Electronic Supplementary Material (ESI) for Digital Discovery.
This journal is © The Royal Society of Chemistry 2023

Supplementary Information for: Reinforcement Learning in Crystal Structure Prediction

Supplementary Note 1. RL-CSP hyperparameters.

The following list contains the hyperparameters of RL-CSP with the proposed default values:

- alpha = 0.0005; the learning rate (see formula 1);
- h_type = ‘linear’; the type of action preferences function;
- free_term = 1; free term coefficient in h function;
- beta = 1; β from updating rule 1 reflects how much the updating rule relies on entropy regularization;
- e_threshold = 0.8; the entropy threshold for the entropy regularization mechanism; if the entropy is higher then it does not used in updating rule 1;
- scale_reward = True; the option turning on reward normalization;
- epsilon = 0.1; the proportion with which the uniform policy is used instead of the RL-CSP policy;
- smart_penalty = True; the penalty is calculated as described in the Methods section;
- zero_reward_penalty = 0; if smart_penalty = False this hyperparameter’s value is used instead of the reward if the reward is calculated as zero;
- non_unique_penalty = 0; if smart_penalty = False this hyperparameter is used instead of the reward if the new observed state was met before and this parameter is not zero;
- non_converge_penalty = 0; if smart_penalty = False this hyperparameter is used instead of the reward if the energy calculation did not converge;
- step_reward_limit = 5000; this parameter is used if scale_reward = True and states the number of rewards utilized for the reward scaling;

RL-CSP in FUSE set the hyperparameters as we described above. For MC-EMMA along with increasing the learning rate up to 0.01 the other hyperparameters were simplified, specifically, scale_reward and all kind of penalties were turned off and free_term was set zero. Even in this simplified setting RL-CSP is still able to demonstrate the improvement that is reported in the paper. Potentially, the performance gains will be higher by optimising these parameters further, but such investigation goes beyond the scope of this work.

Supplementary Note 2. Local structure optimisations.

In this work all local structure optimisations were performed using the GULP code[4]. For the optimisation, all energies were calculated using pairwise interatomic potentials, with the Buckingham potential for the short-range interactions, with all potentials having a cut-off radius of 12 Å, unit cell parameters and atomic positions were optimised until the norm of gradient on forces was lower than 0.001(a.u.). Supplementary Tables 1 and 2 give details of the individual runs performed in this work using respectively FUSE and MC-EMMA. Interaction parameters for these calculations with FUSE and MC-EMMA are shown in Supplementary Tables 3 and 4 respectively. The final optimized structures for all runs are placed on GitHub:
https://github.com/lrcfmd/FUSE_RL/blob/main/lowest_energy_structures.zip
https://github.com/lrcfmd/MC-EMMA-RL/blob/master/YBa2Ca2Fe5O13_lowet_energy_structures.zip .

Supplementary Note 3. The Flexible Unit Structure Engine (FUSE).

FUSE assembles crystal structures from small building blocks of atoms called submodules, derived based on chemical knowledge, each of the eight position motifs contains positions for 1 cation (the cation position can be assigned as vacant) and 0–3 anions arranged in a planar grid. When submodules are generated, the species on each specific site is randomly allocated. The submodules are grouped together to assemble single atom thick layers called modules, which are in turn stacked on top of each other to create a full crystal structure. Potential structures are then explored using a modification of a Monte Carlo search (Fig. 3):

1. Instead of using one random structure as a start point, FUSE generates an initial, randomly generated population of structures, which are all then locally optimised. The lowest energy structure from the initial population is then carried forwards as the evolving structure.
2. The BH routine includes a stage inspired by simulated annealing; to avoid the search getting trapped in local minimum, once FUSE has visited 500 structures since its last downhill step it initiates a "melt", where the temperature parameter (*kT*) in the MC accept stage is steadily increased to allow for steep increases in energy, the temperature parameter is then reset once a subsequent downhill step is taken.

When FUSE generates a structure, either randomly or through one of the actions listed below, it will "error check" the proposed new structure for the co-ordination chemistry of all of the cation – anion environments. If greater than 25% of the cations are found to have an incorrect co-ordination, FUSE will not accept the random structure and will repeat the process starting from the action selection. Note: this error checking procedure can then often result in the actual percentage of each action (see below) selected when running the MC search deviating away from that set in the input file by any fixed policy making it difficult to ensure a uniform probability distribution among actions.

