PLOS Computational Biology

METHODS

APDCA: An accurate and effective method for predicting associations between RBPs and AS-events during epithelial-mesenchymal transition

Yangsong He[1], Zheng-Jian Bai[2], Wai-Ki Ching[3], Quan Zou[4], Yushan Qiu[1]*

1 School of Mathematical Sciences, Shenzhen University, Shenzhen, People's Republic of China, **2** School of Mathematical Sciences, Xiamen University, Fujian, People's Republic of China, **3** Department of Mathematics, The University of Hong Kong, Hong Kong, People's Republic of China, **4** Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, People's Republic of China

* yushan.qiu@szu.edu.cn

Check for updates

OPEN ACCESS

Citation: He Y, Bai Z-J, Ching W-K, Zou Q, Qiu Y (2025) APDCA: An accurate and effective method for predicting associations between RBPs and AS-events during epithelial-mesenchymal transition. PLoS Comput Biol 21(11): e1013665. https://doi.org/10.1371/journal.pcbi.1013665

Editor: Jeremiah James Zartman, University of Notre Dame, UNITED STATES OF AMERICA

Received: August 10, 2025

Accepted: October 24, 2025

Published: November 6, 2025

Copyright: © 2025 He et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data availability statement: The code and datasets of APDCA are available from https://github.com/Yangsong-He/APDCA and https://codeocean.com/capsule/0859877/tree.

Abstract

Motivation: Epithelial-mesenchymal transition (EMT) plays a key role in cancer metastasis by promoting changes in adhesion and motility. RNA-binding proteins (RBPs) regulate alternative splicing (AS) during EMT, enabling a single gene to produce multiple protein isoforms that affect tumor progression. Disruption of RBP-AS interactions may disrupt the progress of diseases like cancer. Despite the importance of RBP-AS relationships in EMT, few computational methods predict these associations. Existing models struggle in sparse settings with limited known associations. To improve performance, we incorporate both sparsity constraints and heterogeneous biological data to infer RBP–AS associations.

Result: We propose a new method based on Accelerated Proximal DC Algorithm (APDCA) for predicting RBP–AS associations. In particular, APDCA combines sparse low-rank matrix factorization with a Difference-of-Convex (DC) optimization framework and uses extrapolation to improve convergence. A key feature of APDCA is the use of a sparsity constraint, which filters out noise and highlights key associations. In addition, integrating multiple related data sources with direct or indirect relationships can help in reaching a more comprehensive view of RBPs and AS events and to reduce the impact of false positives associated with individual data sources. we prove that our proposed algorithm is convergent under some conditions and the experimental results have illustrated that APDCA outperforms six baseline methods in both AUC and AUPR. A case study on the RBP QKI shows that the top predictions are verified by the OncoSplicing database. Thus, APDCA provides a fast, interpretable, and scalable tool for discovering post-transcriptional regulatory interactions.

Funding: YQ was supported by grants from the National Natural Science Foundation of China [grant number 62372303, 62002234], Guangdong Basic and Applied Basic Research Foundation [grant number 2024A1515010113] and Shenzhen Science and Technology Program [grant number RCYX20231211090244048]. ZJB was supported by grants from the National Natural Science Foundation of China [grant number 12371382] and Fujian Provincial Natural Science Foundation of China [grant number 2025J01030]. QZ was supported by grants from the National Natural Science Foundation of China [grant number 62131004]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Epithelial–mesenchymal transition (EMT) is a fundamental biological process closely linked to cancer metastasis, where RNA-binding proteins (RBPs) play a critical role by regulating alternative splicing (AS) events. Accurately identifying RBP–AS associations during EMT is essential for understanding post-transcriptional regulation, yet existing computational approaches often struggle with the sparsity of known interactions and the heterogeneity of biological data. To address these challenges, we propose APDCA, an Accelerated Proximal Difference-of-Convex Algorithm, which integrates heterogeneous biological data and formulates the prediction task as a difference-of-convex optimisation problem. APDCA introduces a proximal optimisation scheme enhanced by Nesterov extrapolation, enabling both noise suppression and efficient convergence. The algorithm jointly enforces low-rank structure to capture shared regulatory patterns and sparsity constraints to filter out spurious associations. Moreover, it incorporates adaptive weight updating and error controlled regularisation, eliminating the need for pre-constructed similarity graphs or manual parameter tuning. This unified framework enables APDCA to learn biologically meaningful and interpretable associations with high computational efficiency. Through extensive experiments, APDCA demonstrates superior performance over six existing methods and rapidly identifies high-confidence RBP–AS pairs for downstream biological validation.

1 Introduction

Alternative splicing (AS) is a crucial post-transcriptional regulatory mechanism in eukaryotes that allows a single gene to produce multiple transcript isoforms by variably splicing precursor mRNA [1]. This process greatly enhances the diversity of the proteome. RNA-binding proteins (RBPs) play a central role in regulating AS by recognizing specific sequence motifs and influencing splice site selection, exon inclusion or skipping, and transcript stability [2,3].

AS has been extensively studied in various cellular contexts. However, its regulation during epithelial–mesenchymal transition (EMT) remains underexplored [4]. EMT is a process in which epithelial cells lose their adhesion properties and acquire migratory traits typical of mesenchymal cells [5]. This transition is important for cancer metastasis, and disruption of AS during EMT has a significant impact on tumor progression. RBPs are central to the regulation of AS during EMT, making their interaction with AS events crucial to understanding tumor progression [6]. Thus, understanding the relationship between RBPs and AS events during EMT is essential to uncovering new cancer therapeutic targets.

Studies have shown that disruption of RBP–AS event interactions is closely associated with various diseases, including cancers, neurodegenerative disorders, and immune dysfunction [7,8]. For example, the RNA-binding protein QKI modulates the alternative splicing of the NUMB gene in lung cancer by binding to specific RNA

elements in its pre-mRNA. This regulation suppresses tumor cell proliferation and inhibits the activation of the Notch signaling pathway, highlighting a critical role of QKI-mediated splicing control in tumor suppression [9]. Similarly, PTBP1 regulates splicing isoforms of the PKM gene, thereby impacting the glycolytic pathway and playing a key role in multiple types of tumors [10]. These examples underscore the functional diversity and specificity of RBPs in AS regulation. Therefore, systematically identifying and predicting regulatory relationships between RBPs and AS events is critical for elucidating transcriptomic regulatory mechanisms and discovering novel biomarkers or therapeutic targets.

Although recent years have seen growing attention to the regulatory relationships between alternative splicing (AS) events and RNA-binding proteins (RBPs), computational methods specifically designed to predict such associations remain limited [11]. In contrast, numerous models have been widely developed to predict potential associations between lncRNAs and diseases [12–15]. These methods are generally based on universal molecular network modeling strategies with strong generalizability, making them suitable for exploring associations among other biological entities. Current mainstream association prediction algorithms can be broadly divided into four categories. The first category includes machine learning–based methods. Zeng et al. proposed SDLDA, which constructs feature vectors based on similarities between biological entities and applies a support vector machine (SVM) classifier to determine potential associations [16]. Lan et al. introduced LDAP, which employs label-guided linear discriminant analysis (LDA) to project features into a discriminative subspace, improving the separability between associated and non-associated pairs [17]. The second category consists of graph-based methods. Wu et al. developed GAERF, which constructs a heterogeneous graph and uses a graph autoencoder to learn node embeddings, followed by a random forest classifier to predict associations [18]. The third category comprises matrix factorization methods [19]. Fu et al. proposed MLFDA, which performs multi-view low-rank matrix factorization to capture latent features from multiple association matrices and integrates structural information from various perspectives to enhance expressiveness [20]. Lu et al. designed SIMCLDA, which incorporates similarity-based regularization into non-negative matrix factorization, encouraging similar entities to share latent features and thereby improving robustness and generalization, especially under sparse data conditions [21]. The fourth category includes network propagation methods. Gu et al. introduced GrwLDA, which applies a random walk with restart strategy to simulate information diffusion across a heterogeneous network and infer potential associations [22].

Existing computational models for molecular association prediction still have limitations when applied to the identification of regulatory relationships between RNA-binding proteins (RBPs) and alternative splicing (AS) events during EMT. Many machine learning methods rely on predefined similarity features and lack the ability to integrate heterogeneous biological data. Graph-based and network-based models often depend on known molecular interaction networks, which may not capture hidden or novel associations. Matrix factorization techniques have been widely used due to their effectiveness in latent structure discovery, but most are limited to modeling single-type associations and fail to leverage the complementary information from diverse biological entities. Furthermore, existing methods rarely incorporate sparsity constraints, which are important for improving model interpretability, reducing noise, and enhancing generalization to unseen data. Biologically, only a small subset of RBPs tends to interact with any given AS event. So enforcing sparsity better reflects this selective regulatory landscape.

