Towards Crystalline Porosity Estimators for Porous Molecules

SUPPORTING INFORMATION

Ismael Gómez García and Maciej Haranczyk

1. Assessing performance of classification models

To assess the performance of the classification models trained during this work, both Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves were studied, along with their areas under the curve (ROC-AUC and PR-AUC, respectively). In this section, we discuss the selection of these methods for validation.

ROC curves represent the true positive rate against the false positive rate over a moving classification threshold. This type of curve represents the model's classification power in terms of misclassification of positive cases (i.e., a model that has a lot of false positives will have poor ROC-AUC). Models are considered adequate classifiers if they have ROC-AUC higher than 0.7, and a perfect classifier will have ROC-AUC of 1. The PR curves are added as recent advances in model analysis have shown that imbalance in data testing (i.e. predominance of the negative class over the positive) can artificially increase the ROC-AUC statistic[1]. The PR curves measure how well the model recovers positive cases in a pool where the negative cases predominate, as is the case in the classification models we present in here (in which the skew is 0.1 for all the models). These plots represent precision (i.e., proportion of true positives over all positives predicted by the model) versus recall (i.e., proportion of true positives over the actual positives) for a moving classification threshold. A perfect model would have a squared shape touching the top-right corner and a PR-AUC of 1, whereas a random model would have a PR curve under the baseline, which is a horizontal line at the random precision of the population (shown as a dashed line in the PR diagrams). This horizontal line is placed at the precision of a random classifier, which coincides with the skew of the dataset[2]. The area under the precision recall curve (PR-AUC) indicates that the model presents an improvement over random structure selection if the value exceeds the skew. Models with higher PR-AUC perform better. A classifier must then have PR-AUC at least superior to the skew of the dataset to improve the prediction capacity of randomly selecting models. In this work, we consider the models as to be good if they show at least 4-fold improvement over the prediction capacity of a random selection (i.e. PR-AUC of 0.4 or more).

2. Correlations

In this work, we studied the pairwise Spearman correlations for all pairs of molecular-molecular, material-material, and molecular-material descriptors. All pairwise correlations are presented in the correlogram of Fig. S1. It can be seen that the highest correlation values correspond to either molecular-molecular pairs of descriptors, or to material-material pairs of descriptors. The correlations for molecular-material pairs of descriptors are generally moderate to low.

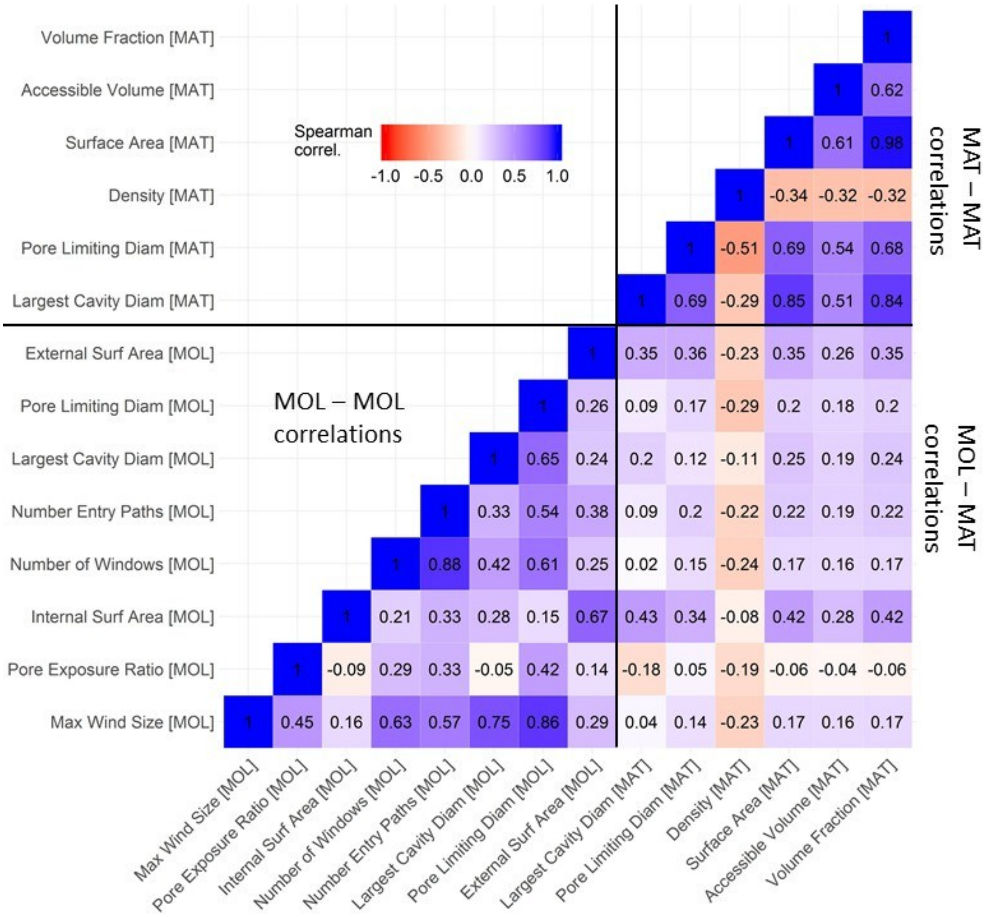

Figure S1. Pairwise correlations for molecular and material descriptors analysed. Positive correlations are reflected in purple and blue colour, whereas negative correlations are reflected in orange and red. Higher colour intensity reflects higher associations.

3. Correlation evolution

To extend the analysis of basic correlations, we considered the correlation evolution for different material properties. In this section, we present the complete list of molecular-molecular pairs of correlations with a ΔK > 0.4 and final absolute correlation value of at least 0.5. It can be seen from Table S1 that the two properties with more emerging correlations are LCD and SA.

