**Supplementary Information:**

Further details on the constraint satisfaction and run times for different phase mapping algorithms, CRYSTAL , NMF [1](#_ENREF_1), [2](#_ENREF_2), AgileFD [3](#_ENREF_3), [4](#_ENREF_4), and NMFK are provided in Table S1 for the Pd-Rh-Ta system. NMF is a non-negative matrix factorization approach using iterative alternating multiplicative rules. Given its lightweight nature without combinatorial inference ability, the computation cost of NMF is low. The convolutive-NMF-based solver AgileFD extends NMF by creating multiple shifted copies for phases and activations, applied to the logarithmically transformed data, and therefore gains the ability to handle shifting to better interpret the original XRD patterns. Comparison was also made with a recently proposed variant of NMF that uses customized clustering for capturing shifting called NMFK.[5](#_ENREF_5) We compare these four methods with respect to five metrics: runtime and satisfaction of four constraints: the Gibbs constraint, the Gibbs-Alloy constraint, and the Phase Connectivity constraint (computed separately for both pure phases and phase fields). Runtime is the wall clock time spent by the solver producing results using 12 cores. For the Gibbs and Gibbs-Alloy constraints, we measure the average proportion of the sample points that satisfy the physical rule. A phase is considered activated at some sample point if the activation is beyond the threshold 10-9. For the Connectivity constraint, solutions were evaluated with respect to two aspects: pure-phase connectivity and phase-field connectivity, and the average ratio of connected phases is taken as the metric. With regard to pure-phase connectivity, all the samples containing a pure phase should be connected according to Delaunay triangulation. Furthermore, all the samples containing the same set of pure phases should form a connected phase field. Extended Data Table S1 shows that generally solutions can be created by NMF, AgileFD or CRYSTAL within 2 minutes, the solution produced by NMFK takes considerably longer in part due to its Matlab implementation, and that some fraction of the sample within these solutions meet each of the physical constraints. However only CRYSTAL meets all physical constraints for any given phase diagram as 0% of the solutions produced by NMF, AgileFD and/or NMFK satisfy all the constraints. The same solutions are evaluated for their reconstruction of the measured data in Table S2. The additional criterion imposed by human analysts is that a single fcc phase is utilized in the solution. As noted in the manuscript, complete solubility of Pd and Rh is observed, yet many solutions utilize both fcc-Rh and fcc-Pd, artificially increasing the number of present phases.

**Table S1. Performance of CRYSTAL, AgileFD, NMF, and NMFK with regard to runtime and validity.** This expanded version of Table 1 includes the average run times and number of successful runs. CRYSTAL is implemented in C++. The mixed integer programming bots are implemented using ILOG CPLEX. **NMFK** is implementedin Matlab, which requires an average of 9s per NMF run. This algorithm involves many runs over a range of K to identify a single solution, in this case 1400 NMF runs from K = 3 to 16, making the total run time per solution quite large. The % of samples satisfying the Gibbs constraint, % of samples satisfying Gibbs-Alloy constraint, and % of phases or phase fields satisfying the Connectivity constraint and % of solutions meeting all four of these constraints are shown for CRYSTAL, AgileFD, NMF, and **NMFK** solutions for the Pd-Rh-Ta system. AgileFD, NMF, and **NMFK** fail to satisfy all the constraints. AgileFD and NMF perform slighty worse than **NMFK** with respect to satisfying the physical constraints. CRYSTAL is the only approach that produces solutions satisfying all the constraints in less than 110s, on average, for all values of K. A time limit of 20 minutes is set for CRYSTAL.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| K (# basis patterns) | 3 | 4 | 5 | 6 | 7 |
| **CRYSTAL** | # succ runs / total runs | 498/500 | 497/500 | 494/500 | 497/500 | 494/500 |
| Avg runtime (s) | 45.45 | 63.98 | 79.08 | 93.69 | 106.55 |
| Gibbs | 100% | 100% | 100% | 100% | 100% |
| Gibbs-Alloy | 100% | 100% | 100% | 100% | 100% |
| Pure phase connectivity | 100% | 100% | 100% | 100% | 100% |
| Phase field connectivity | 100% | 100% | 100% | 100% | 100% |
| Entire phase diagram valid | 100% | 100% | 100% | 100% | 100% |
| **AgileFD** | # succ runs / total runs | 500/500 | 500/500 | 500/500 | 500/500 | 500/500 |
| Avg runtime (s) | 13.71 | 20.06 | 28.01 | 35.65 | 40.70 |
| Gibbs | 100% | 74.93% | 60.20% | 47.46% | 37.46% |
| Gibbs-Alloy | 52.66% | 32.36% | 19.05% | 11.71% | 7.88% |
| Pure phase connectivity | 63.67% | 49.95% | 16.68% | 7.03% | 2.31% |
| Phase field connectivity | 33.99% | 32.29% | 28.58% | 34.55% | 49.38% |
| Entire phase diagram valid | 0% | 0% | 0% | 0% | 0% |
| **NMF** | # succ runs / total runs | 500/500 | 500/500 | 500/500 | 500/500 | 500/500 |
| Avg runtime (s) | 2.60 | 2.87 | 3.83 | 5.61 | 8.85 |
| Gibbs | 100% | 79.01% | 64.35% | 55.01% | 47.24% |
| Gibbs-Alloy | 100% | 79.01% | 64.35% | 55.01% | 47.24 |
| Pure phase connectivity | 51.60% | 29.70% | 23.68% | 11.17% | 1.94% |
| Phase field connectivity | 28.73% | 26.37% | 24.73% | 39.22% | 56.42% |
| Entire phase diagram valid | 0% | 0% | 0% | 0% | 0% |
| **NMFK** | # succ runs / total runs | NA | NA | 1/1 | NA | NA |
| Avg runtime (s) | NA | NA | 12916 | NA | NA |
| Gibbs | NA | NA | 87% | NA | NA |
| Gibbs-Alloy | NA | NA | 77% | NA | NA |
| Pure phase connectivity | NA | NA | 40% | NA | NA |
| Phase field connectivity | NA | NA | 50% | NA | NA |
| Entire phase diagram valid | NA | NA | 0% | NA | NA |