In FUSE, there are nine possible action types for possible structures, labelled with numbers within the code. In the FUSE study[1] three of the actions are united by number 1 due to their similarity, and the default probability of each action being selected is stated in brackets, rounded to the nearest percent:

1: (in the FUSE study[1]) (51%) Swap the position of two or more atoms (of different species) within the structure. There is an 8: 3: 1 probability of swapping; two atoms, a random number between three and *n-1* (where *n* is the total number of atoms in the structure) and all of the atoms in the structure.

2: (14%) Swap the positions of two randomly chosen submodules within the structure.

3: (9%) Each structure in FUSE contains a set of "building instructions" which indicate how the submodules are to be assembled: the dimensions of the unit cell (in submodules) and a translation applied to each full module to prevent atoms being placed atop each other. This action generates a new set of instructions while retaining the current set of sub-modules.

4: (9%) Swap the positions of two full modules.

5: (9%) Double the structure along one of the three crystallographic axes (chosen at random). This action can only be used when the total number of atoms is under half of the maximum permitted in the calculation.

6: (6%) Generate a random new structure, limiting the maximum number of atoms to be equal to the current structure.

7: (3%) Generate a random new structure.

Additional updates were made to the code of FUSE to work correctly with reinforcement learning.

Update 1. As stated above, when an action has been applied and the proposed structure fails the error check stage, FUSE restarts the process to select and perform an action on a structure, which can result in the actual actions used differing significantly from the probabilities chosen in the input file. For RL-CSP and Uniform, the code has been updated such that when a proposed structure is rejected, the same action is performed again (starting from the current structure) until the action obtains a structure which passes the error check.

Update 2. In order to learn the preferences among different types of swaps in action 1 this action was split into and replaced by three separate actions:

1 (in this study): (34%) Swap the positions of two atoms.

9: (13%) Swap random number between three and *N-1* atoms where *N* is the number of atoms in a unit cell.

10: (4%) Swap the positions of all atoms.

The default probabilities of the updated and new actions being selected are stated in brackets (rounded to the nearest percent) and are equal to those for the corresponding branch of action 1 before update. FUSE with the Original policy uses the default probabilities (Fig. 4a) whereas FUSE with the Uniform policy works and RL-CSP policy starts with equal probabilities for all nine actions.

Note: the action numbering reflects that labelling used within the code. At the time of writing there is an action 8 in development, but not fully functional, so it was not used within this study. To keep this work as accessible as possible, we decided to retain the numbering as used in the version of the code associated the FUSE study[1].

Update 3. In order to avoid wasting compute time, we created an input option to force the BH to stop once a target energy has been obtained, allowing us to stop a run if the global minimum structure had been located, which we know in advance due to the previous work on the benchmark examples.

RL-CSP is embedded into FUSE BH routine in two places highlighted in Fig. 3. First, the RL-CSP policy is used to select a new action. Then, after the energy calculation for the trial structure, RL-CSP updates its policy according to the trial structure energy.

Supplementary Note 4. Monte Carlo Extended Module Materials Assembly (MC-EMMA).[2]

MC-EMMA assembles crystal structures using full modules as the fundamental building block. Once a set of modules is chosen, they are then assembled along one stacking direction. Unlike FUSE, in MC-EMMA, modules are user-defined, and are typically chosen based on fragments of known crystal structures, with each provided in the input file as either a .cif file or ase[3] Atoms object. In this work all modules are used as presented previously.[2] Structures are then permuted in MC-EMMA using an MC search, with six possible actions used to permute the structure (Supplementary Fig. 2). For clarity, these are listed as labelled within the code, and the default probability of each action being selected stated in brackets, rounded to the nearest percent:

T1: (29%) Select two modules of different types in the structure and swap their location.

T2: (21%) Randomise the stacking sequence of all the modules in the structure.

T3: (21%) Attempt to swap a module in the structure for one of a different type.

T4: (14%) Generate a random new structure, retaining the number of modules in the structure that are currently being used.

T5: (7%) Double the length of the structure, this action can only be used when the current structure contains under half of the maximum number of modules permitted in the calculation (set in the input file).

T6: (7%) Generate an entirely new random structure.

