Composition and structure analyzer/featurizer for explainable machine-learning models to predict solid state structures

Emil I. Jaffal,[1,2,‡] Sangjoon Lee,[3,‡,*] Danila Shiryaev,[1] Alex Vtorov,[1] Nikhil Kumar Barua,[4] Holger Kleinke,[4] Anton O. Oliynyk[1,2,*]

[1]Department of Chemistry, Hunter College, City University of New York, New York 10065, NY, United States
[2]Ph.D. Program in Chemistry, The Graduate Center of the City University of New York, New York, NY 10016, USA
[3]Department of Applied Physics and Applied Mathematics, Columbia University, New York 10027, NY, United States
[4]Department of Chemistry, University of Waterloo, 200 University Ave W, Waterloo, ON, Canada
‡ – these authors contributed equally to this work
* – corresponding author

Supplementary Table S1. User experience evaluation for available featurizers used in solid state chemistry. The scores are in a 5-point scale (1-below average experience – 5-excellent experience). If applicable, vanilla version evaluation is in parentheses.

MAGPIE (CBFV)	Undergrad with no experience	Undergrad with 2-year experience	Post Bac	Grad student level
Installation procedure (library requirement)	5	5	5 (5)	4
Instructions/how easy to use	5	5	3 (2)	3
Documentation	5 (tutorial) 2 (GitHub)	3	3 (2)	3
Test samples/examples	5	5	5 (N/A)	5
Output (readability of files)	5	5	5 (N/A)	5
Correctness (how well-defined values are)	5	5	5 (N/A)	5
Run time	5	4	5 (N/A)	5
Overall experience	4.5	4	4 (2)	4

JARVIS	Undergrad no experience	Undergrad 2-year experience	Post Bac	Grad student level
Installation procedure (library requirement)	4	4	5	5
Instructions/how easy to use	3	4	4	5
Documentation	2	3	3	4
Test samples/examples	4	4	5	5
Output (readability of files)	4	5	5	5
Correctness (how well-defined values are)	5	5	5	5
Run time	5	4	5	5

Overall experience	4	4	4.5	5

JARVIS (CBFV)	Undergrad no experience	Undergrad 2-year experience	Post Bac	Grad student level
Installation procedure (library requirement)	5	5	5	N/A
Instructions/how easy to use	5	3	3	N/A
Documentation	2	1	2	N/A
Test samples/examples	4	5	4	N/A
Output (readability of files)	3	4	5	N/A
Correctness (how well-defined values are)	?	4	5	N/A
Run time	5	4	5	N/A
Overall experience	4	4	3	N/A

Supplementary Table S2. Detailed comments on user experience using featurizers.

Grad level student comments (solid prior experience)

JARVIS	Score	Comment
Installation procedure (library requirement)	5	The installation process was simple as Conda downloaded all the Python dependencies without any issues.
Instructions/how easy to use	5	The document provided a straightforward tutorial on creating elemental-composition features.
Documentation	4	As the JARVIS tool is a large project, ,the GitHub README.md has no particular emphasis on generating chemical composition-based features. However, the documentation provided a clear tutorial. https://pages.nist.gov/jarvis/tutorials/.
Test samples/examples	5	JARVIS provides samples that can be imported without the need to download the source code. This was good because other projects such as Atom2Vec, CBFV do not provide this feature.
Overall experience	5	JARVIS is actively developed and is one of the most well-maintained and documented repositories so far.

MAGPIE (CBFV)	Score	Comment
Installation procedure (library requirement)	4	The installation via pip was straightforward, but it was initially unclear how to switch the database from the default Oliynyk database to Magpie. I had to look into the source code to figure it out although it was just a quick change to the function parameter from X, y, formulae, skipped = composition.generate_features(df) to X, y, formulae, skipped = composition.generate_features(df, elem_prop='magpie')
Instructions/how easy to use	3	The barrier to entry could be lowered by including a code block in README.md that generates features from pre-existing files and save files. This would be more user-friendly than expecting users to locate and execute the example code within the source code. Additionally, the meaning of "target" value was confusing at first and could have been explained in the documentation.
Documentation	3	The documentation would benefit from added context about the databases. Philosophies behind each database could have been outlined rather than simply employing them. Furthermore, the documentation does not offer a straightforward method to switch the featurizer database as mentioned. The source code was searched to find example compounds.
Test samples/examples	5	There were numerous examples in the source code (with each folder containing around 3,000 files). But as mentioned, new users without programming background cannot use these tools since they are in the source code. A simple function that can import the example by simply copying/pasting one or two lines onto a Jupyter notebook would be ideal.
Output (readability of files)	5	It would have been nicer if they provided an automatic way to save the featurized files into a folder so that we can look at the result via Excel, etc.
Run time	5	It was fast. Within a few seconds to generate features.
Overall experience	4	It was nice to have a Python script that can generate compositional features from various databases.

