Mixed-Precision 'Memcomputing'

Manuel Le Gallo,[1,2,a)] Abu Sebastian,[1,b)] Roland Mathis,[1] Matteo Manica,[1,2] Tomas Tuma,[1] Costas Bekas,[1] Alessandro Curioni,[1] and Evangelos Eleftheriou[1]
[1)] *IBM Research - Zurich, 8803 Rüschlikon, Switzerland*
[2)] *ETH Zurich, 8092 Zurich, Switzerland*

(Dated: 18 May 2022)

To process the ever-increasing amounts of data, computing technology has relied upon the laws of Dennard[1] and Moore[2] to scale up the performance of conventional von Neumann machines. As these laws break down due to technological limits, a radical departure from the processor-memory dichotomy is needed to circumvent the limitations of today's computers. 'Memcomputing' is a promising concept in which the physical attributes and state dynamics of nanoscale resistive memory devices are exploited to perform computational tasks with collocated memory and processing.[3,4] The capability of 'memcomputing' for performing certain logical[5–7] and arithmetic[8–11] operations has been demonstrated. However, device variability and non-ideal device characteristics pose technical challenges to reach the numerical accuracy usually required in practice for data analytics and scientific computing. To resolve this, we propose the concept of mixed-precision 'memcomputing' that combines a von Neumann machine with a 'memcomputer' in a hybrid system that benefits from both the high precision of conventional computing and the energy/areal efficacy of memcomputing. Such a system can achieve arbitrarily high computational accuracy with the bulk of the computation realized as low-precision 'memcomputing'. We demonstrate this by addressing the problem of solving systems of linear equations and present experimental results of solving accurately a system of $10,000$ equations using $959,376$ phase-change memory devices. We also demonstrate a practical application of computing the gene interaction network from RNA expression measurements. These results illustrate that an interconnection of high-precision arithmetic and 'memcomputing' can be used to solve problems at the core of today's computing applications.

Nanoscale resistive memory devices, also referred to as memristive devices, can store information in their conductance states and can remember the history of the current that has flowed through them[12–14]. They form the basis of 'memcomputing': a promising computing paradigm where both information processing and storing the computational data are performed on the same physical devices.[3] Various physical mechanisms such as Ohm's law and Kirchhoff's circuit laws[11], chemically driven phase transformations[9], the rich pattern dynamics exhibited by ferroelectric domain switching[15] or the physics of crystallization[6] and melting[7] in phase-change materials can be used to perform a range of arithmetic and logical operations. Massively parallel, memory-centric hardware accelerators based on this concept are now becoming a prominent subject of research with applications ranging from image processing to health-care[16–18]. However, building a 'memcomputer' that can solve practical problems in a reliable and accurate way is challenging. Memristive devices suffer from significant inter-device variability and inhomogeneity across an array[19]. Moreover, there is intra-device variability and randomness intrinsic to the way these devices operate[20,21]. While this randomness could be exploited for certain types of computational tasks[22,23], for the majority of practical applications the lack of precision associated with 'memcomputing' is prohibitive.

In this letter, we introduce the concept of mixed-precision 'memcomputing' to address this problem. The concept is motivated by the observation that many computational tasks can be formulated as a sequence of two distinct parts. In the first part, an approximate solution is obtained. In the second part, the resulting error in the overall objective is calculated accurately. Then, based on this, the approximate solution is adapted (by repeating the first part). The first part typically has a high computational load whereas the second part has a light computational load. By repeating this sequence several times, it is often possible to arrive at a solution with arbitrarily high accuracy.[24] In a mixed-precision 'memcomputing' system, the idea is to use a low-precision 'memcomputing' unit to obtain the approximate solution of the first part and a high-precision processing unit to realize the second part (Fig. 1a). The expectation is that in this way we will be able to retain an overall high areal and energy efficiency because the bulk of the computation is still realized in a non-von Neumann manner, while at the same time not sacrificing the desired computational accuracy.

