 outcomes.
AS3 and AS4	Models AS3 and AS4 use same datasets as AS1 and AS2, respectively, with addition of solvent descriptors and experimental factors classification.
Selected Solvents Models	Best AS3 and AS4 models, each rebuilt for only one of five specific solvents: methanol, ethanol, chloroform, toluene and acetonitrile. Solvent descriptors were not explicitly used for these models.

Table SM-2: List of 20 most important API descriptors selected by the RFR algorithm for the best AO model

Name	Type	Importance†	Note
Chirality	Physicochemical	0.512	Presence of chiral centers
MW	Physicochemical	0.410	Molecular weight
cPFlogD	Physicochemical	0.144	logD
Accptdonr	Physicochemical	0.141	Difference between number of Hydrogen Bond Donors and Acceptors
ROTBND	Physicochemical	0.138	Number of Rotatable Bonds
PSA	Physicochemical	0.137	Polar Surface Area
HBACCPT	Physicochemical	0.130	Number of hydrogen bond acceptors
PEOE_VSA_FPOL*	MOE 2D	0.129	Fractional polar van der Waals surface area
PEOE_VSA_FPOS*	MOE 2D	0.126	Fractional polar positive van der Waals surface area
PEOE_VSA.0.2*	MOE 2D	0.123	Direct Electrostatic Interactions
Gao_chi4cav*	MOE 2D	0.110	Kier & Hall Connectivity
SMR_VSA3*	MOE 2D	0.107	Partial Charge
kS_ssCH2*	MOE 2D	0.104	Sum of E-state values
Gao_chi4pcv*	MOE 2D	0.098	Kier & Hall Connectivity
Rings*	MOE 2D	0.095	Topological
SlogP_VSA4*	MOE 2D	0.089	Subdivided surface areas
PEOE_VSA_PPOS*	MOE 2D	0.086	Total polar positive van der Waals surface area
HBDONR	Physicochemical	0.082	Number of hydrogen bond donors
NOCNT	Physicochemical	0.081	Number of Nitrogen and Oxygen atoms
ClogP	Physicochemical	0.078	Measurement of Partition Coefficient

† Reference [6]
* References [7-14]

Table SM-3: List of single solvents with their corresponding IDs, as used in AS models.

Solvent ID	Single Solvent
1	1-Butanol
2	1-Methyl pyrrolidone
3	1-Propanol
4	1;1;1-Trichloroethane
5	1;2-Dichloroethane
6	1;4-Dioxane
7	2-Butanol
8	2-Methyl tetrahydrofuran
9	2-Methyl-1-butanol
10	2-Propanol
11	2-Propyl ether
12	2;2;2-Trifluoroethanol
13	Acetic acid
14	Acetone
15	Acetonitrile
16	Benzyl alcohol
17	Butyl acetate
18	Chloroform
19	Cyclohexane
20	Dichloromethane
21	Diethyl ether
22	Diisopropyl ether
23	Dimethyl sulfoxide
24	Dimethylacetamide
25	Dimethylformamide
26	Ethanol
27	Ethyl acetate
28	Formamide
29	Heptane
30	Hexafluoro isopropanol
31	Hexane
32	Isopropyl acetate
33	Methanol
34	Methyl ethyl ketone
35	Methyl isobutyl ketone

36	Methyl tert-butyl ether
37	Nitromethane
38	Pyridine
39	Tert-butanol
40	Tetrahydrofuran
41	Toluene
42	Water

Table SM-4: List of 20 most important descriptors selected by the RFR algorithm for the best AS model

Name	**Type**	**Importance†**	**Note**
Chirality	Physicochemical	0.543	Presence of chiral centers
ROTBND	Physicochemical	0.405	Number of Rotatable Bonds
PSA	Physicochemical	0.142	Polar Surface Area
HBACCPT	Physicochemical	0.139	Number of hydrogen bond acceptors
MW	Physicochemical	0.138	Molecular weight
cPFlogD	Physicochemical	0.135	Solubility
PEOE_VSA_FPOS*	MOE 2D	0.128	Fractional polar positive van der Waals surface area
MW	Physicochemical (solvents)	0.126	Molecular weight of solvents
Melting Point	Experimental (solvents)	0.125	Physical Property
HBDONR	Physicochemical	0.119	Number of hydrogen bond donors
logP.o.w*	MOE 2D	0.115	Log octanol/water partition coefficient
Accptdonr	Physicochemical	0.111	Difference between number of Hydrogen Bond Donors and Acceptors
NOCNT	Physicochemical	0.108	Number of Nitrogen and Oxygen atoms
Dielectric Constant	Experimental (solvents)	0.097	Solvent property
balabanJ*	MOE 2D	0.088	Topological parameter
Dipole moment	Experimental	0.084	Polarity
PEOE_VSA.0.1*	MOE 2D	0.080	Direct electrostatic interaction
Q_VSA_FPNEG*	MOE 2D	0.065	Partial Charge Descriptors
Log S*	MOE 2D	0.061	Log solubility in water
ClogP	Physicochemical	0.059	Calculated Partition Coefficient

† Reference [6]
* References [7-14]