The only change to the original MC-EMMA code that was made before embedding RL-CSP is the same update as Update 3 for FUSE; the additional input parameter indicates the target energy achieving which MC-EMMA stops calculations. RL-CSP was embedded into MC-EMMA in two stages reflected on the general BH Fig. 3.

The policy sharing between 40 runs of MC-EMMA with RL-CSP is organised via the common policy vector θ. All runs start simultaneously but choose the initial structure independently. Each run follows its own trajectory (sequence of structures and actions) and does not consider the trajectories of other runs. Given the current structure, a run selects an action according to the shared policy θ, observes a new structure and updates θ according to the reward obtained. When the run finds the global minimum, it stops while other working runs continue following their own trajectories in which they did not find the global minimum yet. Fig. 7e demonstrates the shared policy learned by 40 runs of MC-EMMA with RL-CSP.

Supplementary Note 5. Numbers of steps in runs.

This note explains the numerical results of the RL-CSP policy benchmarking provided in Supplementary Tables 1 and 2, which were utilized to calculate the numbers displayed in Figs. 5a and 7a respectively.

For FUSE, 3 policies were tested on 5 compositions each. The three tested policies are as follows:

- Original; the fixed policy tuned by hand for composition $Y_2Ti_2O_7$ and used in the original FUSE study[1];
- Uniform; the fixed policy with equal probabilities for each of the 9 actions to be selected;
- RL-CSP; the dynamic policy, that starts from the Uniform policy and improves during the CSP run via Reinforcement Learning.

Each policy was tested for the five compositions below:

- $Sr_4Ti_3O_{10}$
- Y_2TiO_5
- $Y_2Ti_2O_7$
- SrY_2O_4
- Y_2O_3

Each policy-composition pair was tested in 40 independent runs of FUSE. Each run was allotted a minimum of 90,000 steps to discover the global minimum but was stopped earlier if it was located. Supplementary Table 1 contains the numbers of steps in all runs of FUSE in ascending order for each policy-composition pair. If the global minimum was not found in a specific run, the number is coloured in red. The bottom row of Supplementary Table 1 shows the mean number of steps among 40 runs (12 fastest for Y_2O_3) rounded to the nearest integer. These numbers represent N_{mean}, the mean number of steps required to find the global minimum, which is used in the main text to compare the policies for all policy-composition pairs except SrY_2O4 with the Original policy and Y_2O_3 with the Uniform policy where only the lower bound is provided.

The sums of N_{mean} for all compositions of FUSE under different policies are provided in the bottom row of the table in Fig. 5a. These sums are utilized to evaluate the performance of policies for the task of exploration across compositions of a phase field and are highlighted in red if only the lower bound is provided. The last two columns of the table in Fig. 5a contain the percentage of the steps used in the policy RL-CSP compared to the Original and Uniform policies. Green and red colours indicate percentages when RL-CSP takes respectively fewer or more steps than the policy it is being compared to.

For MC-EMMA, 3 following policies were tested for composition $YBa_2Ca_2Fe_5O_{13}$:

- Original; the fixed policy tuned by hand and used in the MC-EMMA study[2];
- Uniform; the fixed policy with equal probabilities for each of 6 actions to be selected;
- RL-CSP; the dynamic policy learned on the fly.

The Original and Uniform policies were tested in 40 independent runs of MC-EMMA. As for RL-CSP, 40 runs of this policy functioned independently as well but were initiated simultaneously and utilized and improved upon the same policy vector, sharing the learning process. Supplementary Table 2 displays the lengths of all runs, sorted in ascending order for each policy. The bottom row represents the mean among 40 entries rounded to the nearest integer and displayed in Fig. 7a as well. These numbers were used to evaluate the performance of RL-CSP in comparison to the fixed policies in MC-EMMA.

