Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints

Supporting Information

Olexandr Isayev,[†] Denis Fourches,[†] Eugene N. Muratov,[†] Corey Oses,[‡] Kevin Rasch,[‡] Alexander Tropsha,[*,†] and Stefano Curtarolo[*,‡,¶]

Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599, United States, Center for Materials Genomics, Duke University, Durham, North Carolina 27708, United States, and Materials Science, Electrical Engineering, Physics and Chemistry, Duke University, Durham, North Carolina 27708, United States

E-mail: alex_tropsha@unc.edu; stefano@duke.edu

Table 1 provides additional statistical information of the materials cartograms (see Figure 3). In network theory, a "component" is a group of nodes that are all connected to each other. A "giant component" is a connected component of a given random graph that contains a constant fraction of the entire graph's vertices.[1] Figures in parenthesis are calculated by fitting only the asymptotic portion of the curve in Figure 3(b).

[*]To whom correspondence should be addressed

[†]Laboratory for Molecular Modeling, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599, United States

[‡]Center for Materials Genomics, Duke University, Durham, North Carolina 27708, United States

[¶]Materials Science, Electrical Engineering, Physics and Chemistry, Duke University, Durham, North Carolina 27708, United States

Table 1: Topological properties for constructed materials cartograms

	D-fingerprints Network	B-fingerprints Network
Total number of cases	17420	17420
Giant component	10521 (60.4%)	15535 (89.2%)
Edges	466,000	564,000
Average degree	88.60	72.59
Network diameter (edges)	27	23
Power law γ	2.745	0.916 (2.04)

Table 2: Statistical characteristics of the continuous QMSPR models for superconductivity

Model	N	Q^2(ext)	RMSE	MAE
RF-SiRMS	295	0.64	0.24	0.18%
PLS-SiRMS	295	0.61	0.25	0.20%
Consensus	295	0.66	0.23	0.18%

Table 2 provides additional statistical information of the continuous model (see Methods Section). We define the following abbreviations/metrics: Q^2(ext) refers to the leave-one-out five-fold external cross-validation coefficient, RMSE refers to root-mean-square error, MAE refers to the mean absolute error, RF-SiRMS refers to the application of the Random Forest technique with Simplex descriptors, PLS-SiRMS refers to the application of the Partial Least Squares regression technique with Simplex descriptors, and concensus refers to the average of the RF-SiRMS and PLS-SiRMS results.

Table 3 provides additional statistical information of the classification model (see Methods Section). We define the following abbreviations/metrics: accuracy is determined by the ratio of correct predictions to the total number of predictions, sensitivity is determined by the ratio of correctly predicted $T_c > T_{thr}$ to the number of empirical $T_c > T_{thr}$, specificity is determined by the ratio of correctly predicted $T_c \leq T_{thr}$ to the number of empirical $T_c \leq T_{thr}$, CCR (correct classification rate) is the average of the sensitivity and the specificity, and coverage is determined by the ratio of the total number of predictions to the total number of cases.

Table 4 lists the the fragments that show the greatest impact on T_c variation (see Figure

Table 3: Statistical characteristics of the classification QMSPR models for superconductivity

	No Applicability Domain[2]	With Applicability Domain[2]
Total number of cases	464	464
Total number of predictions	464	451
Number of correct predictions	452	446
Number of wrong predictions	12	5
Number of empirical $T_c > T_{thr}$	29	22
Number of empirical $T_c \leq T_{thr}$	435	429
Number of correctly predicted $T_c > T_{thr}$	19	17
Number of correctly predicted $T_c \leq T_{thr}$	433	429
Number of incorrectly predicted $T_c > T_{thr}$	2	0
Number of incorrectly predicted $T_c \leq T_{thr}$	10	5
$T_c > T_{thr}$ prediction value	0.90	1.00
$T_c \leq T_{thr}$ prediction value	0.98	0.99
Accuracy	0.97	0.99
Sensitivity	0.66	0.77
Specificity	1.00	1.00
CCR	0.83	0.89
Coverage	1.00	0.97

Table 4: Top statistically significant fragments and their contributions to T_c variation

Fragment name	Contribution to $log(T_c)$ Score
O-Cu-O	18%
Periodic groups IB-IIB-IVA[21]	14%
Periodic groups IIA and[22] IB	12%
As, As, Fe fragment count	5%
Periodic groups IIB-IVA[21]	5%
Periodic groups IIA and[22] IVA	5%
Charges[3] (-1.5)(-1.5)(+2.5)	3%
O element count	2%
Cu element count	2%
O, O, O fragment count	2%
Charge[3] (+2.5)	2%
Nb element count	2%
Charge[3] (-1.5)	2%

5). *NB:* Regarding the fragment descriptions, "-" demonstrates that the collection is bonded, while "and" demonstrates that the collection is not bonded.

References

(1) Chung, F.; Lu, L.; of the Mathematical Sciences, C. B.; (U.S.), N. S. F. *Complex Graphs and Networks*; American Mathematical Society, 2006; Chapter 6. The Rise of the Giant Component.

(2) Tropsha, A.; Golbraikh, A. *Curr. Pharm. Des.* **2007**, *13*, 3494–3504.

(3) Bader, R. *Atoms in Molecules: A Quantum Theory*; Oxford University Press, Incorporated, 1994.