Supporting information

DESignSolvents: An Open Platform for Search and Prediction of the Physicochemical Properties of Deep Eutectic Solvents

Valeria Odegova[a], Anastasia Lavrinenko[a], Timur Rakhmanov[a,b], George Sysuev[a], Andrei Dmitrenko[a], Vladimir Vinogradov[*a]

[a] *International Institute "Solution Chemistry of Advanced Materials and Technologies", ITMO University, 191002 Saint Petersburg, Russia.*

[b] *Advanced Engineering School, Almetyevsk, Russia.*

Methods

Descriptors for machine learning were calculated using RDKit (2022.9.3) [1] based on the collected database. The full list of descriptors for each DES property, including their meanings and formulae, are given in Table S1 and Table S2. Several classical machine learning models were employed for prediction tasks, including decision tree, random forest, gradient boosting, k-nearest neighbors, support vector machines, and multilayer perceptron. We used open-source implementations from scikit-learn (0.0.post1) [2], xgboost (1.7.3) [3] and catboost (1.1.1) [4] Python (3.9.11) packages. For hyperparameter optimization, we performed a grid search and Bayesian search from scikit-learn (0.0.post1) [2] and scikit-optimize (0.9.0) [5] Python packages, respectively. The best model was selected by evaluating R^2 and RMSE metrics (Table S3). Hyperparameters delivering the best performance for each prediction task are given in Table S4. The importance of descriptors was assessed using Shapley values calculated with the SHAP (0.41.0) [6] Python package. Finally, the DESignSolvents web service was developed with the following technological stack: Python (Django (4.1.7)), Javascript () (Vanilla, Plotly, noUiSlider), Docker [7], Nginx [8], Ofelia, and PostgreSQL [9].

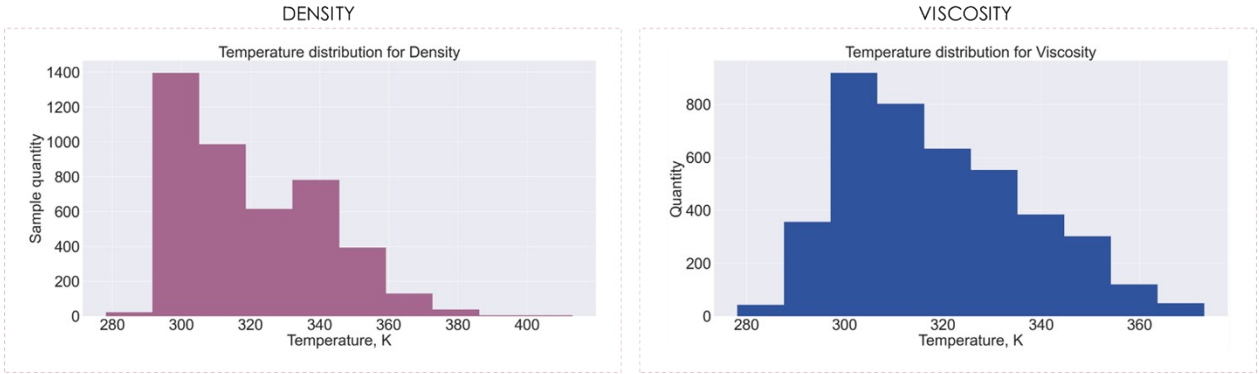

Figure. S1. Distribution for the temperature at which the density and viscosity were measured

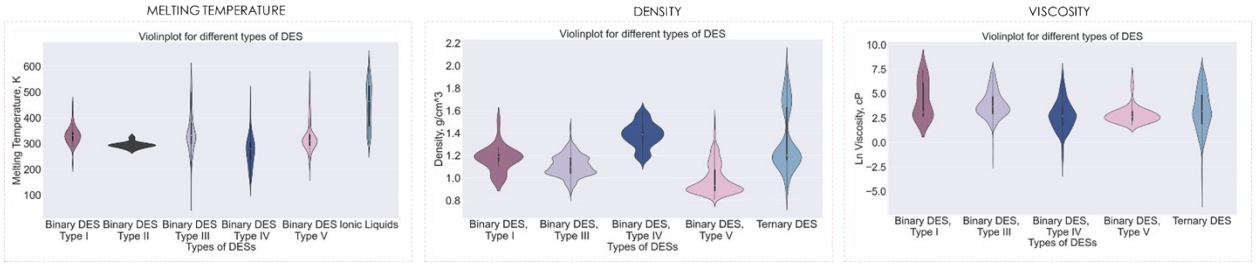

Figure. S2. Distribution for melting point, density and viscosity depending on the type of DES

