The Mendeleev-Meyer Force Project

Sergio Santos[1,8†], Chia-Yun Lai[1†], Carlo A. Amadei[2], Karim R. Gadelrab[3], Tzu-Chieh Tang[4], Albert Verdaguer[5], Victor Barcons[6], Josep Font[6], Jaime Colchero[7], Matteo Chiesa[1]

[1]Laboratory for Energy and NanoScience (LENS), Institute Center for Future Energy (iFES), Masdar Institute of Science and Technology, Abu Dhabi, UAE

[2]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA

[3]Department of Materials Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA

[4] Department of Biological Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA

[5]ICN2 – Institut Catala de Nanociencia i Nanotecnologia, CSIC – Consejo Superior de

Investigaciones Cientificas, ICN2 Building, Campus UAB, 08193 Bellaterra, Barcelona,

Spain

[6]Departament de Disseny i Programació de Sistemes Electrònics, UPC - Universitat Politècnica de Catalunya, Av. Bases, 61, 08242 Manresa (Barcelona), Spain

[7]Instituto de Óptica y Nanociencia (CIOyN), Campus Espinardo, Universidad de Murcia, E-30100 Murcia, Spain

[8]Future Synthesis, Adderbury Banbury, Oxfordshire, OX17 3EP, UK

Corresponding email: mchiesa@masdar.ac.ae

† Equally contributing authors

Abstract

Here we present the Mendeleev-Meyer Force Project which aims at tabulating all materials and substances in a fashion similar to the periodic table. The goal is to group and tabulate substances using nanoscale force footprints rather than atomic number or electronic configuration as in the periodic table. The process is divided into: 1) acquiring nanoscale force data from materials, 2) parameterizing the raw data into standardized input features to generate a library, 3) feeding the standardized library into an algorithm to generate, enhance or exploit a model to identify a material or property. We propose producing databases mimicking the Materials Genome Initiative, the Medical Literature Analysis and Retrieval System Online (MEDLARS) or the PRoteomics IDEntifications database (PRIDE) and making these searchable online via search engines mimicking Pubmed or the PRIDE web interface. A prototype exploiting deep learning algorithms, i.e. multilayer neural networks, is presented.

The starting point in this work for the standardization or tabulation of materials via nanoscale forces consists of force versus distance curves (FDCs) as acquired in force spectroscopy. These curves contain the nanoscale footprint of the substance or material and are typically acquired with an atomic force microscope (AFM) or with a surface force apparatus (SFA). The force between a nanostructure, i.e. the tip of an AFM, and a surface is monitored as a function of separation or distance. In principle, the sensitivity of the AFM should provide information from all the relevant nanoscale force footprints or force contributions exerted between materials in the form of FDCs, as concluded when the AFM was first introduced as a method[1].

On the other hand, the quest to identify and recognize atoms or materials from atomic footprints or FDC data has remained an active field of research up to this date[2, 3], remains challenging and proves elusive particularly when considering the generalization and standardization of measurements and procedures[4, 5]. Furthermore, experiments are typically sophisticated[6] and are reported by carrying out extensive analysis from complex models or fundamental theory[2] rather than via automated processes. Approximately a decade ago a significant advance was reported by invoking a particular form of normalization of the raw FDC data[4] and single atoms were identified via specific atomic footprints. Similar normalization of FDC data was more recently employed to identify more complex heterogeneous systems[2]. Even more recently the influence of the AFM tip and chemical bonding configuration between materials has been reported to influence the force footprint curve to the point of preventing atom identification[7]. Other forms of sample recognition and identification consist of modelling the FDC and parameterizing it with physically relevant parameters such as stiffness[8], adhesion, viscoelasticity[9] or other parametric models[10], and even model free parameters[11-13]. Parameterization typically involves an

intermediate step after acquiring the raw data which consists of quantification and comparison. In this way differences detected in the interaction are exploited as parameter contrast maps that might be employed to discriminate between materials[11]. Standardization and tabulation however is still lagging far behind[14] other fields of research such as proteomics, metabolomics and genomics that are heavily assisted by computer science, large databases, powerful search engines and submission protocols[15] that allow rapid access to the databases. On the other hand, Kalinin, et al. have been early proponents of the exploitation of computer science techniques in probe microscopy[13, 16].