To illustrate this concept, we present the problem of solving systems of linear equations. The problem is to find an unknown vector $x \in \mathbb{R}^N$ that satisfies the constraint

$$Ax = b, \text{ where } A \in \mathbb{R}^{N\times N} \text{ and } b \in \mathbb{R}^N. \tag{1}$$

Here A is a non-singular matrix and b is a known column vector of N observations or measurements. This problem can be solved in the mixed-precision 'memcomputing' framework as shown in Fig. 1b. In a so-called iterative refinement algorithm, an initial solution is chosen as the starting point, and is iteratively updated with a low-precision error-correction term, z. The error-correction term is computed by solving $Az = r$ with an inexact inner solver using the residual $r = b - Ax$, calculated with high

[a)] Electronic mail: anu@zurich.ibm.com
[b)] Electronic mail: ase@zurich.ibm.com

precision.[25] The algorithm runs until the norm of the residual falls below a desired tolerance, tol. For the inner solver, we use an iterative Krylov subspace method, such as the Conjugate Gradient (CG) method or the Generalized Minimum Residual (GMRES) method[26]. These techniques rely on building a basis $\{v^k\}_{k=1}^m$ of the Krylov subspace $\mathscr{K}_m(A,r) = \text{span}\{r, Ar, A^2r, ..., A^{m-1}r\}$. This basis is obtained by performing multiple matrix-vector multiplications $w^k = Av^k$ with the matrix A, w^k being used to compute the next basis vector v^{k+1} following an orthogonalization procedure. From this basis, the error correction term, which is an approximation of $A^{-1}r$, can be obtained. In all Krylov subspace methods, the most computationally intensive operation is the matrix-vector multiplication $w^k = Av^k$. Hence, the key idea is to realize this operation in the 'memcomputing' unit, using a memristive crossbar array in which the matrix A is programmed as conductance values of the memristive devices (Fig. 1(b)). This mode of computing is highly efficient because the matrix-vector product is computed *in situ* within the memristive array, thereby eliminating any intermediate movement of data.[11] Even if the computation realized this way is approximate, the iterative refinement algorithm ensures convergence to a high-accuracy solution even when strong perturbations are introduced in the inner solver.[25] The magnitude of the perturbations that can be tolerated is expected to decrease with increasing condition number of the matrix A (the condition number associated with (1) reflects how much the solution x will change with respect to a change in b).[27]

For our experiments, we implemented the low-precision matrix-vector multiplication using an array of one million phase-change memory (PCM) devices. PCM devices are resistive memory devices that can be programmed to achieve a desired conductance value by altering the amorphous/crystalline phase configuration within the device (Fig. 2a).[28] The array consists of a matrix of 512 word lines $\times$ 2048 bit lines integrated in 90-nm CMOS technology and connected in a crossbar. Each crosspoint consists of a PCM device in series with an access transistor (Supplementary Note I).

First, we investigate the scalar multiplication operation that forms the core of the matrix-vector multiplication performed with the PCM devices. Let $\theta_n = \beta_n \cdot \gamma_n$, where β_n and γ_n are numbers generated uniformly in $[0,1]$. β_n was mapped to an effective conductance value G_n (I/V ratio at $V = 0.2\,\text{V}$) between approximately 0 and $50\,\mu\text{S}$, and γ_n to a voltage V_n between approximately 0.1 V and 0.3 V (see Supplementary Note II). Because the current is a slightly non-linear function of the voltage in our PCM devices, the analogue multiplication was assumed to follow a "pseudo" Ohm's law:

$$I_n \simeq \alpha G_n f(V_n). \tag{2}$$

In this equation, α is an adjustable parameter and f a polynomial function that approximates the current-voltage characteristics of the PCM devices (Supplementary Note II). The devices were programmed to the effective conductance G_n using an iterative program-and-verify procedure and were subsequently read by applying a voltage V_n. The experiment was repeated for $n = 1, \ldots, 1024$ different combinations of $\{\beta_n, \gamma_n\}$ and the results for each value of n were averaged on K devices (thus using $1024 \times K$ devices in total). As shown in Fig. 2b, the computation of Eq. (2) is effectively realized over approximately 2 decades of current. The current, I_n, can then be converted to an approximate value $\hat{\theta}_n$ that represents the final result of the computation (Supplementary Note II), which is plotted in Fig. 2c against the exact result θ_n computed in double-precision floating point. The distributions of the error $\hat{\theta}_n - \theta_n$ get narrower with increasing K (see Fig. 2d), with the standard deviation scaling as $K^{-0.5}$ (see inset) as dictated by the central limit theorem when averaging independent and identically distributed (iid) random variables. It indicates that the predominant part of the error comes from random perturbations in the current I_n. Possible causes for such perturbations are inter-device and intra-device variability[21,23], inherent conductance variations and low-frequency noise arising from the amorphous phase-change material[29].