**Table S2. Performance of CRYSTAL, AgileFD, NMF, and NMFK on the Pd-Rh-Ta system with regard to reconstruction.** Theaverage reconstruction % results for Pd-Rh-Ta system is provided for 4 different solvers and 5 different configurations (K=3,4,5,6,7) using the same runs as Table S1. In addition to the “# succ runs” from Table S1, the “# Phase Dim. bot valid runs” is shown along with “# runs with correct fcc alloying”, which uses the criterion that a single fcc phase is identified in the basis patterns due to the known complete solubility of Pd-Rh. Though for all the four approaches, # physically meaningful runs and # valid runs drop down as K increases, only runs from CRYSTAL when K $\leq $ 6 are accepted in the end. The average reconstruction % is calculated for the three cases as long as at least one run exists, with these numbers being nearly identical. The reconstruction fraction for the NMFK solution (71%) can be compared to the average values for the K=5 AgileFD and NMF solutions (73% and 62%, respectively). The AgileFD and NMFK solutions have higher reconstruction scores due to their use of more than K=5 components, although for AgileFD these additional components are constrained to be alloy-shifted versions of the K=5 basis patterns. The importance of alloy-based shifting in the Ta-Pd-Rh system leads to the higher reconstruction score of this constrained AgileFD compared to NMFK, whose K=5 solution (post-clustering) originated with 7 unconstrained basis patterns. What is perhaps more remarkable is that CRYSTAL with a 73% average reconstruction score meets or exceeds the reconstruction quality of the other solvers while simultaneously meeting the all physical constraints for its K=5 solutions.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| K (# basis patterns) | 3 | 4 | 5 | 6 | 7 |
| **CRYSTAL** | # runs with correct fcc alloying / # Phase Dim. bot valid runs / # succ runs | 498/498/498 | 494/496/497 | 260/413/494 | 4/230/497 | 0/2/494 |
| Avg reconstruction % for succ runs | 66% | 71% | 73% | 75% | 76% |
| Avg recon. % for valid runs (after Phase Dim. Bot) | 66% | 71% | 73% | 75% | 76% |
| Avg recon. % for physically meaningful runs (after constraint checker and Human Input )  | 66% | 71% | 73% | 76% | NA |
| **AgileFD** | # runs with correct fcc alloying / # Phase Dim. bot valid runs / # succ runs | 0/500/500 | 0/498/500 | 0/405/500 | 0/210/500 | 0/57/500 |
| Avg reconstruction % for succ runs | 65% | 71% | 73% | 74% | 76% |
| Avg recon. % for valid runs (no duplicates) | 65% | 71% | 73% | 74% | 76% |
| Avg recon. % for physically meaningful runs (after constraint checker ) | NA | NA | NA | NA | NA |
| **NMF** | # runs with correct fcc alloying / # Phase Dim. bot valid runs / # succ runs | 0/500/500 | 0/209/500 | 0/118/500 | 0/0/500 | 0/0/500 |
| Avg reconstruction % for succ runs | 52% | 57% | 62% | 67% | 69% |
| Avg recon. % for valid runs (no duplicates) | 52% | 57% | 62% | NA | NA |
| Avg recon. % for physically meaningful runs (after constraint checker ) | NA | NA | NA | NA | NA |
| **NMFK** | # runs with correct fcc alloying / # Phase Dim. bot valid runs / # succ runs | NA | NA | 1 | NA | NA |
| Avg reconstruction % for succ runs | NA | NA | 71.45% | NA | NA |
| Avg recon. % for valid runs (no duplicates) | NA | NA | 71.45% | NA | NA |
| Avg recon. % for physically meaningful runs (after constraint checker ) | NA | NA | NA | NA | NA |