Here we propose a full transition in the field towards an integration of force spectroscopy and advanced computer science techniques as done in bioinformatics assisted biology. The objective is to parameterize force curves, or more generally nanoscale footprints arising form force microscopy, by turning the FDC raw data into features with the abstract meaning typically given to the features employed in machine learning algorithms. In this way, we do not impose any restriction to the number of input features to identify a given material or family of materials or substances. We further employ the term substance purposefully to emphasize that we would like to deal with chemistry, mechanics or even phases, i.e. whether a material's surface is hydrated. Features are then employed to construct feature libraries for groups of families or specific families. Finally an algorithm is exploited to generate a model from a given feature library that, as in the periodic table, groups materials according to similarity. We further employ the concept of classification from standard machine learning where the output of the algorithm is zero when a non-match is predicted and one when the algorithm predicts a match. In the prototype that we report here a deep learning algorithm, *i.e.* a multilayer neural network, is trained with the

backpropagation method (see supplementary) in Matlab[17]. We use F-score as a figure of merit to quantify Precision and Recall for the models or classifiers. Precision and Recall are defined as in machine learning where Precision is the ratio between true positives and predicted positives and Recall is the ratio between true positives and actual positives. The F-score parameter combines Precision and Recall as

$$F\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

The advantage of employing the F-score rather than Precision or Recall alone is that high values in F-score will be obtained if and only if both Recall and Precision are high simultaneously. In a more intuitive note, Precision could be defined as specificity and Recall as sensitivity implying that high F-score values include both high specificity and sensitivity. More detail on these figures of merit is given when discussing a practical example below. We then show that raw contrast images can be processed with the learned models to turn them into images that predict the likelihood that a given material has been identified in every pixel of the image.

Raw data acquisition

The initial step for parametrization and tabulation involves acquiring the nanoscale force footprint in the form of an FDC. An example of a raw experimental force curve is shown in Fig. 1a(i) where the tip-sample force F is shown in the y-axis and the distance d in the x-axis. This force arises from the atomic interactions between the atoms on the tip and the atoms on the sample. We note that both net attractive, i.e. F<0 nN, and net repulsive, i.e. F> 0 nN, forces are shown in the figure. It is also typical to associate the point of minima in force with mechanical

contact between the AFM tip and the surface[18]. Only points in the force curves satisfying F<0 nN are considered next since we find that these provide enough information to classify materials. Our method of parameterization is also suited for such range as detailed below.

Parametrizing raw data and transformation into input features

The second step consists of parametrizing the raw FDC. Here we choose to measure the distances in the well of the FDC in a similar fashion to that recently proposed elsewhere by the authors[19, 20]. The steps are as follows:

1) The adhesion force F_{AD} (or minima in force, see Fig. 1a(ii)) is taken as the force reference for a given curve.

2) This reference allows considering all other force-distance pairs with the use of a factor β as $F=\beta F_{AD}$. We note that by varying β from 0 to 1 any arbitrary force curve can be fully parameterized and quantified[19] for F<0 nN.

3) Without loss of generality we limit β to 0.85, 0.75… 0.05 and normalize the distances in the well of the curves with the reference $\beta=0.85$.

4) This is done by first computing the absolute distances $dF_{\beta}=dF_{0.85}, …, dF_{0.05}$ where $\beta=0.85, …, 0.05$ as illustrated in Fig. 1a(ii). This produces 9 distances as input features for each single curve.

5) We next normalize the distances dF_{β} by computing the ratios $dF_i=dF_{\beta}/dF_{0.85}$ where i=1 to 8 resulting in 8 normalized distances as illustrated in Fig. 1a(iii).

6) The distances dF_1 to dF_8 can be now employed as a table of input features for a machine learning algorithm to generate a model. In order to remove noise we averaged the

distances for a given substance or family of substances over 40-100 samples. An example of tabulation of input features to generate a feature library is shown in Table I. In Table I two polymers, polyethylene high-density (PEHD) and Polycaprolactone (PCL), and two materials from the silica family, i.e. glass and silicon, have been employed to generate three sets of input features. The three sets for each family form a feature library for polymers and silica respectively and concluding the second step of the procedure as shown as an illustration in Fig. 1a(iv).