To address these challenges, we propose APDCA (Accelerated Proximal DC Algorithm), a novel matrix factorization framework specifically designed for predicting RBP–AS event associations during the EMT process. Our model integrates multiple types of biological entities to form a heterogeneous molecular network, enabling a more comprehensive representation of the regulatory landscape. To highlight important associations and suppress irrelevant ones, we introduce a sparsity constraint on the reconstructed matrix. This results in a non-convex optimization problem, which we reformulate using Difference-of-Convex (DC) programming to ensure solvability [23]. In addition, APDCA leverages a proximal optimization framework to handle the sparsity-induced non-smoothness, which improves stability and ensures convergence in complex, high-dimensional spaces. Proximal methods are well suited for non-differentiable problems and enable efficient iterative updates by incorporating regularization through proximity operators [24]. To further accelerate convergence,

we adopt a Nesterov extrapolation technique during the iterative optimization process [25]. These innovations collectively allow APDCA to achieve accurate, interpretable, and scalable predictions across complex biological systems.

To evaluate the model's performance, we conducted experiments using 5-fold cross-validation on the constructed dataset. The results show that the proposed APDCA model consistently outperforms six representative baseline methods, achieving the highest scores in both AUC (Area Under the ROC Curve) and AUPR (Area Under the Precision-Recall Curve). To further assess the model's applicability and biological interpretability, we performed a case study based on the OncoSplicing database [26]. Focusing on specific RNA-binding proteins such as QKI, we explored the top-ranked predictions in detail, demonstrating the model's potential to uncover meaningful regulatory associations between RBPs and AS events.

2 Materials and methods

2.1 Problem formulation

Throughout this paper, we use the following notations. Let $\mathbb{R}^{m\times n}$ be the set of all $m\times n$ real matrices, and $\mathbb{R}^n = \mathbb{R}^{n\times 1}$. Let $\mathbb{R}^n$ be equipped with the Euclidean inner product $\langle\cdot,\cdot\rangle$ and its induced norm $\|\cdot\|$. The superscripts "$\cdot^{-1}$", "$\cdot^{\dagger}$", and "$\cdot^{T}$" stand for the inverse, the Moore-Penrose pseudoinverse, and the transpose of a matrix, respectively. Denote by $\|\cdot\|_F$ and $\|\cdot\|_0$ the Frobenius matrix norm and the ℓ_0 norm (i.e., the number of non-zero entries of a vector or matrix). Let $\mathrm{tr}(\cdot)$ and $\mathrm{vec}(\cdot)$ be the trace of a square matrix and a column vector from a matrix by stacking its column vectors below one another, respectively. Let $\mathbf{e}$ be a column vector with all elements equal to one. Finally, $A \geq O$ or $\mathbf{a} \geq \mathbf{0}$ means a matrix or a vector in which all the elements are equal to or greater than zero and $\mathbf{a} \leq \mathbf{e}$ means that $\mathbf{e} - \mathbf{a} \geq \mathbf{0}$.

The problem can be described as follows: suppose there are m types of molecules, either directly or indirectly related to RBPs or AS-events, and a collection of relational data sources R. Each data source R_{ij} represents the interaction between objects of the i-th and j-th types, where $i,j \in \{1,2,\ldots,m\}$. Specifically, $R_{ij} \in \mathbb{R}^{n_i\times n_j}$ stores the inter-relation between n_i objects of the i-th type and n_j objects of the j-th type. Note that R_{ij} can be asymmetric. Additionally, the intra-relation of objects within the i-th type is captured by a constraint matrix $\Theta_i \in \mathbb{R}^{n_i\times n_i}$. Matrix factorization-based data fusion aims to decompose the data matrices $R \in \mathbb{R}^{(\sum_{i=1}^m n_i)\times(\sum_{i=1}^m n_i)}$ or its sub-matrices, while being constrained by the corresponding sub-matrices of $\Theta \in \mathbb{R}^{(\sum_{i=1}^m n_i)\times(\sum_{i=1}^m n_i)}$. The decomposition yields low-rank matrices that explore latent relationships both within the same type of objects and across different types. These low-rank matrices are subsequently used to reconstruct the target association matrix, R_{ij}, for predicting new associations between objects of the i-th type (RBPs) and objects of the j-th type (AS-events).

2.2 Objective function and constraints of APDCA

Biological interactions among different types of entities often exhibit strong sparsity, since many potential interactions never manifest in practice while only a few are functionally relevant. By preserving only the top k largest elements in G_i, the model automatically filters out negligible interaction strengths, thus focusing on the most meaningful signals. This element-level sparsity design not only reduces noise and spurious connections but also helps prevent overfitting, ultimately improving both the interpretability and robustness of the final predictions.

Hence, we consider the following optimization problem:

$$\begin{cases} \min\limits_{\substack{S_{ij}\in\mathbb{R}^{k_i\times k_j},1\leq i,j\leq m\\ G_i\in\mathbb{R}^{n_i\times k_i},1\leq i\leq m}} & F(S,G) := \sum_{i,j}\omega_{ij}\|R_{ij} - G_iS_{ij}G_j^T\|_F^2 \\ & \qquad +\lambda\sum_{r=1}^{\max_i r_i}\sum_{i=1}^m \mathrm{tr}(G_i^T\Theta_iG_i) \\ \text{subject to (s.t.)} & G_i \geq O, \|G_i\|_0 \leq k,\ i=1,\ldots,m, \end{cases} \tag{1}$$

where $\{\omega_{ij} > 0\}_{i,j=1}^m$ are the prescribed weight parameters, and $\lambda > 0$ is a regularized parameter.

$G_i \in \mathbb{R}^{n_i \times k_i}$ denotes the low-rank representation matrix for objects of the i-th type, and $S_{ij} \in \mathbb{R}^{k_i \times k_j}$ encodes the latent interactions between the i-th and j-th types. Multiplying G_i, S_{ij}, and G_j^T reconstructs the observed matrix R_{ij} in a compact, low-dimensional form.

For any $\mathbf{x} = (x_i) \in \mathbb{R}^n$ and $1 \le k \le n$, let $\|\mathbf{x}\|_k := \sum_{t=1}^{k} |x_{\pi(t)}|$ be the largest-k norm of $\mathbf{x}$, where π is a permutation of $[n] := \{1, 2, \ldots, n\}$ such that $|x_{\pi(1)}| \ge |x_{\pi(2)}| \ge \cdots \ge |x_{\pi(n)}|$. Then we have

$$\|\mathbf{x}\|_0 \le k \quad \text{if and only if} \quad \|\mathbf{x}\|_n - \|\mathbf{x}\|_k = 0.$$

In particular, when $x \ge 0$,

$$\|\mathbf{x}\|_0 \le k \quad \text{if and only if} \quad \mathbf{e}^{\mathsf{T}}\mathbf{x} - s_{(k)}(\mathbf{x}) = 0,$$

where $s_{(k)}(\mathbf{x}) = \sum_{t=1}^{k} x_{\pi(t)}$ be the sum of the k largest entries of $\mathbf{x}$, where π is a permutation of $[n]$ such that $x_{\pi(1)} \ge x_{\pi(2)} \ge \cdots \ge x_{\pi(n)}$. We note that the subdifferential of $\|\cdot\|_k$ at a point $\mathbf{x} \in \mathbb{R}^n$ is given by [27,28]

$$\partial \|\mathbf{x}\|_k = \arg\max_{\mathbf{z} \in \mathbb{R}^n} \left\{ \langle \mathbf{x}, \mathbf{z} \rangle \mid \sum_{t=1}^{n} |z_i| = k, -1 \le z_i \le 1, i = 1, \ldots, n \right\}$$

and the subdifferential of $s_{(k)}(\cdot)$ at a point $\mathbf{x} \in \mathbb{R}^n$ is given by [29]

$$\partial s_{(k)}(\mathbf{x}) = \arg\max_{\mathbf{z} \in \mathbb{R}^n} \left\{ \langle \mathbf{x}, \mathbf{z} \rangle \mid \sum_{t=1}^{n} z_i = k, 0 \le z_i \le 1, i = 1, \ldots, n \right\}.$$

As noted in [29], for any $\mathbf{x} = (x_i) \in \mathbb{R}^n$ and $1 \le k \le n$, we can compute a subgradient $\mathbf{z} \in \partial s_{(k)}(\mathbf{x})$ as follows: (i) Sort the entries of $\mathbf{x}$ in decreasing order, i.e., $x_{\pi(1)} \ge x_{\pi(2)} \ge \cdots \ge x_{\pi(n)}$; (ii) Set $z_{\pi(i)} = 1$ for $i = 1, \ldots, k$ and $z_{\pi(i)} = 0$ for $i = k + 1, \ldots, n$. Analogiously, we can compute a subgradient $\mathbf{z} \in \partial \|\mathbf{x}\|_k$ as follows: (i) Sort the entries of $\mathbf{x}$ such that $|x_{\pi(1)}| \ge |x_{\pi(2)}| \ge \cdots \ge |x_{\pi(n)}|$; (ii) Assign $\mathrm{sign}(x_{\pi(i)})$ to $z_{\pi(i)}$ for $i = 1, \ldots, k$ and 0 to $z_{\pi(i)}$ for $i = k + 1, \ldots, n$.