****

**Figure S1:** Comparison of K=5 phase diagrams for the Pd-Rh-Ta. The selected CRYSTAL solution is at left and contains 11 phase fields from different mixtures of the 5 phases. The NMFK solution has 3 of the same 5 phases and 22 phase fields, providing substantially different physical interpretation of the data. The activation of several phases in different composition regions leads to the poor connectivity constraint satisfaction rate and Gibbs violations noted in Table 1 and S1.

**The phase mapping problem**: **problem definition and formulation**

In search of new materials, several elements are deposited onto a substrate from physically separated sources to form a materials library. Different sample locations on the library receive different concentrations of the elements. As a result, distinct and potentially undiscovered materials are formed at different locations. Materials can be characterized by a one-dimensional X-ray diffraction (XRD) pattern, which can be measured using synchrotron radiation from high-energy accelerators. Several pure crystal phases might be present at a single sample location, and therefore, the XRD pattern at that location is a mixture of basis patterns, each corresponding to the basis pattern of a pure crystal phase, not necessarily measured separately (see Fig. 1).

The ***phase mapping problem*** requires the identification of basis patterns (or factors) that correspond to pure crystal phases such that all the XRD patterns at the multiple sampled locations are explained as mixtures of the basis patterns, and the parameters for each sample location that characterize such a mixture.

The phase-mapping problem is an *unsupervised* learning problem. It defies the current state of the art of AI and machine learning techniques because the *solutions have to satisfy intricate constraints* that are based on the underlying solid state physics. Here we provide further details on the **physical constraints** that characterize equilibrium phase mapping for a given synthesis condition (temperature, pressure, and elemental chemical potentials), namely:

**Peak Shifting** - XRD patterns can change as a function of composition due to alloying, i.e. the interstitial or substitutional solution of element(s) within a single phase, which typically yields a composition-dependent scaling of lattice constants. Under the approximation of isometric expansion or contraction of the crystal structure, this peak shifting is adequately modelled as an offset or shift of the logarithmically transformed basis pattern, assuming the basis pattern was measured as a function of the scattering vector magnitude.

**Gibbs constraint** – Gibbs’ phase rule specifies the relationship between the number of phases, number of components, and degrees of freedom for a chemical system at equilibrium. Assuming that temperature and pressure variations are independent (i.e. that measurements rarely take place at temperature or pressure phase boundaries), this implies that a system including *l* components (elements) can have at most *l* phases present at a given sample location, regardless of any additional degrees of freedom.

**Gibbs-Alloy constraint -** alloying, which can be identified when at least one basis pattern shifts between neighboring composition samples, indicates that at least one additional thermodynamic degree of freedom is present in the system at that composition. According to Gibbs’ phase rule, in the presence of alloying the maximum number of coexisting phases is reduced to *l*-1.

**Phase Connectivity constraint -** the set of composition samples containing a given phase must be contiguous in the composition space. The set of composition samples that belong to a given phase field (a unique set of activated phases) must also be connected in the composition space. These two types of connectivity are geometric representations of equilibrium phase behavior and are enforced together.

**Mathematical description of alloying, updating rules, and constraint enforcement**

The AgileFD bot implements a convolutive non-negative matrix factorization algorithm using lightweight multiplicative gradient-descent updating rules operating on **W** and **H** and applied to the logarithmic transformation of the XRD data [6](#_ENREF_6), [7](#_ENREF_7). The alloying-based peak shifting involves shifting the XRD patterns either to the left or right compared to a reference pattern. Let $W$ be an $L×K$ matrix representing the basis patterns in its columns, and column $W\_{\*,k}$ is a length L vector representing the *k*-th basis pattern. Shifting this pattern right so that it begins at index *m* units corresponds to shifting the column vector downward to row index *m*, and filling in 0 for any displaced entries; we denote this downward shift as $W\_{\*,k}^{\downright m}$, and the similar shift upward as $W\_{\*,k}^{\uparrow m}$. The shift operations for the matrix are denoted by $W^{\downright m}$ and $W^{\uparrow m}$, and similar notation is used for other matrices. $m=1$ corresponds to the original value without shift. The mixing parameters are represented by **H**, a collection of M matrices each with dimensions $K×N$, where M is the number of discrete shift steps considered, and $H^{m}$ denotes the mixing weights(activations) corresponding to shift amount *m*. The phase mapping problem entails factorizing the measurement matrix **A** in terms of **W** and **H**:

$$A≈R≡\sum\_{m=1}^{M}W^{\downright ^{m}}H^{m},$$

where **R** is the reconstructed data matrix.

The algorithm alternates between updating $W$ while holding $H$ fixed, and updating **H** while holding **W** fixed until convergence. The update rules used by the are:

$$\tilde{H}^{m}=H^{m}∘\frac{W^{\downright ^{m}}^{T}\left(\frac{A}{R}\right)}{W^{\downright ^{m}}^{T}1\_{L×N}},$$

$$\tilde{W}=W∘\frac{\sum\_{m}^{}\left(\frac{A^{\uparrow ^{m}}}{R^{\uparrow ^{m}}}\right)H^{m}^{T}}{\sum\_{m}^{}1\_{L×N}H^{m}^{T}}.$$

where the ~ denotes the updated matrix, $∘$ denotes element-wise multiplication, division is also performed element-wise and $1\_{L×N}$represents a matrix of $L×N$ ones. The multiplicative updating rules can be derived following similar steps as in Smaragdis[8](#_ENREF_8" \o "Smaragdis, 2004 #2348), using the generalized Kullback-Leibler divergence as the loss function. It can be shown that the loss function never increases when we apply the aforementioned two multiplicative rules. Furthermore, we add a sparsity regularization term into the objective function, extending the objective function to be:

$$C\_{KL}\left(A,R\right)=\sum\_{q,n}^{}(A\_{q,n}log\frac{A\_{q,n}}{R\_{q,n}}-A\_{q,n}+R\_{q,n})+γ\sum\_{m}^{}\left|\left|Γ∘H^{m}\right|\right|\_{1}$$

where the scalar$ γ $and matrix $Γ$ are weight-coefficients that control the importance of the whole sparsity term and the individual activation weights in **H** with regard to sparsity, respectively. As in the simpler case, multiplicative rules can be derived for this regularized objective function

$$H^{m}=H^{m}∘\frac{W^{\downright m}^{T}(\frac{A}{R})}{W^{\downright m}^{T}1\_{L×N}+γΓ}$$

$$W=W∘\frac{\sum\_{m}^{}(\left(\frac{A^{\uparrow m}}{R^{\uparrow m}}\right)H^{m}^{T}+W⋅diag(1\_{1×L}((1\_{L×N}H^{m}^{T})∘W))}{\sum\_{m}^{}(1\_{L×N}H^{m}^{T}+W⋅diag(1\_{1×L}(\left(\frac{A^{\uparrow m}}{R^{\uparrow m}}\right)H^{m}^{T})∘W))}$$

**H** is normalized between multiplicative updates and at the end such that the maximum activation of each phase in **H** equals to 1. This ensures that the sparsity term cannot be improved simply by letting **H** go to zero. This extended objective function is used in the CRYSTAL framework.

Applying the **Gibbs** constraint can then be performed by the following MIP:

Minimize $\left|\left|A-\sum\_{m}^{}W^{\downright m}H^{m}\right|\right|\_{1} ,$

Subject to $0\leq H\_{k,n}^{m}\leq Mδ\_{k,n}, \sum\_{k}^{}δ\_{k,n}\leq l, $

$∀ 1\leq k\leq K,1\leq n\leq N,1\leq m\leq M,1\leq q\leq L,where M=1$.

Here, $δ\_{k,n}$ is an indicator variable which is 1 if phase *k* is present at sample location *n* and the Gibbs constraint is enforced using the big M method.

Applying the **Gibbs-Alloy** constraint requires setting a criterion for the presence of alloying. The mean shift parameter $λ\_{k,n}$ for phase *k* at location *n* is defined as an average: $λ\_{k,n}=(\sum\_{m}^{}H\_{k,n}^{m}β^{m-1})/\sum\_{m}^{}H\_{k,n}^{m}$, where $β$ is the multiplicative shift amount in the original XRD basis between two consecutive shifted versions. Mathematically, the Gibbs-Alloy constraint then becomes: if $\left|λ\_{k,n}-λ\_{k,n^{'}}\right|>ϵ$ holds for any phase *k* in one sample location *n* for any neighboring sample location $n^{'}$, then the number of non-zero elements in vector $\sum\_{m}^{}H\_{\*,n}^{m}$ is at most *l-*1 for a physical system where *l* elements are deposited.