Feeding the standardized input feature library into a learning algorithm

The third step consists of generating a model from a feature library as shown in the diagram in Fig. 1a(v). In the case of Table I, this model should be able to identify or detect whether input features belong to the polymer family or the silica family. In order to generate models we implemented a standard multilayer artificial neural network in Matlab[17] that included a regularization term λ to avoid overfitting. The steps are as follows:
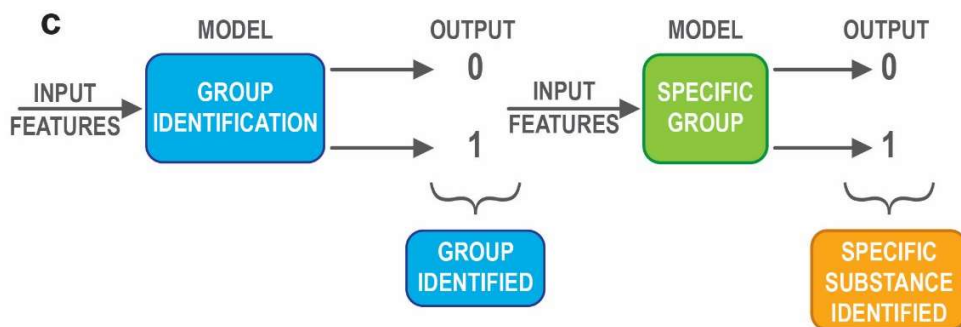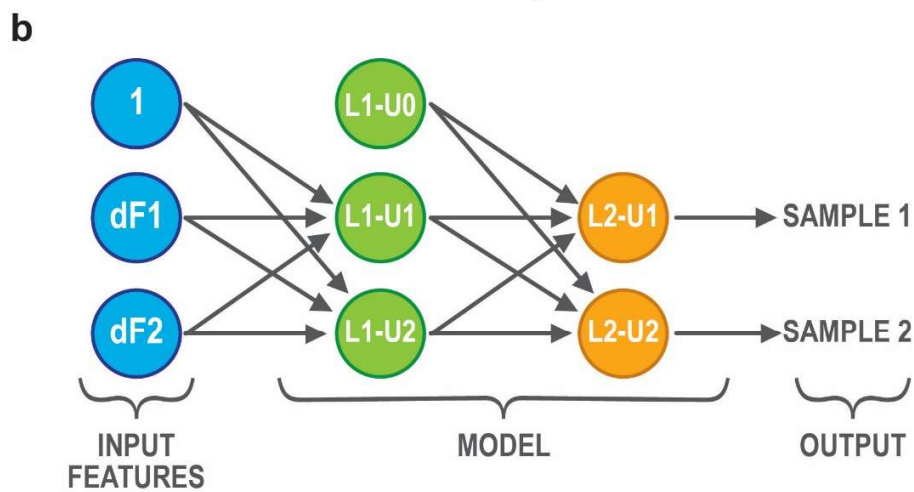
1) Inputting an input feature library, as that shown in Table I and as illustrated in Fig. 1a(iv), into a machine learning algorithm as illustrated in Fig. 1a(v). Here we have chosen an artificial neural network composed of U units per layer L. Units in these algorithms stand for unit cells or neurons and each unit is modelled with a sigmoid function where the inputs are processed by the function and the output fed into the units U in the next layer L as illustrated in Fig. 1b. The very last layer of the system will produce the predicted outputs as shown in the figure. For example, in the case of Table I, the two unit cells in

the last layer L will produce the predictive outcomes for the polymer (one unit) and silica (another unit) families.

2) The model is first trained with a set of input features from a given library where the output is known. For example, in the case of Table I, the last unit cell for polymers should produce ones if and only if data from polymers is fed into the system and similarly for the unit cell of the silica family.

3) Then the model is tested by inputting data into the model generated from the training data and comparing the output to the known values for the output. This is typical from supervised algorithms where the algorithm learns from inputting data for which the outcome is known by the user in advance. Errors in the outcomes are quantified via Recall and Precision and together via the F-score parameter as discussed above. In the experimental section we report errors on testing sets of data via Precision, Recall and F-score. The testing step is discussed in more detail below.

This concludes the procedure of training and testing a model from feature libraries for substance identification. An illustration of the full process is shown as a diagram in Fig. 1b.

a

(i) Raw input

(ii) Normalized input features

$$dF_1 = \frac{dF_{\beta 1}}{dF_{\beta 0}}$$

(iii)