Material property	**Pair of molecular properties**	**ΔK**
Largest Cavity Diameter	ISA - NEP	0.42
	ISA - NW	0.44
	mLCD – ESA	0.43
	mPLD - ESA	0.51
	mPLD - ISA	0.62
	MWS - ISA	0.63
	MWS – ESA	0.53
Pore Limiting Diameter	None	
Density	ISA - NEP	0.41
	ISA - NW	0.49
	MWS - ESA	0.42
	mPLD - ESA	0.44
	mLCD - ESA	0.41
Surface Area	ISA - NEP	0.47
	PLD - ISA	0.59
	PER - NW	1.12
	PER – NEP	1.04
	PER – ISA	0.6
	MWS - ISA	0.59
	LCD – PER	0.66
	LCD - NW	0.46
	LCD – ISA	0.52
	ISA - NW	0.73
Accessible volume	mPLD - ESA	0.44
	MWS – ESA	0.44
Volume fraction	MWS - ESA	0.41

Table S1. Emerging molecular correlations (ΔK > 0.4) for the six material properties considered in this study.

4. Performance of the classifier for structures with solvent

We extend here the results from the main article, exposing the performance of the classification random forest model for the subset of structures with solvent (14659 structures). This model has slightly worse performance than the model for the entire dataset, thus being of potentially less interest, as in the cases where solvent can't be removed, the more general model should produce similar or better results.

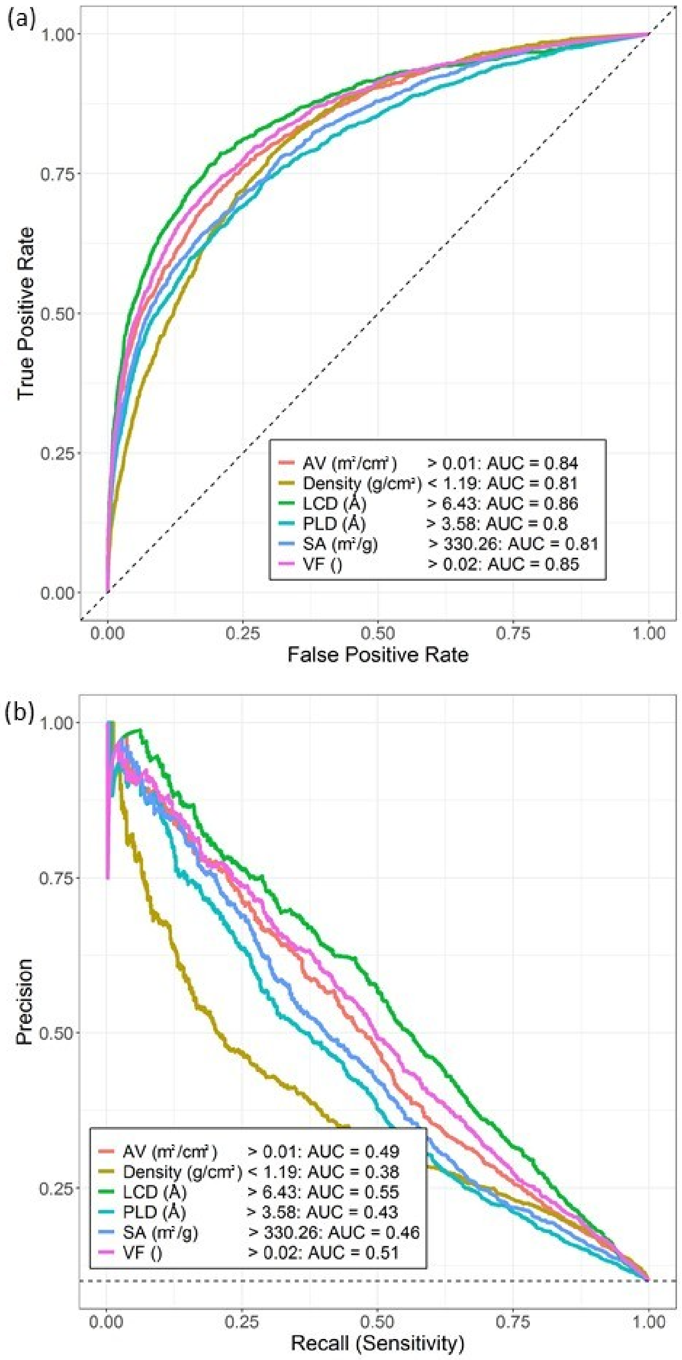

Figure S2. Performance of the random forest model trained for the solvent-only dataset. (a) ROC curves, along with the AUC values for the six models trained (one per material property). (b) PR curves, along with the AUC values for the six models trained (one per material property).

5. Error distribution for the regressor RF model

To better assess the performance of the random forest models for regression, we studied the distribution of the absolute errors for the six models presented in the main text (no solvent, NW > 2, NE/NW = 1). In Fig. S3, the error distribution for the models can be found. It can be seen that most structures group at the left of the mean error.