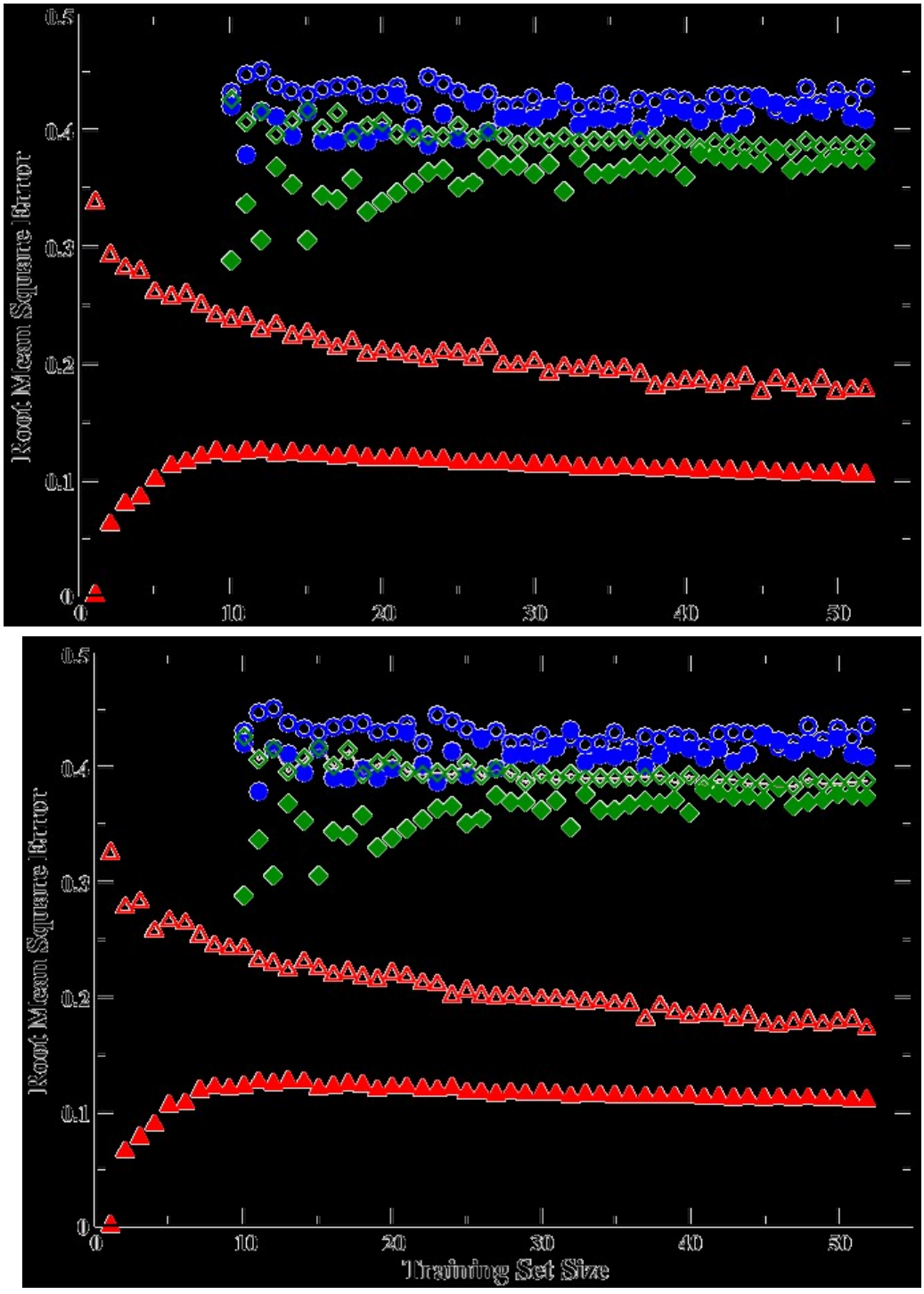

Figure SM-1: Comparative Learning Curves for (a) AS1 and (b) AS3 models constructed utilizing all three machine-learning algorithms.

References:

(1) https://cran.r-project.org/web/packages/randomForest/index.html
(2) Available at http://www.R-project.org.
(3) Algorithms are available at https://CRAN.R-project.org.
(4) https://cran.r-project.org/web/packages/e1071/index.html
(5) https://cran.r-project.org/web/packages/neuralnet/index.html
(6) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Model.* 2003, **43**,1947-1958.
(7) Clark, A.M; Labute, P.; Santavy, M. 2D Structure Depiction. *J. Chem. Inf. Model*. 2006, **46**, 1107-1123.
(8) MOE molecular descriptor package is available at https://www.chemcomp.com/MOE-Cheminformatics_and_QSAR.htm.
(9) Clark, A.M.; Labute, P. Detection and Assignment of Common Scaffolds in Project Database of Lead Molecules. *J. Med. Chem*. 2008, **52**, 469-483.
(10) Wildman, S.A.; Crippen, G.M. Prediction of Physiochemical Parameters by Atomic Contributions. *J.Chem. Inf. Comput. Sci*. 1999, **39**, 868-873.
(11) Labute, P. Binary QSAR: a new method for quantitative structure activity relationships. In *Proceedings of the 1999 Pacific Symposium*. World Scientific Publishing.
(12) Labute, P. A Widely Applicable Set of Descriptors. *J. Mol. Graph. Mod*. 2000, **18**, 464-477.
(13) Labute, P.; Nilar, S.; Williams, C. A Probabilistic Approach to High Throughput Drug Discovery; *Comb. Chem. High Throughput Screen.* 2002, **5**, 135-145.
(15) Labute, P. Derivation and applications of molecular descriptors based on approximate surface area. In *Chemoinformatics*. 2004, 261-278. Humana Press.