From the above analysis, we see that problem Eq (1) is equivalent to the following DC-constrained problem:

$$\begin{cases} \min\limits_{\substack{S_{ij} \in \mathbb{R}^{k_i \times k_j}, 1 \le i,j \le m \\ G_i \in \mathbb{R}^{n_i \times k_i}, 1 \le i \le m}} & F(S, G) := \sum_{i,j} \omega_{ij} \|R_{ij} - G_i S_{ij} G_j^T\|_F^2 \\ & \qquad + \lambda \sum_{r=1}^{\max_i r_i} \sum_{i=1}^{m} \mathrm{tr}(G_i^T \Theta_i G_i) \\ \text{s.t.} & G_i \ge O, \ \mathbf{e}^T G_i \mathbf{e} - s_{(k)}(\mathrm{vec}(G_i)) = 0, \\ & i = 1, \ldots, m, \end{cases} \tag{2}$$

Instead of Eq (2), we consider the following penalized problem:

$$\min_{\substack{S_{ij} \in \mathbb{R}^{k_i \times k_j}, 1 \le i,j \le m \\ G_i \in \mathbb{R}^{n_i \times k_i}, 1 \le i \le m}} F(S, G) := f(S, G) + g(G) \tag{3}$$

where

$$\begin{aligned} f(S,G) &= \sum_{i,j} \omega_{ij} \|R_{ij} - G_i S_{ij} G_j^T\|_F^2 + \lambda \sum_{r=1}^{\max_i r_i} \sum_{i=1}^{m} \mathrm{tr}(G_i^T \Theta_i G_i), \\ g(G) &= I_{\mathcal{C}}(G) + \lambda_G \sum_{i=1}^{m} \left(\mathbf{e}^T G_i \mathbf{e} - s_{(k)}(\mathrm{vec}(G_i))\right) \\ &:= I_{\mathcal{C}}(G) + g_1(G) - g_2(G), \end{aligned}$$

where $\lambda_G > 0$ is a penalty parameter and $I_{\mathcal{C}}$ denotes the indicator function of the set $\mathcal{C} = \{G = \mathrm{diag}(G_1, \ldots, G_n) \mid G_i \in \mathbb{R}^{n_i \times k_i}, G_i \geq O, i = 1, \ldots, m\}$, i.e., $I_{\mathcal{C}}(G) = 0$ if $G \in \mathcal{C}$ and $I_{\mathcal{C}}(G) = +\infty$, otherwise.

2.3 Algorithm for APDCA

We propose an APDCA algorithm to solve the non-convex optimization problem described in Eq (3), incorporating the ideas originally developed for general nonconvex programming [30] (e.g., sparse optimization [31]). The main process of model is shown in Fig 1.

2.3.1 Initialisation. For each molecular type i, we initialize the nonnegative low-rank basis matrix $G_i^{(0)} \in \mathbb{R}^{n_i \times k_i}$ by performing tri-factorization over the relational matrices R_{ij} involving the i-th object type. Once $G_i^{(0)}$ is obtained, we set $G_i^{(1)} = Z_i^{(1)} = G_i^{(0)}$ to ensure that the first extrapolation step is well-defined. The corresponding interaction matrix $S^{(1)}$ is initialized by minimizing the squared Frobenius norm:

$$\min_S \|R_{ij} - G_i S_{ij} G_j^T\|_F^2.$$

Taking the derivative with respect to S and setting it to zero yields the closed-form solution:

$$S = \left((G^{(1)})^T G^{(1)}\right)^{\dagger} (G^{(1)})^T R G^{(1)} \left((G^{(1)})^T G^{(1)}\right)^{\dagger}.$$

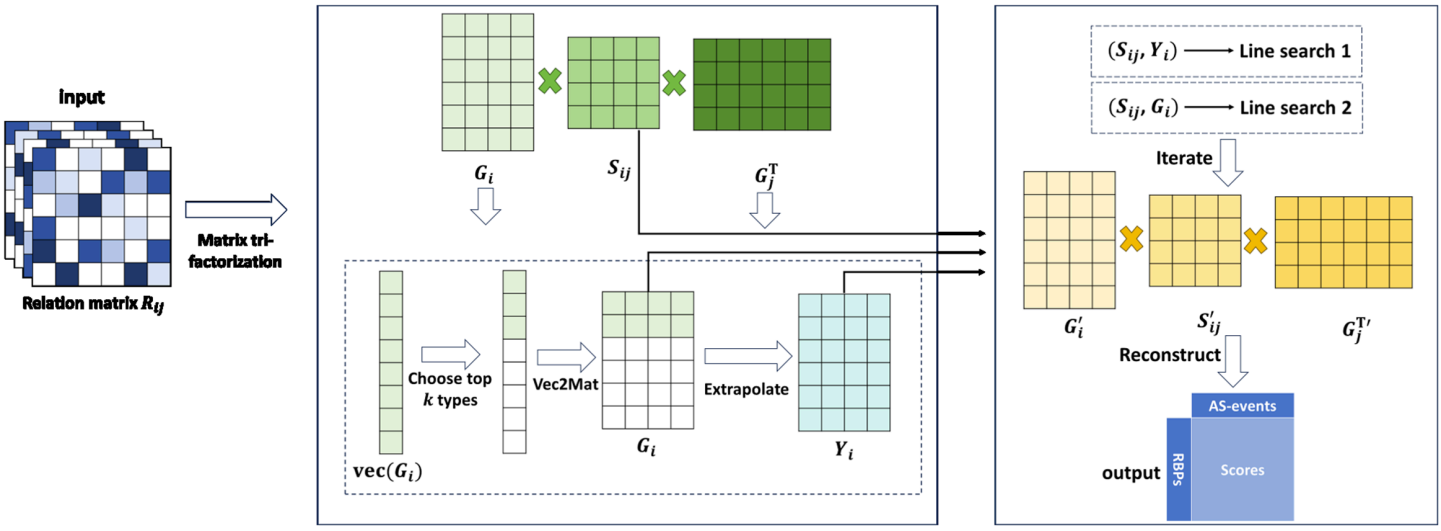

Fig 1. The operating principle of APDCA. APDCA iteratively optimizes the low-rank matrices (G_i) of multiple relational data matrices through matrix tri-factorization, and updates (G_i) by selecting the top k relationships based on sparsity. Then construct an extrapolated point Y_i. Then iterate S_{ij}, Y_i by LS-1 if the condition is satisfied, otherwise iterate S_{ij} and G_i by LS-2. Finally reconstruct the relation matrix and get association scores and rank.

https://doi.org/10.1371/journal.pcbi.1013665.g001

We record the initial objective $c_1 = F(S^{(1)}, G^{(1)})$, set the FISTA parameter $\theta^{(1)} = 1$ and the non-monotone weight $q_1 = 1$. The Armijo decrease threshold $\delta > 0$, the history decay factor $\tau \in (0, 1)$. We list the initialization of all variables in Table 1.