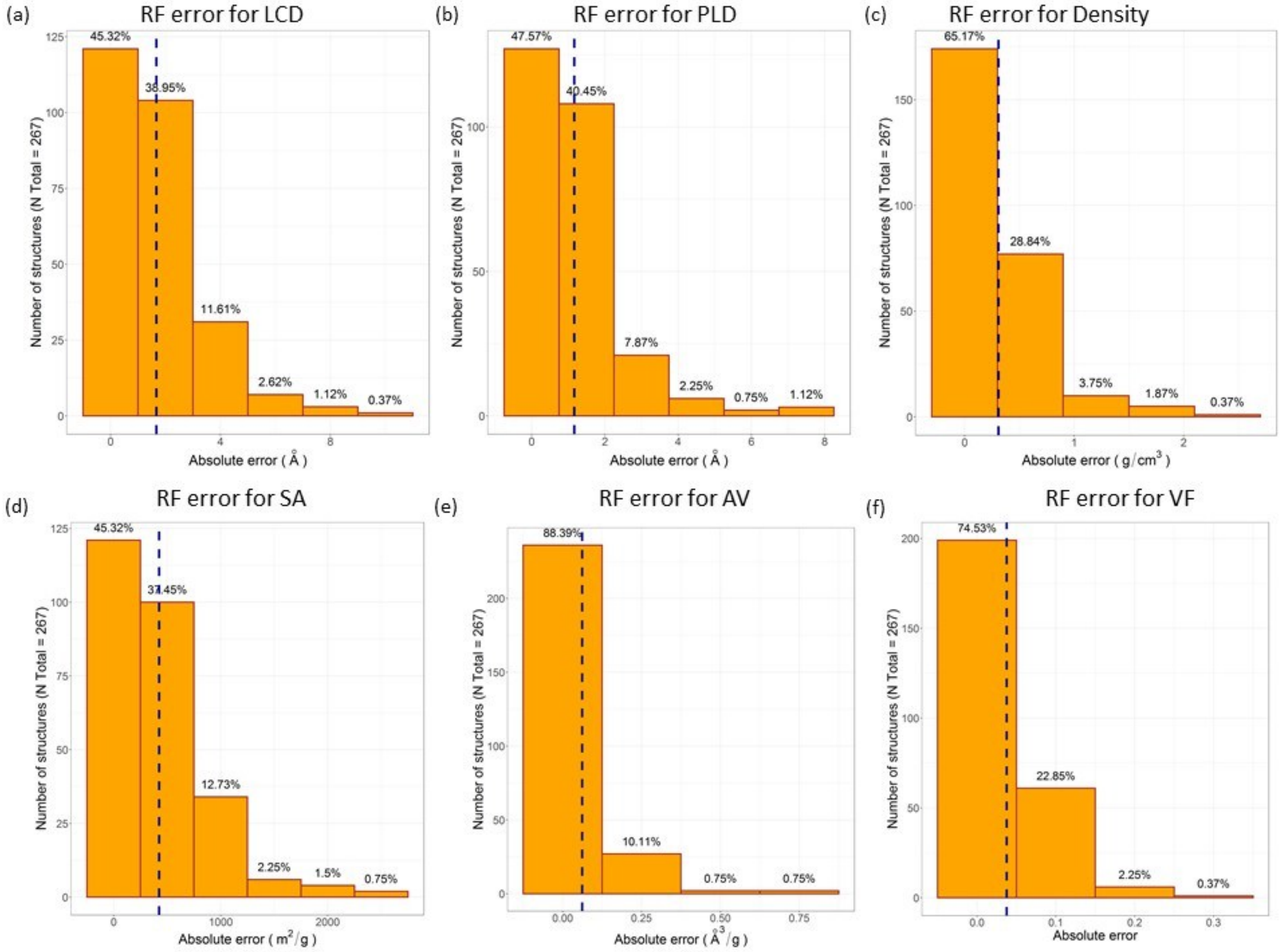

Figure S3. Absolute error distribution plots for the six random forest regression models (histograms). The blue dashed line indicates the mean absolute error, whereas orange bars indicate the number of structures placed in a given interval of error.

6. Other RF regression models

During the development of this project, several random forest models for regression were considered. In Figure S4, the performance for the six RF models trained for the entire dataset can be found. The models have moderate performance (R^2 between 0.3 and 0.4), being generally worse than the models presented in the main text. For some of the models, this can be explained by the strong predominance of zero-valued points, which may bias the model. The presence of solvent is also a factor that may difficult the prediction of porosity. Note that these models generally outperform the linear models (see Section 6), even using out-of-bag predictions compared with the internal validation used for linear models.

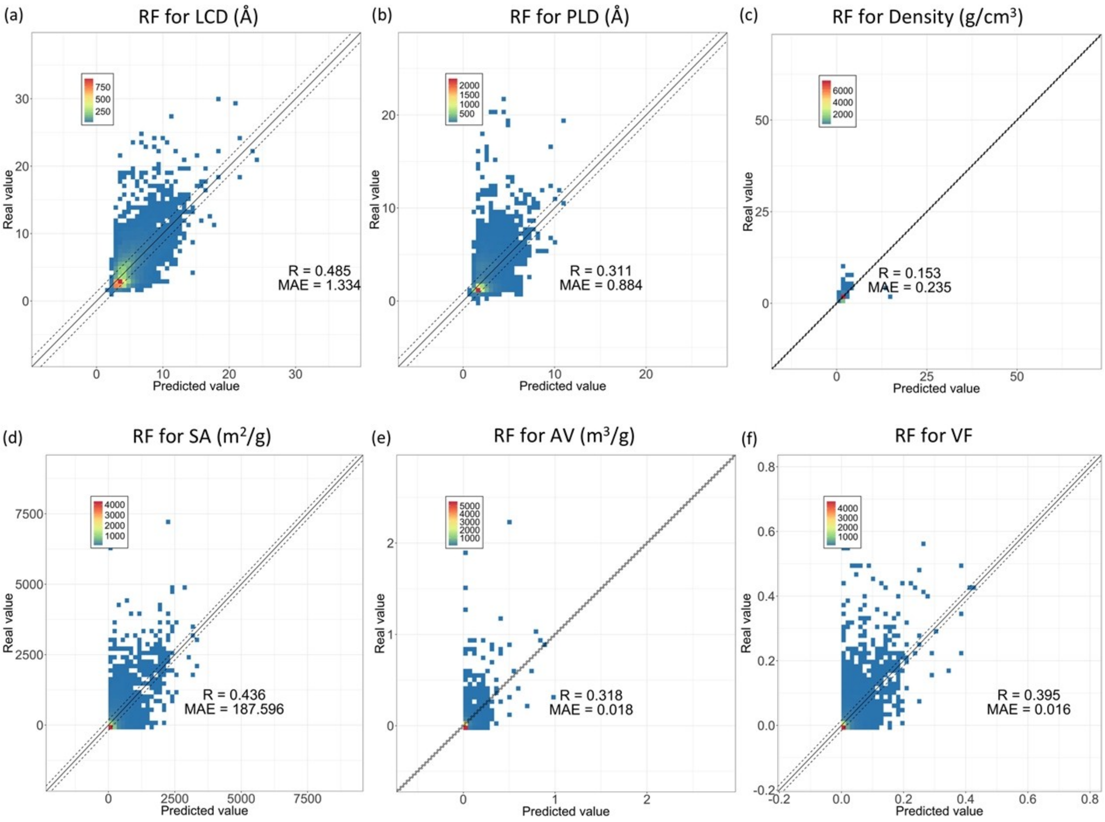

Figure S4. Predicted versus real values for regression models for the six material porosity measures predicted: LCD (a), PLD (b), density (c), SA (d), AV (e) and VF (f). Color-coding reflects density (i.e. number of structures placed in that region of the diagram). The models were trained over the 17832 structures that form our dataset. The counts per colour are presented in the colour bar within each graph. Diagonal lines represent the line of perfect prediction (continuous) and margins at one mean absolute error distance (dashed).