Table S1. Descriptors for melting temperature prediction

Descriptor	Formula / A source	Meaning
Molar fraction of component	Experimental data	Molar fraction of individual component #1 (since only binary DES are used)
Type of DES	Experimental data	Type of DES (I, II, III, IV, V, IL)
Melting temperatures	Experimental data	Melting temperatures of individual components (in kelvin)
Molecular weight	$MW = MW1 * x_1 + MW2 * x_2 + MW * x_3$, where MW is the total molecular weight, MW_m is the molecular weight in individual component m, and x_m is the molar fraction of the individual component m.	The molecular weight a molecule calculated for the entire molecule, taking into account the molar fractions of individual components [1]
Number of hydrogen bond donors	$HD = HD1 * x_1 + HD2 * x_2 + HD3 * x_3$, where HD is the total number of H-bond donors, HD_m is the number of H-bond donors in individual component m, and x_m is the molar fraction of the individual component m.	Number of hydrogen bond donors calculated for the entire molecule, taking into account the molar fractions of individual components [1]
Number of different groups	$N_{Groups} = x_1 * n_1 + x_2 * n_2 + x_3 * n_3$, where N_{Groups} is the total number of different groups, x_m is the molar fraction of the individual component, and n_m is the number of different groups in the individual component.	The number of elements in DES, taking into account the mole fractions of individual components [1] Number of aliphatic carboxylic acids Number of aromatic carboxylic acid Number of aromatic nitrogens Number of aromatic hydroxyl groups Number of Tertiary amines Number of Secondary amines Number of amides
Number of aromatic rings	$AROM = AROM1 * x_1 + AROM2 * x_2 + AROM3 * x_3$, where $AROM$ is total number of aromatic rings, $AROM\_m$ is the number of aromatic rings in individual component m, and x_m is the molar fraction of the individual component m.	Number of aromatic rings in DES, taking into account the mole fractions of individual components [1]
Chemical toxicity evaluation	$ALERTS = ALERTS1 * x_1 + ALERTS2 * x_2 + ALERTS3 * x_3$, where $ALERTS$ is the total chemical toxicity evaluation, $ALERTS\_m$ is the chemical toxicity evaluation in individual component m, and x_m is the molar fraction of the individual component m.	Chemical toxicity evaluation of DES, taking into account the mole fractions of individual components [1]
Number of heavy atoms	$HM = HM1 * x_1 + HM2 * x_2 + HM3 * x_3$, where HM is the total number of heavy atoms, HM_m is the number of heavy atoms in individual component m, and x_m is the molar fraction of the individual component m.	Number of heavy atoms in DES, taking into account the mole fractions of individual components [1]