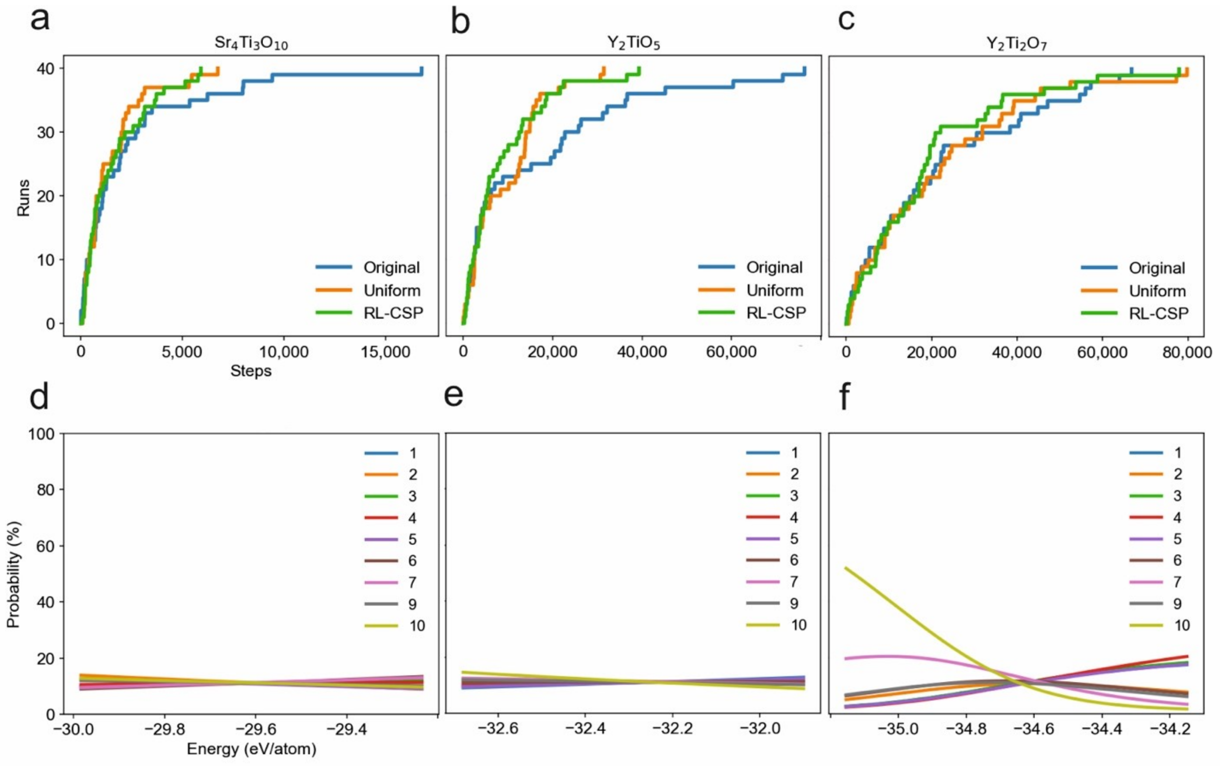
Supplementary Note 6. Computational setup.

The RL-CSP code was written in Python and requires MySQL server to store the learning process variables. The University of Liverpool parallel Linux cluster nodes (two 20-cores Intel(R) Xeon(R) Gold (6138 CPU @ 2.00GHz) processors and 384 Gb memory each node) connected by a 100 gbit/s OmniPath network were used to run the code.

FUSE with different policies and compositions was run 600 times (40 runs for each of 3 policies and 5 compositions) allowing minimum 90,000 steps of BH search and was stopped earlier if the global minimum was found. The typical run of the fastest composition ($Sr_4Ti_3O_{10}$) took ~24 hours whereas for the slowest composition (Y_2O_3), where only 30% (12 / 40) of runs on average manage to find the global minimum in 90,000 steps, the remaining runs took ~45 days till they passed 90,000 limit.