In Figure S5, a set of RF models trained for the subset of structures without solvent can be found. It can be seen that the models have a reasonable performance (except for the model for density), although inferior to the ones presented in the main paper. The high number of structures placed around values close to 0 in several models seem to bias the predictions, reducing the model accuracy.

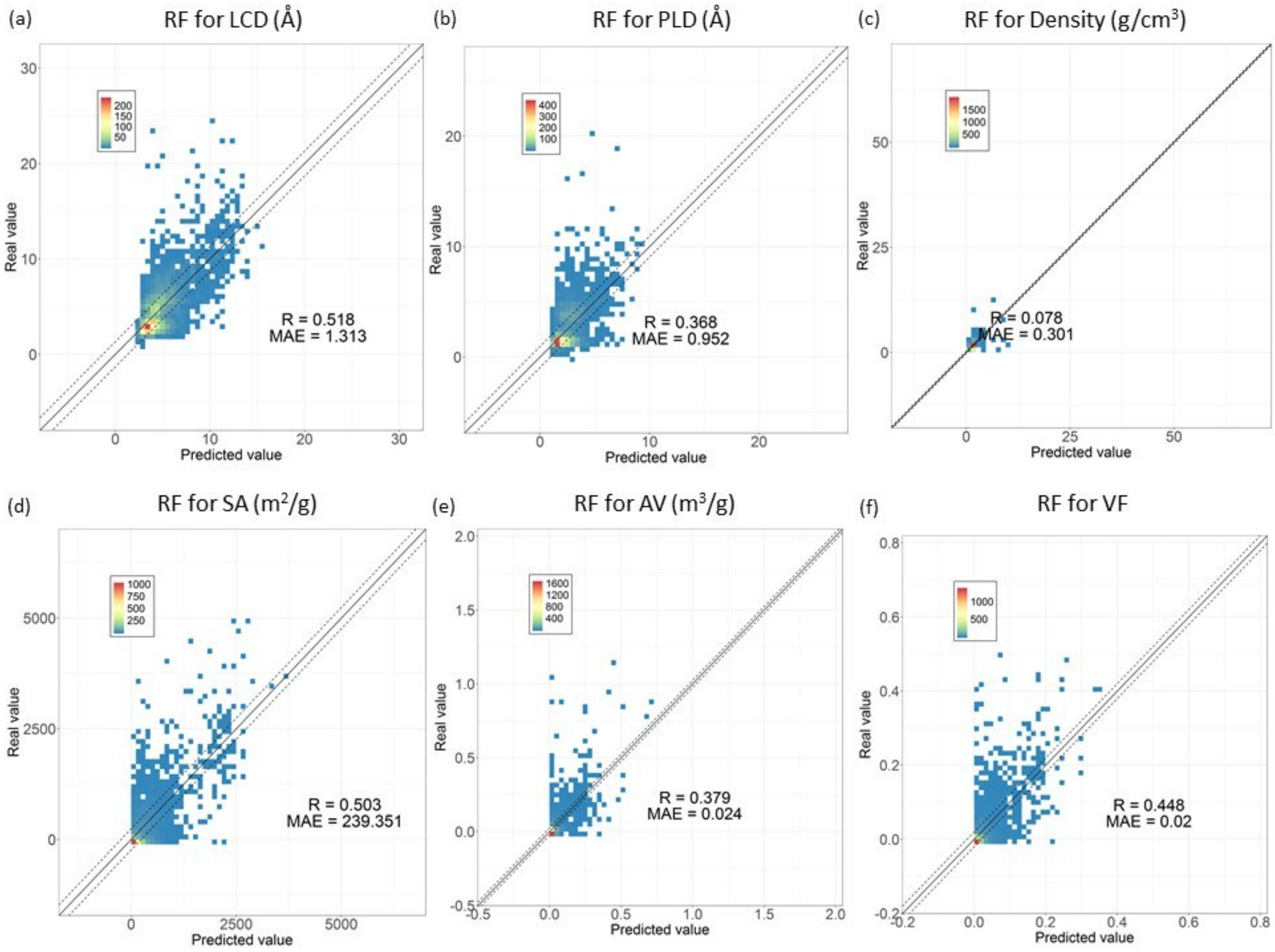

Figure S5. Predicted versus real values for regressor models for the six material porosity measures predicted: LCD (a), PLD (b), density (c), SA (d), AV (e) and VF (f). Colour coding reflects density (i.e. number of structures placed in that region of the diagram). The models were trained over the 3173 structures without solvent. The counts per colour are presented in the colour bar within each graph. Diagonal lines represent the line of perfect prediction (continuous) and margins at one mean absolute error distance (dashed).

7. Linear Regression Models

As a first attempt to train classification models to predict material properties from molecular descriptors, we used linear models. This kind of models has the advantage of being simpler than random forest, with the possibility to summarize them in terms of a simple equation. Additionally, they are more interpretable in terms of the contribution of each molecular descriptor to the material property. To avoid collinearity, we selected a subset of molecular porosity descriptors as predictor variables, avoiding the combinations of variables that were too strongly related (i.e. those with absolute correlation higher than 0.7). Correlation among molecular predictors can be found in Section 1 of this document. We computed multiple regression models for the six material properties analysed during this study. The parameters of these models can be found in Table S2, including the effect sizes (α_i's) and significance values under Wald’s t-test. In the models, most or all predictors are strongly significant, indicating the existence of an association between molecular and material properties. Effect sizes are not scaled and should be considered in reference to their average value. The parameters of the model can be inserted in the linear model equation to obtain an estimate of the material porosity parameter of interest:

$$y_{MAT} = \alpha_0 + \alpha_1 \cdot MWS + \alpha_2 \cdot PER + \alpha_3 \cdot ISA + \alpha_4 \cdot NW + \alpha_5 \cdot ESA \qquad (1)$$

The value of this estimate, however, should be used conservatively, as the moderate R^2 of the linear models indicate that the variance of the set of molecules is high.