MatMiner

- The featurization module of the project is no longer maintained and it fails to import modules. The details are provided here: https://matsci.org/t/two-issues-with-importing-elementproperty/53738

Atom2Vec

- The initial issue with the Atom2Vec project was a pip installation problem, which was resolved after contacting the project owner. However, the project still lacks enriched documentation or tutorials, which complicates usage. Moreover, it requires manual importation of structural files from an external source, namely pymatgen, for which API keys are necessary. Users must register via an email or Google account to obtain these keys. Regrettably, the project utilizes the legacy pymatgen API, further complicating its use. Despite suggestions from the project owner, the code did not perform as expected. Requests to provide a working script adapted to specific API keys were implicitly declined.

Post Bac level student comments (limited prior experience, with bachelor's degree but no further education)

JARVIS	Scores	Comment
Installation procedure (library requirement)	5	Installation was easy. The only problem was installing lightgbm, but a conda install solved it.
Instructions/how easy to use	4	Easy to follow through. Very in-depth tutorial
Documentation	3	Certain things are incorrect/need to be updated. i.e., jarvis.ai.descriptors.elemental.get_element_fraction_desc(formula='SiO2', max_nelements=103) giving error: module 'jarvis.ai.descriptors' has no attribute 'elemental'
Test samples/examples	5	Good and easy to follow through.
Overall experience	4.5	Very good featurizer. It has much more capability than just being a descriptor generator.

JARVIS CBFV: Only difference with CBFV is that you will have to refer to JARVIS documentation. Therefore, a lack of documentation and loss of other usage besides just generating descriptors.

Mat2vec	Scores	Comment
Installation procedure (library requirement)	2	Installation was difficult. Lots of outdated modules and can only install with python 3.8 due to DAWG not being maintained anymore. Even with all that fixed, it seemed that Unidecode, monty, gensim, pymatgen were unable to be imported from the mat2vec folder so I had to use pip install instead.
Instructions/how easy to use	2	Not detailed at all. Compared to elemnet, very limited tutorials as to what you can do.
Documentation	2	Lacking documentation, somewhat barebones.
Test samples/examples	2	Same as above.
Overall experience	1	Not maintained, old dependencies restricted testing. Ran into a myriad of problems with both installation and evaluation.

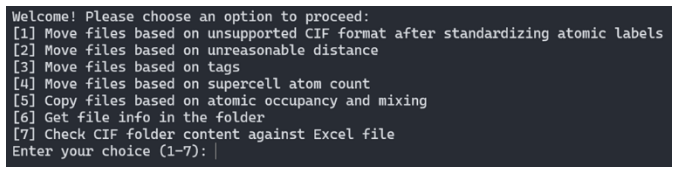
Mat2Vec CBFV: Installation went well as before but here we have no idea what the featurizers mean due to lack of proper naming. There is no original documentation to refer to either.

Elemnet	Score	Comment
Installation procedure (library requirement)	2	Ran into a lot of problems, Matminer isn't maintained, older versions of models not updating due to age, etc.
Instructions/how easy to use	2	Installation documentation is lacking on the repository, it gets confusing separating between the different folders.