$$dF_N = \frac{dF_{\beta N}}{dF_{\beta 0}}$$

(iv) Feature library

FEATURE	SET 0	...	SET N

(v) Learining algorithm

Output

Model generation

b

INPUT FEATURES

MODEL

OUTPUT

SAMPLE 1

SAMPLE 2

c

INPUT FEATURES → MODEL: GROUP IDENTIFICATION → OUTPUT 0 / 1

GROUP IDENTIFIED

INPUT FEATURES → MODEL: SPECIFIC GROUP → OUTPUT 0 / 1

SPECIFIC SUBSTANCE IDENTIFIED

9
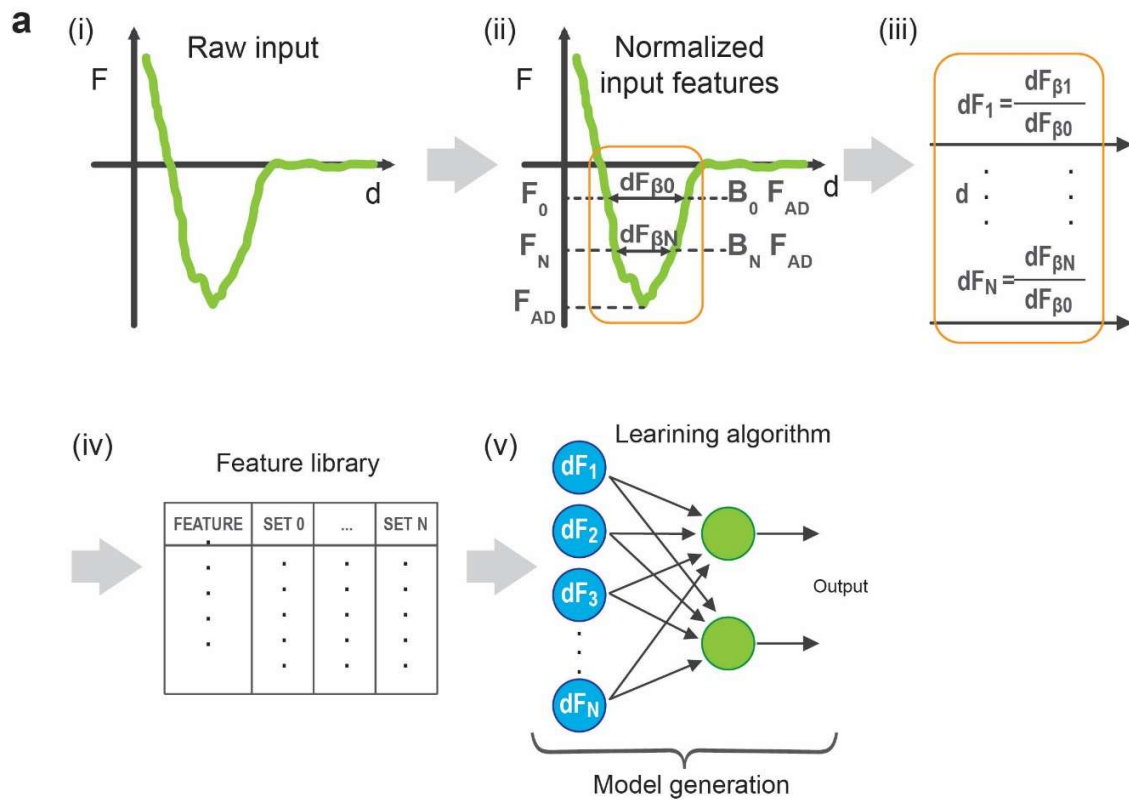
Figure 1| Illustrations of (**a**) i) raw data, ii) input features, iii) normalized input features, to iv) generation of a feature library and v) model generation from a feature library set. **b**, Scheme of the hierarchy and ordering of the number of layers L and number of unit cells U in the artificial neural networks. **c**, Process of identifying and classifying data first into families or groups and then into specific substances or samples.

β	Set 1 Pol	Set 2 Pol	Set 3 Pol	Set 1 Silica	Set 2 Silica	Set 3 Silica
0.75	0.95	0.95	0.94	0.89	0.89	0.9
0.65	0.88	0.88	0.88	0.8	0.78	0.81
0.55	0.8	0.8	0.81	0.71	0.69	0.71
0.45	0.7	0.72	0.73	0.61	0.6	0.61
0.35	0.59	0.62	0.6	0.52	0.5	0.5
0.25	0.45	0.48	0.48	0.4	0.38	0.37
0.15	0.3	0.31	0.31	0.27	0.24	0.26
0.05	0.1	0.1	0.13	0.11	0.09	0.1

	Precision		Recall		F1 Score	
2 L 2 U	0.8		0.66		0.72	
3 L 2 U	0.78		0.71		0.74	
4 L 4 U	0		0		0	

Table I. Example of libraries employed as the input data to generate models for two groups of materials: polymers family (Pol) and silica (Silica) family. The features employed to parameterize the raw data consist of the normalized dF values at the β points given in the first column. Each value for each set consists of averages of 40-100 data points or force curves each. At the bottom, figures of merit for 2L-2U, 3L-2U and 4L-4U models are shown. The models were produced with the input data shown in this table and then tested on separate data from the silica and polymer families.