	Predictors												**Model performance**	
	Intercept		MWS		PER		ISA		NW		ESA			
Outcome	α_0	p-value	α_1	p-value	α_2	p-value	α_3	p-value	α_4	p-value	α_5	p-value	MAE	Adj. R^2
LCD	1.88	$<10^{-5}$	0.32	$<10^{-2}$	-2.7	$<10^{-3}$	0.0012	$<10^{-7}$	0.28	$<10^{-4}$	$9 \cdot 10^{-5}$	$<10^{-7}$	1.98	0.38
PLD	0.77	$<10^{-2}$	0.1	0.18	-1.32	0.02	$3 \cdot 10^{-3}$	0.07	0.16	$<10^{-3}$	$6 \cdot 10^{-5}$	$<10^{-7}$	1.24	0.21
Density	1.9	$<10^{-16}$	$-8 \cdot 10^{-3}$	0.5	-0.33	0.001	$8.7 \cdot 10^{-5}$	0.73	$-3 \cdot 10^{-2}$	$<10^{-7}$	$-2 \cdot 10^{-5}$	$<10^{-15}$	0.31	0.15
SA	-671.3	$<10^{-16}$	107.9	$<10^{-16}$	-236.9	0.1	0.15	0.68	114.4	$<10^{-16}$	0.023	$<10^{-15}$	395.0	0.39
AV	-0.12	$<10^{-15}$	$2 \cdot 10^{-2}$	$<10^{-8}$	$-5 \cdot 10^{-2}$	0.09	$2 \cdot 10^{-5}$	0.76	10^{-2}	$<10^{-10}$	$4 \cdot 10^{-6}$	$<10^{-11}$	0.06	0.26
VF	$-6 \cdot 10^{-2}$	$<10^{-14}$	10^{-2}	$<10^{-8}$	$-4 \cdot 10^{-2}$	0.02	$8 \cdot 10^{-5}$	0.04	$9 \cdot 10^{-3}$	$<10^{-16}$	$2 \cdot 10^{-6}$	$<10^{-9}$	0.04	0.35

Table S2. Parameters of linear regression models.

From the table, it can be seen that PER has a negative association with material porosity, whereas the rest of molecular porosity descriptors considered have a positive association. This responds to the intuition of low PER value to be associated with more encapsulated molecules (which are expected to form more porous materials). The rest of parameters are naturally expected to associate positively with material porosity. This results also demonstrate how linear models are easier to interpret, although their prediction capacity is generally lower than that provided by random forest. Even when restricting to internal validation in linear models, out-of-bag predictions from RF outperforms the simpler approach.