Documentation	3	Installation documentation is lacking, but very detailed notebooks that show you how to do pretty much every step of the actual machine learning portion.
Test samples/examples	5	Lots of notebooks on their repository detail various processes such as data processing, creating holdouts, making and evaluating predictions.
Overall experience	2.5	If it was maintained, and pymatgen and Matminer was deprecated or at least updated, this would have run pretty well. Feature generators that were Matminer or pymatgen based were able to be generated after fixing scripts, but magpie being deprecated leads to us only generating 9 of the 145 features. Had to update composition.py, data.py, conversions.py from matminer... code not maintained in accordance with updated submodules. Fixed code below: *#from pymatgen.core.composition import Composition* *#from pymatgen.ext.matproj import MPRester*

Table S3. Additional features

(a) The list of descriptors also can be modified to tailor the output for a specific problem the user wants to solve. For example, we include the option of having the user choose whether to hot-encode (convert of categorical information into a format that may be fed into machine learning algorithms) their data, a representation of categorical variables as binary vectors. The presence or absence of an element will mark it with a 1 or 0, respectively. To maximize the useful data output, we prepared binary and ternary featurizers along with a universal featurizer which is agnostic to the number of elements in a compound. The last two options are for the final check with the list of compounds against the folder containing CIFs and to add features to the file from other files (*e.g.*, generated from other featurizers). To supplement CAF, we also developed a standalone CIF Cleaner code (https://github.com/bobleesj/cif-cleaner). Often, CIFs, even from reputable databases, require some editing, due to typos or missing entries which prevents them from being parsed. A simple solution would be to delete these files and proceed without them. However, scientific data is scarce and every single datapoint is a result of hard work done by experimentalists, therefore we developed a tool that takes care of formatting and ensures that simple cases are fixed



User prompt options in CIF cleaner code

The databases often do not specify (and tag) modulated structures or split position structures, reporting CIFs in reduced three dimensions or an averaged structure. This results in a case when interatomic distances are unreasonably short. We propose an option to weed out such cases. Next, crystallographic data could be reported at non-ambient conditions. To make it possible to have a consistent list of structure reports, including the interatomic distances in these structures, the conditions should be uniform. Unfortunately, the reported temperatures are not sufficient to identify the phase as the ambient condition phases, given that the reported temperatures are often synthesis temperatures or default readings from the diffractometers that give either normal (0°C) or standard (rt) temperature values. Well established databases provide the tags (ht, lt, rt – high temperature, low temperature, room temperature; hp – high pressure, *etc.*), which allow for efficient sorting.

(b) Some of the features could be copied directly as the entries from the database. Each database is different, but most information, like cell parameters, is present in any database. This could be used for manual inspection to study the phases or to select which columns will work for as features for use in ML. We list an example of what could be easily extracted from Pearson's Crystal Database (PCD) database without any additional parsing required. Some of the columns could be helpful for ML modeling but should be used with some caution. For instance, the space group number itself is not a very helpful feature, however, it bears some useful information, given that the space group numbers are ordered from the least symmetric to most. Unit cell volume is helpful when used with other properties, for example, number of electrons which results in electron density. Other crystallographic parameters, such as crystal system, crystal class, Laue class are originally not numerical values, but numerical labels could be given to represent the changing trend in these. For instance, crystal system labels could be monoclinic = 1 ... cubic = 7, this way representing the change in symmetry. Similarly, the presence or absence of something could be numerically represented with 1/0, like in the case of inversion center or polar axis. The database entry number should be used to match the data with the data from other featurizers, if the entry number was carried over.

List of helpful information that could be extracted directly from the database.

Column	Comment
Formula	Formula, often sorted alphabetically (*label for ML*)
Formula (phase)	Formula with a tag, phase sorting criterion
Entry prototype	Structure type (*class for ML*)
SGR symbol (std.)	Space group symbol
SGR no. (std.)	Space group number (*feature for ML*)
Wyckoff seq.	Wyckoff sequency, used to check for new structure type