The matrix-vector multiplication is a natural extension of the scalar multiplication where the elements of the matrix are coded into the conductance states of PCM devices. Since our experimental hardware only allows serial access to each individual crosspoint, only the element-by-element multiplications of the matrix-vector product were performed in hardware and the sum was performed outside of the chip (Supplementary Note III). The accumulated effect of errors in this mode of computing is fundamentally different from that of rounding errors arising for example from fixed-point data conversions[27] (Supplementary Note IV). The structural relaxation of the amorphous phase to an energetically more favorable "ideal glass" state and its manifestation as a temporal evolution of the conductance values also poses challenges that need to be accounted for (Supplementary Note V).

Next, we present the solution of (1) for model covariance matrices of different sizes defined as $A_{ij}^{i\neq j} = |i-j|^{-1}, A_{ij}^{i=j} = 1+\sqrt{i}$ for $i = 1, ..., N$ and $j = 1, ..., N$. Such matrices exhibit a decaying behavior that simulates decreasing correlation of features away from the main diagonal.[24] The elements of b were generated uniformly in $[0,1]$. The inner solver was chosen to be Conjugate Gradient with a diagonal scaling as preconditioner (Supplementary Note VI). We coded a reduced banded version of the matrix in the memristive array with 12 entries on each side of the main diagonal using $K = 4$ devices averaged per matrix element. The banding allowed us to code a matrix of maximum size $10,000 \times 10,000$ with $959,376$ total PCM devices. In this way, the inner solver works on an *inexact* version of the matrix A which is coded in the memristive array while the outer iterative refinement loop works towards finding the *exact* solution of (1) by using the full matrix A for the computation of the residuals. The evolution of the error between computed and exact solution as a function of the number of iterative refinements is shown Fig. 3. The algorithm converged exponentially to the desired precision after 11 iterative refinements with convergence rate independent of N. Accurate solving of problem (1) was thus possible despite the inaccurate computations in the 'memcomputing' unit and even when most elements of A were actually not coded in the memristive array (for $N = 10,000$ only 0.24% of the matrix elements were coded because of the banding used).

Finally, we tested the mixed-precision 'memcomputing' algorithm on a practical problem for which the matrix A was built from real-world data. For this, we used RNA expression measurements of genes obtained from cancer patients, publicly available from The Cancer Genome Atlas (TCGA) project (see Supplementary Note VII). We focused our investigation on 40 genes reported in the manually curated autophagy pathway of the Kyoto Encyclopedia of Genes and Genomes (KEGG). Autophagy plays opposing roles in cancer by both acting as a tumor suppressor by degrading damaged proteins and organelles, as well as enabling tumors to tolerate metabolic stress[30,31]. To infer and compare the networks of gene interactions (interactomes) from normal and cancer tissues, we calculated the partial correlations between the genes by computing the inverse covariance matrix Σ from 946 normal tissue samples and from 946 cancer tissue samples (Supplementary Note VII). Given the covariance matrix A of the 40 genes, Σ can be obtained by solving $Ax^n = e^n$ for $n = 1, ..., 40$, where e^n has all entries equal to zero except the n-th one, which is 1, and x^n is the resulting n-th column of Σ. We coded the 40×40 covariance matrix in the memristive array and used GMRES as the inner solver to solve the 40 linear equations. The procedure was repeated for both cancer and normal tissues. The algorithm converged to the desired precision for all 40 linear systems solved (see Fig. 4a) and the resulting Σ matrix was sufficiently accurate for computing the interactome (the interactomes obtained with the exact and computed Σ are identical). The computed partial correlations of the 40 genes studied and their distributions are shown in Fig. 4b. While part of the interactions between the cancer and normal tissues are preserved, the cancer network exhibits a different connectivity pattern (see Fig. 4c and 4d). In the normal tissue, the upstream signals INS, AMPK, ULK, ATG13, ATG17, IFNA and IFNG (dark colored) correlate with many of the downstream targets (light colored) known to be involved in the formation of autophagosomes, the molecular agents of autophagy. The partial correlations computed on cancerous tissue yield a sparsely connected network, implying an altered regulation pattern, as is commonly observed in cancer[32–34].