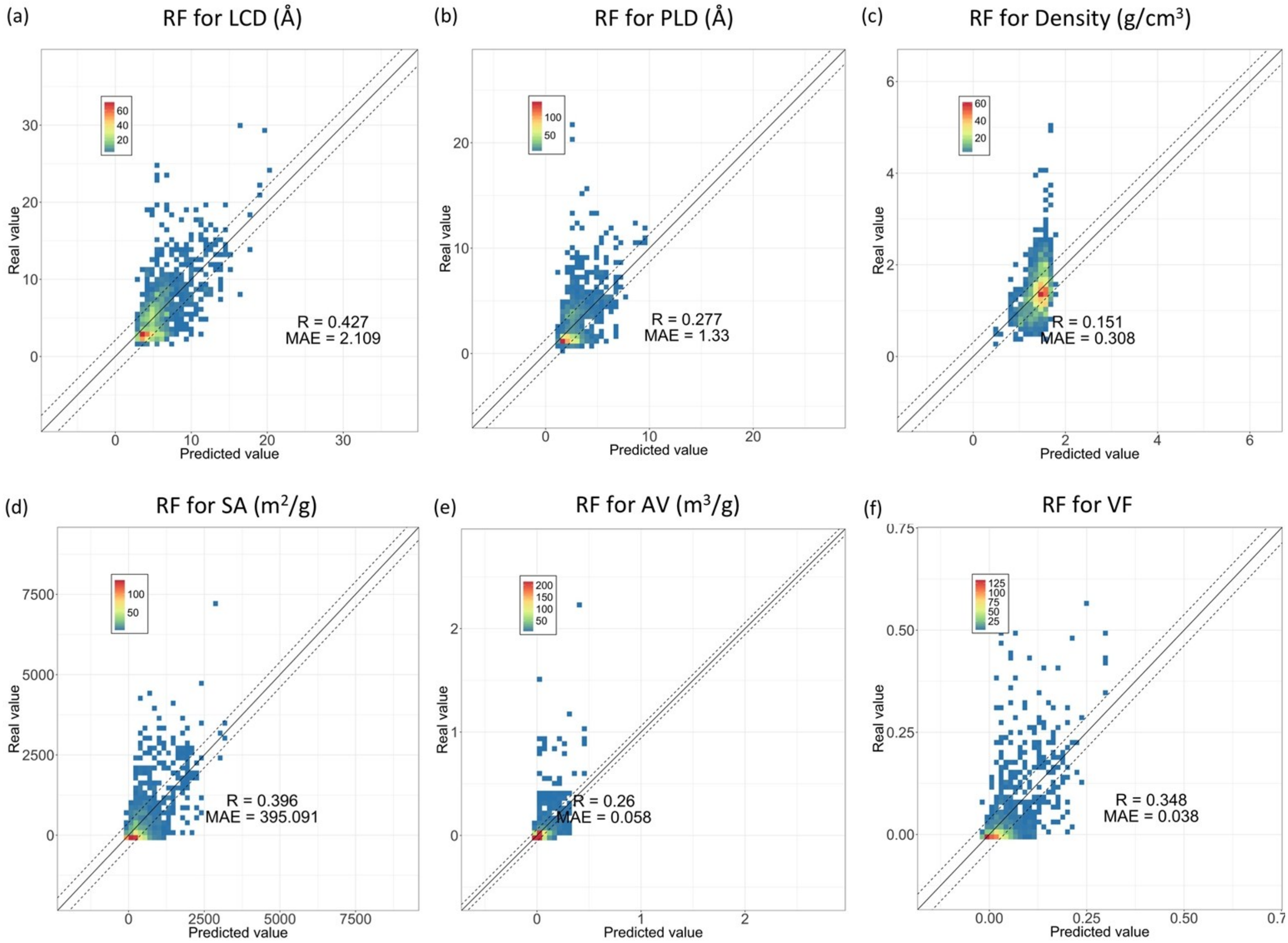

Figure S6. Predicted versus real values (internal validation) for linear regression models for the six material porosity measures predicted: LCD (a), PLD (b), density (c), SA (d), AV (e) and VF (f). Colour coding reflects density (i.e. number of structures placed in that region of the diagram). The models were trained over the 17832 structures that form our dataset. The counts per colour are presented in the colour bar within each graph. Diagonal lines represent the line of perfect prediction (continuous) and margins at one mean absolute error distance (dashed). In the figure, the unadjusted R^2 value is shown. The adjusted R^2 for the models can be found in Table S2.

8. Logistic Regression Models

As a first attempt to train classification models to identify the structures with largest material porosity parameters, we used logistic regression models. This kind of models have the advantage of being simpler than random forest, with the possibility to summarize them in terms of a simple equation. Thus, with a good predictive logistic regression model we would be able to provide simple guidelines for chemists in order to evaluate the potential of a given molecule to build up into a porous molecular material with outstanding properties. We trained a set of logistic regression models to identify structures within the best 10% for each of the six material properties analysed during this study. Logistic regression models provide with a way to compute the probability of a material (formed by a certain molecule) being over the threshold for the selected material property. To do so, the equation below should be applied: first, the logistic score can be calculated as follows:

$$W_{mol} = e^{\beta_0 + \beta_1 \cdot MWS + \beta_2 \cdot PER + \beta_3 \cdot ISA + \beta_4 \cdot NW + \beta_5 \cdot ESA} \qquad (2)$$

Then, the probability of the material parameter being over the given threshold is:

$$p(y_{mat} > threshold) = \frac{1}{1 + W_{mol}} \qquad (3)$$

The parameters of the equation correspond (for each material porosity property) with the odds ratios of the model, presented in Fig. S7. These descriptors can be computed with help of Molipor.

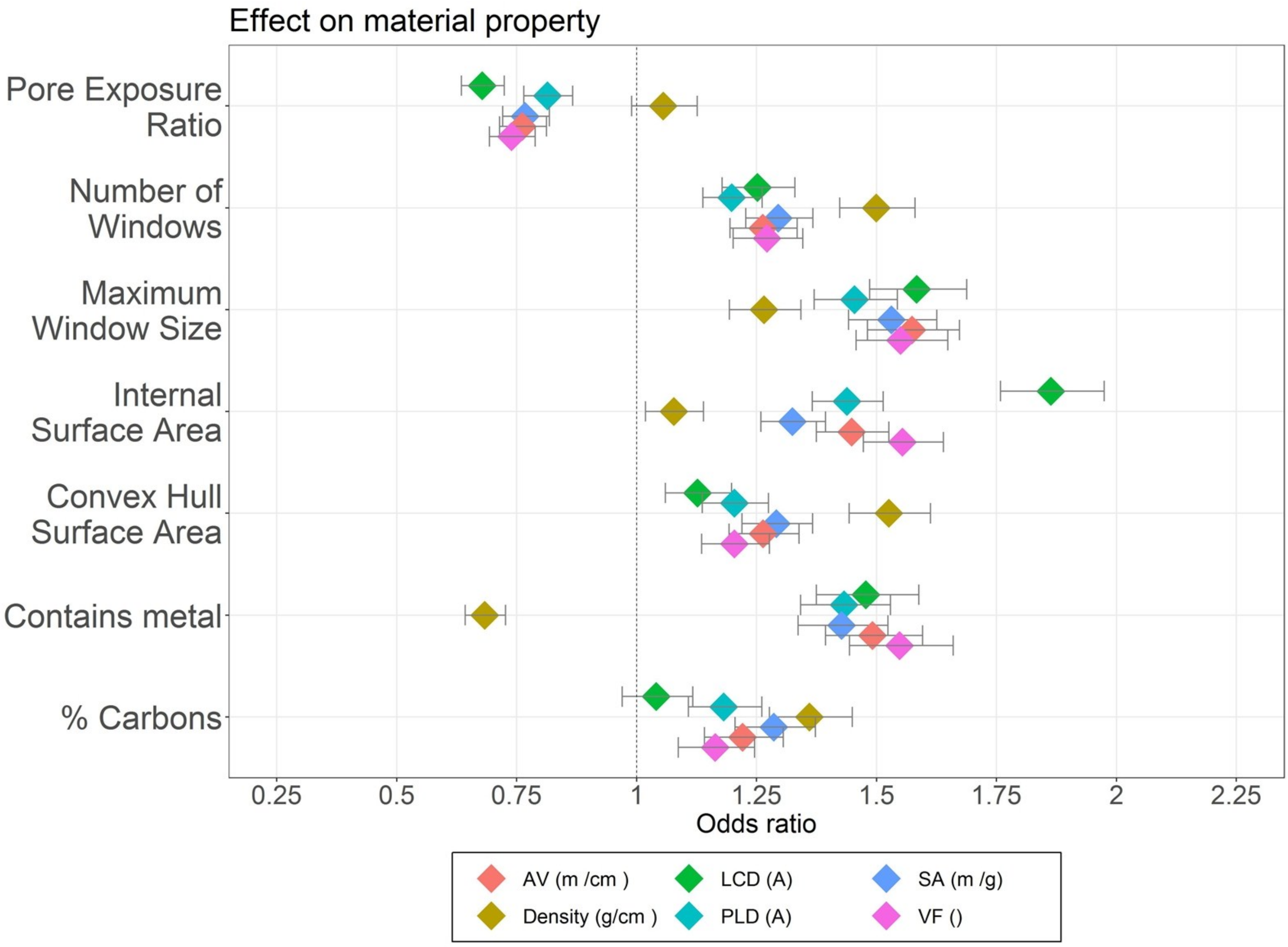

Figure S7. Odds ratios for the logistic regression models. All the predictors are scaled by their own standard deviation to make them comparable. In total, six models are represented in the plot, with the odd-ratios of each model marked with a different colour. Error bars (in grey) indicate that the parameter is statistically significant if they do not cross the dashed vertical line in 1.