a [nm]	Cell parameter (*should be converted to Å*)
b [nm]	Cell parameter (*should be converted to Å*)
c [nm]	Cell parameter (*should be converted to Å*)
alpha [°]	Cell parameter
beta [°]	Cell parameter
gamma [°]	Cell parameter
V (std.) [nm…]	Cell volume (*feature for ML, should be converted to Å^3*)
Density (calc.) [Mg m‰…]	Density (*feature for ML*)
Crystal system (std.)	Crystal system (label should be given, *feature for ML*)
Laue class	Laue class (label should be given, *feature for ML*)
Crystal class	Crystal class (label should be given, *feature for ML*)
Inv. center	Centrosymmetric or not (label should be given, *feature for ML*)
Polar axis	Polar or not (label should be given, *feature for ML*)
Temperature [K]	Rarely reported
Pressure [GPa]	Rarely reported
Elem. cnt.	Number of elements, used to determine binary *vs* ternary (*feature for ML, but only for universal features*)
Chem. system	System, alphabetically sorted
Compound class	Very general description of compound class
Structure class	Structure class, could be used to outline the scope of the study
Calc. struct.	Could be used as sorting criterion, if structure is hypothetical
Level struct. studies	Powder *vs* single crystal structure determination
Title	Fast screening for properties
Year	To check how reliable the result might be
First release date	To check new phases for update
Entry number	PCD number to match with other features

(c) More information is usually available in the databases, but this information is typically not readily available for the user to extract. For example, entry reports might contain useful data to extract more features or properties, *e.g.*, color for specific property prediction. The information would require additional parsing (multiple atomic sites) and conversion of the text to numbers for use as features. This might also be helpful as a fast way to get coordination number (CN), however, often CN reports contain mistakes and should be calculated by the user. We provide an example of what could be extracted with more mining.

Additional info report analysis.

Additional report info	Comment
Entry number	Used to match data with the data from other featurizers
Atomic environment info	Atomic environment info line that has Site, Coordination number, Environment type, Composition of the coordination sphere.
Site	Site label, might have some inconsistencies
Coordination Number	Coordination Number
Environment Type	Shape of the polyhedron
Composition	Composition of the polyhedron
Synthesis	Information about synthesis procedure with temperature and time.

	It is helpful to include if the reaction conditions need to be predicted.
Crystal growth info	Information about crystal growth (annealing, arc-melting, *etc*.) It is helpful to include if the reaction conditions need to be predicted.
Stability info	Information about the stability region (temperature, pressure) or the form (thin films, nanoparticles) It is helpful to include if the reaction conditions need to be predicted.
Color info	Indication if the compound has color. Could be used as a property for prediction.

(d) The list of properties used for calculating features could be enhanced with novel size or electronegativity scales defined by the user. For instance, size scale is sensitive to the class of materials and the presence of other elements. It is advised to calculate the user's own scale for effective modeling. For example, if one wants to define a new size scale to correspond to the size of all element CIF reports. Once the coordination geometry files are extracted as *.txt file, the first line is used as the shortest homoatomic distance, which is divided by 2 to get the CIF radius. We suggest creating the output with the mean value, standard deviation, and a histogram that could be created for visual inspection. Later this CIF radius scale we can use as a property for the feature definition, *e.g.*, radius ratio, on par with covalent radius, ionic radius, and others.

Supplementary Table S4. Binary structural features

	Feature	omment
1	Entry	Used to match with CAF data or any other data to merge in one single file
2	formula	Label
3	A	Element A matched with the group
4	B	Element B matched with the group
5	INT_distAA	Interatomic distances from CIF
6	INT_distBB	Interatomic distances from CIF
7	INT_distAB	Interatomic distances from CIF
8	INT_Asize	CIF radius for atom A
9	INT_Bsize	CIF radius for atom B
10	INT_Asize_by_Bsize	Radius ratio (CIF radius)
11	INT_distAA_by2_byAsize	Default size scale fit
12	INT_distBB_by2_byBsize	Default size scale fit
13	INT_distAB_by2_byAsizebyBsize	Default size scale
14	INT_Asize_ref	Refined radius
15	INT_Bsize_ref	Refined radius
16	INT_percent_diff_A_by_100	How much different from default scale
17	INT_percent_diff_B_by_100	How much different from default scale
18	INT_distAA_minus_ref_diff	Refined scale fitness (0 means the distance used for refinements)
19	INT_distBB_minus_ref_diff	Refined scale fitness (0 means the distance used for refinements)
20	INT_distAB_minus_ref_diff	Refined scale fitness (0 means the distance used for refinements)
21	INT_refined_packing_eff	Packing efficiency of the structure
22	INT_R_factor	R factor for the refinement (least square difference)
23	INT_UNI_shortest_homoatomic_dist	Shortest homoatomic distance
24	INT_UNI_shortest_heteroatomic_dist	Shortest heteroatomic distance
25	INT_UNI_shortest_homoatomic_dist_by_2_by_atom_size	Shortest homoatomic distance by 2 by atom size
26	INT_UNI_shortest_heteroatomic_dist_by_sum_of_atom_sizes	Shortest heteroatomic distance by sum of atom sizes
27	INT_UNI_shortest_homoatomic_dist_by_2_by_refined_atom_size	Shortest homoatomic distance by 2 by refined atom size
28	INT_UNI_shortest_heteroatomic_dist_by_sum_of_refined_atom_sizes	Shortest heteroatomic distance by sum of refined sizes
29	INT_UNI_highest_refined_percent_diff_abs	Highest refined percent difference by 100 (abs value)
30	INT_UNI_lowest_refined_percent_diff_abs	Lowest refined percent difference by 100 (abs value)
31	INT_UNI_packing_efficiency	Packing efficiency
32	WYC_A_lowest_wyckoff_label	Lowest Wyckoff number for element A
33	WYC_B_lowest_wyckoff_label	Lowest Wyckoff number for element B
34	WYC_identical_lowest_wyckoff_count	Number of sites with the lowest Wyckoff number
35	WYC_A_sites_total	Number of crystallographic sites for element A
36	WYC_B_sites_total	Number of crystallographic sites for element B