The above demonstration highlights the importance of linear analysis in problems associated with cognitive computing and data analytics. The fact that such computation can be performed partly with 'memcomputing' without sacrificing the overall computational accuracy opens up exciting new avenues towards energy-efficient large-scale data analytics, in which the massive data transfers inherent to the traditional von Neumann architecture have become the most energy-hungry part. Such solutions are much-needed because analyzing the ever-growing datasets we produce will quickly increase the computational load to the exascale level if standard techniques are to be used.[24] The problems tackled in this work were well-conditioned and of relatively small scale because of the limited size and precision of our hardware. Strategies to scale up include building larger arrays and/or operating several of them in parallel. To improve the precision of the 'memcomputing' unit and address problems with a broader range of condition numbers, possible avenues are in coding single matrix elements on multiple devices (Supplementary Note III) or using advanced memristive device concepts[29]. We expect a reduction in energy-to-solution when using mixed-precision 'memcomputing' compared to simply using high-precision computing if the gain in performance achieved by the 'memcomputing' unit sufficiently offsets the additional resources spent on data conversions and computation of residuals. Such gains are expected because the matrix-vector multiplication can be realized in a single time step without any intermediate data transfer of the matrix in a memristive crossbar[11]. We finally stress that mixed-precision 'memcomputing' can be used in applications that extend beyond the solution of linear equations to other relevant computational tasks arising in automatic control, optimization problems, machine learning and signal processing.

[1] R. H. Dennard, F. H. Gaensslen, V. L. Rideout, E. Bassous, and A. R. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," IEEE Journal of Solid-State Circuits **9**, 256–268 (1974).

[2] G. E. Moore, "Cramming more components onto integrated circuits, Reprinted from Electronics, volume 38, number 8, April 19, 1965, pp.114 ff." IEEE Solid-State Circuits Society Newsletter **11**, 33–35 (2006).

[3] M. Di Ventra and Y. V. Pershin, "The parallel approach," Nature Physics **9**, 200–202 (2013).

[4] F. L. Traversa and M. Di Ventra, "Universal memcomputing machines," IEEE Transactions on Neural Networks and Learning Systems **26**, 2702–2715 (2015).

[5] J. Borghetti, G. S. Snider, P. J. Kuekes, J. J. Yang, D. R. Stewart, and R. S. Williams, "'Memristive' switches enable 'stateful' logic operations via material implication," Nature **464**, 873–876 (2010).

[6] M. Cassinerio, N. Ciocchini, and D. Ielmini, "Logic computation in phase change materials by threshold and memory switching," Advanced Materials **25**, 5975–5980 (2013).

[7] D. Loke, J. M. Skelton, W.-J. Wang, T.-H. Lee, R. Zhao, T.-C. Chong, and S. R. Elliott, "Ultrafast phase-change logic device driven by melting processes," Proceedings of the National Academy of Sciences **111**, 13272–13277 (2014).

[8] C. D. Wright, Y. Liu, K. I. Kohary, M. M. Aziz, and R. J. Hicken, "Arithmetic and biologically-inspired computing using phase-change materials," Advanced Materials **23**, 3408–3413 (2011).

[9] H. Xu, Y. Xia, K. Yin, J. Lu, Q. Yin, J. Yin, L. Sun, and Z. Liu, "The chemically driven phase transformation in a memristive abacus capable of calculating decimal fractions," Scientific reports **3** (2013).

[10] P. Hosseini, A. Sebastian, N. Papandreou, C. D. Wright, and H. Bhaskaran, "Accumulation-based computing using phase-change memories with FET access devices," IEEE Electron Device Letters **36**, 975–977 (2015).

[11] M. Hu, J. P. Strachan, Z. Li, E. M. Grafals, N. Davila, C. Graves, S. Lam, N. Ge, J. J. Yang, and R. S. Williams, "Dot-product engine for neuromorphic computing: Programming 1T1M crossbar to accelerate matrix-vector multiplication," in *Proceedings of the 53rd Annual Design Automation Conference*, DAC '16 (ACM, 2016) pp. 19:1–19:6.

[12] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found," Nature **453**, 80–83 (2008).

[13] L. Chua, "Resistance switching memories are memristors," Applied Physics A **102**, 765–783 (2011).

[14] H.-S. P. Wong and S. Salahuddin, "Memory leads the way to better computing," Nature Nanotechnology **10**, 191–194 (2015).

