nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-19597-w OPEN

On-the-fly closed-loop materials discovery via Bayesian active learning

A. Gilad Kusne[1,2,10✉], Heshan Yu[2,10], Changming Wu[3], Huairuo Zhang[4,5], Jason Hattrick-Simpers[1], Brian DeCost[1], Suchismita Sarker[6], Corey Oses[7], Cormac Toher[7], Stefano Curtarolo[7], Albert V. Davydov[4], Ritesh Agarwal[8], Leonid A. Bendersky[4,5], Mo Li[3], Apurva Mehta[6] & Ichiro Takeuchi[2,9✉]

Active learning—the field of machine learning (ML) dedicated to optimal experiment design—has played a part in science as far back as the 18th century when Laplace used it to guide his discovery of celestial mechanics. In this work, we focus a closed-loop, active learning-driven autonomous system on another major challenge, the discovery of advanced materials against the exceedingly complex synthesis-processes-structure-property landscape. We demonstrate an autonomous materials discovery methodology for functional inorganic compounds which allow scientists to fail smarter, learn faster, and spend less resources in their studies, while simultaneously improving trust in scientific results and machine learning tools. This robot science enables science-over-the-network, reducing the economic impact of scientists being physically separated from their labs. The real-time closed-loop, autonomous system for materials exploration and optimization (CAMEO) is implemented at the synchrotron beamline to accelerate the interconnected tasks of phase mapping and property optimization, with each cycle taking seconds to minutes. We also demonstrate an embodiment of human-machine interaction, where human-in-the-loop is called to play a contributing role within each cycle. This work has resulted in the discovery of a novel epitaxial nanocomposite phase-change memory material.

[1] Materials Measurement Science Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA. [2] Materials Science and Engineering Department, University of Maryland, College Park, MD 20742, USA. [3] Electrical & Computer Engineering Department, University of Washington, Seattle, WA 98195, USA. [4] Materials Science and Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA. [5] Theiss Research, Inc., La Jolla, CA 92037, USA. [6] Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA. [7] Mechanical Engineering and Materials Science Department and Center for Autonomous Materials Design, Duke University, Durham, NC 27708, USA. [8] Materials Science and Engineering Department, University of Pennsylvania, Philadelphia, PA 19104, USA. [9] Maryland Quantum Materials Center, University of Maryland, College Park, MD 20742, USA. [10] These authors contributed equally: A. Gilad Kusne, Heshan Yu. ✉email: aaron.kusne@nist.gov; takeuchi@umd.edu

Technologies drive the perpetual search for novel and improved functional materials, necessitating the exploration of increasingly complex multi-component materials[1]. With each new component or materials parameter, the space of candidate experiments grows exponentially. For example, if investigating the impact of a new parameter (e.g., introducing doping) involves approximately ten experiments over the parameter range, N parameters will require on the order of 10^{N+} possible experiments. High-throughput synthesis and characterization techniques offer a partial solution: with each new parameter, the number of candidate experiments rapidly escapes the feasibility of exhaustive exploration. The search is further confounded by the diversity and complexity of materials composition-structure-property (CSP) relationships, including materials-processing parameters and atomic disorder[2]. Coupled with the sparsity of optimal materials, these challenges threaten to impede innovation and industrial advancement.

Structural phase maps, which describe the dependence of materials structure on composition, serve as blueprints in the design of functional and structural materials, as most materials properties are tied to crystal-structure prototypes. For example, property extrema tend to occur within specific phase regions (e.g., magnetism and superconductivity) or along phase boundaries (e.g., caloric-cooling materials and morphotropic phase-boundary piezoelectrics). Structural phase maps, and more specifically equilibrium phase diagrams, were traditionally generated over years with point-by-point Edisonian approaches guided by expert knowledge and intuition and involving iterative materials synthesis, diffraction-based structure characterization, and crystallographic refinement.

Machine learning (ML) is transforming materials research before our eyes[3], and yet direct coupling of ML with experiments remains a formidable challenge. Closed-loop autonomous system for materials exploration and optimization (CAMEO) offers a new materials research paradigm to truly harness the accelerating potential of ML, setting the stage for the 21st-century paradigm of materials research—the autonomous materials research lab run under the supervision of a robot scientist or artificial scientist[4]. Furthermore, CAMEO embraces one embodiment of human-in-the-loop autonomous systems[5,6], where the human provides their expertise while ML presides over decision making steps. Live visualization of data analysis and decision making (including uncertainty quantification) provides interpretability of the autonomous process for the human expert in the human-machine research team (see Supplementary Fig. 11). CAMEO also exploits (as of yet) non-automated capabilities of the human expert in the closed loop, thus elevating the capabilities of both human and machine.

Active learning[7]—the ML field dedicated to optimal experiment design (i.e., adaptive design), is key to this new paradigm. Active learning provides a systematic approach to identify the best experiments to perform next to achieve user-defined objectives. Scientific application can be traced as far back as the 18th century to Laplace's guided discoveries of celestial mechanics[8]. Bayesian optimization (BO) active learning techniques have been used more recently to guide experimentalists in the lab to optimize unknown functions[9–14]. BO methods balance the use of experiments to explore the unknown function with experiments that exploit prior knowledge to identify extrema. However, these past studies only advised researchers on the next experiment to perform, leaving experiment planning, execution, and analysis to the researcher. Recent advances in robotics have shifted the burden of materials synthesis from human experts to automated systems, accelerating materials discovery[15,16]. Concurrently, active learning has been demonstrated to accelerate property optimization by guiding simulations of known phases[17]. More recently, autonomous systems and machine learning driven research have been demonstrated for optimizing process and system operation[18–20] sample characterization[21], and tuning chemical reactions of known polymers and organic molecules for technological applications[22–24], using off-the-shelf optimization schemes. Taking another step and placing active learning in real-time control of solid-state materials exploration labs promises to accelerate materials discovery while also rapidly and efficiently illuminating complex materials-property relationships. Such potential innovation has been discussed in recent prospectives[25,26], with a primary focus on autonomous chemistry[27–29].

Here we demonstrate CAMEO (see Fig. 1) in real-time control of X-ray diffraction measurement experiments over composition spreads at the synchrotron beamline and in the lab. The algorithm accelerates phase mapping and materials discovery of a novel solid-state material, with a 10-fold reduction in required experiments, each iteration taking tens of seconds to tens of minutes depending on the experimental task.

Results

CAMEO uses a materials-specific active-learning campaign that combines the joint objectives of maximizing knowledge of the phase map $\mathrm{P}(\mathbf{x})$ with hunting for materials $\mathbf{x}_*$ that correspond to property $\mathrm{F}(\mathbf{x})$ extrema. Here $\mathbf{x}\in\mathbb{R}^{\mathrm{d}}$ is the set of d materials-composition parameters. In particular, subsequent phase mapping measurements are driven by Bayesian graph-based predictions combined with risk minimization-based decision making, ensuring that each measurement maximizes phase map knowledge (see "Methods" section "M1c–e"). CAMEO accelerates both tasks by exploiting their mutual information via function g (see Eq. 1). Further acceleration is achieved through integration of physics knowledge (e.g., Gibbs phase rule) as well as prior experimental and theory-based knowledge of the target material system.

These features allow CAMEO to target its search in specific phase regions or to search near phase boundaries where significant changes in the target property are likely to occur, thus exploiting the dependence of materials property on structure. An example of how phase map knowledge can accelerate materials discovery and optimization is shown in Fig. 2a. Off-the-shelf BO methods ignore material structure and assume material properties are a function of only the synthesis parameters, while CAMEO incorporates knowledge that significant changes in properties may occur at phase boundaries. Phase mapping knowledge thus improves property prediction estimate and uncertainty where it matters most.

For this work a simplified implementation of g is used, switching from phase mapping to materials optimization once phase mapping converges. Materials optimization is focused in the most promising phase region, with a greater importance given to compounds near the phase boundaries. CAMEO thus provides two benefits—a significant reduction of the search space and an improved functional property prediction due to phase map knowledge. Discussion of other approaches to g appears in the "Methods" section. We demonstrate that this physics-informed approach accelerates materials optimization compared to general optimization methodologies that focus on directly charting the high dimensional, complex property function.

$$\mathbf{x}_* = \operatorname{argmax}_{\mathbf{x}}[\mathrm{g}(\mathrm{F}(\mathbf{x}), \mathrm{P}(\mathbf{x}))] \qquad (1)$$

Here, we explored the Ge–Sb–Te ternary system to identify an optimal phase-change memory (PCM) material for photonic switching devices[30]. PCM materials can be switched between the amorphous and crystalline states with an associated change in resistance and optical contrast which can be accessed on the nanosecond scale or shorter. Various Ge–Sb–Te based PCMs,