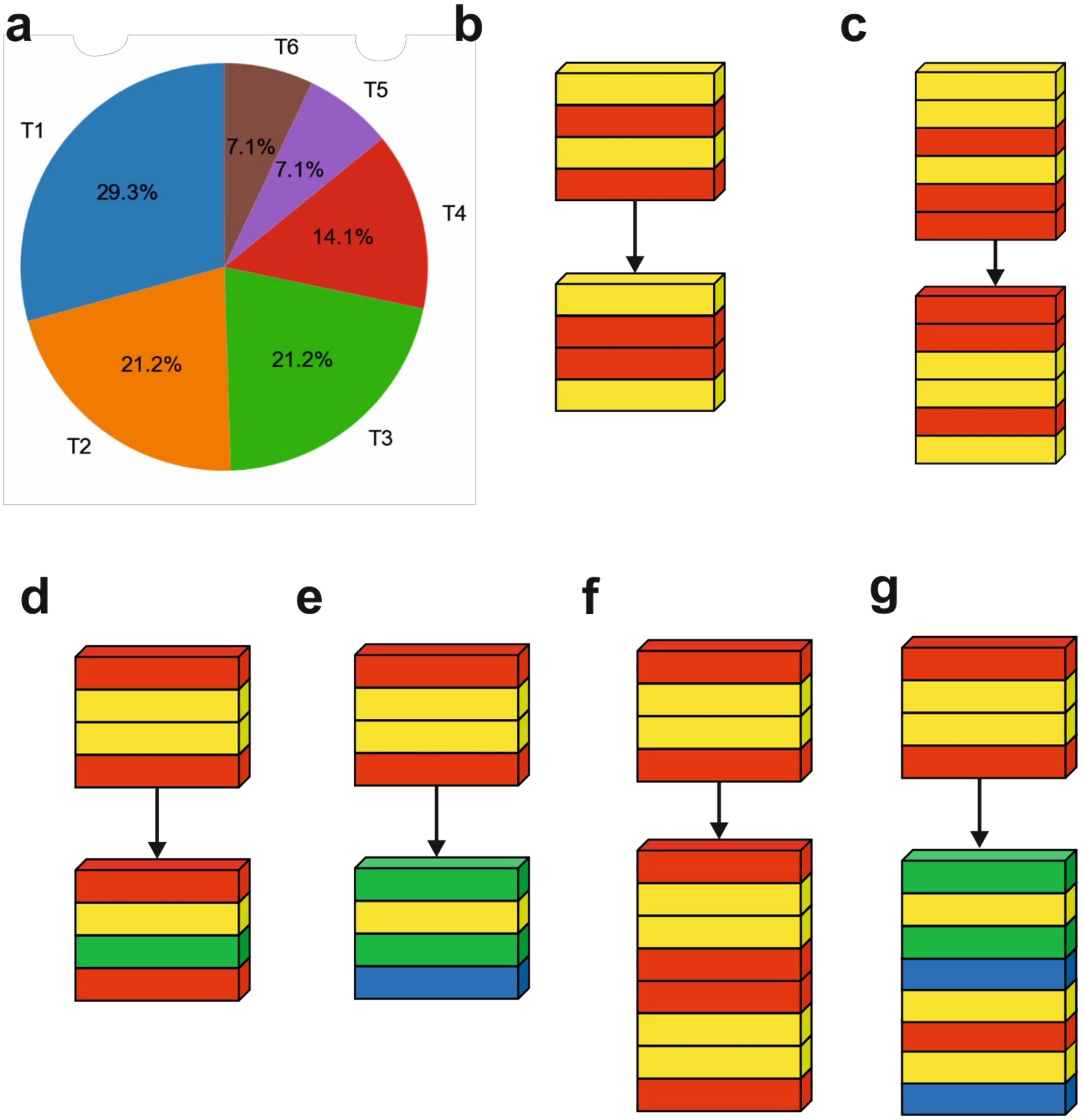
MC-EMMA with different policies and one composition was run 120 times (40 runs for each of 3 policies) and all 120 runs stopped when the global minimum was found.

Supplementary Figure 1. Efficiency of different FUSE policies at the compositions $Sr_4Ti_3O_{10}$, Y_2TiO_5, and $Y_2Ti_2O_7$.



a, b, and **c** show how many runs have found the global minimum in the given number of steps. **b, e,** and **f** show the final policies learnt by RL-CSP for $Sr_4Ti_3O_{10}$ (**d**), Y_2TiO_5 (**e**), and $Y_2Ti_2O_7$ (**f**) at the end of a typical RL-CSP run.

Supplementary Figure 2. Actions for MC-EMMA.



a, The Original fixed policy probabilities for six possible actions.[2] **b**, “T1”, The position of two modules within the structure are switched. **c**, “T2”, The module sequence for the current structure is randomized. **d**, “T3”, One module within the structure is swapped for one of a different type from the input file. **e**, “T4”, The current structure is replaced by a new random structure containing the same number of modules. **f**, “T5”, The structure is doubled along the stacking direction, so long as the current structure contains half or fewer of the maximum number of modules permitted from the input file. **g**, “T6”, The current structure is replaced by a new random structure by the maximum number of modules permitted in the input file.