[15] A. Ievlev, S. Jesse, A. Morozovska, E. Strelcov, E. Eliseev, Y. Pershin, A. Kumar, V. Y. Shur, and S. Kalinin, "Intermittency, quasiperiodicity and chaos in probe-induced ferroelectric domain switching," Nature Physics **10**, 59–66 (2014).

[16]M. N. Bojnordi and E. Ipek, "Memristive Boltzmann machine: A hardware accelerator for combinatorial optimization and deep learning," in *2016 IEEE International Symposium on High Performance Computer Architecture (HPCA)* (2016) pp. 1–13.
[17]P. M. Sheridan, C. Du, and W. D. Lu, "Feature extraction using memristor networks," IEEE Transactions on Neural Networks and Learning Systems **27**, 2327–2336 (2016).
[18]S. Choi, P. Sheridan, and W. D. Lu, "Data clustering using memristor networks," Scientific reports **5** (2015).
[19]S. Ambrogio, S. Balatti, A. Cubeta, A. Calderoni, N. Ramaswamy, and D. Ielmini, "Statistical fluctuations in HfO_x resistive-switching memory: Part I-set/reset variability," IEEE Transactions on Electron Devices **61**, 2912–2919 (2014).
[20]A. Fantini, L. Goux, R. Degraeve, D. Wouters, N. Raghavan, G. Kar, A. Belmonte, Y.-Y. Chen, B. Govoreanu, and M. Jurczak, "Intrinsic switching variability in HfO_2 RRAM," in *2013 5th IEEE International Memory Workshop* (IEEE, 2013) pp. 30–33.
[21]M. Le Gallo, T. Tuma, F. Zipoli, A. Sebastian, and E. Eleftheriou, "Inherent stochasticity in phase-change memory devices," in *Proc. of the European Solid-State Device Research Conference (ESSDERC)* (IEEE, 2016) pp. 373–376.
[22]S. Gaba, P. Sheridan, J. Zhou, S. Choi, and W. Lu, "Stochastic memristive devices for computing and neuromorphic applications," Nanoscale **5**, 5872–5878 (2013).
[23]T. Tuma, A. Pantazi, M. Le Gallo, A. Sebastian, and E. Eleftheriou, "Stochastic phase-change neurons," Nature Nanotechnology **11**, 693–699 (2016).
[24]C. Bekas, A. Curioni, and I. Fedulova, "Low cost high performance uncertainty quantification," in *Proceedings of the 2nd Workshop on High Performance Computational Finance* (ACM, 2009) pp. 8:1–8:8.
[25]P. Klavík, A. C. I. Malossi, C. Bekas, and A. Curioni, "Changing computing paradigms towards power efficiency," Phil. Trans. R. Soc. A **372**, 20130278 (2014).
[26]Y. Saad, *Iterative methods for sparse linear systems* (Siam, 2003).
[27]N. J. Higham, *Accuracy and stability of numerical algorithms* (Siam, 2002).
[28]G. W. Burr, M. J. Brightsky, A. Sebastian, H.-Y. Cheng, J.-Y. Wu, S. Kim, N. E. Sosa, N. Papandreou, H.-L. Lung, H. Pozidis, *et al.*, "Recent progress in phase-change memory technology," IEEE Journal on Emerging and Selected Topics in Circuits and Systems **6**, 146–162 (2016).
[29]W. W. Koelmans, A. Sebastian, V. P. Jonnalagadda, D. Krebs, L. Dellmann, and E. Eleftheriou, "Projected phase-change memory devices," Nature communications **6** (2015).
[30]R. Mathew, V. Karantza-Wadsworth, and E. White, "Role of autophagy in cancer," Nature Reviews Cancer **7**, 961–967 (2007).
[31]Z. J. Yang, C. E. Chee, S. Huang, and F. A. Sinicrope, "The role of autophagy in cancer: therapeutic implications," Molecular cancer therapeutics **10**, 1533–1541 (2011).
[32]J. West, G. Bianconi, S. Severini, and A. E. Teschendorff, "Differential network entropy reveals cancer system hallmarks," Scientific reports **2** (2012).
[33]G. Schramm, N. Kannabiran, and R. König, "Regulation patterns in signaling networks of cancer," BMC systems biology **4**, 1 (2010).
[34]S. Hong, X. Chen, L. Jin, and M. Xiong, "Canonical correlation analysis for RNA-seq co-expression networks," Nucleic acids research **41**, e95–e95 (2013).