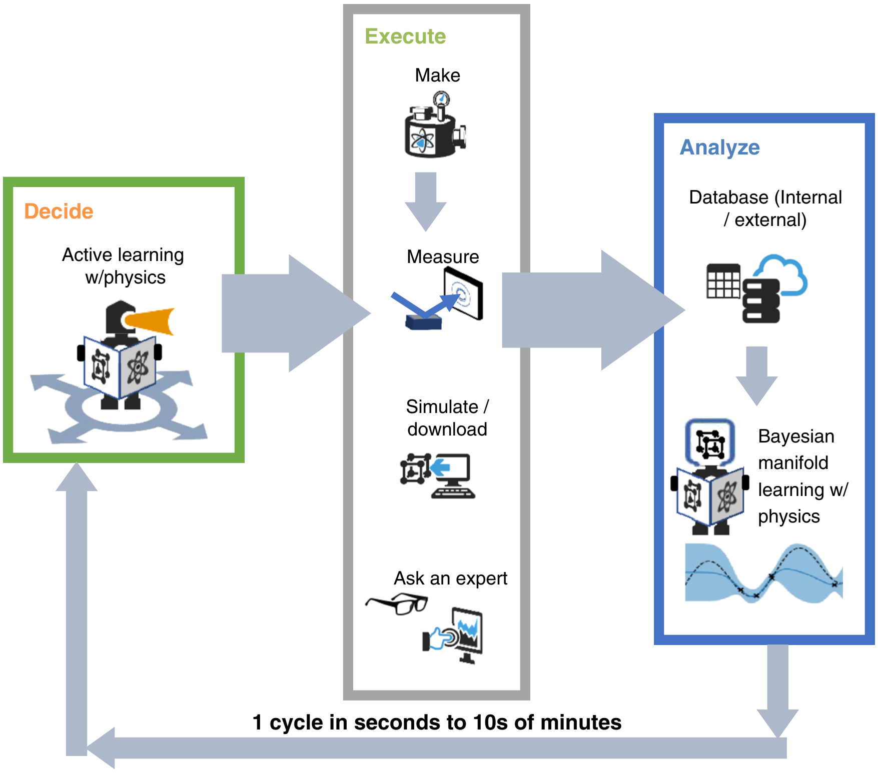
Execute
Make
Measure
Simulate / download
Ask an expert
Decide
Active learning w/physics
Analyze
Database (Internal / external)
Bayesian manifold learning w/ physics
1 cycle in seconds to 10s of minutes

Fig. 1 Closed-loop autonomous materials exploration and optimization (CAMEO). The autonomous cycle begins with loading data from databases including composition data for the materials on the composition spread and computed materials data from the AFLOW.org[41] density functional theory database. The collected data is then used to begin analysis of the data using physics-informed Bayesian machine learning. This process extends knowledge of structure and functional property from materials with data to those without, predicting their estimated structure and functional property, along with prediction uncertainty. Physics-informed active learning is then used to identify the most informative next material to study to achieve user-defined objectives. For this work, active learning can select the next sample to characterize through autonomous control of the high-throughput X-ray diffraction system at a synchrotron beamline or it can (optionally) request specific input from the human-in-the-loop. Future implementations will include autonomous materials synthesis and simulation. The data collected from measurements and from human input are added to the database and used for the next autonomous loop. For more information, see “Methods”.

especially $Ge_2Sb_2Te_5$ (GST225), have been used in DVD-RAM and nonvolatile phase-change random-access memory. We have implemented our strategy for identifying the optimal composition within the ternary for high-performance photonic switching with an eye toward neuromorphic memory applications[31]. Our goal was to find a compound with the highest optical contrast between amorphous and crystalline states in order to realize multi-level optical switching with a high signal-to-noise ratio. The composition range mapped was selected based on the lack of detailed phase distribution and optical property information near known PCM phases. We tasked CAMEO to find the composition with the largest difference in the optical bandgap ΔE_g and hence optical contrast between amorphous and crystalline states. We have discovered a naturally-forming stable epitaxial nanocomposite at a phase boundary between the distorted face-centered cubic Ge–Sb–Te structure (which we refer to as FCC–Ge–Sb–Te or simply GST) phase region and phase co-existing region of GST and Sb–Te whose optical contrast is superior to the well-known GST225 or other compounds within the Ge–Sb–Te ternary. In a direct comparison, a photonic switching device made of the newly discovered material outperforms a device made of GST225 with a significant margin.

CAMEO satisfies many attributes of a robot scientist, as diagrammed in Fig. 1. The modular algorithm has ‘at its fingertips’ a collection of information—knowledge of past experiments both physical and computational, materials theory, and measurement-instrumentation science. The algorithm uses this knowledge to make informed decisions on the next experiment to perform in the pursuit of optimizing a materials property and/or maximizing knowledge of a materials system. For example, at each iteration the set of possible phase maps are identified and ranked by Bayesian likelihood, given analysis results of the measured materials. Phase map and functional property likelihoods establish scientific hypotheses and drive further phase mapping and materials optimization and are also presented to the human-in-the-loop who can then (optionally) provide guidance. Specifics are presented in Methods. CAMEO controls lab-based characterization equipment in real-time to orchestrate its own experiments, update its knowledge, and continue its exploration. The more specific implementation of Eq. 1 is shown in Fig. 2.

CAMEO is based on the fundamental precept that in navigating compositional phase diagrams, enhancement in most functional properties is inherently tied to the presence of particular structural phases and/or phase boundaries. The strategy is, therefore, broadly applicable to a variety of topics with disparate physical properties. The method was first benchmarked and its hyperparameters tuned using a previously characterized composition spread Fe–Ga–Pd, where an entirely different physical property—remnant magnetization, was optimized (see “Methods” for benchmarking method and performance analysis). It was then successfully used to discover a new photonic PCM composition whose ΔE_g (between crystalline and amorphous states) is up to 3 times larger than that of the well-known $Ge_2Sb_2Te_5$.

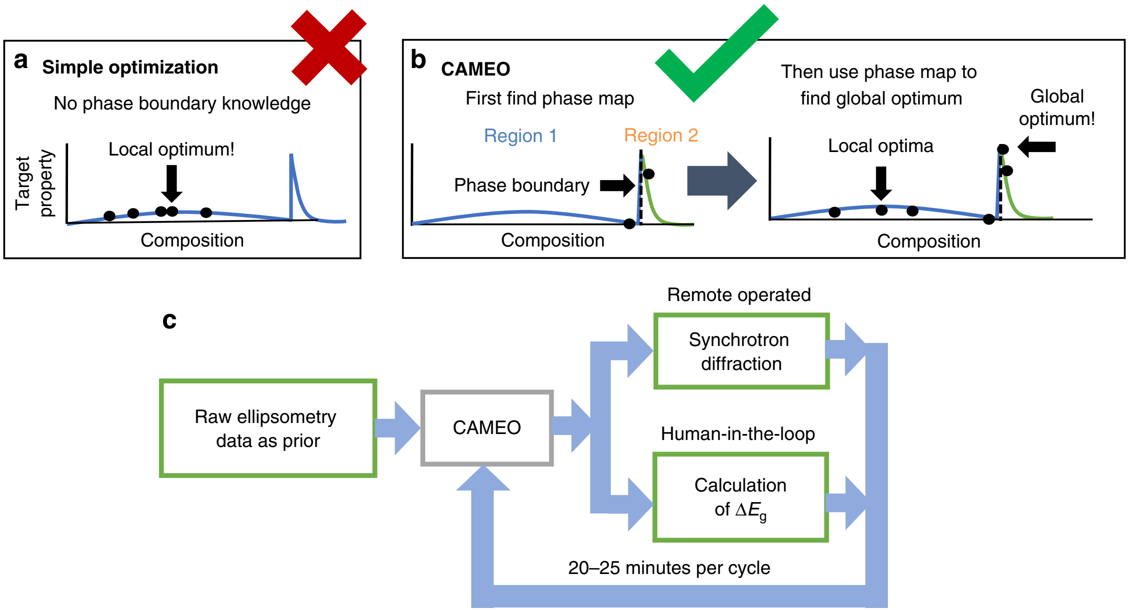

Fig. 2 Comparison of materials optimization schemes. a Simple optimization seeks to identify the property optimum with a mixture of exploration and exploitation without knowledge of the composition-structure-property (CSP) relationship. These methods are more likely to get caught in local optima. **b** The phase-map-informed optimization scheme exploits CSP relationship by recognizing that the property is dependent on phase, thus including phase mapping in the search for the optimum. (i) Phase-mapping steps and ii) materials optimization step that exploits knowledge of the phase boundaries. This allows a search for phase region dependent optima. **c** The Ge-Sb-Te CAMEO workflow began with incorporating raw ellipsometry data as a phase-mapping prior. On each iteration, CAMEO selects a material to measure for X-ray diffraction and concurrently requests an expert to calculate ΔE_g for that material. Each cycle takes 20–25 min.