2.3.2 Extrapolation. At iteration ℓ, we construct an extrapolated point $Y^{(\ell)}$ by

$$Y^{(\ell)} = G^{(\ell)} + \frac{\theta^{(\ell-1)}}{\theta^{(\ell)}}\left(Z^{(\ell)} - G^{(\ell)}\right) + \frac{\theta^{(\ell-1)} - 1}{\theta^{(\ell)}}\left(G^{(\ell)} - G^{(\ell-1)}\right), \tag{4}$$

The step size pair (l_1, l_2) used in the gradient and proximal updates is initialized at this stage. Rather than using fixed values, l_1 and l_2 are dynamically computed based on the iterates and gradients to better reflect local curvature and improve convergence. Specifically, l_1 is calculated by:

$$l_1 = \frac{\langle S^{(\ell)} - S^{(\ell-1)}, S^{(\ell)} - S^{(\ell-1)}\rangle}{\langle S^{(\ell)} - S^{(\ell-1)}, \nabla_S f(S^{(\ell)}, Y^{(\ell)}) - \nabla_S f(S^{(\ell-1)}, Y^{(\ell)})\rangle}, \tag{5a}$$

and l_2 is similarly computed from the G-update as:

$$l_2 = \frac{\langle Y^{(\ell)} - G^{(\ell-1)}, Y^{(\ell)} - G^{(\ell-1)}\rangle}{\langle Y^{(\ell)} - G^{(\ell-1)}, \nabla_G f(S^{(\ell)}, Y^{(\ell)}) - \nabla_G f(S^{(\ell)}, G^{(\ell-1)})\rangle}. \tag{5b}$$

2.3.3 Line Search Algorithm 1 (LS-1). With the extrapolated point $Y^{(\ell)}$ fixed, APDCA performs a forward–backward update to generate the candidate pair $(S^{(\ell+1)}, Z^{(\ell+1)})$. The update for S is a standard gradient descent step with respect to the smooth part of the objective:

$$S^{(\ell+1)} = S^{(\ell)} - \frac{1}{l_1}\nabla_S f\left(S^{(\ell)}, Y^{(\ell)}\right), \tag{6}$$

Table 1. Initialization of variables in the APDCA algorithm.

Variable	Initialization Description
$Z_i^{(1)}$	$Z_i^{(1)} = G_i^{(0)} = G_i^{(1)} \geq 0$ for $i = 1, \ldots, m$
$G_i^{(0)}, G_i^{(1)}$	≥ 0
ω_{ij}	$\omega_{ij} > 0$ for $i, j = 1, \ldots, m$
λ	> 0
λ_G	> 0
$\theta^{(0)}$	0
$\theta^{(1)}$	1
δ	> 0
τ	$0 < \tau < 1$
q_1	1
c_1	$F(S^{(1)}, G^{(1)})$
$S^{(1)}$	$\left((G^{(1)})^T G^{(1)}\right)^{\dagger} (G^{(1)})^T R G^{(1)} \left((G^{(1)})^T G^{(1)}\right)^{\dagger}$
$S^{(0)}$	$S^{(0)} = S^{(1)}$
$\omega_{ij}^{(1)}$	$\frac{1}{2\|R_{ij} - G_i^{(1)} S_{ij}^{(1)} (G_j^{(1)})^T\|_F^2}$
l_1, l_2	> 0
η	> 1
ℓ	1

https://doi.org/10.1371/journal.pcbi.1013665.t001

For the Z update, we perform a proximal step as:

$$Z^{(\ell+1)} = \mathbf{prox}_{(g_1+I_{\mathcal{C}})/l_2}\left(Y^{(\ell)} - \frac{1}{l_2}\nabla_G f(S^{(\ell)}, Y^{(\ell)}) + \frac{1}{l_2}W(Y^{(\ell)})\right), \tag{7}$$

where $W(Y^{(\ell)})$ is a subgradient of g_2 at $Y^{(\ell)}$, and

$$\mathbf{prox}_{(g_1+I_{\mathcal{C}})/l_2}(X) := \arg\min_Z\left(\frac{1}{l_2}(g_1(Z) + I_{\mathcal{C}}(Z)) + \frac{1}{2}\|Z - X\|_F^2\right).$$

The proximal operator handles the nonsmooth components, including the sparsity constraint and nonnegativity projection. Once $(S^{(\ell+1)}, Z^{(\ell+1)})$ is obtained, LS-1 evaluates whether it provides sufficient descent according to an Armijo-type condition:

$$F(S^{(\ell+1)}, Z^{(\ell+1)}) \leq c_\ell - \delta(\|S^{(\ell+1)} - S^{(\ell)}\|_F^2) + \|Z^{(\ell+1)} - Y^{(\ell)}\|_F^2) \tag{8}$$

If this inequality is satisfied, then the extrapolated update is accepted by setting $G^{(\ell+1)} = Z^{(\ell+1)}$, and the algorithm proceeds to the next iteration via the fast track.

2.3.4 Line Search Algorithm 2 (LS-2). When the condition in Eq (8) is repeatedly violated even after reducing the extrapolation weight, the algorithm switches to a secondary back-tracking scheme without extrapolation. In this case, the update is recomputed directly from the current iterate $G^{(\ell)}$, using an alternative gradient-proximal update.

The step sizes l_1 and l_2 are first estimated by:

$$l_1 = \frac{\langle S^{(\ell)} - S^{(\ell-1)}, S^{(\ell)} - S^{(\ell-1)}\rangle}{\langle S^{(\ell)} - S^{(\ell-1)}, \nabla_S f(S^{(\ell)}, G^{(\ell)}) - \nabla_S f(S^{(\ell-1)}, G^{(\ell)})\rangle}, \tag{9a}$$

$$l_2 = \frac{\langle G^{(\ell)} - G^{(\ell-1)}, G^{(\ell)} - G^{(\ell-1)}\rangle}{\langle G^{(\ell)} - G^{(\ell-1)}, \nabla_G f(S^{(\ell)}, G^{(\ell)}) - \nabla_G f(S^{(\ell)}, G^{(\ell-1)})\rangle} \tag{9b}$$

With these step sizes, the new candidate pair $(S^{(\ell+1)}, V^{(\ell+1)})$ is computed as:

$$S^{(\ell+1)} = S^{(\ell)} - \frac{1}{l_1}\nabla_S f(S^{(\ell)}, G^{(\ell)}), \tag{10}$$

$$V^{(\ell+1)} = \mathbf{prox}_{(g_1+I_{\mathcal{C}})/l_2}\left(G^{(\ell)} - \frac{1}{l_2}\nabla_G f(S^{(\ell)}, G^{(\ell)}) + \frac{1}{l_2}W(G^{(\ell)})\right) \tag{11}$$

The proximal operator handles the nonsmooth sparsity and nonnegativity constraints, as in Eq (7). LS-2 then checks the same Armijo-type descent condition:

$$F(S^{(\ell+1)}, V^{(\ell+1)}) \leq c_\ell - \delta(\|S^{(\ell+1)} - S^{(\ell)}\|_F^2) + \|V^{(\ell+1)} - G^{(\ell)}\|_F^2) \tag{12}$$

If this condition is not satisfied, the step sizes l_1 and l_2 are scaled by a factor $\eta > 1$ and clipped within prescribed bounds:

$$\begin{aligned} l_1 &\leftarrow \min\left\{\max\{l_{\min}^1, \eta l_1\},\ l_{\max}^1\right\}, \\ l_2 &\leftarrow \min\left\{\max\{l_{\min}^2, \eta l_2\},\ l_{\max}^2\right\}, \end{aligned} \tag{13}$$

and the updates in Eqs (10)–(11) are recomputed accordingly. The process is repeated until the condition in Eq (12) is met, ensuring convergence even when extrapolation fails.

2.3.5 Final updates and convergence. After each iteration, the algorithm updates the weights $\omega_{ij}^{(\ell+1)}$ based on the reconstruction error to down-weight unreliable data sources, and refreshes the FISTA parameter $\theta^{(\ell+1)}$ along with the non-monotone envelope $c_{\ell+1}$.

Algorithm 1 APDCA Accelerated Proximal DC Algorithm.

Initialize $G_i^0, G_i^1, S_i^1, Z_i^1$
Compute Y by Eq (4)
Compute $(S^{(\ell+1)}, Z^{(\ell+1)})$ by LS-1.
If Eq (8) is satisfied **then**

$$G^{(\ell+1)} = Z^{(\ell+1)}.$$

Else Compute $(S^{(\ell+1)}, V^{(\ell+1)})$ by LS-2 and set

$$(S^{(\ell+1)}, G^{(\ell+1)}) = \begin{cases} (S^{(\ell+1)}, Z^{(\ell+1)}), & \text{if } F(S^{(\ell+1)}, Z^{(\ell+1)}) \\ & \leq F(S^{(\ell+1)}, V^{(\ell+1)}), \\ (S^{(\ell+1)}, V^{(\ell+1)}), & \text{otherwise.} \end{cases}$$

End (If)
$\omega_{ij}^{(\ell+1)} = 1/(2\|R_{ij} - G_i^{(\ell+1)} S_{ij}^{(\ell+1)} (G_j^{(\ell+1)})^T\|_F^2)$
$\theta^{(\ell+1)} = \dfrac{\sqrt{4(\theta^{(\ell)})^2 + 1} + 1}{2}.$
$q_{\ell+1} = \tau q_\ell + 1.$
$c_{\ell+1} = \dfrac{\tau q_\ell c_\ell + F(S^{(\ell+1)}, G^{(\ell+1)})}{q_{\ell+1}}.$
Replace ℓ by $\ell + 1$ and go to step 2.