In order to test the performance of the models, raw data obtained by different users and with different cantilever-sample systems are acquired and fed into trained neural networks (model) produced from the feature libraries. Here we illustrate this procedure with the help of the data from Table I employed to generate the models for the polymer and silica families as described

above. Our first testing sample was silicon. Approximately 1000 data points were collected in approximately 30 minutes and then fed into the trained model. The performance of the models was calculated by computing Precision, Recall and F-score. We defined true positives=TP, false positives=FP and false negatives=FN. Then Precison=TP/(TP+FP), Recall=TP/(TP+FN) and F-score=2(Precison×Recall)/ (Precison+Recall) as described above in Eq. (1). The results of the test are shown in Table I for a 2 L 2 U model, a 3 L 3 U model and a 4 L 4 U model. Values of F-score above 0.5 are indicative of significant predictive power of the model to successfully identify and discriminate between the silica and polymer families. Values of 0.5 or below imply that the models lack sufficient predictive power. We have added data for three models in Table I to exemplify that arbitrarily increasing L or U might not result in a better model. That is, relatively simple models, consisting of a few number of layers L and units U, might produce models with enough predicting power and might not be improved by increasing the complexity of the model arbitrarily. This is typical behavior in machine learning algorithms[21]. Finally, zero values in the figures of merit indicate overfitting, *i.e.* the output from the unit cells in the last layer of the model is always 0.5 independently of the input.

In Table II we show a training feature library produced to discriminate between the PEHD and PCL polymers. The data was employed to train and produce models and the models were then tested. In the table we show the results of these tests, in terms of figures of merit, after feeding data (approximately 200 data points) from a PEHD sample into the generated model. Again we see that a single layer and two unit cells suffice to produce a model with enough predicting power to discriminate between the two samples, i.e. F-score > 0.5. In summary, these two examples illustrate how by using multiple processes or steps, specific samples can be identified.

That is, a given model can be employed to first discriminate between families, as in the case of the silica and polymer families described above (Table I). Then, in a second step, a more specific model can be exploited to discriminate between samples in the same family (Table II). A scheme of such flow is illustrated in Fig. 1c.

β	Set 1 PCL	Set 2 PCL	Set 3 PCL	Set 1 PEHD	Set 2 PEHD	Set 3 PEHD
0.75	0.949	0.95	0.947	0.929	0.927	0.936
0.65	0.9	0.889	0.893	0.855	0.865	0.862
0.55	0.839	0.823	0.819	0.772	0.785	0.783
0.45	0.752	0.724	0.736	0.689	0.689	0.692
0.35	0.655	0.606	0.624	0.572	0.573	0.585
0.25	0.523	0.467	0.487	0.45	0.455	0.462
0.15	0.349	0.322	0.317	0.292	0.309	0.31
0.05	0.124	0.133	0.111	0.11	0.11	0.113
	Precision		Recall		F-Score	
1 L 2 U	0.71		0.59		0.64	
2 L 2 U	0.7		0.64		0.67	
3 L 4 U	0		0		0	

Table II. Example of libraries employed as input data to generate models for PCL and PEHD samples. The features consist of the normalized dF values at the β points as given in the first column. Each value for each set consists of averages of 40-100 data points or force curves each. At the bottom, figures of merit for 1L-2U, a 2L-2U and a 3L-4U models obtained when feeding a test sample (PEHD) into the trained model are given.

For completeness we also report a feature library in Table III that includes an otherwise disparate collection of samples; barium fluoride (BaF_2), calcium fluoride (CaF_2), silicon, PCL, graphite and glass. In this case we report the F-score in Fig. 2 as a function of regularization parameter λ. Three different models are also shown for 1L-2U, 3L-5U and 4L-5U. The F-score is 1 for the lowest values of λ for the 4L-5U models. Low values of λ imply more over-fitting and therefore

less capacity to generalize to data other than that of the training and testing sets. In this respect the largest values of λ giving high F-score values should be preferred. For the 1L-2U model the F-score is never 1 implying that the model is never 100% sensitive (Recall) and specific (Precision) independently of the selected λ value. Since F-score=1 in Fig. 2 for some λ values, this figure illustrates that it is possible to discriminate between substances or materials belonging to different groups or families directly, that is, without first employing a family discrimination model. This is provided the model is complex enough however, *i.e.* in this case we never obtained F-score=1 with models consisting of less than 3 layers.