Supplementary Table 1. The number of steps in FUSE by policy and composition. For a given policy-composition pair, each column contains the numbers of steps required to find the global minimum in 40 runs of FUSE in ascending order. The bottom row highlighted in grey shows the mean number of steps among all 40 runs for compositions $Sr_4Ti_3O_{10}$, Y_2TiO_5, $Y_2Ti_2O_7$, SrY_2O_4 and the mean among the first 12 entries (the 12 shortest runs) for Y_2O_3. The numbers in red indicate the runs over 90,000 steps that did not reach the global minimum.

$Sr_4Ti_3O_{10}$			Y_2TiO_5			$Y_2Ti_2O_7$			SrY_2O_4			Y_2O_3		
Original	Uniform	RL-CSP	Original	Uniform	RL-CSP	Original	Uniform	RL-CSP	Original	Uniform	RL-CSP	Original	Uniform	RL-CSP
7	87	76	159	25	280	81	607	223	7782	104	569	3348	28399	7864
7	160	145	457	217	509	280	963	326	35702	920	815	4543	34049	16585
102	178	162	473	420	659	595	1212	569	40066	3677	1234	5013	34900	18850
110	205	198	885	787	794	1140	1475	979	41769	4034	1490	6698	43653	23581
144	212	204	1013	1168	978	1221	1961	1927	43459	4591	1593	14007	67431	25372
156	214	204	1308	1188	1077	1742	1983	2847	46613	4799	2345	15183	77912	25952
170	230	296	1846	2254	1125	2530	2379	3228	48876	6720	4658	32085	80567	26343
226	242	304	1956	2409	1296	2802	2439	3763	51354	6737	4872	38018	83045	39231
282	367	372	1977	2479	1621	3609	3886	5733	65258	8148	4977	46699	86823	57737
291	385	441	2246	2489	2146	4506	5202	6906	66167	8203	5352	53940	90361	63980
453	410	446	2493	2527	2382	5431	6765	7003	78121	9992	8552	62408	90700	73070
463	467	481	2726	2567	2639	5450	6797	7294	116869	10457	9498	67441	91195	74907
550	655	481	2947	3330	3340	8172	9032	7600	96619	11613	13945	71678	91838	90106
705	664	527	2990	3497	3448	8779	9034	8198	100898	12813	15368	79608	92081	90191
711	670	607	3004	3558	3733	8829	9621	9422	101234	13496	15954	87940	92223	90394
748	703	663	4097	4278	3879	10209	10230	9910	101947	13903	16102	90012	92552	90445
863	740	695	4111	4416	3918	10491	11047	12229	103248	14395	16553	90028	93827	91078
932	752	697	4213	4426	4501	13436	12669	13326	103289	14601	17660	90031	93833	91784
1039	755	786	4752	5910	5019	13484	14662	13989	104324	15566	18435	90047	93990	92206
1072	762	847	6114	6021	5076	14887	15578	15914	105375	16616	18544	90048	94079	93160
1105	930	954	6232	8315	5422	15819	17777	16994	105891	23648	19181	90058	95967	93879
1243	1000	1095	7045	10185	5644	16661	18250	17022	106563	23683	21189	90063	96794	94007
1260	1036	1212	8866	11687	5808	19730	18860	17240	108430	30355	21540	90065	97272	94211
1611	1041	1336	12219	12337	6717	20339	21940	17704	109445	32518	21905	90070	97752	94487
1879	1100	1501	15184	12526	7631	20867	22181	18311	109473	33041	22155	90080	98606	96126
1923	1510	1599	19562	12813	8212	22181	23005	19009	109481	33751	22245	90090	98969	96413
1974	1560	1732	20434	13684	9105	22262	24169	19590	115286	35527	29958	90097	99175	96807
2210	1968	1879	21748	13816	10081	22806	24708	19604	115417	35534	30757	90111	99266	97014
2307	1993	1907	22034	13875	11972	30003	27805	20532	116258	42128	31551	90170	99445	97184
2694	1993	2165	22719	14009	12477	30543	31859	20847	124046	44084	34670	90170	100373	97276
2816	2069	2575	25808	14874	13030	38346	31881	22099	125698	47211	36233	90261	101097	97830
3146	2084	2899	26338	14929	13307	40379	35843	30676	131737	48705	36424	90281	102429	98441
3157	2209	3070	31211	15379	15754	40862	36417	32577	132034	55177	37291	90450	102849	98583
3504	2356	3140	32164	15666	17462	44872	39032	33224	134841	58157	41977	90559	103398	99270
5356	2836	3631	36233	16315	18364	47136	39245	36328	136952	61456	46715	90971	103488	99455
6235	3007	3718	36674	17182	18665	54665	44215	36614	137727	61732	53051	91514	104295	99501
7968	3140	4090	45187	21290	21680	56010	45511	46339	142249	75624	56744	91826	106834	99697
8006	5322	5135	60372	22704	22399	57263	52451	53747	143392	109865	62498	92824	109221	99929
9422	5457	5773	71459	30664	36582	63949	77270	58797	143555	125524	68261	93238	112244	99991
16775	6743	5910	76324	31448	39312	66795	79769	77909	148075	197100	90905	96426	115839	100471
2341	1455	1599	16189	9442	8701	21229	20993	18664	98888	33905	24094	29115*	67420*	37789*