Algorithm 2 LS-1 algorithm (Compute $(S^{(\ell+1)}, Z^{(\ell+1)})$ with line search).

Initialize l_1, l_2, η
Update l_1, l_2 by Eq (5a) and Eq (5b)
Compute $S^{(\ell+1)}$ by Eq (6) and $Z^{(\ell+1)}$ by Eq (7)
Repeat until Eq (8) is satisfied.
Replace l_1 , l_2 by Eq (13), compute $S^{(\ell+1)}$, $Z^{(\ell+1)}$
end (Repeat)
$l_1^{(\ell)} = l_1$ and $l_2^{(\ell)} = l_2$.

Algorithm 3 LS-2 algorithm (Compute $(S^{(\ell+1)}, V^{(\ell+1)})$ with line search).

Initialize l_1, l_2, η
Update l_1, l_2 by Eq (9a) and Eq (9b)
Compute $S^{(\ell+1)}$ by Eq (10) and $V^{(\ell+1)}$ by Eq (11)
Repeat until Eq (12) is satisfied.
Replace l_1 , l_2 by Eq (13), compute $S^{(\ell+1)}$, $V^{(\ell+1)}$
end (Repeat)
$l_1^{(\ell)} = l_1$ and $l_2^{(\ell)} = l_2$.

2.4 Convergence proof

In this section we introduce the essential logic of convergence proof for the algorithm above, the detailed version see S1 Text.

First, the back-tracking criteria embedded in LS-1 and LS-2 create a descent envelope c_ℓ that is quasi-Fejér monotone: every accepted update decreases c_ℓ by at least a multiple of the squared iterate gap. Second, because the smooth term f is coercive while the regularization g_1 and the non-negativity indicator I_C confine $G^{(\ell)}$ to a closed cone, all iterates reside in a compact level set of F.

With boundedness secured, closedness of the subdifferential mapping $G \mapsto \partial(g_1 + I_C)(G)$ permits passage to the limit in the optimality condition of each proximal step. Hence every accumulation point $(S^\star, G^\star)$ obeys the stationarity condition $0 \in \partial F(S^\star, G^\star)$. Whether the accepted updates come from LS-1 (on extrapolated points) or LS-2 (on current points) is immaterial. The argument treats the two cases symmetrically and covers all subsequences.

If the objective F moreover satisfies the Kurdyka–Łojasiewicz (KL) inequality—which is automatic for the semi-algebraic penalties used in Section 1—the standard DC-programming framework further yields full sequence convergence together with an explicit sub-linear rate.

3 Results and discussions

3.1 Experimental setup

To investigate the performance of APDCA, we consider six core biological object types: RBPs (Type 1), miRNAs (Type 2), genes (Type 3), AS events (Type 4), diseases (Type 5), and drugs (Type 6). We integrated nine distinct relational data sources from public biological databases, covering both inter-relational and intra-relational associations. The relation between different entities is shown in Fig 2.

For RBP-related associations, we collected gene expression data of 1,532 RBPs from a recent census on BRCA samples [32,33] and obtained AS event features by integrating junction reads from The Cancer Genome Atlas with the Mixture-of-Isoforms database [19,34,35]. Associations between RBPs and AS events were established based on Pearson correlation analysis, with more detailed preprocessing procedures provided in S2 Text. This yielded three binary matrices describing relationships among RBPs, genes, and AS events: RBP-genes (R_{13}), RBP–AS events (R_{14}) and gene–AS events (R_{34}).

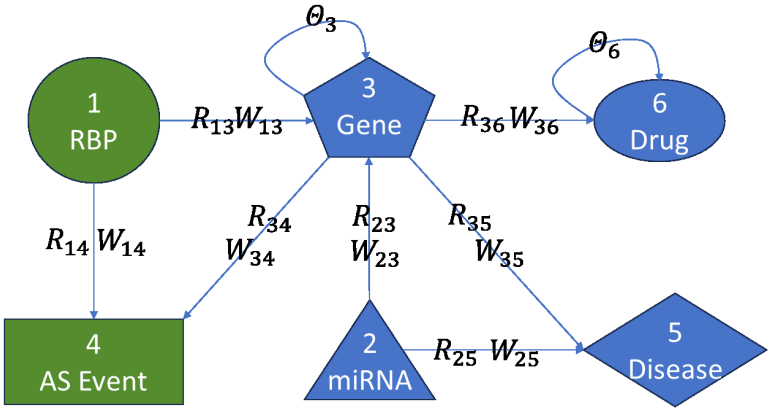

Fig 2. Schematic illustration of the multi-type biological network used in APDCA. Six types of biological entities are represented: RNA-binding proteins (RBPs), miRNAs, genes, alternative splicing (AS) events, diseases, and drugs. Edges between entities denote observed associations R_{ij} with corresponding weights W_{ij}. Self-loop parameters θ_3 and θ_6 are used for gene and drug regularization, respectively.

https://doi.org/10.1371/journal.pcbi.1013665.g002

In addition, we incorporated miRNA–gene and miRNA–disease associations from miRTarBase [36] and HMDD [37] respectively; gene–disease and gene–drug associations from DisGeNET [38] and DrugBank [39]; and gene–gene and drug–drug interactions from BioGrid [40] and DrugBank [39]. All datasets were retrieved from the latest versions of these databases.

We perform 5-fold cross-validation on the RBP–AS event association matrix (R_{14}). In each fold, the known associations were partitioned into training and testing sets along the row dimension, while all other relational matrices remained unchanged throughout the evaluation.

3.2 Performance comparison with existing methods

To evaluate the predictive performance of APDCA, we conduct experiments on benchmark datasets using 5-fold cross-validation, and compared it against six representative models: SDLDA [16], LDAP [17], GAERF [18], MLFDA [20], SIMCLDA [21], and GrwLDA [22]. Performance is assessed using two widely adopted metrics—AUC (area under the ROC curve) and AUPR (area under the Precision-Recall curve). As shown in Fig 3, APDCA consistently achieves the best results across both metrics. This reflects its ability to integrate heterogeneous biological relationships in a unified optimization framework while adaptively filtering noise via sparse constraints and low-rank factorization. Such a structure enhances both discriminative capability and generalization under sparse label scenarios.

In terms of AUC performance, APDCA obtains the highest score of 0.9784, outperforming all baselines. The ROC curve in Fig 3(a) shows that the closest result comes from GAERF (0.9642), followed by GrwLDA (0.9063) and MLFDA (0.8928). In contrast, SIMCLDA and LDAP exhibit relatively poor performance, with AUCs of 0.6923 and 0.5942, respectively. The strong AUC of GAERF can be attributed to its graph-enhanced learning structure, but its reliance on static graph construction may limit adaptability across data types. GrwLDA, based on topic-aware random walks, captures global associations but lacks precision in local structure modeling. SIMCLDA, despite using inductive matrix completion, is constrained by Gaussian kernel similarity and fixed feature propagation, making it less effective in multi-relational tasks. In contrast, APDCA dynamically learns embedding spaces with joint low-rank representations, enabling more expressive modeling across different biological object types.

The AUPR comparison further highlights APDCA's robustness under label imbalance. As shown in the Precision-Recall curve in Fig 3(b), APDCA achieves an AUPR of 0.8166, significantly higher than SDLDA (0.7503), MLFDA (0.6861), and

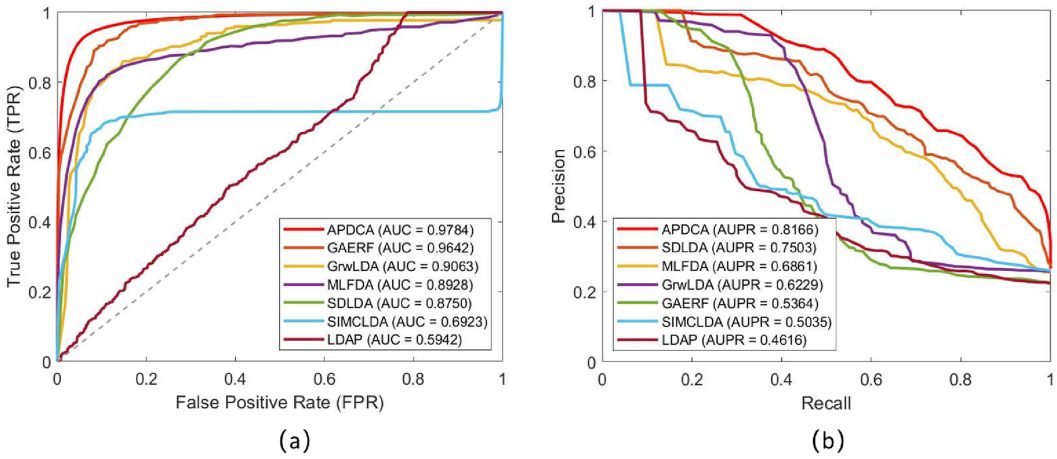

Fig 3. Comparison of predictive performance under 5-fold cross-validation. (a) ROC curves with AUC values. (b) Precision-Recall curves with AUPR values. APDCA achieves the highest scores across both metrics.

https://doi.org/10.1371/journal.pcbi.1013665.g003

GrwLDA (0.6229). GAERF, despite its strong AUC, drops sharply to 0.5364 in AUPR, indicating a large number of false positives. LDAP and SIMCLDA also yield low AUPRs of 0.4616 and 0.5035, respectively, likely due to their limited capacity to prioritize true positives under sparse supervision. The weak AUPR of SDLDA suggests sensitivity to hyperparameter tuning and difficulty in handling noisy unlabeled samples. In contrast, APDCA's proximal DC framework and envelope update mechanism promote stable convergence and strong discriminative structure, allowing it to maintain high precision and recall even in highly imbalanced settings.