For this task, scanning ellipsometry measurements were performed on the spread wafer with films in amorphous (initial) and crystalline states ahead of the CAMEO run, and the raw ellipsometric spectra data were incorporated as a phase-mapping prior. This was performed by increasing graph edge weights between samples of similar raw ellipsometry spectra during the phase mapping operation (see "Methods" section "M1c"). Thus, the algorithm makes use of information regarding phase distribution across the spread that is "hidden" in the unreduced complex spectroscopic data, which vary non-trivially across the spread. At each iteration, CAMEO identifies the next material to query, indicates the material to the experimentalist (human-in-the-loop) who performs the intensive task of processing the raw optical data to extract ΔE_g. In parallel, CAMEO remotely controls scanning of the synchrotron beam to collect X-ray diffraction data from the spread wafer with films in the crystalline state. CAMEO first seeks knowledge of the phase map until 80 % convergence, and then switches to material optimization (see "Methods" section "M1"). This procedure identified the material with the largest ΔE_g over 19 iterations taking approximately 10 h, compared to over 90 continuous hours for the full set of 177 composition spots. After data collection CAMEO was benchmarked against common active learning schemes with each run for 100 simulations. CAMEO provides an approximate maximum average 35-iteration lead over the best alternative Gaussian process- upper confidence bounds (GP-UCB) focusing on ΔE_g optimization in the composition space. The use of the ellipsometry prior to accelerate phase mapping provides a 25-iteration lead out of the 35. Furthermore, over the 100 runs, CAMEO gets within 1% of the optimal in the first 20 runs 31% of the time compared to GP-UBC's 10% (see Supplementary Fig. 10).

As seen in Fig. 3b, the optimal composition identified here lies at the boundary between the FCC–Ge–Sb–Te (GST) phase region and the region where there is co-existence of GST and Sb–Te phases. The average composition of the region is $Ge_4Sb_6Te_7$, and henceforth we refer to the region as GST467. Its ΔE_g is found to be 0.76 ± 0.03 eV, which is nearly 3 times that of GST225 (0.23 ± 0.03 eV). To investigate the origin of the enhanced ΔE_g of GST467 at the phase boundary, we have performed high-resolution transmission microscopy of this composition (Fig. 4a) which revealed a complex nanocomposite structure consisting of GST and Sb–Te phases. As seen in the figure, the phases have grown coherently with the relationship GST Fm-3m (111)//SbTe (001). (see "Methods" for details.)

This boost in ΔE_g indeed directly leads to large enhancement in optical contrast as captured in $\Delta k = k_c - k_a$, the difference in the extinction coefficient (between amorphous (k_a) and crystalline states (k_c)) extracted from the ellipsometry data at different wavelengths (Fig. 4c). Δk for GST467 is 60–125% larger than that of GST225 in the 1000–1500 nm wavelength range. The superior physical properties of GST467 shown here were reproduced on multiple composition spread wafers.

We have fabricated photonic switching PCM devices based on the discovered GST467 nanocomposite. With a sequence of laser pulses (energy and pulse width) with varying amplitude sent through the device, the material can be switched between the crystalline and amorphous phases (Fig. 4b). The device made of the nanocomposite GST467 thin film was found to be stable up to at least 30,000 cycles indicating the high reversibility of the crystallization and quenching processes of the coherent nanocomposite. The one-to-one comparison between the devices fabricated with our GST225 and GST467 films here (Fig. 4d and e) shows that GST467 device exhibits a much-enhanced switching contrast resulting in up to 50% more in the number of interval states, important for photonic memory and neuromorphic devices[32,33].

Recent reports of nanostructured PCM materials, including multilayer and superlattice thin films have highlighted the crucial roles interfaces and defects play in their switching mechanisms leading to faster switching speed and lower switching energies[34,35].

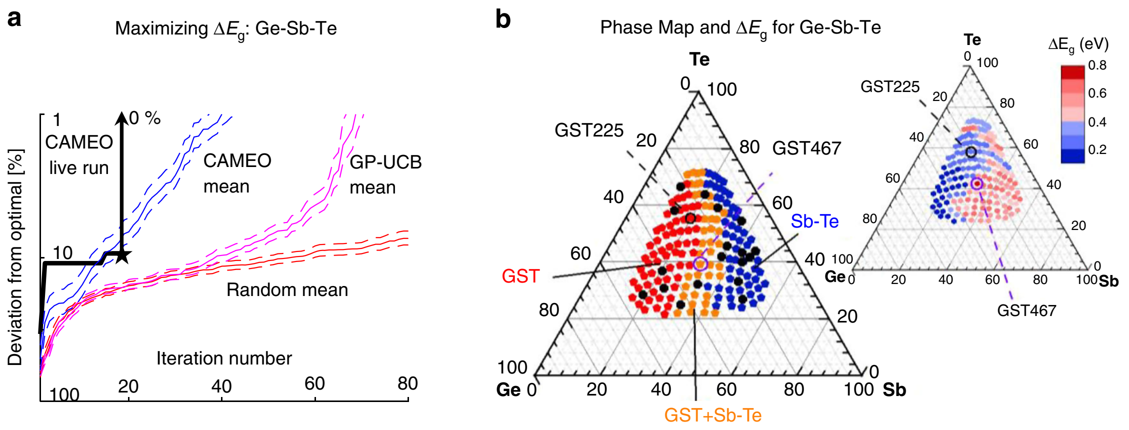

Fig. 3 Discovery of $Ge_4Sb_6Te_7$ (GST467). a Optimization of the Ge-Sb-Te system, with the objective of identifying the material with the largest ΔE_g, bandgap difference between amorphous and crystalline states. Performance is shown for: the CAMEO live run (black) with GST467 discovered on iteration 19 (black star); for the mean and 95% confidence interval of 100 CAMEO runs computed post data collection (blue), 100 runs of Gaussian process-upper confidence bounds (GP-UBC, magenta), and random sampling (red). For the 100 runs, CAMEO gets within 1% of the optimal in the first 20 runs 31% of the time compared to GP-UBC's 10%. **b** Left: structural phase map for the crystalline Ge–Sb–Te composition spread. Complete phase map constructed after all the diffraction measurements (beyond the live CAMEO run described in **a**)) is shown. Structural phase regions are color-coded as single-phase FCC-Ge-Sb-Te (GST) structure region (red, GST), single-phase Sb–Te region (blue), and region where GST and Sb–Te phases co-exist (orange). The materials measured during the live CAMEO run are indicated in black. Right: Complete mapping of ΔE_g following analysis of entire ellipsometry data (beyond the live CAMEO run). The discovered GST467 and GST225 are indicated in both maps.

Our finding of GST467 exhibiting significant boost in ΔE_g, and consequently larger optical contrast, underscores the effectiveness of naturally-forming nanocomposites as another approach to enhancing performance of PCM materials, especially for optical switching devices. It is the presence of epitaxial nano-pockets of the SbTe phase in GST467 which is locally modifying the resonant bonding in the GST matrix resulting in the lowered optical bandgap in the crystalline state, which in turn leads to the larger ΔE_g.

The discovery of a novel PCM material demonstrates that systems similar to CAMEO will fulfill the primary goals of materials design by accelerating the discovery and collection of materials knowledge, streamlining the experiment cycle, improving control over experimental variability, and improving reproducibility, thus improving trust in scientific results. They will also generate reference and benchmark datasets—automatically processed, analyzed, and converted to actionable knowledge with all associated metadata, for developing and improving trust in machine learning tools. Further benefits include automatic knowledge capture to maintain institutional knowledge, maximizing information gain, and reducing use of consumable resources, such as expert time, freeing up experts to work on higher level challenges. Research at the synchrotron exemplifies these resource demands and limitations, where obtaining scientist and equipment time is difficult or expensive. And potentially most impactful, placing labs under the control of AI may greatly reduce the technical expertise needed to perform experiments, resulting in a greater 'democratization' of science[36]. In turn, this may facilitate a more distributed approach to science, as advocated by the materials collaboratory concept[37].

Methods

M0 Outline of sections. A description of the closed-loop, autonomous system for materials exploration and optimization (CAMEO) scheme can be found in M1 beginning with a detailed description of results and a description of materials and device synthesis and characterization in section M2. The description of CAMEO is broken down into the subsections: M1a detailed results, M1b Initialization and data pre-processing, M1c Phase mapping, M1d Knowledge propagation, M1e Active learning, M1f Statistics and performance measures. The materials and device section is broken down into the subsections: M2a Sample fabrication, M2b Mapping of phase-change temperature, M2c Structural mapping, M2d High-angle annular dark-field scanning transmission electron microscopy (HAADF-STEM) of $Ge_4Sb_6Te_7$ (GST467), M2e Modeling and calculation of the ellipsometry spectra, M2f GST467 photonic device fabrication and measurement.

M1 CAMEO. CAMEO's methodology follows the diagrams of Figs. 1 and 2c (main text), where active learning drives measurement and expert input. Active-learning-driven synthesis and simulation are excluded for this work. The materials are pre-synthesized as a composition spread and, if desired, the AFLOW.org density functional theory (DFT) simulations are run prior to CAMEO's control of the X-ray diffraction measurement. For this work, DFT calculations were run to build a prior for the Fe–Ga–Pd phase mapping. DFT data were not used for the Ge–Sb–Te material system, instead ellipsometry data was used to build a phase mapping prior. The combinatorial library is physically loaded into the high-throughput X-ray diffraction system, and any data captured from external or internal databases is automatically imported into CAMEO. All preliminary data is analyzed to build the first estimated phase map along with uncertainty quantifications. This kicks off the iterative autonomous process where the phase map and material property estimate and estimate uncertainty are used to inform the active-learning-driven selection of the material to query next. At each iteration, CAMEO selects a material to study and requests and obtains structure and functional property data for the query material, with automatic X-ray diffraction pre-processing. In parallel, results and predictions are presented to the expert user and pertinent knowledge is captured from the expert. All gathered knowledge is then stored in a database. A description of the capabilities of CAMEO are presented in Supplementary Table 1.