Violations of the Gibbs-Alloy constraint are rectified using a small-scale MIP, which again is computed independently for each sample point:

Minimize $\left|\left|A-\sum\_{m}^{}W^{\downright m}⋅H^{m}\right|\right|\_{1} ,$

Subject to $H\_{k,n}^{m}\leq Mδ\_{n,k}, \sum\_{k}^{}δ\_{n,k}+τ\_{n}\leq l, $

 $\left|\sum\_{m}^{}H\_{k,n}^{m}β^{m-1}-λ\_{k,j}\sum\_{m}^{}H\_{k,n}^{m}\right|\leq ϵ\sum\_{m}^{}H\_{k,n}^{m}+Mτ\_{n}+M\left(1-δ\_{k,n}\right),$

 $∀ 1\leq k\leq K,1\leq n\leq N,1\leq m\leq M,1\leq q\leq L, j\in N\left(n\right),$

$where M=1.15M$, $N(n)$ is defined as the neighbors of sample n given Delaunay triangulation and $ϵ$ is typically set to 1e-6.

The MIP decouples naturally between sample points, which leads to a very small scale mathematical programming which can be solved very quickly in parallel. We conservatively assume that the largest shifting will not exceed 15%, which is larger than any alloying-based lattice constant shift of which we are aware. This constraint means $1\leq β^{M-1}, λ\_{k,n}\leq 1.15$. Therefore, the upper bound for the left side of the aforementioned inequality constraint could be lowered to 1.15 M where M is the number of shifted versions.

Given phase k at sample point n, we want to minimize the following objective function:

Minimize $\left|\left|A-\sum\_{m}^{}W^{\downright m}⋅H^{m}\right|\right|\_{1} ,$

 Subject to $M\left(ζ\_{k,n}^{m-1}+ζ\_{k,n}^{m}\right)\geq H\_{k,n}^{m}, Mζ\_{k,n}^{1}\geq H\_{k,n}^{1}, \sum\_{m}^{}ζ\_{k,n}^{m}\leq 1,$

 $∀ 1\leq k\leq K,1\leq n\leq N,2\leq m\leq M, 1\leq q\leq L, where M=1.$

Here, $ζ\_{k,n}^{m}$ is an indicator variable which is 1 if the m-th shifted version of phase k at sample n is activated. The (*m*+1)-th shifted version is also allowed to be activated if the *m*-th shifted version is activated. Then the weighted mean of the *m*-th and (*m*+1)-th shifted version could approximately simulate a continuous shifted version of phase k. $M$ is the total number of shifted versions while$ M$ is a large constant. However, since **H** is normalized after each iteration, $M$ could be lowered down to just 1.

The Phase and Phase Filed Connectivity bot rectifies the connectivity constraints in a lazy and interactive manner. Experimentally we observed that in most cases the connectivity constraint is violated with a large connected component, and only a few additional small components whose intensity is near the background noise. If this is the case, the Phase Connectivity bot zeros out these small connected components in the **H** matrix, to ensure a physically meaningful solution.

**IAFD parameters and integration into CRYSTAL**

Parallel runs of IAFD are performed in 2 distinct stages of solution generation. In the first stage, CRYSTAL performs a user-specified number (typically 500) of runs for each configuration (value of K, typically 5 configurations) with random initialization to generate a total of approximately 2500 phase diagrams. From this large set of phase diagrams, 100 representative solutions are selected to initialize IAFD runs in the second stage, in which 100 parallel IAFD runs yield refined phase diagrams.

Within each IAFD run, p is typically set to be 3 to gradually enforce the Gibbs constraint in the AgileFD-Gibbs loop and q is typically set to 2 to repair the solutions according to the Gibbs-Alloy and Phase Connectivity constraints and then relax and repair again. We have heuristically determined that these parameter values provide a nice balance between computationally efficiency and phase map refinement. Since the Gibbs-Alloy and Phase Connectivity bots determine the global minima in the objective function, there are no termination parameters for these bots. The AgileFD-Gibbs loop requires a convergence criterion to trigger termination of iterative updating, which occurs when the relative difference between the reconstruction errors from the current iteration and previous iteration drops below the convergence criterion value. This value is set to 10-5 is stage 1 and to a smaller value of 10-6 in stage 2 to promote further refinement of the phase maps.