3.3 Algorithm convergence speed

We further assess the convergence speed of APDCA by comparing it with SIMCLDA, MLFDA, and SDLDA to show its accelerated convergence. The number of iterations is set to 100 for all methods. To evaluate convergence, we use relative error computed via max–min normalization. As shown in Fig 4, APDCA demonstrates the fastest convergence, with its relative error dropping below 0.05 within the first 10 iterations and reaching zero around the 20th iteration. SIMCLDA also shows a steady convergence trend, although at a much slower rate. In contrast, the convergence curve of SDLDA is unstable and exhibits noticeable oscillations, making it less reliable. MLFDA converges more smoothly than SDLDA but still requires more iterations than APDCA. These results suggest that APDCA not only achieves convergence more quickly but also does so more consistently.

3.4 Case studies of APDCA

To assess the biological validity of the predicted associations from APDCA, We report the top 100 pairs of RBPsand AS events associations during EMT which are illustrated by Fig 5. Fig 5 shows that many RBPs form groups to regulate AS events. We further conducted a case study on the RNA-binding protein QKI, which has been implicated in regulating alternative splicing during epithelial–mesenchymal transition (EMT) and cancer progression [9]. We extracted the top 10 AS events most strongly associated with QKI as predicted by our model. The OncoSplicing database was used for evaluation, as it aggregates clinically relevant AS events across 33 human cancers, many of which are closely related to EMT progression, and provides regulatory support including eCLIP peaks, motif mappings, and co-expression data [26].

Among the top 10 predicted QKI–AS event pairs shown in Table 2, 7 out of 10 were supported by at least one type of regulatory evidence in OncoSplicing. Specifically, 6 AS events showed eCLIP peak support, such as IST1 with 9 peaks

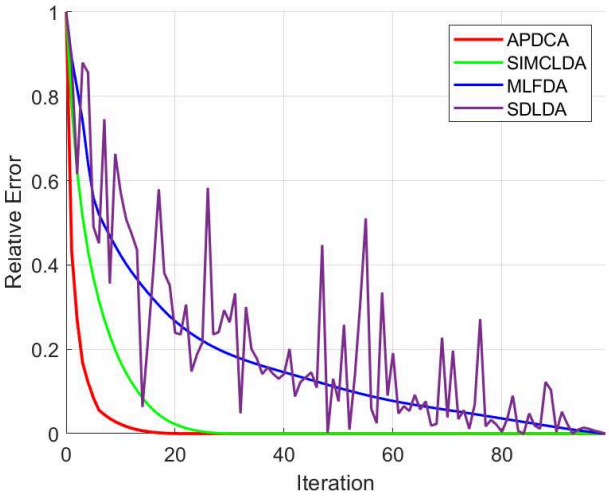

Fig 4. Convergence comparison of APDCA, SIMCLDA, MLFDA, and SDLDA over 100 iterations using relative error as the evaluation metric.

https://doi.org/10.1371/journal.pcbi.1013665.g004

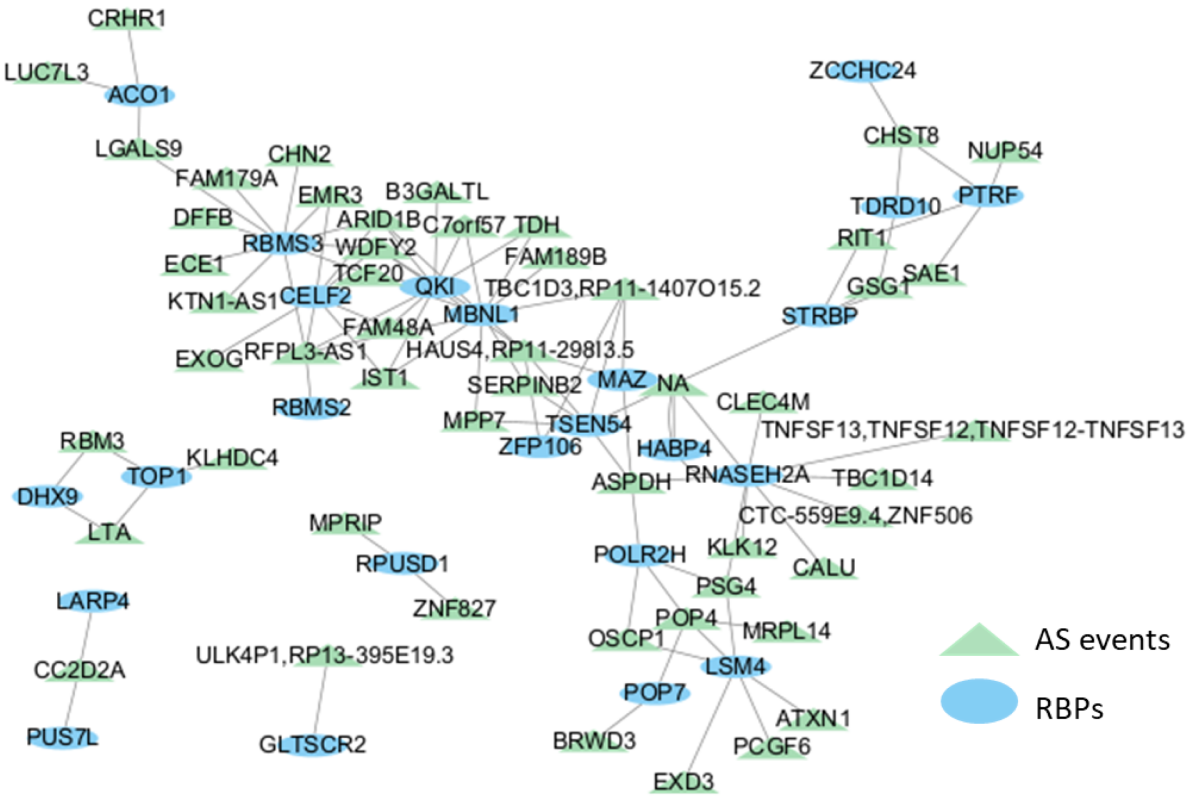

Fig 5. Network showing the RBP–AS event associations during EMT.

https://doi.org/10.1371/journal.pcbi.1013665.g005

and ARID1B with 5 peaks, and 5 events exhibited significant co-expression with QKI in cancer samples, such as IST1 with correlations observed in 5 cancer types. In terms of regulatory confidence, 2 AS events IST1 and ARID1B were labeled as high confidence (high rate), while several others were labeled mild based on the integrated evidence. Only three events AS1, FAM48A, and TDH had no evidence recorded in the current database release. As a representative example, the association between QKI and the exon skipping event in IST1 is supported by 9 eCLIP peaks and correlated expression in 5 distinct TCGA cancer types. This multi-modal support strongly indicates a regulatory role of QKI over this AS event, consistent with prior biological studies. Overall, the validation results suggest that APDCA is capable of prioritizing biologically meaningful RBP AS pairs.

We also evaluated APDCA on an independent lncRNA–disease dataset to further assess its generalizability, and the detailed results are provided in S3 Text.

3.5 Parameter analysis

To assess the robustness and stability of APDCA, we performed parameter sensitivity analyses on the regularization parameters λ and λ_G, as well as the sparsity control parameter k. We aim to evaluate how different settings affect model performance, measured by AUC.

3.5.1 Effect of λ and λ_G. We first explored the impact of the regularization parameters λ and λ_G with 5-fold cross-validation, which control the low-rank representation and sparsity constraints, respectively. A grid search is performed over $\lambda, \lambda_G \in 1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$, and the AUC values for each parameter pair are summarized in Fig 6.