CAMEO's specific implementation of Eq. 1 is shown in Eq. 2:

$$g(\mathbf{x}) = \begin{cases} \mathrm{P}(\mathbf{x}), & c < 80\% \\ \mathrm{F}(\mathbf{x}_r) = \mu(\mathbf{x}_r) + \beta\sigma(\mathbf{x}_r) + \gamma \mathrm{d}(bfx_r), & else \end{cases} \quad (2)$$

Recent use of active learning in materials science seeks to optimize functional material properties as a function of only the material synthesis parameters, e.g., composition. However, a material's properties are clearly not just dependent on its composition. Fundamental to functional properties are a material's lattice structure, microstructure, stress, etc. The general function g provides a framework for incorporating these different pieces of knowledge as well as an active learning scheme for acting on the knowledge. Here we take the simplest action scheme, switching between one phase of optimization (phase mapping) to another (materials optimization). However, g is a general function, flexible for other implementations, such as combining knowledge of composition, lattice structure, and functional property (as well as microstructure, and other information) in one integrated acquisition function. Future work will explore the pros and cons of these other implementations.

For this work, the first set of iterations maximize phase map knowledge until the estimated phase map convergences to the user defined threshold c, at which point the system switches to materials property optimization. A separate Gaussian Process is fit to each individual phase region for the functional property, allowing for phase region dependent hyperparameter optimization. This exploits the CSP relationship to improve functional property prediction accuracy, accelerate materials optimization, and provide potential computational resource savings. The

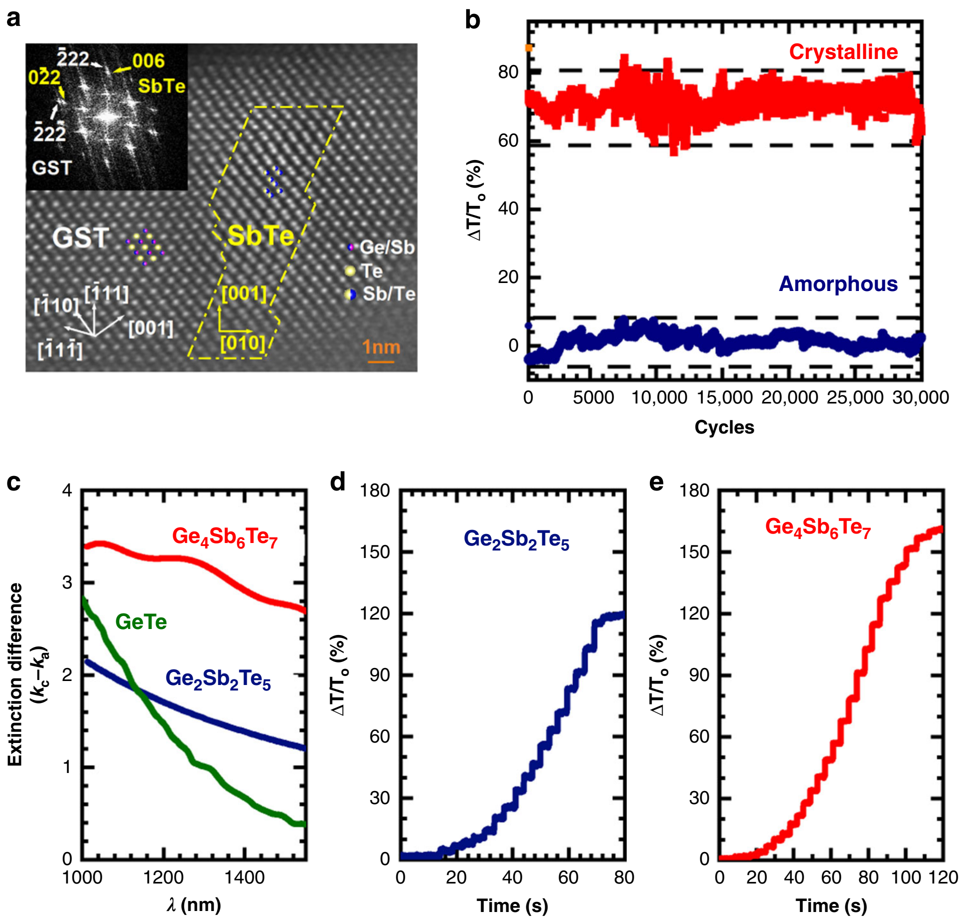

Fig. 4 Nanostructure and device performance of ($Ge_4Sb_6Te_7$) GST467. Newly identified phase-change memory material GST467 shows large optical contrast ideal for photonic-switching device applications such as neuromorphic computing. **a** High resolution transmission electron microscopy image reveals formation of coherent nanocomposite of GST structure matrix and SbTe. The dotted lines denote the atomically sharp interface. The FFT (inset) of this region indicates structural similarity of the adjacent phases; **b** endurance of the GST467: it is stable over 30,000 cycles indicating the robustness of the nanocomposite structure defined by local composition variation. The dotted lines indicate the range of each state in relative optical transmission $\Delta T/T_o$ at 1500 nm. Laser pulses were 50 ns with 183 pJ for quenching and 500 ns with 3.3 nJ for crystallization. The fluctuations in $\Delta T/T_o$ are due to the thermal fluctuation of the device measurement set-up; **c** comparison of the optical contrast here indicated by difference in the extinction coefficient k between crystalline and amorphous phases ($k_c - k_a$) for the wavelength range of 1000–1500 nm for various compositions within Ge-Sb-Te system. GST225 and GST467 data are from this work. The GeTe data are from ref. [55]. GST467 shows higher extinction difference over other known compositions; one-to-one comparison of **d** GST225 (left) and **e** GST467 (right) for multi-level switching in optical transmission at 1500 nm ($\Delta T/T_o$) using 500 ns, 6 pJ pulses: GST467 having larger optical contrast results in substantially more states than GST225. Device fabrication and characterization details are in Methods.

phase regions are then ranked by the maximum expected functional property value and the top R regions are selected for optimization, with R a user defined variable. Here R is set to 1. Optimization balances exploitation and exploration through the mean $\mu(\mathbf{x}_r)$ and weighted variance $\beta\sigma(\mathbf{x}_r)$ (the iteration dependent β follows ref. [38] and is described below). The optimization acquisition function also allows the user to target points closer or further from phase boundaries via $\gamma d(\mathbf{x}_r)$, where $d(\mathbf{x}_r)$ is the distance from point $\mathbf{x}_r$ to the nearest phase boundary and γ is a user-defined parameter—negative (positive) to emphasize points near the edge (center) of the phase region. Here the value is set to 10. Myopia to particular phase regions can be removed with an additional exploration policy.

Pre-synthesized (pseudo) ternary combinatorial spreads are used to provide a pool of hundreds of materials to investigate. While for this demonstration the autonomous system must select samples from the given pool of pre-synthesized samples, this is only a limit of the current physical experimental system and not a limit of the presented ML methodology.

M1a CAMEO detailed results. CAMEO was benchmarked on a material system previously extensively studied by the authors[39]. Efficacy was compared to a range of alternative methods as shown in Supplementary Fig. 1 with phase mapping performance measured with the Fowlkes-Mallows index (FMI) and Bayesian optimization performance measured by minimum percent deviation from optimal. The mean performance and 95 % confidence intervals over 100 iterations are plotted in Supplementary Fig. 1. The algorithm provides significant accuracy improvement and lower variance in phase mapping. Additionally, each level of increased physical knowledge further accelerates phase mapping. The benchmark optimization challenge was to maximize a functional property that is a simple function of composition with one broad, dominant peak in one phase region and a smaller peak with a maximum in another phase region. For this simple challenge, CAMEO provides improved performance compared to the next best optimization scheme—Gaussian process-upper confidence bounds (GP-UCB). For more information about the benchmarking process see (Active learning—materials optimization: benchmark system).

Once tuned, CAMEO was placed in active control over the high-throughput X-ray diffraction system at SLAC and a commercial in-house diffraction system. Here, the material optimization goal was to identify an optimal phase change material in the Ge–Sb–Te system, characterized by maximizing ΔE_g—the difference between the amorphous and crystalline optical bandgap. Scanning ellipsometry measurements were performed on the spread wafer in amorphous and crystalline states ahead of the CAMEO run, and we fed the unprocessed ellipsometric spectra as a prior for building the phase map model. At each iteration, the query material was indicated to the experimentalist (human-in-the-loop) who then performed the intensive task of processing the raw optical data to obtain ΔE_g and provided this data to CAMEO (see section "M2e" for full description). This procedure identified the material with the largest ΔE_g over 19 iterations taking approximately 10 h, compared to 90 h for the full set of 177 materials. A post-analysis is shown in Fig. 4 (main text), where 100 runs are performed comparing CAMEO to alternative methods. CAMEO provides an approximate maximum average 35-iteration lead over GP-UCB. More importantly, the algorithm is able to mine and make use of information regarding phase distribution across the spread hidden in the complex raw spectroscopic data.