**Phase Matching bot**

Due to inability to directly query the database, the expert user first assembles the set of known ICDD patterns for the composition space under consideration and removes duplicate ICDD entries such that each unique phase known for the composition space appears once in the ICDD library provided to the Phase Matching bot. Each ICDD pattern is a collection of sticks with locations and intensities. To make the comparison between basis patterns produced by CRYSTAL and the ICDD patterns, we fit Gaussian distributions to the discrete ICDD sticks in order to form a continuous signal that is compared against the basis patterns produced by CRYSTAL. The loss between a basis pattern generated by the IAFD module and an ICDD-derived pattern is defined as the L2 distance between the phase and the ICDD pattern after normalization. We minimize via gradient descent the L2 loss with respect to the following parameters: the peak width of the Gaussians, intensity scaling of the ICDD pattern, and peak shifting parameter that is akin to the alloying-based shifting modelled in **W**, which is necessary because the basis pattern may correspond to a given ICDD phase but with a different lattice constant. After the fit is complete, the L2 loss between the basis pattern and each ICDD-derived pattern is used to rank the ICDD match from best (lowest L2 loss) to worst, which provides the additional opportunity to provide a threshold loss such that if no ICDD-derived pattern sufficiently matches the basis pattern, then the basis pattern may be describing a new phase.

If the system configuration parameter K is lower than the actual number of phases present in the sample, a basis pattern may contain multiple phases and the Phase Matching bot may flag that pattern as possibly containing a new phase. However, in runs performed at higher K values, this erroneous “new” phase will likely be split into the appropriate basis patterns. Therefore, the possibility of new phase discovery is generally only considered for the highest K value yielding valid solutions. In the two systems described in the present work, ICDD matches were identified for all basis patterns. Subtle signatures of new phases such as ordering peaks that arise due to long range order at select alloying compositions, will not be sensitive to the L2­ loss, so human-expert analysis of the CRYSTAL-generated reports is necessary to apply prior knowledge in order to determine whether each basis pattern is sufficiently described by the ICDD fit.

**Clustering bot**

As mentioned above, the Clustering bot compares a pair of phase diagrams according to sample-wise dissimilarity of the phase behavior from the two different solutions. To perform this comparison for a given composition sample, we first assemble the set of ICDD phases that may be present in the XRD pattern according to each phase diagram. For each of the k activated basis patterns in sample *n* of solution *i*, the set of possible phases is determined by choosing the best-match ICDD pattern and any additional phases within a threshold fraction increase in the L1­ loss compared to the best match, and the union of these matched phases is defined as *Si,n*. This threshold is needed because often the library of ICDD patterns contains phases with similar XRD patterns and the Phase Match bot cannot adequately differentiate the phase match, so all inadequately differentiated phases are included in *Si,n*. For example, due to the broad features in the Pd-Rh-Ta dataset this threshold was set to 20% while a lower 5% value was used for the Nb-Cu-V system. Assuming all the phases were matched, the number of phases in *Si,n* can be no smaller than k, and the task now becomes the comparison of *Si,n* and *Sj,n*, i.e. the ICDD phase sets from the same sample in two different solutions (*i* and *j*), to determine if this sample is in the same phase field in the two solutions. If *Si,n*∩*Sj,n* is smaller than k, if different numbers of phases were activated in the solutions, or if at least one phase was not matched to any ICDD pattern, the phase fields are assumed to be different and the distance is the maximum value of 1. Otherwise, the Jaccard distance between *Si,n* and *Sj,n* is used. The average over the samples of this distance metric is then an estimate of the fraction of samples that are in different phase fields in the two solutions.

Clustering on this pairwise dissimilarity using the average linkage method (UPGMA, Unweighted Pair Group Method with Arithmetic Mean) provides clusters of phase diagrams such that dissimilarities within the cluster are smaller than a threshold value, and the representative phase diagram (medoid) can be chosen as the representative phase diagram for each cluster. It is worth noting that solutions with different configurations k can be clustered together, e.g. if a 4-phase solution has nearly the same phase diagram as a 3-phase solution except for 1 additional phase that appears only in a small fraction of the samples, the solutions will have low dissimilarity.