β	Set BaF2	Set CaF2	Set Silicon	Set PCL	Set graphite	Set glass
0.75	0.79	0.97	0.90	0.94	0.94	0.87
0.65	0.61	0.92	0.82	0.88	0.87	0.76
0.55	0.46	0.88	0.72	0.81	0.80	0.67
0.45	0.34	0.82	0.65	0.71	0.70	0.57
0.35	0.25	0.75	0.53	0.60	0.56	0.46
0.25	0.16	0.65	0.39	0.45	0.42	0.34
0.15	0.11	0.44	0.22	0.29	0.24	0.22
0.05	0.05	0.17	0.07	0.11	0.08	0.09

Table III. Example of libraries employed as input data to generate models for a set of samples. The features employed to parameterize the raw data consist of the normalized dF values at the β points given in the first column. Each value for each set consists of averages of 40-100 data points each.
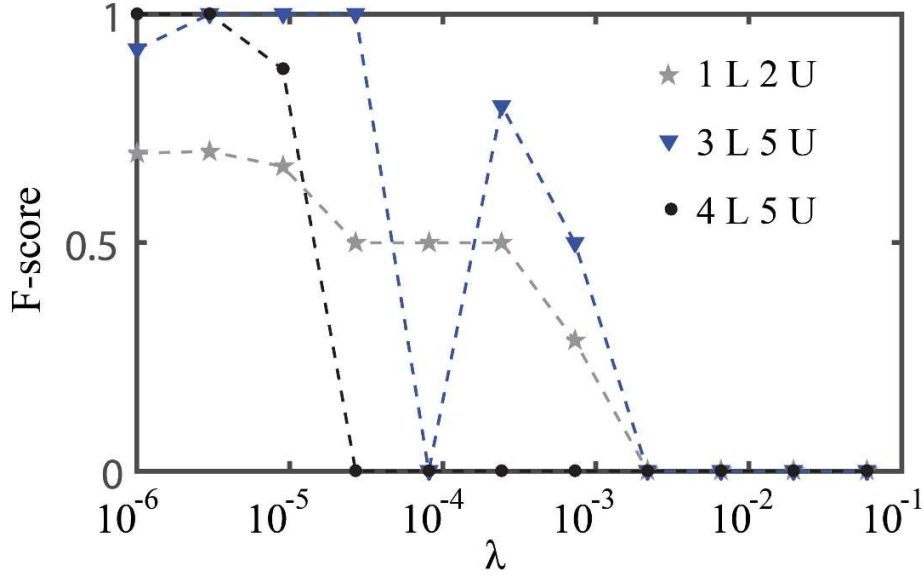
Figure 2| F-score reported as a function of regularization parameter λ and for three different models consisting of 1L-2U (stars), 3L-5U (triangles) and 4L-5U (circles).

Next we employ a feature library capable to differentiate between calcite and CaF_2. The surface of calcite can further be divided into two different substances or phases, namely an unhydrated and a hydrated phase. These phases can be routinely imaged with an AFM[22]. For the purpose of this example we term the two phases calcite P1 and calcite P2 and these are shown as a standard phase contrast image in Fig. 3a. The third substance of the feature library consists of CaF_2. We constructed a model with 4L and 6U that gave us an F-score=1 when tested. Then we acquired FDCs on a 80 nm^2 area of the calcite sample. The mapping was carried out with a 32 by 32 pixel resolution producing pixels with areas of $\approx$ 6nm^2 and 1024 force curves or FDC data points per image. Then we tested the model consisting of calcite P1, calcite P2 and CaF_2 against the 1024 data points collected. The results are shown in Fig. 3b. The black pixels imply that the model could not assign or identify (107 out of 1024 pixels) any of the three samples to that pixel. The blue pixels stand for positive identification of calcite P1 (402 out of 1024 pixels), the green

stands for positive identification of calcite P2 (459 out of 1024 pixels), and the red pixels stand for positive identification of CaF_2 (56 out of 1024 pixels). We observed thermally-induced drift when collecting the force data. This can be observed as a relative displacement of the surface in in Fig. 3b relative to 3a. The patch of calcite P2 in Fig. 3a has been circled in order to relate it to the displaced patch predicted by the model in Fig. 3b, also circled. A feature library for calcite P1 and calcite P2 was then employed to generate models to discriminate between these two substances only. The results are shown in Figs. 3c-e for a model of 1L-2U (3c) and 2L-3U (3d) where black pixels stand for ambiguity, blue for calcite P1 and green for calcite P2. The fact that Figs. 3c-d look almost identical implies that increasing the complexity of the model from 1L-2U to 2L-3U did not significantly improve predicting power – a figure including 3L-4U is showed in the Supplementary. The implication is that the 1L-2U model is superior since it is equally predictive and computationally much simpler.