Table S2. Descriptors for density and viscosity prediction

Descriptor	Formula / A source	Meaning
Molar fractions	Experimental data	Molar fractions of individual components in DES
Type of DES	Experimental data	Type of DES (I, II, III, IV, V, Ternary)
Temperature	Experimental data	The temperature at which the density or viscosity measurement was carried out
VdWVolume	$VvdW = \sum all\ atom\ contributions - 5.92 * NB - 14.7 * RA - 3.8 * RNA$, where NB is the number of bonds, RA is the number of aromatic rings, and RNA is the number of nonaromatic rings. To calculate the contributions of atoms, we use the ball volume formula: $\bar{V}_{ball} = \frac{4}{3} * \pi * \bar{R}^3$, where *R* is Van der Waals radius of atom. To calculate the number of bonds, we use the following formula: $NB = N - 1 + RA + RNA$, where N is the total number of atoms.	Van der Waals volumes of individual components calculated using Van der Waals radii obtained from RDKit [1,10]
NumHeteroatoms	$NHet = NHet1 * x_1 + NHet2 * x_2 + NHet3 * x_3$, where $NHet$ is the total number of heteroatoms, $NHet\_m$ is the number of heteroatoms in individual component m, and $x_m$is the molar fraction of the individual component m.	The number of heteroatoms in a molecule calculated for the entire molecule, taking into account the molar fractions of individual components [1]
RingCount	$RingC = RingC1 * x_1 + RingC2 * x_2 + RingC3 * x_3$, where $RingC$ is the total number of rings, $RingC\_m$ is the number of rings in individual component m, and x_m is the molar fraction of the individual component m.	The number of rings in a molecule calculated for the entire molecule, taking into account the molar fractions of individual components [1]
InertialShapeFactor	$Inertial\ Shape\ Factor = \frac{}{pm_1 * pm_3}$, where pm_n is the principal moment of inertia *n* of DES calculated by the following formula: $pm_n = PMI\_1_n * x_1 + PMI\_2_n * x_2 + PMI\_3_n * x_3$, where PMI_m_n is the principal moment of inertia n of DES of individual component m and x_m is the molar fraction of the individual component m.	Inertial shape factor (it is related to the moment of inertia of the molecule and the characteristics of its shape) of DES taking into account the molar fractions of individual components [1,11]
SpherocityIndex	$Spherocity\ Index = 3 * \frac{}{pm_1 + pm_2 + pm_3}$, where pm_n is the principal moment of inertia *n* of DES calculated by the following formula: $pm_n = PMI\_1_n * x_1 + PMI\_2_n * x_2 + PMI\_3_n * x_3$, where PMI_m_n is the principal moment of inertia n of DES of individual component m and x_m is the molar fraction of the individual component m.	Spherocity Index (related to the shape of molecules and aspects that can be associated with sphericity or elongation) of DES taking into account the molar fractions of individual components [1,11]
Mass fraction of metal	$MetalMassFraction = \frac{}{x_1 * mw_1 + x_2 * mw_2 + x_3 * mw_3}$, where x_m is the molar fraction of the individual component, n_m is the number of metals in the individual component, $MW(Me)$ is the molar mass of metal, and mw_n is the molar mass of the individual component.	Mass fraction of metal in DES, taking into account the mole fractions of individual components
Number of elements	$N_{Elem} = x_1 * n_1 + x_2 * n_2 + x_3 * n_3$, where x_m is the molar fraction of the individual component and n_m is the number of elements in the individual component m.	The number of elements in DES taking into account the mole fractions of individual components [1] Density: Li, C, N, O, F, Na, Mg, Al, P, S, Cl, K, Ca, Cr, Mn, Fe, Co, Ni, Cu, Zn, Br

		Viscosity: Li, C, N, O, F, Mg, Al, P, S, Cl, K, Cr, Mn, Fe, Co, Cu, Zn, Br

Table S3. Metrics for evaluating the accuracy of machine learning models

Metric	Meaning	Formula
Coefficient of Determination (R^2)	A metric that shows the proportion of the variance in the response variable of a regression model that can be explained by the predictor variables[12].	$\sum_i$ where y_i is the true value of the parameter, $\hat{y}_i$ is the predicted value of the parameter, and $\underline{y_i}$ is the average value of the variable.
Root Mean Square Error (*RMSE*)	A metric that tells how far apart the predicted values are from the observed values in a dataset, on average [12].	$\sqrt{\sum_{i=1}^{n}}$ where y_i is the true value of the parameter and $\hat{y}_i$ is the predicted value of the parameter.
Average absolute relative deviation (*AARD*)	A metric used to measure the accuracy of a model by calculating the average percentage difference between predicted and actual values [13].	$\frac{1}{n}\sum \frac{\quad}{\lvert y_i \rvert}$ where n is the total number of data points or observations, y_i is the true value of the parameter, and $\hat{y}_i$ is the predicted value of the parameter.

Table S4. Optimal hyperparameters for the best models

Property	**ML model**	**Hyperparameters**
Melting point	Cat Boosting Regression	iterations: 123 learning_rate: 0.061 depth: 6
Density	Cat Boosting Regression	iterations: 600 learning_rate: 0.050 depth: 4
Viscosity	Cat Boosting Regression	iterations: 600 learning_rate: 0.050 depth: 6