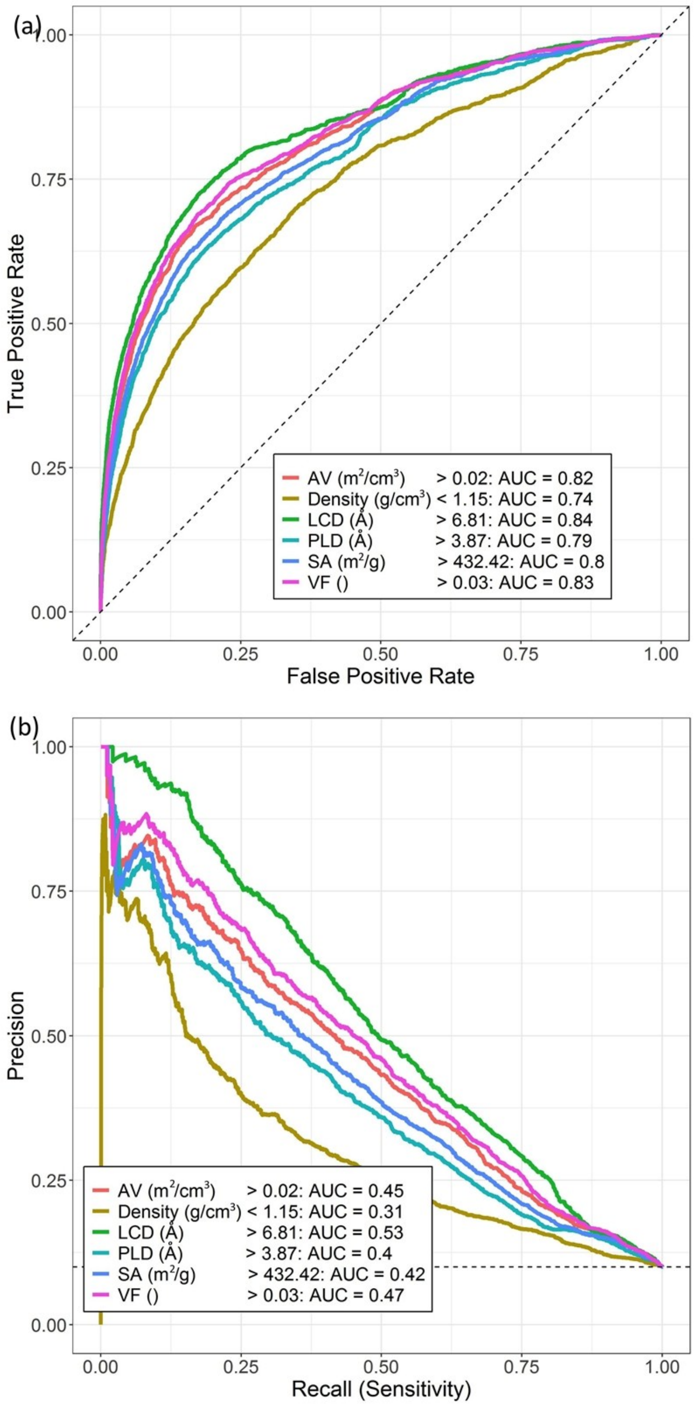

Figure S8. Performance of the logistic regression models trained for the entire dataset. (a) ROC curves, along with the AUC values for the six models trained (one per material property). (b) PR curves, along with the AUC values for the six models trained (one per material property).

The performance (internal validation) of the six logistic regression models is analysed with help of ROC and PR curves, just as random forest model's performance was assessed. The results can be found in Figure S8. From the figure, it can be seen that logistic regression models are an improvement with respect to a random classifier, but that of random forest classification models generally outclasses their performance.

9. Error prediction models

During the process of exploration, several iterations of regression models were considered. We found a satisfying performance for the regression models trained for a subset of structures formed by molecules considered to be more regular (NW > 2, NE/NW = 1), and without solvent. These models were not yet perfectly accurate, although their prediction capacity improved significantly with respect to the baseline models (linear regression for the whole population). During this work, we investigated how molecular structure can help to predict the model to make a big error during prediction. This was achieved by training logistic regression models, where two classes were considered: a positive class formed by the structures with absolute error bigger than the mean, and a negative class formed by the structures having less (or equal) absolute error than the mean. These models were then assessed with help of the ROC AUC (the skew in these populations was generally higher than 0.3). In Table S3, the ROC-AUCs for each model, along with the list of significant parameters (i.e., molecular properties that predict high error in the regression model) can be found. Generally speaking, the presence of metal seems to produce a higher error in the prediction of the RF regression model. Higher numbers in the descriptors also lead to less predictable scenarios (except for PER, that works in the opposite direction).