Table 2. Top 10 QKI–AS Event Associations Predicted by APDCA and Their Evidence in OncoSplicing.

AS Event	Peaks	Corr Cancer	Rate
IST1	9	5	High
AS1	–	–	–
ARID1B	5	3	High
TCF20	3	2	Mild
HAUS4	4	1	Mild
WDFY2	1	0	Mild
FAM48A	–	–	–
TDH	–	–	–
C7orf57	0	0	Mild
B3GALTL	1	0	Mild

https://doi.org/10.1371/journal.pcbi.1013665.t002

Table 3. AUC values under different combinations of regularization parameters λ and λ_G.

$\lambda_G \backslash \lambda$	**1**	**10^{-1}**	**10^{-2}**	**10^{-3}**	**10^{-4}**	**10^{-5}**
1	0.8523	0.8520	0.8518	0.8586	**0.9784**	0.9784
10^{-1}	0.8521	0.8518	0.8554	0.9695	0.9695	0.9695
10^{-2}	0.8517	0.8516	0.9666	0.9695	0.9694	0.9695
10^{-3}	0.8521	0.8514	0.9694	0.9694	0.9694	0.9694
10^{-4}	0.8519	0.8521	0.9695	0.9694	0.9695	0.9694
10^{-5}	0.8521	0.8514	0.9694	0.9695	0.9695	0.9694

https://doi.org/10.1371/journal.pcbi.1013665.t003

As shown in Table 3, the highest AUC value of 0.9784 is achieved when $\lambda = 0.0001$ and $\lambda_G = 1$. And when λ is within the range of 10^{-2},10^{-3},10^{-4},10^{-5} and λ_G is within the range of 10^{-2},10^{-3},10^{-4},10^{-5}, the model consistently obtains high AUC values above 0.969. In contrast, when both parameters are set above 0.01, the AUC drops to around 0.85, indicating a clear degradation in performance. Based on these results, we recommend choosing $\lambda \in 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$ and $\lambda_G \in 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$ to ensure strong and stable model performance.

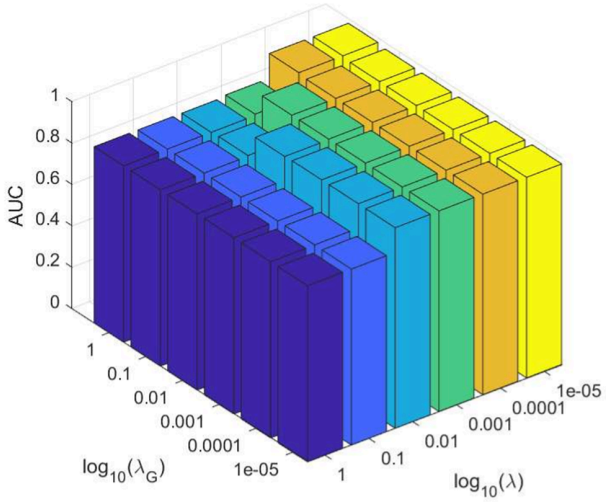

Fig 6. AUC surface with respect to regularization parameters λ and λ_G. Performance improves as both parameters decrease and stabilizes at low values.

https://doi.org/10.1371/journal.pcbi.1013665.g006

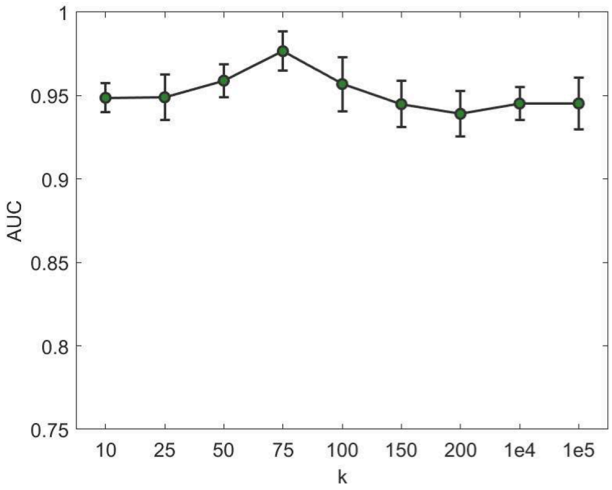

Fig 7. AUC sensitivity of APDCA to the sparse control parameter *k* under 5-fold cross-validation. Performance peaks at $k = 75$ and remains stable across a wide range.

https://doi.org/10.1371/journal.pcbi.1013665.g007

3.5.2 Effect of *k*. We conducted a sensitivity analysis on the sparse control parameter k, which regulates the sparsity level of the reconstructed relationship matrix in our APDCA framework. As shown in Fig 7, we varied k over a wide range from 10 to 10^5 and evaluated the model performance using the AUC metric under 5-fold cross-validation.

The results demonstrate that the performance of APDCA remains robust across a broad range of k values. Specifically, AUC values consistently exceed 0.94 for all settings, with the highest performance observed around $k = 75$. This suggests that moderate sparsity levels provide an optimal balance between noise suppression and information preservation in the reconstructed association matrix. When k is too small, important associations may be missed, while excessively large k may introduce redundant or noisy connections, slightly degrading performance.

Overall, this analysis confirms that APDCA is not overly sensitive to the choice of k, and a reasonably chosen k (e.g., 50–100) yields stable and high-quality predictions.

4 Conclusions

Identifying regulatory associations between RNA-binding proteins (RBPs) and alternative splicing (AS) events during epithelial–mesenchymal transition (EMT) is essential for understanding cell state transitions and cancer metastasis. However, experimental validation remains costly and context-dependent. To address this, we proposed APDCA (Accelerated Proximal DC Algorithm), a novel computational method designed to predict RBP–AS associations during EMT. APDCA integrates heterogeneous biological data using a sparse low-rank matrix factorization framework. A sparsity constraint is applied to reflect the selective nature of RBP–AS regulation. The resulting non-convex problem is reformulated via Difference-of-Convex (DC) programming and solved using a proximal optimization scheme with Nesterov-accelerated updates, which improves both convergence speed and stability. Experimental results show that APDCA outperforms six baseline models in AUC and AUPR across 5-fold cross-validation. A case study involving the EMT-related RBP QKI demonstrates the biological relevance of the predicted associations, supported by external evidence from the OncoSplicing database. While APDCA achieves strong predictive performance, its effectiveness depends on the quality and completeness of input data. Beyond this specific application, APDCA offers a generalizable optimization framework that can be readily extended to other molecular association prediction tasks, such as lncRNA–disease or miRNA–disease relationships. Future work will incorporate dynamic transcriptomic datasets and extend the model to additional types of

AS events. Moreover, integrating APDCA with deep learning architectures or probabilistic inference schemes could further enhance its scalability and interpretability, broadening its impact across multi-omics integration and systems biology research.

Supporting information

S1 Text. Convergence proof of APDCA.
(PDF)

S2 Text. Data preprocessing procedures.
(PDF)

S3 Text. Evaluation of APDCA on lncRNA–disease prediction.
(PDF)

Author contributions

Conceptualization: Yushan Qiu.

Data curation: Yangsong He, Yushan Qiu.

Formal analysis: Yushan Qiu.

Funding acquisition: Zheng-Jian Bai, Quan Zou, Yushan Qiu.

Investigation: Yushan Qiu.

Methodology: Yangsong He, Zheng-Jian Bai, Yushan Qiu.

Project administration: Quan Zou, Yushan Qiu.

Software: Yangsong He.

Supervision: Zheng-Jian Bai, Quan Zou, Yushan Qiu.

Visualization: Yangsong He.

Writing – original draft: Yangsong He, Zheng-Jian Bai, Wai-Ki Ching, Quan Zou, Yushan Qiu.

Writing – review & editing: Yangsong He, Zheng-Jian Bai, Wai-Ki Ching, Quan Zou, Yushan Qiu.