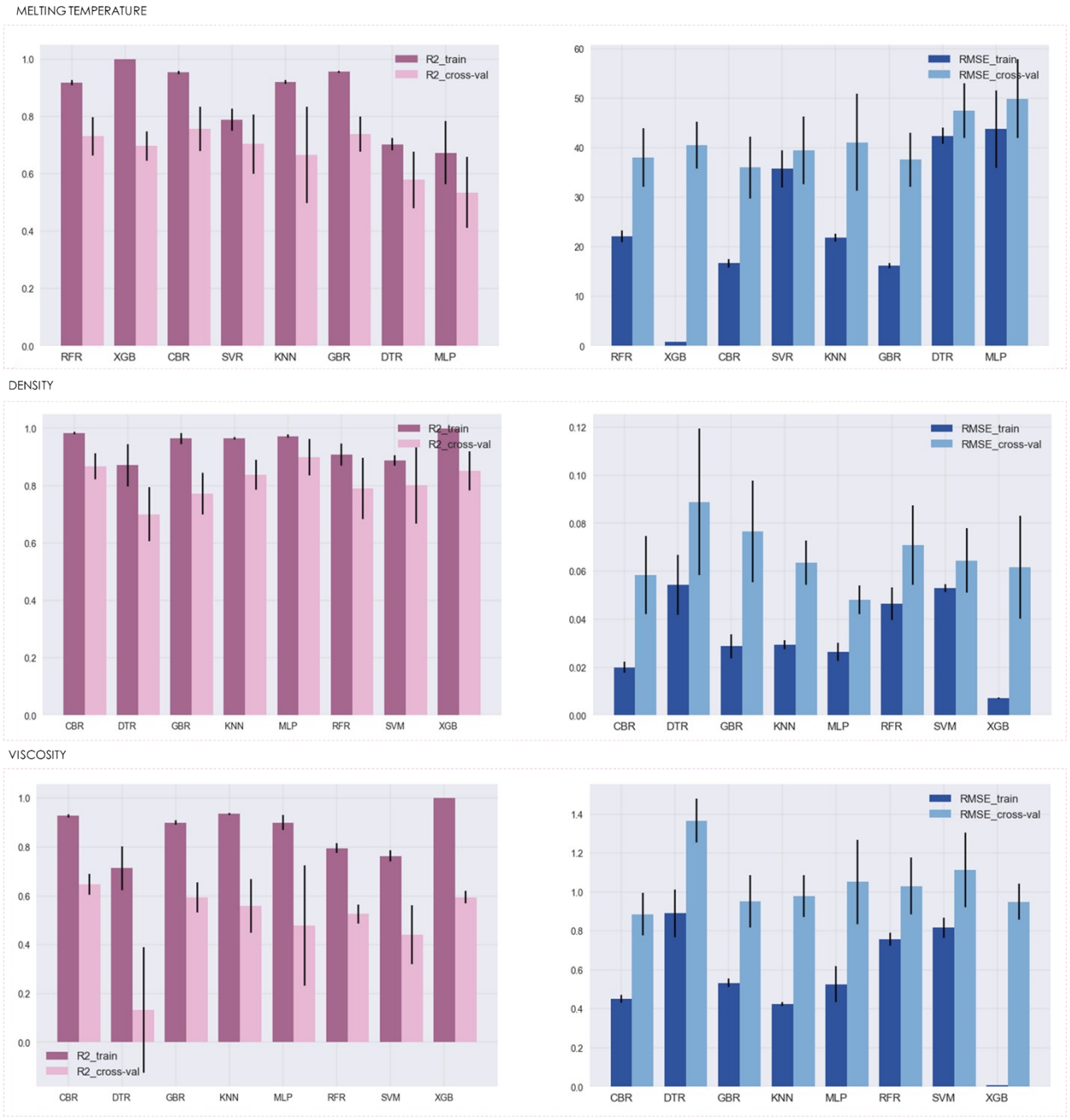

Figure. S3. Performance of predictive machine learning models