For the initial clustering of ~2500 solutions, the dissimilarity threshold is set such that 100 clusters are obtained, which corresponds to an assumption that there are no more than 100 inadequately differentiated interpretations of the data. For final report generation, the number of inadequately differentiated interpretations of the data is automatically determined through analysis of the hierarchical clustering results. The lowest dissimilarity threshold is chosen such that the representative solutions (1 from each cluster) sufficiently represent the span of different physically-meaningful phase diagrams among all 100 refined solutions, which is defined as follows. For a given solution *i*, a comparison is made with each solution *j* that has the same number of phases (k). If the number of match ICDD phases between these 2 solutions (*Si*∩*Sj* where these sets describe all ICDD matches in the solution, not in a single sample as defined above) is at least k, then solution *i* is sufficiently well represented by solution *j*, and every solution must be represented this way. Intuitively, this algorithm corresponds to ensuring that every phase diagram with a unique set of ICDD phases is included in the list of representative solutions that are reported to the human expert. The rank of the cluster from which a representative solution originates corresponds to the number of refined phase diagrams that provide a similar interpretation of the data.

**Phase Mapper: Visualizer & Interface**

Phase Mapper (see Extended Data Fig. 2) is a web based application that allows for the visualization, generation, and refinement of solution files with an interface designed to provide data and solution visualizations that are intuitive to materials scientists [4](#_ENREF_4). A user can load in different material systems by uploading their instance files. A user can then interactively plot the XRD patterns of different sample points as well as select slices of sample points to plot. In Extended Data Fig. 2 a slice is shown by the red data points enclosed in the grey rectangle in the top left plot. Those points’ XRD patterns are then shown in a heat map plot in the top right. When the user loads a solution file they can select an individual sample point and Phase Mapper will plot the XRD pattern of the point and reconstructed signal as shown in the upper-middle left plot. Users can also see each point’s reconstruction accuracy, given by the upper-middle right plot, where a perfectly reconstructed point is bright green, and a point with no reconstruction is bright red. This allows an expert user to identify potential areas of the system that are not being reconstructed well. The lower-middle left plot allows the user to plot each basis pattern of the solution, while the lower-middle right plot shows where the basis pattern is used in the system, how concentrated it is at each data point and which shifted copy is used for each data point. Additionally, the lower-middle right plot shows the shifts of basis pattern 1. The user can also load ICDD patterns (or known patterns from other databases such as ICSD) as shown in the bottom right plot and compare them to a given solution’s basis patterns, as shown in Extended Data Fig. 2. In this way the expert can see how particular ICDD patterns fit the basis patterns the solver found. Additionally, Phase Mapper is connected to a back end server that can generate new solution files as well as allowing the user to refine the phase field of a loaded solution, point by point.

**Methods for Pd-Rh-Ta experiments**

The deposition used a Si substrate with a 12 nm Ti coating that served as an adhesion layer and diffusion barrier. The substrate was heated to 400 °C before and during deposition with sputtering occurring from elemental targets (>99.9% purity) in separate magnetron sputter sources (Angstrom Sciences) with an atmosphere of 0.66 Pa Ar in system with 10-5 Pa base pressure. The thin films all have thickness on the order of 200 nm, which was not explicitly measured but calibrated with deposition rate monitor measurements of the individual sources and corroborated with the XRF measurements. The asymmetric composition region contained in the Pd-Rh-Ta library and in particular the skewing of the compositions to Ta-rich compositions is due to resputtering of Pd, and to a lesser extent Rh, from the thin film by Ar reflected from the Ta target, as explained by resputtering models [9](#_ENREF_9).

As described in Ref. [10](#_ENREF_10), the as-deposited film, after synchrotron measurements, is prepared for catalyst testing by attempting to reduce surface oxides that are formed in air, which entails a linear voltage sweep from 0 V to -0.8 V (vs Ag/AgCl) in an aqueous solution of 0.01 M H2SO4 and 1 M Na2SO4.

As previously discussed [11](#_ENREF_11), the pH of the near-neutral solution drops precipitously to below pH 5 in the vicinity of active catalysts, and since this is the pH below which quinine fluoresces and the catalyst must continue to be active to result in positive detection by the screening experiment, this pH value can be used to provide an upper limit on the onset potential vs the relative hydrogen electrode (RHE), which is similar to the overpotential for methanol oxidation.