37	WYC_A_multiplicity_total	Sum of Wyckoff numbers for element A
38	WYC_B_multiplicity_total	Sum of Wyckoff numbers for element B
39	ENV_A_shortest_dist_count	Number of atoms that are at the shortest distance from atom A
40	ENV_B_shortest_dist_count	Number of atoms that are at the shortest distance from atom B
41	ENV_A_avg_shortest_dist_count	Average number of atoms that are at the shortest distance from atom A (case if multiple sites present)
42	ENV_B_avg_shortest_dist_count	Average number of atoms that are at the shortest distance from atom B (case if multiple sites present)
43	ENV_A_shortest_tol_dist_count	Number of atoms that are at the shortest distance from atom A (with some distance tolerance applied, default 5%)
44	ENV_B_shortest_tol_dist_count	Number of atoms that are at the shortest distance from atom B (with some distance tolerance applied, default 5%)
45	ENV_A_avg_shortest_dist_within_tol_count	Average number of atoms that are at the shortest distance from atom A (case if multiple sites present, with some distance tolerance applied, default 5%)
46	ENV_B_avg_shortest_dist_within_tol_count	Average number of atoms that are at the shortest distance from atom B (case if multiple sites present, with some distance tolerance applied, default 5%)
47	ENV_A_second_by_first_shortest_dist	2^{nd} shortest distance/1^{st} shorted distance for atom A, measures distortion of polyhedron
48	ENV_B_second_by_first_shortest_dist	2^{nd} shortest distance/1^{st} shorted distance for atom B, measures distortion of polyhedron
49	ENV_A_avg_second_by_first_shortest_dist	2^{nd} shortest distance/1^{st} shorted distance for atom A, measures distortion of polyhedron (case if multiple sites present)
50	ENV_B_avg_second_by_first_shortest_dist	2^{nd} shortest distance/1^{st} shorted distance for atom B, measures distortion of polyhedron (case if multiple sites present)
51	ENV_A_second_shortest_dist_count	2^{nd} shortest distance count for atom A
52	ENV_B_second_shortest_dist_count	2^{nd} shortest distance count for atom B
53	ENV_A_avg_second_shortest_dist_count	Average 2^{nd} shortest distance count for atom A
54	ENV_B_avg_second_shortest_dist_count	Average 2^{nd} shortest distance count for atom B
55	ENV_A_homoatomic_dist_by_shortest_dist	A-A distance / shortest distance
56	ENV_B_homoatomic_dist_by_shortest_dist	B-B distance / shortest distance
57	ENV_A_avg_homoatomic_dist_by_shortest_dist	Average A-A distance / shortest distance
58	ENV_B_avg_homoatomic_dist_by_shortest_dist	Average B-B distance / shortest distance
59	ENV_A_count_at_A_shortest_dist	Number of A atoms next to the A atoms at the shortest distance
60	ENV_A_count_at_B_shortest_dist	Number of A atoms next to the B atoms at the shortest distance
61	ENV_A_avg_count_at_A_shortest_dist	Average number of A atoms next to the A atoms at the shortest distance
62	ENV_A_avg_count_at_B_shortest_dist	Average number of A atoms next to the B atoms at the shortest distance
63	ENV_B_count_at_A_shortest_dist	Number of B atoms next to the A atoms at the shortest distance
64	ENV_B_count_at_B_shortest_dist	Number of B atoms next to the B atoms at the shortest distance
65	ENV_B_avg_count_at_A_shortest_dist	Average number of B atoms next to the A atoms at the shortest distance