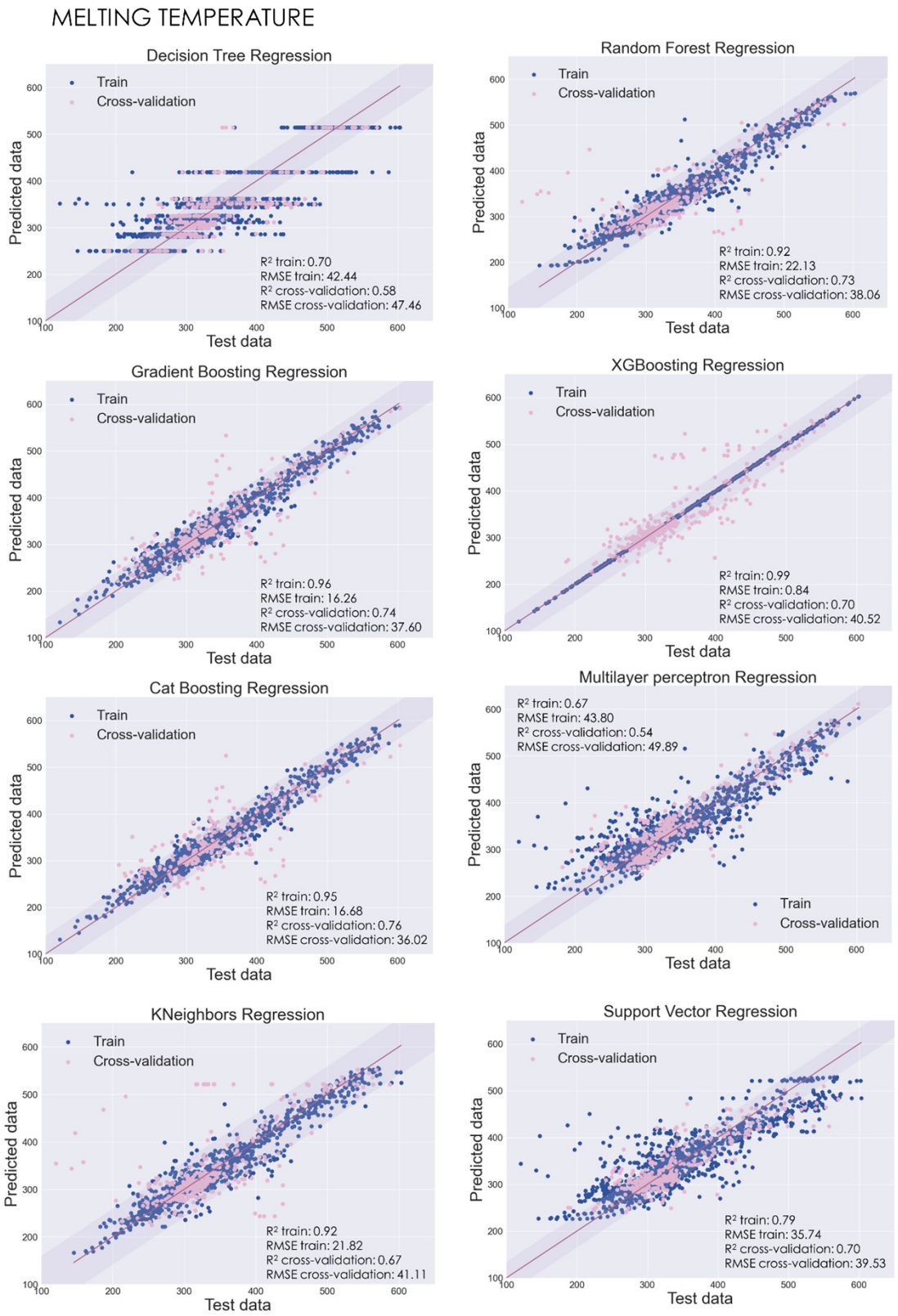

Figure. S4. Metrics for different machine learning models for melting temperature