Acknowledgments We thank C. Malossi and M. Rodriguez for discussions; N. Papandreou, A. Athmanathan and U. Egger for experimental help; T. Delbruck for reviewing the manuscript; and C. Bolliger for help with the preparation of the manuscript. A. S. would like to acknowledge partial financial support from the ERC grant 682675.

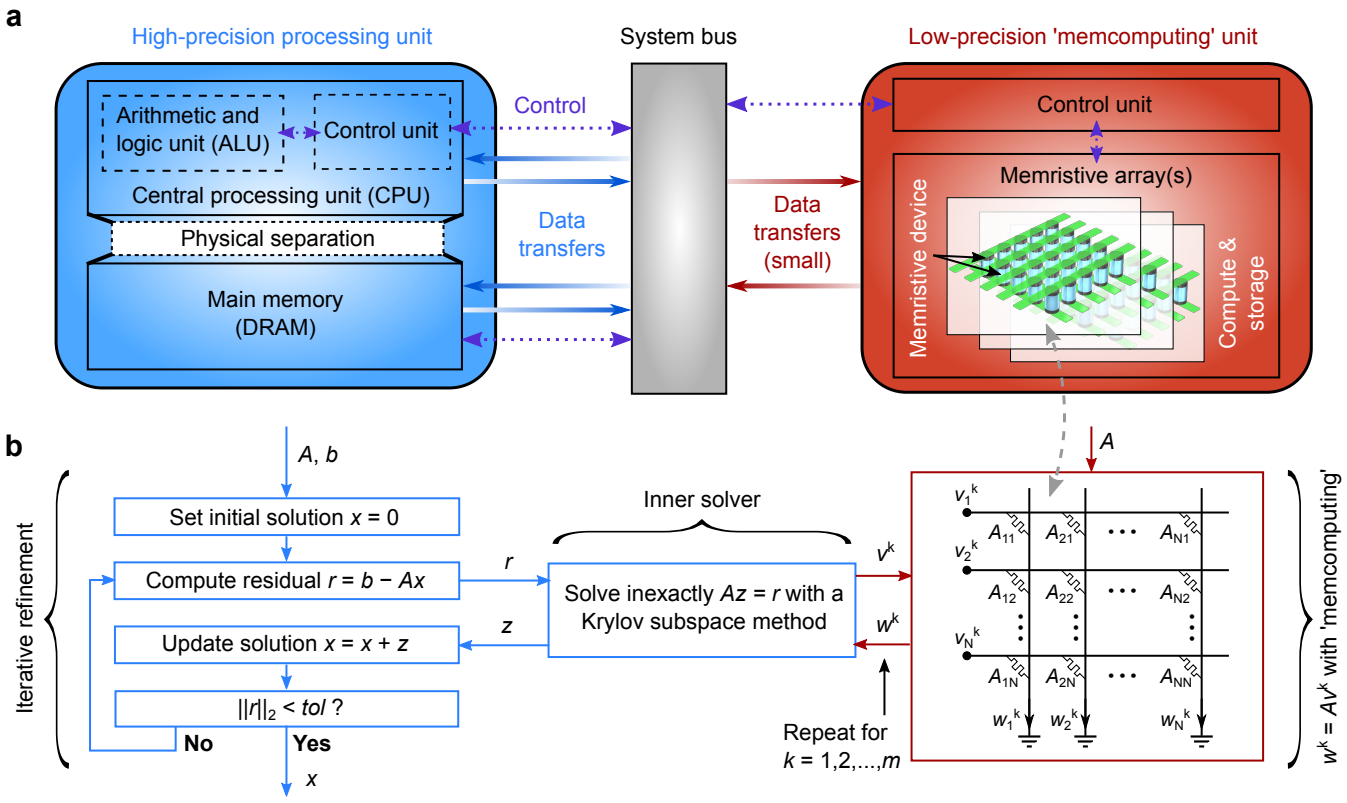

FIG. 1. **Concept of mixed-precision 'memcomputing' a**, Possible architecture of a mixed-precision 'memcomputing' system. The high-precision processing unit (left) performs digital logic computation and is based on the standard von Neumann computing architecture. The low-precision 'memcomputing' unit (right) performs analog in-memory computation using one or multiple memristive arrays. The system bus (middle) implements the overall management (control, data, addressing) between the two units. The purple dotted arrows indicate control communication and the plain arrows (red, blue) indicate data transfers. **b**, Algorithm for solving a system of linear equations $Ax = b$ using the mixed-precision 'memcomputing' system of **a**. The blue boxes show the steps implemented in the high-precision processing unit and the red box shows the matrix-vector multiplication step implemented in the low-precision 'memcomputing' unit.