M1b System initialization and data pre-processing

Physical system initialization. The system is initialized by loading the composition spread into the X-ray diffraction system, either the Bruker D8* or the SSRL diffraction synchrotron beamline endstation. For the SSRL system, a network connection is used for sending commands to the X-ray diffraction system via the SPEC interface[40]. Exposure time for each point measurement was 15 s.

Importing external data: ICSD and AFLOW.org. The user first indicates the material system of interest. A database of known stable phases, derived from past phase diagrams, is then used to automatically identify pertinent phases. Structure data is then automatically assembled for these phases from the Inorganic Crystal Structure Database (ICSD)—a database of critically evaluated experimental structures, and the AFLOW.org[41] density functional theory database. All retrieved structures are then used to generate simulated diffraction patterns through a call to Bruker's Topas[42]. After data is collected from the databases, the pool of material samples is updated to contain both the samples on the composition spread and those derived from databases. Previously it was shown that external structure data improved phase mapping performance in the case of exhaustive data collection[42]. For this work, the AFLOW.org computed ternary energy hull is imported and converted to region labels which are used as phase region (i.e. cluster) priors, see Supplementary Fig. 2 and Phase mapping prior.

Initialize phase mapping. Phase mapping is initialized with a user-selected expected number of phase regions for the material system, 5 for Fe–Ga–Pd and 10 for Ge–Sb–Te. While this number is used to initialize the phase map model, the phase mapping technique will converge to either a larger or smaller number of phase regions as described in the GRENDEL (graph-based endmember extraction and labeling) section. All other phase mapping hyperparameters were optimized on the benchmark system, and these values were used without modification for the Ge–Sb–Te system. Other default parameters include: graph distance multiplier is 1.2 and max number of iterations is 100.

Selection of first sample to seed processes. If prior material structure data is imported, such as data from AFLOW.org, that knowledge is used to initialize phase mapping (see Phase mapping prior), with the active learning criterion used to select the most informative material to query next. However, if no such prior data is used, the first sample queried can be selected randomly or using some other informative process. For benchmarking, the initial material was selected at random with uniform probability. For the live application to the Ge–Sb–Te system, the first sample was selected to be the one at the composition center of the materials on the composition spread. This sample was selected as it is potentially the most informative, given no other knowledge of the samples. The live run for the Ge–Sb–Te system completed after all the materials were measured, allowing for later analysis of active learning methodologies. To compare these methods, the initial material was again selected at random with a uniform prior.

Measurement and data pre-processing: collection, integration, background subtraction. Once the next query material has been identified, the system then measures the query material for X-ray diffraction using a programmatically generated script via SPEC for the SLAC high-throughput system or a GADDS script for the Bruker system. For the Bruker system, the diffraction image is integrated into a diffraction pattern automatically, and for the SLAC system, integration is performed as well[43]. The background signal is then automatically identified and subtracted.

The background signal from sample to sample can vary significantly, requiring a background subtraction method capable of handling these variations. For both the SLAC and Bruker diffraction measurements, Matlab's envelope function with the 'peak' setting and the parameter value of 50 was used to identify and remove the background curve.

M1c Phase mapping

Main method: GRENDEL—list of physical constraints. Phase mapping was performed using the physics-informed phase region and constituent phase identification method GRENDEL[44]. This method represents the material composition space as a graph, where each material is represented by a vertex that is connected by edges to neighboring materials in composition space (or wafer x–y coordinate). Neighboring materials are defined by Voronoi tessellating the composition space[44]. Mathematically, $G = \{V, E\}$, where V is the set of vertices, E is the set of edges with all edge weights set to 1. G is used to define a Markov Random Field (MRF)[45] where materials identified with the same vertex label belong to the same phase region, and each phase region is described by a set of constituent phases. This method encodes a list of physical constraints through the methods listed in Supplementary Table 2.

This method identifies a phase map for hundreds of samples in tens of seconds, on the same order of X-ray diffraction measurements at SSRL which typically takes 30 s, and measurements on the Bruker D8 which takes over 10 min.

GRENDEL hyperparameters include the MRF segmentation (i.e., graph cut) weight and the balance between the material-phase region matrix based on clustering and that based on phase mixture[44]. As the graph cut weight is increased, a greater number of clusters becomes possible, increasing the phase mapping performance using the measured described in the text, while also increasing cluster complexity. For the Fe–Ga–Pd a graph cut weight of $w_{\mathrm{gc}} = 100{*}n^3/5^3$ was found to output the desired number of clusters n. The full set of phase mapping parameters are described in the text.

During the GRENDEL process, if the number of clusters drops below 90 % of the number of clusters used when starting the process, GRENDEL is terminated and the computed phase map labels and constituent phases from the previous internal GRENDEL iteration are output.

Phase mapping prior. Material property data is incorporated into the MRF model as a prior through the edge weights of the composition graph G, where the original edge weights of G are modified by a functional property graph G_{p} with edge weights of 0 (disconnected) or 1 (connected) and $\mathrm{f} : E, E_{\mathrm{p}} \to E$. If $e \in E \cap E_{\mathrm{p}}$ then $e = 1 + \epsilon_{\mathrm{e}}$ else $e = 1 - \epsilon_{\mathrm{e}}$. The value of ϵ_{e} was varied for the benchmark material system and the value of $\epsilon_{\mathrm{e}} = 0.5$ selected as it demonstrates clear improved phase mapping performance during the first active learning selected measurements and worse performance near the end of the run. This is to be expected as prior knowledge can benefit initial analysis but can overwhelm knowledge gained from data if the prior is weighted too heavily. A smaller (larger) value of ϵ_{e} demonstrates a smaller (larger), similar effect.

For the benchmark system an AFLOW.org based phase map prior was used, as shown in Supplementary Fig. 2, where the AFLOW.org tie-lines are used to define regions. Points that fall in the same region are given the same label, resulting in a prior for phase mapping. For materials that share a graph edge and a clustering label, the edge weight in E_{p} is set to 1, otherwise the edge connecting them is removed from E_{p}.

For the Ge–Sb–Te material system, the prior was determined based on optical data collected. For each material, the complex reflectance ratio amplitude ψ and phase difference Δ for the amorphous and crystalline phases were collected for the set of angles $\theta = \{50°, 55°, 60°, 65°, 70°\}$ relative to the laser's plane of incidence, creating 20 spectral measurements for each material consisting of different measurement types $m \in \{\psi^{\mathrm{crystalline}}, \psi^{\mathrm{amorphous}}, \Delta^{\mathrm{crystalline}}, \Delta^{\mathrm{amorphous}}\}$. Example optical data used for the prior is shown in Supplementary Fig. 3.

To define a prior for the phase diagram, the set of all spectra are reduced into a set of similarity weights defining a similarity of 0 or 1 for each pair of samples, which can then be used to evaluate E_{p}. The following equations are used for mapping of spectra to similarity values. First the Euclidean difference between each set of materials (i, j) is computed for each spectral measurement type and angle $\{m, \theta\}$. These differences are then averaged for each pair of samples (i, j) over the set of angles θ and then normalized to between 0 and 1 for each measurement type m. These values are then averaged again over measurements m, resulting in a final dissimilarity value for each pair (i, j). A threshold is then used to convert the continuous dissimilarity values to 0 or 1, defining whether an edge between (i, j) exists (1) or does not (0). The threshold of $D_{\mathrm{Threshold}} = 0.07$ was selected to achieve a ratio of $|E_{\mathrm{p}}|/|E| = 0.49 \approx 0.5$, i.e., the prior removes approximately half the edges from the initial graph.

$$D_{i,j}^{m} = \operatorname*{mean}_{\theta}\left[d_{\mathrm{Euclidean}}(m_{i,\theta}, m_{j,\theta})\right] \tag{3}$$

$$D\prime_{i,j}^{m} = \left[D_{i,j}^{m} - \min D_{i,j}^{m}\right] / \left[\max D_{i,j}^{m} - \min D_{i,j}^{m}\right] \tag{4}$$

$$\bar{D}_{i,j} = \operatorname*{mean}_{m}\left[D_{i,j}^{\prime m}\right] \tag{5}$$

$$G_{\mathrm{p}} = \bar{D}_{i,j} < D_{\mathrm{Threshold}} \tag{6}$$

M1d Knowledge propagation

Phase mapping knowledge propagation. Once the phase map has been identified for the given data, the phase region labels must be propagated to the materials that have yet to be measured for structure. To exploit the graph-described data manifold, the semi-supervised learning technique Gaussian random field harmonic energy minimization[46] (HEM) is used. HEM computes the likelihood of each material belonging to each phase region and then assigns each material to the phase region with the greatest likelihood, thus defining the most likely phase map for the full set of materials on the composition spread. Using this information, alternative phase maps can also be identified along with their likelihoods. The edge weights E_{p} define the similarity matrix used to define the graph Laplacian.

Phase mapping knowledge propagation—comparison method: nearest neighbor (NN). The phase mapping knowledge propagation harmonic energy minimization method is compared to the use of 1-nearest neighbor, where any material without a phase region label takes on the label of its 1st nearest neighbor with a label. First nearest neighbor was implemented using MATLAB's knnsearch function with default parameters.

Functional property knowledge propagation: Gaussian process regression. GPR was implemented using MATLAB's 'fitrgp' function with default parameters.