References

1. Chen M, Manley JL. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. Nat Rev Mol Cell Biol. 2009;10(11):741–54. https://doi.org/10.1038/nrm2777 PMID: 19773805
2. Fu X-D, Ares M Jr. Context-dependent control of alternative splicing by RNA-binding proteins. Nat Rev Genet. 2014;15(10):689–701. https://doi.org/10.1038/nrg3778 PMID: 25112293
3. Fredericks AM, Cygan KJ, Brown BA, Fairbrother WG. RNA-binding proteins: splicing factors and disease. Biomolecules. 2015;5(2):893–909. https://doi.org/10.3390/biom5020893 PMID: 25985083
4. Shapiro IM, Cheng AW, Flytzanis NC, Balsamo M, Condeelis JS, Oktay MH, et al. An EMT-driven alternative splicing program occurs in human breast cancer and modulates cellular phenotype. PLoS Genet. 2011;7(8):e1002218. https://doi.org/10.1371/journal.pgen.1002218 PMID: 21876675
5. Thiery JP, Acloque H, Huang RYJ, Nieto MA. Epithelial-mesenchymal transitions in development and disease. Cell. 2009;139(5):871–90. https://doi.org/10.1016/j.cell.2009.11.007 PMID: 19945376
6. Bebee TW, Cieply BW, Carstens RP, et al. Genome-wide activities of RNA-binding proteins that regulate cellular changes in the epithelial to mesenchymal transition. Systems Biology of RNA Binding Proteins. 2014. p. 267–302.
7. Tao Y, Zhang Q, Wang H, Yang X, Mu H. Alternative splicing and related RNA binding proteins in human health and disease. Signal Transduct Target Ther. 2024;9(1):26. https://doi.org/10.1038/s41392-024-01734-2 PMID: 38302461

8. Kelaini S, Chan C, Cornelius VA, Margariti A. RNA-binding proteins hold key roles in function, dysfunction, and disease. Biology (Basel). 2021;10(5):366. https://doi.org/10.3390/biology10050366 PMID: 33923168
9. Zong F-Y, Fu X, Wei W-J, Luo Y-G, Heiner M, Cao L-J, et al. The RNA-binding protein QKI suppresses cancer-associated aberrant splicing. PLoS Genet. 2014;10(4):e1004289. https://doi.org/10.1371/journal.pgen.1004289 PMID: 24722255
10. Hou J-Y, Wang X-L, Chang H-J, Wang X-X, Hao S-L, Gao Y, et al. PTBP1 crotonylation promotes colorectal cancer progression through alternative splicing-mediated upregulation of the PKM2 gene. J Transl Med. 2024;22(1):995. https://doi.org/10.1186/s12967-024-05793-5 PMID: 39497094
11. Huo L, Jiao Li J, Chen L, Yu Z, Hutvagner G, Li J. Single-cell multi-omics sequencing: application trends, COVID-19, data analysis issues and prospects. Brief Bioinform. 2021;22(6):bbab229. https://doi.org/10.1093/bib/bbab229 PMID: 34111889
12. Sheng N, Huang L, Lu Y, Wang H, Yang L, Gao L, et al. Data resources and computational methods for lncRNA-disease association prediction. Comput Biol Med. 2023;153:106527. https://doi.org/10.1016/j.compbiomed.2022.106527 PMID: 36610216
13. Yan J, Wang R, Tan J. Recent advances in predicting lncRNA-disease associations based on computational methods. Drug Discov Today. 2023;28(2):103432. https://doi.org/10.1016/j.drudis.2022.103432 PMID: 36370992
14. Chen Y, Wang J, Wang C, Liu M, Zou Q. Deep learning models for disease-associated circRNA prediction: a review. Brief Bioinform. 2022;23(6):bbac364. https://doi.org/10.1093/bib/bbac364 PMID: 36130259
15. Park R, Winnicki M, Liu E, Chu W-M. Immune checkpoints and cancer in the immunogenomics era. Brief Funct Genomics. 2019;18(2):133–9. https://doi.org/10.1093/bfgp/ely027 PMID: 30137232
16. Zeng M, Lu C, Zhang F, Li Y, Wu F-X, Li Y, et al. SDLDA: lncRNA-disease association prediction based on singular value decomposition and deep learning. Methods. 2020;179:73–80. https://doi.org/10.1016/j.ymeth.2020.05.002 PMID: 32387314
17. Lan W, Li M, Zhao K, Liu J, Wu F-X, Pan Y, et al. LDAP: a web server for lncRNA-disease association prediction. Bioinformatics. 2017;33(3):458–60. https://doi.org/10.1093/bioinformatics/btw639 PMID: 28172495
18. Wu Q-W, Xia J-F, Ni J-C, Zheng C-H. GAERF: predicting lncRNA-disease associations by graph auto-encoder and random forest. Brief Bioinform. 2021;22(5):bbaa391. https://doi.org/10.1093/bib/bbaa391 PMID: 33415333
19. Qiu Y, Ching W-K, Zou Q. Prediction of RNA-binding protein and alternative splicing event associations during epithelial-mesenchymal transition based on inductive matrix completion. Brief Bioinform. 2021;22(5):bbaa440. https://doi.org/10.1093/bib/bbaa440 PMID: 33517359
20. Adhikari B, Hou J, Cheng J. DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. Bioinformatics. 2018;34(9):1466–72. https://doi.org/10.1093/bioinformatics/btx781 PMID: 29228185
21. Bullock JMA, Thalassinos K, Topf M. Jwalk and MNXL web server: model validation using restraints from crosslinking mass spectrometry. Bioinformatics. 2018;34(20):3584–5. https://doi.org/10.1093/bioinformatics/bty366 PMID: 29741581
22. Gu C, Liao B, Li X, Cai L, Li Z, Li K, et al. Global network random walk for predicting potential human lncRNA-disease associations. Sci Rep. 2017;7(1):12442. https://doi.org/10.1038/s41598-017-12763-z PMID: 28963512
23. LTH A, Pham DT, et al. The DC programming and DCA revisited with DC models of real-world non-convex optimization problems. Ann Oper Res. 2005;133(1):23–46. https://doi.org/10.1007/s10479-005-2077-0
24. Parikh N. Proximal algorithms. FNT in Optimization. 2014;1(3):127–239. https://doi.org/10.1561/2400000003
25. Nesterov Y. A method for solving the convex programming problem with convergence rate O(1/k2). Sov Math Dokl. 1983;27(2):372–6.
26. Xiao S, Guo S, Han J, Sun Y, Wang M, Chen Y, et al. High-Throughput-Methyl-Reading (HTMR) assay: a solution based on nucleotide methyl-binding proteins enables large-scale screening for DNA/RNA methyltransferases and demethylases. Nucleic Acids Res. 2022;50(2):e9. https://doi.org/10.1093/nar/gkab989 PMID: 34718755
27. Watson GA. Linear best approximation using a class of polyhedral norms. Numer Algor. 1992;2(3):321–35. https://doi.org/10.1007/bf02139472
28. Wu B, Ding C, Sun D, Toh K-C. On the Moreau–Yosida regularization of the vector k-norm related functions. SIAM J Optim. 2014;24(2):766–94. https://doi.org/10.1137/110827144
29. Lenstra JK, Shmoys DB, Tardos É. Approximation algorithms for scheduling unrelated parallel machines. Mathematical Programming. 1990;46(1–3):259–71. https://doi.org/10.1007/bf01585745
30. Li H, Lin Z. Accelerated proximal gradient methods for non-convex programming. Adv Neural Inf Process Syst. 2015. 379–87.
31. Tono K, Takeda A. Efficient DC algorithm for constrained sparse optimization. arXiv prerpint 2017. https://doi.org/arXiv:1701.08498
32. Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Xiao R, et al. A large-scale binding and functional map of human RNA-binding proteins. Nature. 2020;583(7818):711–9. https://doi.org/10.1038/s41586-020-2077-3 PMID: 32728246
33. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nat Rev Genet. 2014;15(12):829–45. https://doi.org/10.1038/nrg3813 PMID: 25365966
34. Qiu Y, Lyu J, Dunlap M, Harvey SE, Cheng C. A combinatorially regulated RNA splicing signature predicts breast cancer EMT states and patient survival. RNA. 2020;26(9):1257–67. https://doi.org/10.1261/rna.074187.119 PMID: 32467311
35. Shen S, Park JW, Lu Z, Lin L, Henry MD, Wu YN, et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc Natl Acad Sci U S A. 2014;111(51):E5593-601. https://doi.org/10.1073/pnas.1419161111 PMID: 25480548
36. Huang H-Y, Lin Y-C-D, Li J, Huang K-Y, Shrestha S, Hong H-C, et al. miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. Nucleic Acids Res. 2020;48(D1):D148–54. https://doi.org/10.1093/nar/gkz896 PMID: 31647101

37. Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, et al. HMDD v3.0: a database for experimentally supported human microRNA-disease associations. Nucleic Acids Res. 2019;47(D1):D1013–7. https://doi.org/10.1093/nar/gky1010 PMID: 30364956

38. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017;45(D1):D833–9. https://doi.org/10.1093/nar/gkw943 PMID: 27924018

39. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Research. 2017;46(D1):D1074–82. https://doi.org/10.1093/nar/gkx1037

40. Oughtred R, Rust J, Chang C, Breitkreutz B-J, Stark C, Willems A, et al. The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. 2021;30(1):187–200. https://doi.org/10.1002/pro.3978 PMID: 33070389