*The mean of the first 12 entries in the column.

Supplementary Table 2. The number of steps in MC-EMMA by policy. Each column shows the numbers of steps required to find the global minimum in 40 runs of MC-EMMA for $YBa_2Ca_2Fe_5O_{13}$ with a given policy. The bottom row highlighted in grey shows the mean number of steps among all 40 runs rounded to the nearest integer.

Original	Uniform	RL-CSP
20	160	1
91	162	206
115	178	230
150	180	267
156	555	293
179	651	349
187	667	358
247	685	422
253	704	435
253	707	443
259	731	450
292	747	464
308	782	611
366	828	624
539	862	666
734	902	682
764	904	693
794	1003	873
929	1080	899
1077	1336	1059
1125	1363	1181
1148	1582	1251
1190	1725	1310
1365	1734	1323
1372	1893	1381
1787	2062	1680
1898	2252	1711
1995	2429	1757
2244	2639	1828
2266	2859	1828
2329	2904	1926
2515	3200	1957
2695	3531	2038
2957	3787	2091
2984	4185	2693
2995	4333	2831
3221	5454	3362
3594	5936	4214
4081	6807	4492
7434	11656	10273
1473	2154	1529

Supplementary Table 3. **Buckingham potential parameters used for calculations with FUSE**, as used in previous work[1], all potentials use a cut-off distance of 12 Å.

Interaction	A (eV)	ρ (Å)	C (eV Å^{-6})
O^{2-} - O^{2-}	1388.77	0.36262	175
Y^{3+} - O^{2-}	23 000	0.24203	0
Sr^{2+} - O^{2-}	1952.39	0.33685	19.22
Ti^{4+} - O^{2-}	4590.7279	0.261	0

Supplementary Table 4. Buckingham potential parameters used for calculations with MC-EMMA, as used in previous work[2], all potentials use a cut-off distance of 12 Å. For these potentials, Ba, Ca and O atoms were modelled using a shell model, where a massless shell is attached to the core by a spring constant, and the charges between the shell and core set to give a combined total equal to the formal charge on the atom. The spring constants were 34.05, 34.05 and 42 eVÅ^{-2} for Ba, Ca and O respectively. The charges on the shells were set to 1.831, 1.281 and -2.24 for Ba, Ca and O respectively.

Interaction	A (eV)	ρ (Å)	C (eV Å^{-6})
O^{2-} - O^{2-}	22764.3	0.149	42
Y^{3+} - O^{2-}	20717.5	0.24203	0
Ba^{2+} - O^{2-}	4818	0.3067	0
Fe^{3+} - O^{2-}	1244.5	0.3299	0
Ca^{2+} - O^{2-}	2272.741	0.2986	0

References

1. C. M. Collins, G. R. Darling and M. J. Rosseinsky, *Faraday Discuss*, 2018, **211**, 117-131.
2. C. Collins, M. Dyer, M. Pitcher, G. Whitehead, M. Zanella, P. Mandal, J. Claridge, G. Darling and M. Rosseinsky, *Nature*, 2017, **546**, 280-284.
3. A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer and C. Hargus, *Journal of Physics: Condensed Matter*, 2017, **29**, 273002.
4. J. D. Gale and A. L. Rohl, *Molecular Simulation*, 2003, **29**, 291-341.