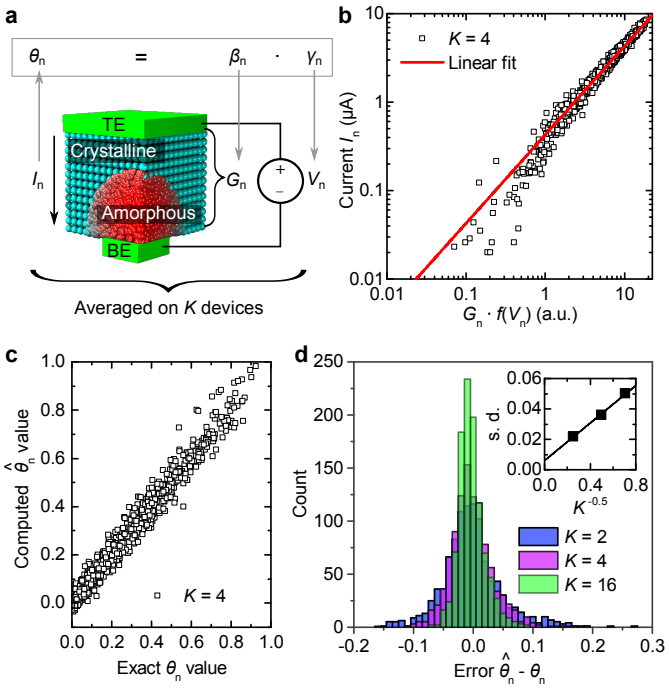

FIG. 2. **Scalar multiplication a**, Schematic of a PCM device and the scalar multiplication implementation based on Ohm's law. TE (BE) denotes top (bottom) electrode. The grey arrows indicate mappings from one variable to another. **b**, Plot showing the proportionality between I_n and $G_n f(V_n)$ (Eq. (2)) for the 1024 different combinations of $\{\beta_n, \gamma_n\}$. **c**, Final result of the computed scalar multiplication $\hat{\theta}_n$ plotted against the exact result θ_n. **d**, Error distributions for different numbers of averaged devices K. The inset in **d** shows the standard deviation (s.d.) of the distributions versus $K^{-0.5}$.

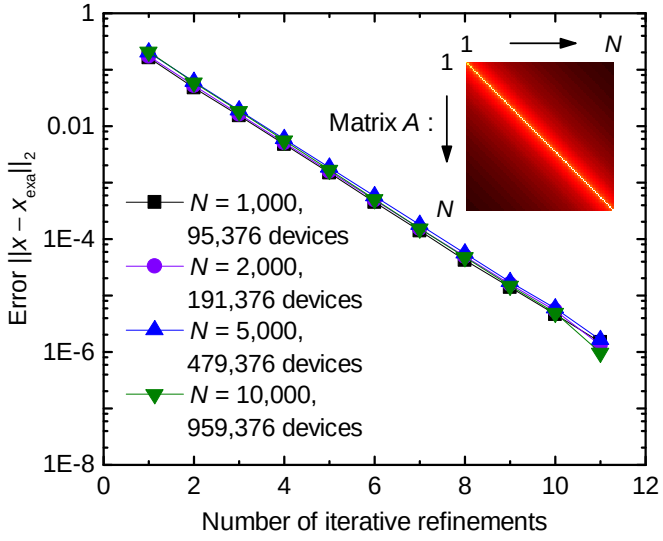

FIG. 3. **Solution of a system of linear equations involving a model covariance matrix** Norm of error between the computed x and exact x_{exa} solution of Eq. (1) as a function of the number of iterative refinements for different covariance matrix sizes. x_{exa} was computed by direct inversion of Eq. (1) in double-precision floating point. The inset shows a heat map (colormap in log-scale) of the model covariance matrix A used for $N = 1,000$.