66	ENV_B_avg_count_at_B_shortest_dist	Average number of B atoms next to the B atoms at the shortest distance
67	CN_AVG_coordination_number	Average coordination number
68	CN_AVG_A_atom_count	Average atom A number within CN
69	CN_AVG_B_atom_count	Average atom B number within CN
70	CN_AVG_polyhedron_volume	Average volume of polyhedra
71	CN_AVG_central_atom_to_center_of_mass_dist	Average distance from the central atom to the center of mass of polyhedron
72	CN_AVG_number_of_edges	Average number of edges of polyhedron
73	CN_AVG_number_of_faces	Average number of faces of polyhedron
74	CN_AVG_shortest_distance_to_face	Average shortest distance from central atom to center of face of polyhedron
75	CN_AVG_shortest_distance_to_edge	Average shortest distance from central atom to middle edge of polyhedron
76	CN_AVG_volume_of_inscribed_sphere	Average volume of inscribed sphere that could be fit in polyhedron
77	CN_AVG_packing_efficiency	Average packing efficiency of polyhedron
78	CN_MIN_coordination_number	Minimum coordination number
79	CN_MIN_A_atom_count	Minimum atom A number within CN
80	CN_MIN_B_atom_count	Minimum atom B number within CN
81	CN_MIN_polyhedron_volume	Minimum volume of polyhedra
82	CN_MIN_central_atom_to_center_of_mass_dist	Minimum distance from the central atom to the center of mass of polyhedron
83	CN_MIN_number_of_edges	Minimum number of edges of polyhedron
84	CN_MIN_number_of_faces	Minimum number of faces of polyhedron
85	CN_MIN_shortest_distance_to_face	Minimum shortest distance from central atom to center of face of polyhedron
86	CN_MIN_shortest_distance_to_edge	Minimum shortest distance from central atom to middle edge of polyhedron
87	CN_MIN_volume_of_inscribed_sphere	Minimum volume of inscribed sphere that could be fit in polyhedron
88	CN_MIN_packing_efficiency	Minimum packing efficiency of polyhedron
89	CN_MAX_coordination_number	Maximum coordination number
90	CN_MAX_A_atom_count	Maximum atom A number within CN
91	CN_MAX_B_atom_count	Maximum atom B number within CN
92	CN_MAX_polyhedron_volume	Maximum volume of polyhedra
93	CN_MAX_central_atom_to_center_of_mass_dist	Maximum distance from the central atom to the center of mass of polyhedron
94	CN_MAX_number_of_edges	Maximum number of edges of polyhedron
95	CN_MAX_number_of_faces	Maximum number of faces of polyhedron
96	CN_MAX_shortest_distance_to_face	Maximum shortest distance from central atom to center of face of polyhedron
97	CN_MAX_shortest_distance_to_edge	Maximum shortest distance from central atom to middle edge of polyhedron
98	CN_MAX_volume_of_inscribed_sphere	Maximum volume of inscribed sphere that could be fit in polyhedron
99	CN_MAX_packing_efficiency	Maximum packing efficiency of polyhedron