The generality of the model produced to discriminate between calcite P1 and calcite P2 above was tested with a second set of calcite data collected using a different AFM tip and a different calcite sample. The results are shown in Figs. 4a-b. In this case the thermal drift was much smaller. A standard phase contrast image is shown in Fig. 4a and the prediction of the model against force data for the same spot in Fig. 4b. The central pattern in Fig. 4a is relatively well reconstructed by the model even though the model was generated with data from a different cantilever and calcite sample.
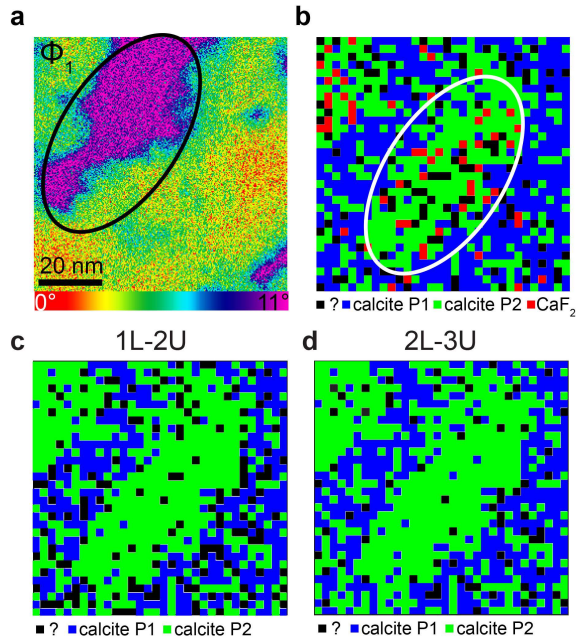
Figure 3| **a**, Two phases of calcite P2 (pink-purple) and calcite P1 (rest of the image) acquired as a standard phase image in dynamic AFM. **b**, Prediction of the model produced from a feature library consisting of calcite P1 (blue), calcite P2 (green) and CaF_2 (red). The black pixels refer to pixels where the model could not predict any output unambiguously. The two images where generated in approximately the same spot (80 nm^2) but some thermal-drift is observed. **c-d**, Predictions of the models produced from a feature library consisting of calcite P1 (blue), and calcite P2 (green) only for the same raw data as b. The black pixels refer to pixels where the model could not guess any output unambiguously. The results are shown for models consisting of (**c**) 1L-2U and (**d**) 2L-3U.
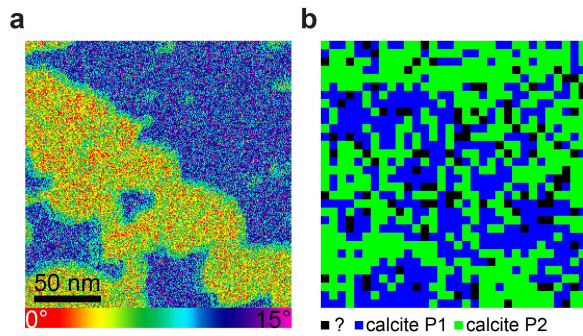

Figure 4| **a**, Two phases of calcite acquired with different tip and sample phase image in dynamic AFM.

b, Guess of the model produced from another data set. The blue pixels refer to calcite P1, green pixels

refer to calcite P2 and black pixels refer to pixels where the model could not predict any output unambiguously.

Conclusion

While we acknowledge that there are obstacles to overcome for The Mendeleev-Meyer Force project, we conclude that the use of libraries and models generated from libraries can be exploited to automatically identify substances from force data alone. The final goal is to generate standardized open source databases mimicking either the recent Materials Genome Initiative, the Medical Literature Analysis and Retrieval System Online (MEDLARS) or the PRoteomics IDEntifications database (PRIDE)[14]. These databases should be searchable via dedicated search engines mimicking, for example, the Pubmed or the PRIDE web interface[23]. The generation of models should not be restricted to artificial neural networks either, but could be enhanced, or even replaced, by other methodologies. We suggest to implement well known methodologies in machine learning such as support vector machines or Bayesian networks and to exploit them in parallel to improve predictive power. These methods are standard in the machine learning field and packages can be found in Matlab, python and the R languages. Finally, we would like to point out that the ordering or fabrication of libraries does not need to be limited to the air environment, but can be expanded to liquid and vacuum environments and to the use of probes other than silicon. Furthermore the classification into families does not have to be restricted to material properties but can be enhanced to, for example, identifying the presence or absence of atomic irregularities or dislocations or identifying biological patterns or behavior of systems for which distinct features might be produced[13, 24]. The number and type of input features might be further increased into linear or non-linear combinations of the features presented here or any

other input feature, such as temperature, relative humidity, tip radius, geometry or chemistry, that might enhance identification or recognition. Arguably the intuition of researchers working in a particular field will suggest the number and type of features that will make a given feature library preferable to produce a given model. In summary, the application of models and the massive testing of data should ultimately tell us what the limit of the proposed method of standardization is in the field of force spectroscopy. In this sense, we anticipate that the tip radius, the tip geometry and the relative humidity might have to be treated as input features to improve predictability and this process might be very challenging.