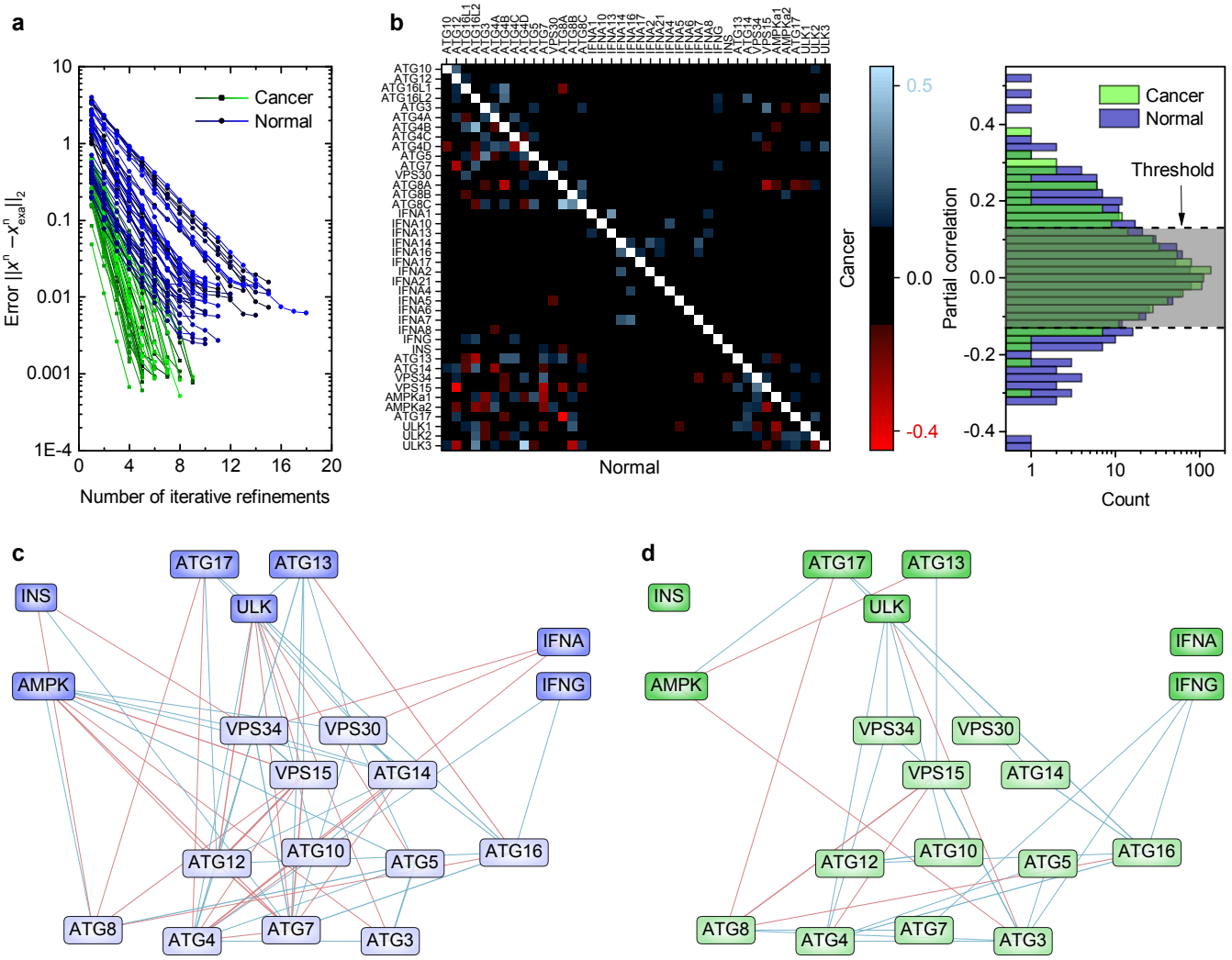

FIG. 4. **Estimation of autophagy-related gene interactions from RNA measurements a**, Convergence of the mixed-precision 'memcomputing' algorithm for the 40 linear equations solved for the cancer and normal tissues. x^n_{exa} was computed by direct inversion of Eq. (1) in double-precision floating point. **b**, Matrix of computed partial correlations of the 40 genes studied for cancer and normal tissues (left) and their distributions (right). For visualization purposes, only the interactions for which the magnitude of the partial correlations is larger than a threshold of 0.13, corresponding to the 90-th percentile of the normal tissue, are displayed. **c**, Interactome obtained from normal tissue. **d**, Interactome obtained from cancer tissue. In **c** and **d**, the upstream nodes are dark colored and the downstream targets are light colored. The blue edges denote positive interactions and the red edges denote negative interactions.