Supplementary Table S5. List of ternary Structural features

	Feature
1	Entry
2	formula
3	R
4	M
5	X
6	INT_distRR
7	INT_distMM
8	INT_distXX
9	INT_distRM
10	INT_distMX
11	INT_distRX
12	INT_Rsize
13	INT_Msize
14	INT_Xsize
15	INT_Rsize_by_Msize
16	INT_Msize_by_Xsize
17	INT_Rsize_by_Xsize
18	INT_distRR_by2_byRsize
19	INT_distMM_by2_byMsize
20	INT_distXX_by2_byXsize
21	INT_distRM_byRsizebyMsize
22	INT_distMX_byMsizebyXsize
23	INT_distRX_byRsizebyXsize
24	INT_Rsize_ref
25	INT_Msize_ref
26	INT_Xsize_ref
27	INT_percent_diff_R_by_100
28	INT_percent_diff_M_by_100
29	INT_percent_diff_X_by_100
30	INT_distRR_minus_ref_diff
31	INT_distMM_minus_ref_diff
32	INT_distXX_minus_ref_diff
33	INT_distRM_minus_ref_diff
34	INT_distMX_minus_ref_diff
35	INT_distRX_minus_ref_diff
36	INT_refined_packing_eff
37	INT_R_factor
38	INT_UNI_shortest_homoatomic_dist
39	INT_UNI_shortest_heteroatomic_dist
40	INT_UNI_shortest_homoatomic_dist_by_2_by_atom_size
41	INT_UNI_shortest_heteroatomic_dist_by_sum_of_atom_sizes
42	INT_UNI_shortest_homoatomic_dist_by_2_by_refined_atom_size
43	INT_UNI_shortest_heteroatomic_dist_by_sum_of_refined_atom_sizes

44	INT_UNI_highest_refined_percent_diff_abs
45	INT_UNI_lowest_refined_percent_diff_abs
46	INT_UNI_packing_efficiency
47	WYC_R_lowest_wyckoff_label
48	WYC_M_lowest_wyckoff_label
49	WYC_X_lowest_wyckoff_label
50	WYC_identical_lowest_wyckoff_count
51	WYC_R_sites_total
52	WYC_M_sites_total
53	WYC_X_sites_total
54	WYC_R_multiplicity_total
55	WYC_M_multiplicity_total
56	WYC_X_multiplicity_total
57	ENV_R_shortest_dist_count
58	ENV_M_shortest_dist_count
59	ENV_X_shortest_dist_count
60	ENV_R_avg_shortest_dist_count
61	ENV_M_avg_shortest_dist_count
62	ENV_X_avg_shortest_dist_count
63	ENV_R_shortest_tol_dist_count
64	ENV_M_shortest_tol_dist_count
65	ENV_X_shortest_tol_dist_count
66	ENV_R_avg_shortest_dist_within_tol_count
67	ENV_M_avg_shortest_dist_within_tol_count
68	ENV_X_avg_shortest_dist_within_tol_count
69	ENV_R_second_by_first_shortest_dist
70	ENV_M_second_by_first_shortest_dist
71	ENV_X_second_by_first_shortest_dist
72	ENV_R_avg_second_by_first_shortest_dist
73	ENV_M_avg_second_by_first_shortest_dist
74	ENV_X_avg_second_by_first_shortest_dist
75	ENV_R_second_shortest_dist_count
76	ENV_M_second_shortest_dist_count
77	ENV_X_second_shortest_dist_count
78	ENV_R_avg_second_shortest_dist_count
79	ENV_M_avg_second_shortest_dist_count
80	ENV_X_avg_second_shortest_dist_count
81	ENV_R_homoatomic_dist_by_shortest_dist
82	ENV_M_homoatomic_dist_by_shortest_dist
83	ENV_X_homoatomic_dist_by_shortest_dist
84	ENV_R_avg_homoatomic_dist_by_shortest_dist
85	ENV_M_avg_homoatomic_dist_by_shortest_dist
86	ENV_X_avg_homoatomic_dist_by_shortest_dist
87	ENV_R_count_at_R_shortest_dist
88	ENV_R_count_at_M_shortest_dist