Material property	**Error ROC-AUC**	**Significant parameters (effect size)**
LCD	0.7	PER (0.15), ISA (1.0), ESA (1.0), HasMetal (1.7), LCD (1.2) PLD (1.3), NE/NW (1.2)
PLD	0.76	PER (0.04), ISA (1.0), NW (0.9), ESA (1.0), CP (2.6), HasMetal (2.2), LCD (1.2), PLD (1.5), NE/NW (1.1)
Density	0.61	MWS (1.3), PER (0.3), ESA (1.0), HasMetal (1.5), CP (0.7), LCD (1.1), PLD (0.8), NE/NW (1.2)
SA	0.86	MWS (0.9), PER (0.04), ESA (1.0), HasMetal (2.2), CP (7.9), LCD (1.3), PLD (1.7), NE/NW (1.1)
AV	0.86	MWS (0.8), PER (0.03), ISA (1.0), ESA (1.0), HasMetal (2.0), CP (11.9), LCD (1.6), PLD (1.7)
VF	0.86	MWS (0.8), PER (0.04), ISA (1.0), ESA (1.0), HasMetal (2.4), CP (3.4), LCD (1.7), PLD (1.7)

Table S4. Error prediction models. Logistic regression models to predict which structures will have an error bigger than the mean absolute error for the six regression models presented in the main text. ROC-AUCs are shown as performance assessment for the error models. All significant parameters, along with their effect size (scaled by the standard error for each variable) are shown. The effect size is shown between brackets. Values higher than 1 favour errors, whereas values lower than 1 favour the model to be right. Values of 1 represent small effects (lost due to rounding).

10. Prediction of material property over PubChem

To demonstrate the utility of the models presented in this work, we performed material property prediction over the PubChem subset of porous molecules, first introduced in a previous work from our group[3]. This set consists of 6020 porous molecules mined from PubChem. For this analysis, we use the classification model for structures without solvent, obtaining the probability of molecules forming a porous crystal with material porosity property within the best 10% according to the values obtained from CSD. In Figure S9 we present a bar plot with the number of molecules expected to form such a crystal based on their molecular properties.

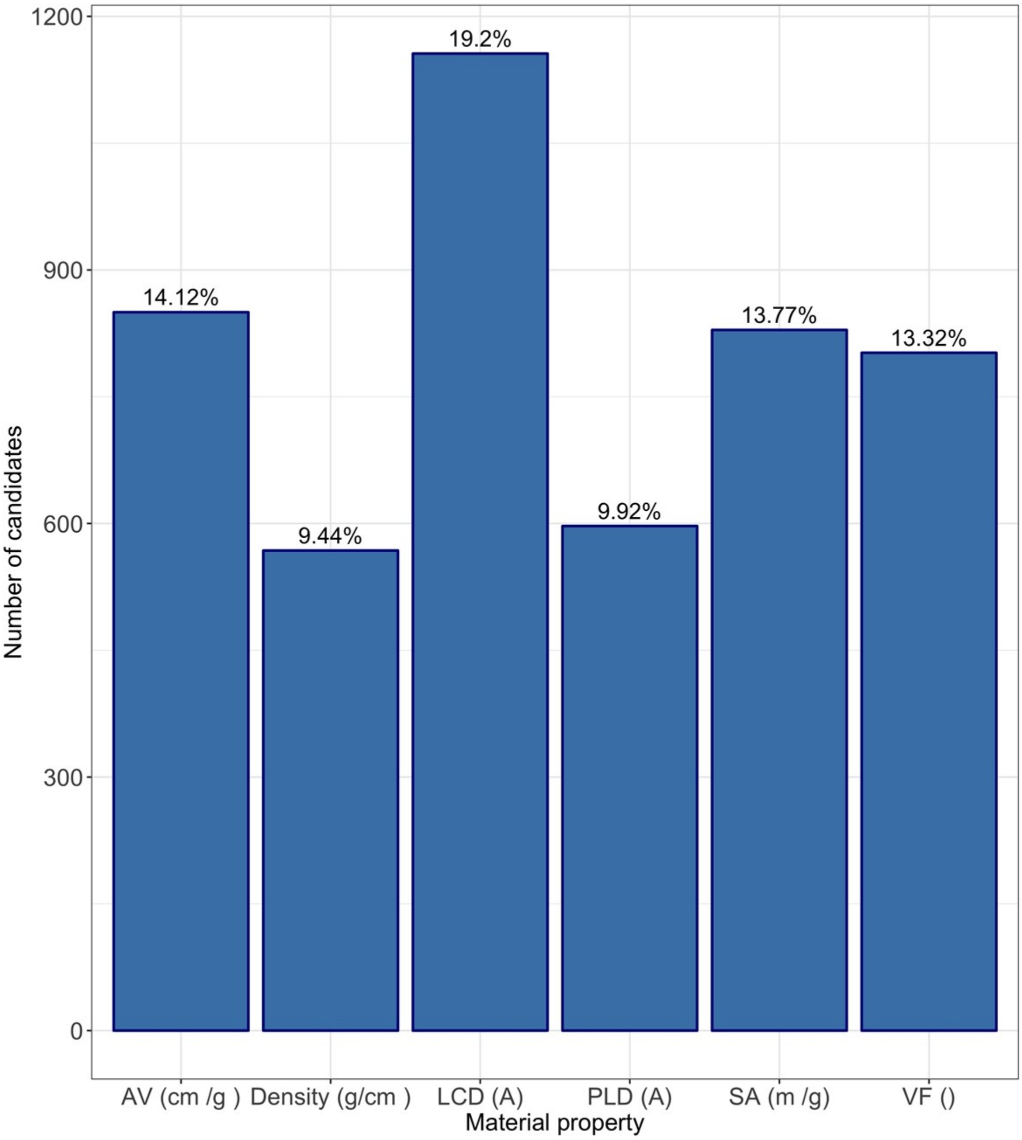

Figure S9. Bar plot representing the number of PubChem molecules expected to form a porous crystal with material porosity higher than the best 10% threshold for CSD

(percentage with respect to the total 6020 porous molecules found in PubChem in previous screenings).

It can be seen, from Fig S9, that a high number of molecules are expected to form porous crystals with interesting properties, according to the models. These numbers, however, have to be considered cautiously, as some of the molecules considered in this database are only predicted computationally, and also this values are predictions based on the expectation that solvent will be successfully removed, which is not always possible.

Bibliography

1. Jesse Davis & Mark Goadrich. The Relationship Between Precision-Recall and ROC Curves. *Planning* **73,** 55 (2007).
2. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10,** 1–21 (2015).
3. Gómez García, I., Bernabei, M. & Haranczyk, M. Toward Automated Tools for Characterization of Molecular Porosity. *J. Chem. Theory Comput.* **15,** 787–798 (2019).