**Non-Equilibrium Phase Behavior and Alloying in Pd-Rh-Ta**

In many materials syntheses and especially sputter deposition, non-equilibrium phase behavior may be obtained. While the Pd-Rh-Ta ternary phase diagram has not been previously explored, we can compare the CRYSTAL phase diagram with reported phase behavior in the binary systems. The lattice constants of fcc-Rh and fcc-Pd are 0.3803 nm and 0.3980 nm, respectively, and these fcc metals exhibit substantial alloying with the equilibrium Pd-Rh phase diagram indicating complete solid solubility for temperatures in excess of 900 °C and incomplete solubility at lower temperatures. Our results indicate that in the athermal sputtering process, thin film alloys are formed for all Pd-Rh compositions. Similar the equilibrium bulk solubility of Ta in these structures is approximately 10 atom% but is observed to be close to 20 atom% in our thin films. The alloying within the intermetallic structures has not previously been explored but we suspect is also far from equilibrium. It is important to note that the solution revealing this nonequilibrium phase behavior was obtained through CRYSTAL’s imposition of equilibrium thermodynamic constraints. The imposed constraints were chosen as a result of our deep experience exploring phase behavior of sputtered thin films. To explore whether constraints such as those based on the Gibbs’ phase rule may be violated in the sputter-deposited films, CRYSTAL enables exploration of solutions with altered or removed constraint enforcement bots in IAFD. However, we find that the nonequilibrium phase behavior generally still adheres to the IAFD constraints. From a different perspective, CRYSTAL with full constraint enforcement provides the closest constraint-adherent solution with a composition map of the reconstruction error such that a composition region with poor reconstruction can be further analyzed for constraint-violating phase behavior.

**References**

1. C.J. Long, D. Bunker, X. Li, V.L. Karen and I. Takeuchi: Rapid identification of structural phases in combinatorial thin-film libraries using x-ray diffraction and non-negative matrix factorization. *Rev. Sci. Instrum.* **80**, 103902 (2009).

2. A.G. Kusne, D. Keller, A. Anderson, A. Zaban and I. Takeuchi: High-throughput determination of structural phase diagram and constituent phases using GRENDEL. *Nanotechnology* **26**, 444002 (2015).

3. S.K. Suram, Y. Xue, J. Bai, R. Le Bras, B. Rappazzo, R. Bernstein, J. Bjorck, L. Zhou, R.B. van Dover, C.P. Gomes and J.M. Gregoire: Automated Phase Mapping with AgileFD and its Application to Light Absorber Discovery in the V–Mn–Nb Oxide System. *ACS Comb. Sci.* **19**, 37 (2017).

4. Y. Xue, J. Bai, R. Le Bras, R. Bernstein, J. Bjorck, L. Longpre, S.K. Suram, R.B. van Dover, J. Gregoire and C.P. Gomes: Phase-Mapper: An AI Platform to Accelerate High Throughput Materials Discovery, in The Twenty-Ninth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI2017).

5. V. Stanev, V.V. Vesselinov, A.G. Kusne, G. Antoszewski, I. Takeuchi and B.S. Alexandrov: Unsupervised phase mapping of X-ray diffraction data by nonnegative matrix factorization integrated with custom clustering. *npj Computational Materials* **4**, 43 (2018).

6. J.R. Hattrick-Simpers, J.M. Gregoire and A.G. Kusne: Perspective: Composition–structure–property mapping in high-throughput experiments: Turning data into knowledge. *APL Mater.* **4**, 053211 (2016).

7. M.L. Green, C.L. Choi, J.R. Hattrick-Simpers, A.M. Joshi, I. Takeuchi, S.C. Barron, E. Campo, T. Chiang, S. Empedocles, J.M. Gregoire, A.G. Kusne, J. Martin, A. Mehta, K. Persson, Z. Trautt, J.V. Duren and A. Zakutayev: Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. *Applied Physics Reviews* **4**, 011105 (2017).

8. P. Smaragdis: Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs, in Independent Component Analysis and Blind Signal Separation: Fifth International Conference, ICA 2004, Granada, Spain, September 22-24, 2004. Proceedings, edited by C. G. Puntonet and A. Prieto (Springer Berlin Heidelberg, Berlin, Heidelberg, 2004), pp. 494.

9. J.M. Gregoire, M.B. Lobovsky, M.F. Heinz, F.J. DiSalvo and R.B. van Dover: Resputtering phenomena and determination of composition in codeposited films. *Physical Review B* **76**, 195437 (2007).

10. J.M. Gregoire, M.E. Tague, S. Cahen, S. Khan, H.c.D. Abruña, F.J. DiSalvo and R.B. van Dover: Improved Fuel Cell Oxidation Catalysis in Pt1−xTax. *Chem. Mater.* **22**, 1080 (2009).

11. J.M. Gregoire, M. Kostylev, M.E. Tague, P.F. Mutolo, R.B. van Dover, F.J. DiSalvo and H.D. Abruña: High-throughput evaluation of dealloyed Pt–Zn composition-spread thin film for methanol-oxidation catalysis. *Journal of the Electrochemical Society* **156**, B160 (2009).