Acknowledgement

References

1. G. Binnig, C. F. Quate and C. Gerber, *Phys. Rev. Lett.*, 1986, 56, 930-933.
2. M. Setvín, P. Mutombo, M. Ondráček, Z. Majzik, M. Švec, V. Cháb, I. Ošt'ádal, P. Sobotík and P. Jelínek, *ACS nano*, 2012, 6, 6969-6976.
3. B. Schuler, G. Meyer, D. Peña, O. C. Mullins and L. Gross, *J. Am. Chem. Soc.*, 2015, 137, 9870-9876.
4. Y. Sugimoto, P. Pou, M. Abe, P. Jelinek, R. Perez, S. Morita and O. Custance, *Nature*, 2007, 446, 64-67.
5. S. V. Kalinin, B. G. Sumpter and R. K. Archibald, *Nature materials*, 2015, 14, 973-980.
6. J. Zhang, P. Chen, B. Yuan, W. Ji, Z. Cheng and X. Qiu, *Science*, 2013, 342, 611-614.
7. J. Welker, A. J. Weymouth and F. J. Giessibl, *ACS nano*, 2013, 7, 7377-7382.

8.   E. T. Herruzo, A. P. Perrino and R. Garcia, *Nat Commun*, 2014, 5, 3126.
9.   D. B. Haviland, C. A. van Eysden, D. Forchheimer, D. Platz, H. G. Kassa and P. Leclere, *Soft Matter*, 2016, 12, 619-624.
10.   C. Soussen, D. Brie, F. Gaboriaud and C. Kessler, 2008.
11.   D. Forchheimer, R. Forchheimer and D. B. Haviland, *Nat Commun*, 2015, 6.
12.   M. P. Nikiforov, V. V. Reukov, G. L. Thompson, A. A. Vertegel, S. Guo, S. V. Kalinin and S. Jesse, *Nanotechnology*, 2009, 20, 405708.
13.   A. Kumar, O. Ovchinnikov, S. Guo, F. Griggio, S. Jesse, S. Trolier-McKinstry and S. V. Kalinin, *Physical Review B*, 2011, 84, 024203.
14.   A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Mater.*, 2013, 1, 011002.
15.   J. Griss, A. R. Jones, T. Sachsenberg, M. Walzer, L. Gatto, J. Hartler, G. G. Thallinger, R. M. Salek, C. Steinbeck, N. Neuhauser, J. Cox, S. Neumann, J. Fan, F. Reisinger, Q.-W. Xu, N. del Toro, Y. Pérez-Riverol, F. Ghali, N. Bandeira, I. Xenarios, O. Kohlbacher, J. A. Vizcaíno and H. Hermjakob, *Mol. Cell. Proteomics*, 2014, 13, 2765-2775.
16.   A. Belianinov, R. Vasudevan, E. Strelcov, C. Steed, S. M. Yang, A. Tselev, S. Jesse, M. Biegalski, G. Shipman, C. Symons, A. Borisevich, R. Archibald and S. Kalinin, *Advanced Structural and Chemical Imaging*, 2015, 1, 1-25.
17.   R. a. Matlab and Simulink, The MathWorks, Inc. Natick The MathWorks, Inc. Natick 2010.
18.   R. Garcia and A. San Paulo, *Physical Review B*, 1999, 60, 4961.
19.   C.-Y. Lai, T. Olukan, S. Santos, A. Al Ghaferi and M. Chiesa, *Chem. Commun.*, 2015, 51, 17619-17622.
20.   C.-Y. Lai, T.-C. Tang, C. A. Amadei, A. J. Marsden, A. Verdaguer, N. Wilson and M. Chiesa, *Carbon*, 2014, 80, 784-792.
21.   A. Boschetti and L. Massaron, *Python data science essentials*, Packt Publishing Ltd, 2015.
22.   C.-Y. Lai, S. Perri, S. Santos, R. Garcia and M. Chiesa, *Nanoscale*, 2016, DOI: 10.1039/C6NR00496B.
23.   R. J. Roberts, *Proceedings of the National Academy of Sciences*, 2001, 98, 381-382.
24.   M. E. Dokukin, N. V. Guz, R. M. Gaikwad, C. D. Woodworth and I. Sokolov, *Phys. Rev. Lett.*, 2011, 107, 028101.