89	ENV_R_count_at_X_shortest_dist
90	ENV_R_avg_count_at_R_shortest_dist
91	ENV_R_avg_count_at_M_shortest_dist
92	ENV_R_avg_count_at_X_shortest_dist
93	ENV_M_count_at_R_shortest_dist
94	ENV_M_count_at_M_shortest_dist
95	ENV_M_count_at_X_shortest_dist
96	ENV_M_avg_count_at_R_shortest_dist
97	ENV_M_avg_count_at_M_shortest_dist
98	ENV_M_avg_count_at_X_shortest_dist
99	ENV_X_count_at_R_shortest_dist
100	ENV_X_count_at_M_shortest_dist
101	ENV_X_count_at_X_shortest_dist
102	ENV_X_avg_count_at_R_shortest_dist
103	ENV_X_avg_count_at_M_shortest_dist
104	ENV_X_avg_count_at_X_shortest_dist
105	CN_AVG_coordination_number
106	CN_AVG_R_atom_count
107	CN_AVG_M_atom_count
108	CN_AVG_X_atom_count
109	CN_AVG_polyhedron_volume
110	CN_AVG_central_atom_to_center_of_mass_dist
111	CN_AVG_number_of_edges
112	CN_AVG_number_of_faces
113	CN_AVG_shortest_distance_to_face
114	CN_AVG_shortest_distance_to_edge
115	CN_AVG_volume_of_inscribed_sphere
116	CN_AVG_packing_efficiency
117	CN_MIN_coordination_number
118	CN_MIN_R_atom_count
119	CN_MIN_M_atom_count
120	CN_MIN_X_atom_count
121	CN_MIN_polyhedron_volume
122	CN_MIN_central_atom_to_center_of_mass_dist
123	CN_MIN_number_of_edges
124	CN_MIN_number_of_faces
125	CN_MIN_shortest_distance_to_face
126	CN_MIN_shortest_distance_to_edge
127	CN_MIN_volume_of_inscribed_sphere
128	CN_MIN_packing_efficiency
129	CN_MAX_coordination_number
130	CN_MAX_R_atom_count
131	CN_MAX_M_atom_count
132	CN_MAX_X_atom_count
133	CN_MAX_polyhedron_volume

134	CN_MAX_central_atom_to_center_of_mass_dist
135	CN_MAX_number_of_edges
136	CN_MAX_number_of_faces
137	CN_MAX_shortest_distance_to_face
138	CN_MAX_shortest_distance_to_edge
139	CN_MAX_volume_of_inscribed_sphere
140	CN_MAX_packing_efficiency

Supplementary Table S6. Universal features list

	Feature
1	Entry
2	INT_UNI_formula
3	INT_UNI_shortest_homoatomic_dist
4	INT_UNI_shortest_heteroatomic_dist
5	INT_UNI_shortest_homoatomic_dist_by_2_by_atom_size
6	INT_UNI_shortest_heteroatomic_dist_by_sum_of_atom_sizes
7	INT_UNI_shortest_homoatomic_dist_by_2_by_refined_atom_size
8	INT_UNI_shortest_heteroatomic_dist_by_sum_of_refined_atom_sizes
9	INT_UNI_highest_refined_percent_diff_abs
10	INT_UNI_lowest_refined_percent_diff_abs
11	INT_UNI_packing_efficiency
12	WYC_identical_lowest_wyckoff_count
13	CN_AVG_coordination_number
14	CN_AVG_polyhedron_volume
15	CN_AVG_central_atom_to_center_of_mass_dist
16	CN_AVG_number_of_edges
17	CN_AVG_number_of_faces
18	CN_AVG_shortest_distance_to_face
19	CN_AVG_shortest_distance_to_edge
20	CN_AVG_volume_of_inscribed_sphere
21	CN_AVG_packing_efficiency
22	CN_MIN_coordination_number
23	CN_MIN_polyhedron_volume
24	CN_MIN_central_atom_to_center_of_mass_dist
25	CN_MIN_number_of_edges
26	CN_MIN_number_of_faces
27	CN_MIN_shortest_distance_to_face
28	CN_MIN_shortest_distance_to_edge
29	CN_MIN_volume_of_inscribed_sphere
30	CN_MIN_packing_efficiency
31	CN_MAX_coordination_number
32	CN_MAX_polyhedron_volume
33	CN_MAX_central_atom_to_center_of_mass_dist
34	CN_MAX_number_of_edges
35	CN_MAX_number_of_faces
36	CN_MAX_shortest_distance_to_face
37	CN_MAX_shortest_distance_to_edge
38	CN_MAX_volume_of_inscribed_sphere
39	CN_MAX_packing_efficiency