M1e Active learning. In the Bayesian optimization[47] literature, the following formalism is used:

$$y = \mathrm{f}(\mathbf{x}) + \varepsilon \tag{7}$$

$$\mathbf{x}_* = \mathrm{argmax}_{\mathbf{x}}(\mathrm{f}(\mathbf{x})) \tag{8}$$

where y is the target property to be maximized, $\mathbf{x} \in \mathbb{R}^d$ is the set of material synthesis and processing parameters to be searched over, $\mathrm{f}(\mathbf{x})$ is the function to optimize, ε is typically independent stochastic measurement noise, and $\mathbf{x}_*$ defines the material synthesis and processing parameters that result in the maximal material property (for the minimum, replace argmax with argmin). When $\mathrm{f}(\mathbf{x})$ is unknown, a surrogate model is used to approximate it based on given data. The surrogate function is then used to identify the best next material to study. Each subsequent material is selected to identify the optimal material $\mathbf{x}_*$ in the smallest number of experiments possible. Identifying extrema of a function involves a balance between exploiting prior data to identify nearby extrema and exploring to identify extrema far from prior data. An alternative active learning objective is to select experiments that will best improve the overall prediction accuracy of the surrogate model, or in other words, select experiments to most efficiently learn the unknown function $\mathrm{f}(\mathbf{x})$. Such a campaign learns the general trends of $\mathrm{f}(\mathbf{x})$, which is highly useful when attempting to quantify anomalous behavior of novel materials.

Active learning for phase mapping: risk minimization. The active learning method used to select the next material to query for phase mapping is based on risk minimization[46]. HEM propagates phase region labels to unmeasured material and identifies the likelihood of each material belonging to each phase region. These likelihoods can be aggregated to define the set of potential phase diagrams and their associated likelihoods. The set of potential phase diagrams form a hypothesis space of phase diagrams. Risk minimization seeks to identify the optimal material to query next for its structure that will most rapidly whittle down the hypothesis set and most rapidly hone-in on the optimal phase map for the full set of materials, i.e., minimize expected total phase region label misclassification error and equivalently maximize knowledge of the phase map.

Active learning comparison methods: random, sequential, and 10% sampling. The risk minimization method is compared to (1) random sampling—selecting each subsequent material at random from the wafer, with a uniform prior, (2) sequential sampling—where each sample is selected in the order it appears on the wafer, and (3) where 10% of the materials are selected in a pre-determined design. Random sampling is expected to provide increasingly poor performance relative to active learning as the search space increases in dimension due to the curse of dimensionality[48]. The pre-determined 10% selection of materials in (3) are chosen to provide maximal coverage of the composition space. However, the use of 10% is not a generalizable benchmark. For a given number of data points, the density of data points decreases as the dimensionality of the composition space increases, with each point describing a larger volume. The optimal number of benchmark materials is thus dependent on the expected size of phase regions. If smaller phase regions are expected, a larger number of materials will be required to identify the phase regions.

The Fe–Ga–Pd composition spread contains 278 samples. For the 10% sampling, the 28 samples are indicated in Supplementary Fig. 4a. They were selected to provide uniform coverage of the composition space described by the spread. For the sequential sampling, the order of samples is shown in Supplementary Fig. 4b.

Active learning—materials optimization: Gaussian process upper confidence bounds. For CAMEO and GP-UCB the iteration dependent weight parameter β is used[38].

$$\beta = 2\log\left(DI^2\pi^2/6\lambda\right) \tag{9}$$

Where D is the total number of samples, I is the current iteration number, and $\lambda =$ 0.1.

Active learning—CAMEO: phase mapping convergence. The phase maps identified at each iteration i is compared to the iteration (i—4) using the FMI performance measure. Convergence is defined as FMI > = 80%.

Active learning—materials optimization: benchmark system. The target optimization for the benchmark system is maximizing remnant magnetization, as measured by scanning SQUID voltage. One modification was made to the remnant magnetization signal: The signal saturates over a large range of the composition spread. For BO benchmarking, it is preferred that one material is identified as the optimum. As such, the saturated values were modified with a squared exponential function, in effect "hallucinating" the remnant magnetization values as if sensor saturation had not occurred, converting the signal from Supplementary Fig. 5a to 5b. The squared exponential function used to modify the voltage was defined in cartesian space. For the ternary composition (α_{Fe}, b_{Ga}, c_{Pd}):

$$x = (b/100) + (c/100) * \sin(30^\circ), y = (c/100) * \sin(60^\circ) \tag{10}$$

$$N_{xy}\left(\mu = (0.19, 0.05), \sigma^2 = 0.001\right) \tag{11}$$

Mapping to ternary space gives $\mu = Fe_{78}Ga_{16}Pd_6$.

M1f Statistics and performance metrics

Confidence interval. The 95 % confidence interval was computed for the variable of interest over 100 experiments at the given iteration with:

$$CI_{95} = \left(\frac{\sigma}{\sqrt{n}}\right)F^{-1}(p, \nu) \tag{12}$$

Where F^{-1} is the inverse of the Student's t cumulative distribution function, σ is the standard deviation, $n = 100$ is the number of experiments, $p = \{2.5\%, 97.5\%\}$, and $\nu = 99$ is the degrees of freedom.

Phase mapping. Phase mapping performance is evaluated by comparing phase region labels determined by experts with those estimated by CAMEO for the entire phase map (after the knowledge propagation step). To evaluate system performance, the Fowlkes-Mallows Index (FMI) is used, which compares two sets of cluster labels. The equations are presented below for the expert labels $l \in L$ and the ML estimated labels $\hat{l} \in \hat{L}$, where the labels are enumerated $L \to \mathbb{N}$ and $\hat{L} \to \mathbb{N}$.

If the number of phase regions is taken to be too large by either the user or the ML algorithm while the phase mapping is correct, some phase regions will be segmented into sub-regions with the dominant phase boundaries preserved. For example, peak shifting can induce phase region segmentation[42]. To ensure that the performance measures ignore such sub-region segmentation, each estimated phase region is assigned to the expert labeled phase region that shares the greatest number of samples. The number of phase regions is monitored to ensure that increases in model accuracy are not driven by increases in model complexity.

$$\text{Fowlkes} - \text{Mallows Index} : FMI = TP/\sqrt{(TP + FP)(TP + FN)} \tag{13}$$

$$TP = \frac{1}{2}\sum\nolimits_i \sum\nolimits_j (l_i = l_j \,\&\, \hat{l}_i = \hat{l}_j) \tag{14}$$

$$FP = \frac{1}{2}\sum\nolimits_i \sum\nolimits_j (l_i \neq l_j \,\&\, \hat{l}_i = \hat{l}_j) \tag{15}$$

$$FN = \frac{1}{2}\sum\nolimits_i \sum\nolimits_j (l_i = l_j \,\&\, \hat{l}_i \neq \hat{l}_j) \tag{16}$$

$$TN = \frac{1}{2}\sum\nolimits_i \sum\nolimits_j (l_i \neq l_j \,\&\, \hat{l}_i \neq \hat{l}_j) \tag{17}$$

M1g Live visualizations. CAMEO provides live visualizations to support algorithm interpretability. Phase mapping is supported by Supplementary Fig. 11a, b, where each filled circle indicates a sample in the combinatorial library. Samples labeled with the same color are identified as likely belonging to the same phase region. Samples with a black border have been measured for X-ray diffraction and the sample indicated by the red diamond is the sample that CAMEO will query next. The size of each sample marker indicates the probability of the sample belonging to the color-indicated phase region. Thus, areas of the composition diagram where filled circles are small indicate a high likelihood region for a phase boundary. Supplementary Fig. 11b presents the CAMEO predicted values of the target functional property. Here each predicted phase region is analyzed using its own Gaussian process (see "Methods" Section "M1"). Supplementary Fig. 11c shows the associated Gaussian process predicted variance which has been mapped using the sigmoid function to values between 0 and 1.

Supplementary Fig. 11d presents the FMI score between each subsequent phase mapping iteration, quantifying the percent change in phase mapping from iteration to iteration. Supplementary Fig. 11e shows convergence in materials property optimization computed as the difference in the max identified property in consecutive iterations.

Bayesian optimization. Bayesian optimization performance is measured with minimum percent deviation from optimal, related to simple regret.

$$\text{Minimum percent deviation from optimal} = 100\% * \min_i \frac{(p_{\text{target}} - p_i)}{p_{\text{target}}} \tag{18}$$

$$\text{Simple Regret} = \min_i (p_{\text{target}} - p_i) \tag{19}$$

M2 Materials synthesis and characterization

M2a Sample fabrication. Amorphous thin-film composition spreads encompassing a region of the Ge–Sb–Te ternary (separated into 177 samples using a gridded physical shadow mask) were fabricated on 3-inch silicon wafers with SiO_2 layers (2 μm) by co-sputtering Ge, Sb, and Te targets at room temperature. Different (average) thickness composition spreads (covering the same composition range) were fabricated for different measurements: they were 20 nm, 100 nm, 200 nm, and 500 nm for optical, structural, resistance, and composition mapping, respectively.

To obtain a crystalline state, some of the wafers were annealed at 300 °C for 10 min following their characterization in the amorphous state.