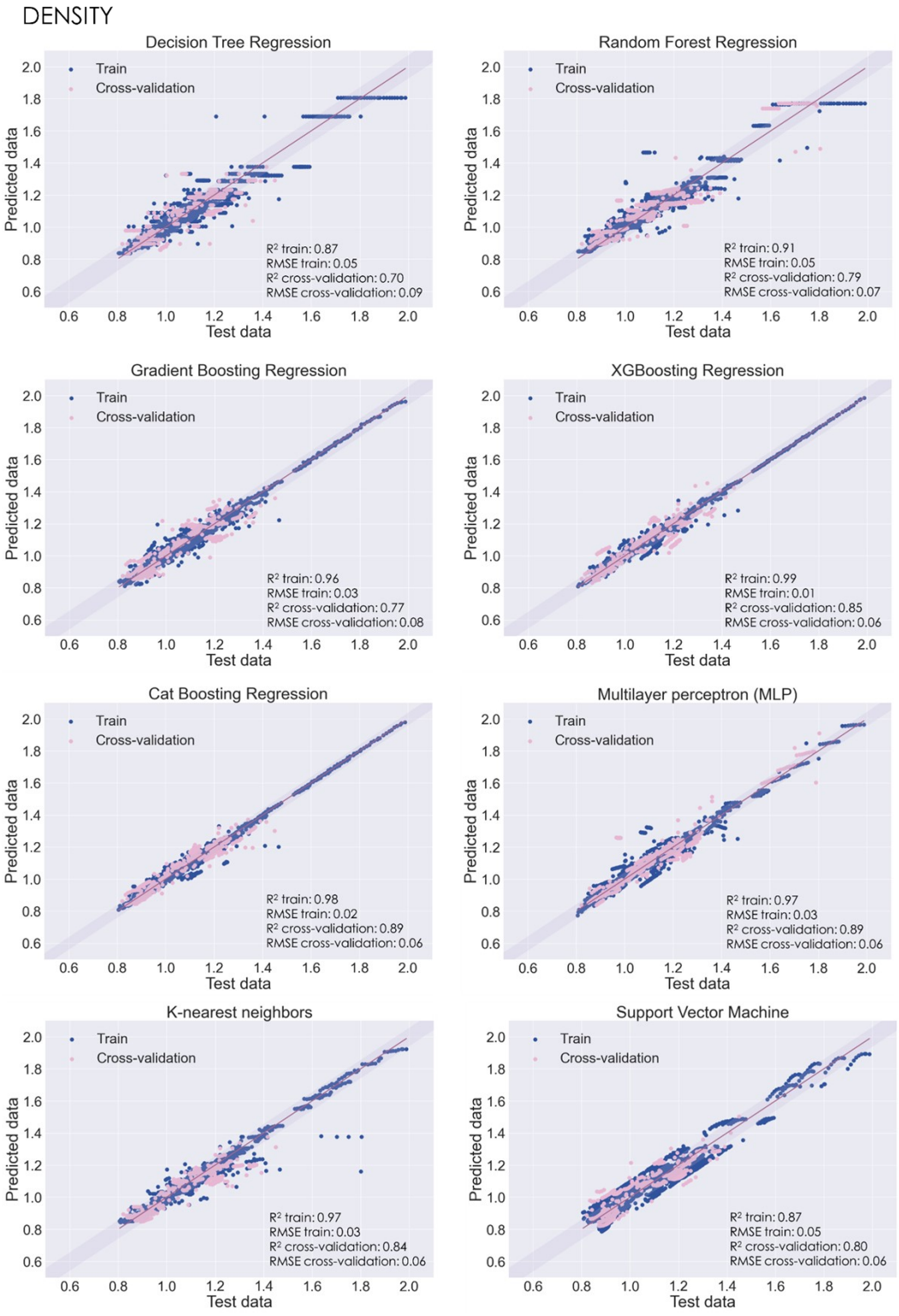

Figure. S5. Metrics for different machine learning models for density

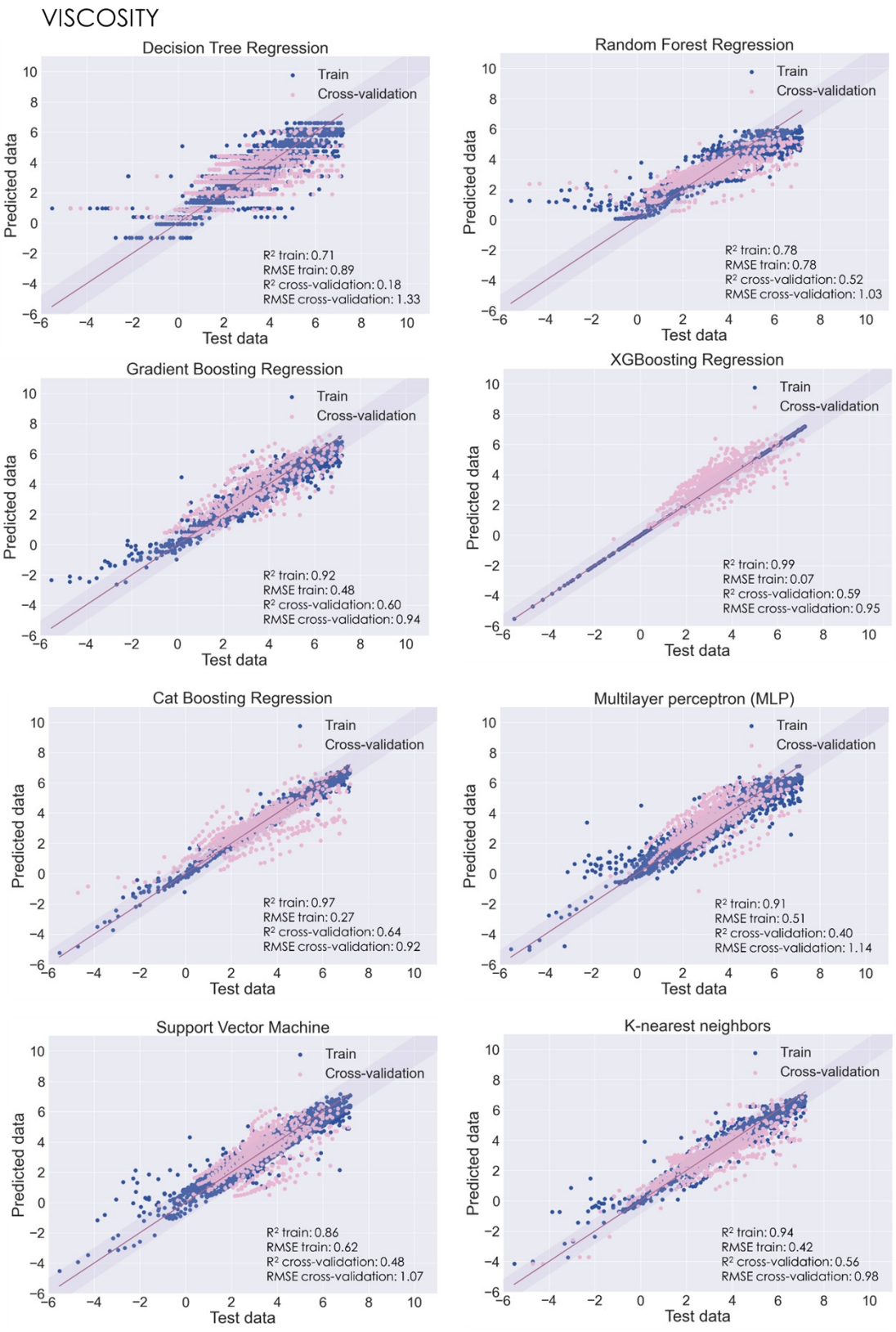

Figure. S6. Metrics for different machine learning models for viscosity

Figure. S7. RMSE for different types of deep eutectic solvents

SI References

1 RDKit: Open-source cheminformatics; http://www.rdkit.org .

2 F. Pedregosa FABIANPEDREGOSA, V. Michel, O. Grisel OLIVIERGRISEL, M. Blondel, P. Prettenhofer, R. Weiss, J. Vanderplas, D. Cournapeau, F. Pedregosa, G. Varoquaux, A. Gramfort, B. Thirion, O. Grisel, V. Dubourg, A. Passos, M. Brucher, M. Perrot andÉdouardand, andÉdouard Duchesnay and Fré. Duchesnay EDOUARDDUCHESNAY, *Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot*, 2011, vol. 12.

3 T. Chen and C. Guestrin, in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, 2016, vol. 13-

17-August-2016, pp. 785–794.
4 L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush and A. Gulin, *CatBoost: unbiased boosting with categorical features*.
5 T. Head, M. Kumar, H. Nahrstaedt and G. Louppe, *scikit-optimize/scikit-optimize (v0.9.0). Zenodo*.
6 S. M. Lundberg and S.-I. Lee, *A Unified Approach to Interpreting Model Predictions*.
7 D. Merkel, *Linux Journal*.
8 W. Reese, *Linux Journal*.
9 *PostgreSQL: The World's Most Advanced Open Source Relational Database*, 2019.
10 Y. H. Zhao, M. H. Abraham and A. M. Zissimos, *Journal of Organic Chemistry*, 2003, **68**, 7368–7373.
11 Peter. Comba and T. W. Hambley, *Molecular modeling of inorganic compounds*, Wiley-VCH, 2001.
12 D. Chicco Corresp, M. J. Warrens, G. Jurman and D. Chicco, *The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE, and RMSE in regression analysis evaluation analysis evaluation 4*, 2021.
13 W. Su, Y. Hwang, Y. shao, S. Deng, L. Zhao, X. Nie and Y. Zhang, *Energy*, 2019, **166**, 414–425.