The composition mapping of the spreads is measured using the wavelength dispersion spectroscopy. For every separated sample region on a spread, three random spots are measured, and the average composition value is used for the actual stoichiometry mapping in Supplementary Fig. 6.

M2b Mapping of phase-change temperature. Upon increasing the temperature, a phase-change memory material undergoes a structural transition from amorphous to crystalline states with up to four orders of magnitude in the change of resistance. The temperature at which the resistance drop takes place can be taken as the phase-change temperature, T_{cp} (Supplementary Fig. 6). The entire spreads were measured in a scanning four-probe station combined with a Keithley 2400 from room temperature up to 300 °C. T_{cp} of GST467 was found to be ~200 °C, which is much higher than that of GST225 (≈140 °C). The higher T_{cp} of GST467 indicates higher stability of the amorphous state of GST467 compared to GST225.

M2c Structural mapping. Synchrotron diffraction on crystallized spreads was carried out at Beamline 10-2 at SLAC. In addition to the remote-controlled CAMEO run, we have also carried out diffraction of entire spreads in order to obtain the complete structural phase mapping of the probed Ge–Sb–Te region and to verify the accuracy of the phase diagram determined by CAMEO. Supplementary Fig. 7 shows an example set of diffraction patterns taken across the spread. Along the marked line in the composition map, the evolution of diffraction patterns indicates phases going from the distorted FCC–Ge–Sb–Te (GST) structure region to the phase co-existence region (GST and Sb–Te) to the Sb–Te region.

SbTe ($R\bar{3}m$), Sb_2Te_2 ($P\bar{3}m$), and Sb_2Te_3 ($R\bar{3}m$) all have very similar diffraction patterns and atomic projections of the [100] zone-axis, except for different lattice periods along the [001] direction. These three phases are present across the Sb–Te region depending on the local composition on our spread. The predominant Sb–Te phase in GST467 is SbTe (below).

M2d HAADF-STEM of GST467. We have performed cross-sectional high-angle annular dark-field scanning transmission electron microscopy (HAADF-STEM) measurements on the GST467 thin film and found that there are nanometer-sized SbTe regions grown coherently inside the distorted cubic GST matrix as shown in Fig. 4a of the main text. To distinguish between similar phases (SbTe, Sb_2Te_2, and Sb_2Te_3), analysis of electron diffraction rings was carried out (not shown here), and Sb–Te phase in the GST467 was identified to be Sb_1Te_1.

M2e Modeling and calculation of the ellipsometry spectra. The experimental ellipsometry data (J.A. Woollam company) of the combinatorial Gs–Sb–Te spread was analyzed in the range from 200 to 1000 nm using the CompleteEASE software. The dielectric function ε(ω) used in the model contains[49] (1) a constant, (2) a Drude-type contribution for free carriers in the case of crystalline state, and (3) a Tauc-Lorentz oscillator to describe the onset of optical transition:

$$\text{amorphous state} : \varepsilon(\omega) = \varepsilon_{const} + \varepsilon_{Tauc-Lorentz}(\omega), \tag{20}$$

$$\text{crystalline state} : \varepsilon(\omega) = \varepsilon_{const} + \varepsilon_{Drude}(\omega) + \varepsilon_{Tauc-Lorentz}(\omega) \tag{21}$$

For the Drude model:

$$\varepsilon_{\mathrm{Drude}}(\omega) = \varepsilon_1(\omega) + i \cdot \varepsilon_2(\omega) = 1(\varepsilon(\infty)) - \frac{\omega_{\mathrm{p}}^2}{\omega^2 + \Gamma^2} + i \cdot \frac{\omega_{\mathrm{p}}^2 \cdot \Gamma}{\omega \cdot (\omega^2 + \Gamma^2)} \tag{22}$$

where $\omega_{\mathrm{p}} = \sqrt{\frac{N \cdot e^2}{m \cdot \varepsilon_0}}$, and ω_{p} is the plasma frequency, Γ is the collision frequency.

For the Tauc-Lorentz model:

$$\begin{aligned} \varepsilon_{\mathrm{Tauc-Lorentz}}(\omega) &= \varepsilon_1(\omega) + i \cdot \varepsilon_2(\omega) \\ &= \varepsilon_1(\infty) + \frac{2}{\pi} P \int_{E_g}^{\infty} \frac{\xi \varepsilon_2(E)}{\xi^2 - E^2} d\xi + i \cdot \frac{A E_0 \Gamma \left(E - E_{\mathrm{g}}\right)^2}{E\left[\left(E^2 - E_{\mathrm{o}}^2\right)^2 + \Gamma^2 E^2\right]} \Theta\left(E - E_{\mathrm{g}}\right) \end{aligned} \tag{23}$$

Where A is the prefactor, E_{o} is the peak in the joint density of the state, E_{g} is the optical bandgap, and Γ is the broadening parameter.

The optical parameters, e.g., refractive index n and extinction coefficient k, as well as the optical bandgap, can be extracted from these models. In order to confirm the accuracy of the optical parameters, one needs to check if the fitting curves as well as a set of fitting parameters, e.g. thickness, carrier density, and surface roughness, can be used to analyze the experimental data of the samples for the entire spread wafer. Typically, a fitting procedure requires repeated steps in order to fine-tune the parameters manually to optimize the results, and some samples require more manual fitting steps for setting the range and the starting values of the parameters than others.

For the CAMEO run, the unprocessed raw ellipsometry data taken at each composition spot (for crystalline and amorphous states) are used as the prior (see section "M1c, Phase mapping prior"). Once a spot is identified as a possible composition with enhanced ΔE_{g} (the difference in the optical bandgap between the amorphous and crystalline state), the fitting procedure above is carried out on the raw data, and the value of ΔE_{g} is computed, the process of which includes manual inspection of fitting parameters. Depending on the number of repeated steps, each computation can take up to 20 min. at a composition spot, and the ΔE_{g} value is then fed back to the CAMEO algorithm.

The complete mapping of the optical bandgap of amorphous and crystalline states measured and calculated from one spread are shown in Supplementary Fig. 8. In the amorphous state, the Ge–Sb–Te based compounds are effectively covalently-bonded semiconductors with large optical bandgaps[50,51]. With changing composition, there is variation in bonding leading to slight shift in the optical bandgap shown in Supplementary Fig. 8. In the crystalline state, the resonantly-bonded p orbitals can delocalize the carriers resulting in the reduced bandgap[49,52–54], leading to the large contrast between the amorphous and crystalline states. In the distorted cubic phase (i.e., GST), with changing composition, the local distortion (i.e., Peierls distortion) due to vacancies[52,55,56] would modify the resonant bonding shifting the optical bandgap. In the Sb–Te phase, the optical bandgap also varies with the changing composition in Supplementary Fig. 8. When the epitaxial nanocomposite with the SbTe phase are coherently and homogeneously grown in the GST matrix as shown in Fig. 4a, the nano SbTe phase can act as the impurity dopant phase in the GST matrix.

M2f $Ge_4Sb_6Te_7$ photonic device fabrication and measurement. Photonic switching devices were fabricated out of GST467 films (Supplementary Fig. 9). The 30 nm thick nanocomposite GST467 thin film was sputtered on a 330 nm thick Si_3N_4 layer on an oxidized silicon wafer. A 10 nm thick SiO_2 protection layer was then coated on the top of the GST467 thin film. Using e-beam lithography and inductively coupled plasma etching, a 1.2 µm wide photonic waveguide was fabricated. Then the GST467 thin film was patterned into disk shaped features 500 nm in diameter on the top of the waveguide, and they were encapsulated with a 200 nm thick Al_2O_3 layer as shown in the inset of Supplementary Fig. 9.

A symmetric multi-level switching of the photonic device was investigated as shown in Supplementary Fig. 9. In order to provide and precisely control the pump pulses to quench or anneal the GST467 thin film in steps, pulses from a CW pump laser were first modulated by an electro-optic modulator and then sent into an erbium-doped fiber amplifier followed by a variable optical attenuator. The output of the optical signal was collected with a photodetector. During the annealing process, a sequence of pump pulse (50 ns, 2 mW) train was applied to the photonic device. In the quenching process, a sequence of 50 ns pump pulses with gradually increased amplitude was sent into the waveguide.

Data availability

Data that supports the findings of this study have been deposited in the github repository and can be found with the following github or DOI link: https://github.com/KusneNIST/CAMEO_NComm. https://doi.org/10.5281/zenodo.3998287.

Code availability

The code can be found at the following github repository or using the following DOI link: https://github.com/KusneNIST/CAMEO_NComm. https://doi.org/10.5281/zenodo.3998287.

Received: 4 June 2020; Accepted: 13 October 2020;
Published online: 24 November 2020

References

1. National Science and Technology Council (US). "Materials Genome Initiative for Global Competitiveness." Executive Office of the President, National Science and Technology Council (2011).
2. Toher, C. et al. Unavoidable disorder and entropy in multi-component systems. *Npj Comput. Mater.* **5**, 1–3 (2019).
3. Hattrick-Simpers, J. R. et al. Perspective: Composition–structure–property mapping in high-throughput experiments: turning data into knowledge. *APL Mater* **4**, 053211 (2016).
4. Shapiro, E. et al. "Toward 2020 Science" Microsoft, (2006).
5. Licklider, J. C. R. Man-computer symbiosis. *IRE Tran. Hum. Facto. Elect.* **1**, 4–11 (1960).
6. Shneiderman, B. Human-centered artificial intelligence: Reliable, safe & trustworthy. *Int. J. Hum. Comput. Inter.* **36**, 495–504 (2020).
7. Settles, B. Active learning literature survey. *Univ. Wis. Madison.* **52**, 11 (2010).
8. Jaynes, E. T. Bayesian methods: general background. (1986).
9. Xue, D. et al. Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.* **7**, 11241 (2016).
10. Yuan, R. et al. Accelerated discovery of large electrostrains in $BaTiO_3$-based piezoelectrics using active learning. *Adv. Mater.* **30**, 1702884 (2018).
11. Balachandran, P. V. et al. Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. *Nat. Commun.* **9**, 1668 (2018).

12. Meredig, B. et al. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. *Mol. Syst. Des. Eng.* **3**, 819–825 (2018).
13. Zhou, Z. et al. Optimization of molecules via deep reinforcement learning. *Sci. Rep.* **9**, 1–10 (2019).
14. Korovina, K. et al. Chembo: Bayesian optimization of small organic molecules with synthesizable recommendations. International Conference on Artificial Intelligence and Statistics. PMLR, 3393-3403 (2020).
15. Li, Z. et al. Robot-accelerated perovskite investigation and discovery. *Chem. Mater.* **32**, 5650 (2020).
16. Pendleton, I. M. et al. Experiment specification, capture and laboratory automation technology (ESCALATE): a software pipeline for automated chemical experimentation and data management. *MRS Commun.* **9**, 846–859 (2019).
17. Zhong, M. et al. Accelerated discovery of CO_2 electrocatalysts using active machine learning. *Nature* **581**, 178–183 (2020).
18. Nikolaev, P. et al. Autonomy in materials research: a case study in carbon nanotube growth. *Npj Comput. Mater.* **2**, 16031 (2016).
19. Roch, L. M. et al. ChemOS: orchestrating autonomous experimentation. *Sci. Robot.* **3**, eaat5559 (2018).
20. Attia, P. M. et al. Closed-loop optimization of fast-charging protocols for batteries with machine learning. *Nature* **578**, 397–402 (2020).
21. Noack, M. M. et al. A kriging-based approach to autonomous experimentation with applications to x-ray scattering. *Sci. Rep.* **9**, 1–19 (2019).
22. Burger, B. et al. A mobile robotic chemist. *Nature* **583**, 237–241 (2020).
23. MacLeod, B. P. et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **6**, eaaz8867 (2020).
24. Langner, S. et al. Beyond ternary OPV: high-throughput experimentation and self-driving laboratories optimize multicomponent systems. *Adv. Mater.* **32**, 1907801 (2020).
25. Reyes, K. G. et al. The machine learning revolution in materials? *MRS Bulletin* **44**, 530–537 (2019).
26. Stein, H. S. et al. Progress and prospects for accelerating materials science with automated and autonomous workflows. *Chem. Sci.* **10**, 9640–9649 (2019).
27. Tabor, D. P. et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat Rev. Mater.* **3**, 5–20 (2018).
28. Jensen, K. F. et al. Autonomous discovery in the chemical sciences part I: Progress (Angewandte Chemie International Edition, 2019).
29. Geiger, A. C. et al. Autonomous science: big data tools for small data problems in chemistry. *Mach. Learn. Chem.* **17**, 450 (2020).
30. Kolobov, A. V. et al. *Springer Handbook of Electronic and Photonic Materials*, (eds Kasap, S. & Capper, P.) (Springer International Publishing, Cham, 2017).
31. Burr, G. W. et al. Neuromorphic computing using non-volatile memory. *Adv. Phys. X* **2**, 89–124 (2016).
32. Li, X. et al. Fast and reliable storage using a 5 bit, nonvolatile photonic memory cell. *Optica.* **6**, 1–6 (2019).
33. Giannopoulos, I. et al. 8-bit precision in-memory multiplication with projected phase-change memory. *IEEE International Electron Devices Meeting (IEDM),* 27.7.1–27.7.4 (2018)
34. Simpson, R. E. et al. Interfacial phase-change memory. *Nat. Nanotechnol.* **6**, 501–505 (2011).
35. Ding, K. et al. Phase-change heterostructure enables ultralow noise and drift for memory operation. *Science* **366**, 210–215 (2019).
36. Kusne, A. G., "Materials Genome Initiative & Artificial Intelligence at NIST", DOE AMO Workshop on Artificial Intelligence Applied to Materials Discovery and Design, Pittsburgh, PA (2017).
37. Green, M. et al. High-throughput experimental materials collaboratory, NIST materials genome initiative. https://mgi.nist.gov/htemc (2020)
38. Srinivas, N. et al. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Trans. Inf. Theory.* **58**, 3250–3265 (2012).
39. Long, C. et al. Rapid structural mapping of ternary metallic alloy systems using the combinatorial approach and cluster analysis. *Rev. Sci. Instrum.* **78**, 072217 (2007).
40. spec Software for Diffraction, Certified Scientific Software. https://certif.com/spec.html. (2017)
41. Curtarolo, S. et al. AFLOWLIB.ORG: a distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **58**, 227–235 (2012).
42. Kusne, A. G. et al. On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Sci. Rep.* **4**, 6367 (2014).
43. Ren, F. et al. On-the-fly data assessment for high-throughput X-ray diffraction measurements. *ACS Combi. Sci.* **19**, 377–385 (2017).
44. Kusne, A. G. et al. High-throughput determination of structural phase diagram and constituent phases using GRENDEL. *Nanotechnology.* **26**, 444002 (2015).
45. Dobruschin, P. L. The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory Probab. Its Appl* **13**, 197–224 (1968).
46. Zhu, X. et al. Semi-supervised learning using gaussian fields and harmonic functions. *Proceedings of the 20th International conference on Machine learning (ICML-03)* (2003).
47. Brochu, E. et al. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. arXiv1012.2599. Preprint at https://arxiv.org/abs/1012.2599 (2010).
48. Hastie, T. et al. The Elements of Statistical Learning 2nd Ed. (Springer, 2009).
49. Shportko, K. et al. Resonant bonding in crystalline phase-change materials. *Nat. Mater.* **7**, 653–658 (2008).
50. Rao, F. et al. Reducing the stochasticity of crystal nucleation to enable subnanosecond memory writing. *Science.* **358**, 1423–1427 (2017).
51. Zhu, M. et al. One order of magnitude faster phase change at reduced power in Ti-Sb-Te. *Nat. Commun.* **5**, 1–6 (2014).
52. Deringer, V. L. et al. Bonding nature of local structural motifs in amorphous GeTe. Angew. *Chem. Int. Ed* **53**, 10817–10820 (2014).
53. Wuttig, M. et al. Phase-change materials for rewriteable data storage. *Nat. Mater.* **6**, 824–832 (2007).
54. Lencer, D. et al. A map for phase-change materials. *Nat. Mater.* **7**, 972–977 (2008).
55. Wuttig, M. et al. Phase-change materials for non-volatile photonic applications. *Nat. Photonics.* **11**, 465 (2017).
56. Pries, J. et al. Phase-change materials: Empowered by an unconventional bonding mechanism. *MRS Bull* **44**, 699–704 (2019).

Acknowledgements

We acknowledge Xiaohang Zhang for assistance with thin film characterization and Fang Ren and Doug Van Campen for assistance in diffraction experiments at SLAC. We also acknowledge Valentin Stanev, Gaurav Modi, and Daniel Samarov for discussions. This work was supported by ONR MURI N00014-17-1-2661, ONR N00014-13-1-0635, and ONR N00014-15-2-222. The work at the University of Maryland was partially supported by the Center for Innovative Materials for Accelerated Compute Technologies (IMPACT), one of the centers in nCORE, a Semiconductor Research Corporation (SRC) program. Diffraction performed at SLAC is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515. H.Z. and L.A.B. acknowledge support from the US Department of Commerce, NIST under financial assistance award 70NANB17H249. A.V.D. acknowledges support from Materials Genome Initiative funding allocated to NIST. NIST disclaimer: Certain commercial equipment, instruments, or materials are identified in this report in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Author contributions

A.G.K. and I.T. initiated and supervised the research. A.G.K. developed and implemented the CAMEO algorithm with guidance from I.T., B.D. helped refine the phase mapping prior algorithm. H.Y. fabricated the Ge–Sb–Te composition spreads and thin films and coordinated their characterization. A.M. and S.S. developed and carried out the high-throughput synchrotron diffraction. C.W. and M.L. characterized the optical properties of Ge–Sb–Te composition spreads and fabricated and performed measurements on photonic devices. H.Z. conducted the transmission electron microscopy (TEM), and A.D. and L.B. helped analyze the TEM results. J.H.S. performed some of the ellipsometry measurements. R.A. discussed the bandgap behavior of GST. C.O., C.T. and S.C. developed the AFLOW framework which was used for some of the phase diagram analysis. A.G.K., I.T., and H.Y. wrote the paper with substantial input from other authors. All authors contributed to the discussion of the results.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-19597-w.

Correspondence and requests for materials should be addressed to A.G.K. or I.T.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.