arXiv:2611.31227v1 [cs.LG] 23 Nov 2026

线性方法的能力边界在哪里？BiPert：单细胞扰动建模的统一框架

Abhishek Sharma[1] 和 Aleksej Zelezniak[1,2,3]

[1]兰德尔细胞与分子生物物理中心，伦敦国王学院，英国伦敦[2]生命科学系，查尔姆斯理工大学，瑞典哥德堡[3]生物技术研究所，生命科学中心，维尔纽斯大学，立陶宛维尔纽斯

abhishek.1.sharma@kcl.ac.uk，aleksej.zelezniak@kcl.ac.uk

摘要

了解基因扰动如何影响细胞行为，对于研究基因功能和调控机制至关重要。大规模CRISPR筛选能够测量数千个基因的这些效应，但覆盖的细胞环境相对有限。因此，预测未见细胞类型中的扰动响应，并根据转录效应识别受扰动基因，仍然颇具挑战。在此，我们表明，这两项任务均可通过将转录响应线性分解为共享成分和扰动特异性成分来解决。基于这一原理，我们提出 BiPert，这是一种用于正向扰动预测和逆向识别的简洁线性框架。在正向预测方面，BiPert 在未见细胞系中预测基因表达变化时，优于不使用外部生物学知识的现有方法。其小样本扩展 BiPert-FS 仅需十个已测量的扰动即可适配新细胞系。在逆向识别方面，BiPert-ID 能够根据单细胞表达谱识别受扰动基因，在未见细胞系和供体中的表现优于深度学习方法。我们进一步表明，即使控制扰动效应大小，在不同细胞环境间效应迁移更一致的扰动也更容易被识别。总之，这些结果揭示了扰动迁移与识别之间的联系，并证明简单的线性模型能够为这两项任务提供强大且易于解释的基线。

关键词： Perturb-seq；遗传扰动预测；扰动识别；线性模型；低秩分解；跨细胞系迁移；可解释模型

1 引言

大规模扰动筛选如今能够测量数千种遗传扰动如何改变细胞的状态[1–3]。对这些变化进行建模有两个互补的方向。在

前向方向中，给定某种扰动在此前研究过的细胞系中的效应，我们旨在预测其在新细胞环境中的效应[4,5]。在逆向方向中，给定一个观测到的细胞状态，我们旨在识别最有可能产生该状态的扰动[6,7]。第一个问题有助于将实验知识迁移到尚未进行分析的细胞系中，从而有望减少额外扰动筛选的需求。第二个问题则通过将此前无法解释的细胞状态与已知扰动响应相匹配，帮助解释这些状态。一个关键观察是，尽管这两项任务的预测目标不同，但都需要理解扰动响应的哪些方面在不同细胞环境中具有共性，哪些方面取决于具体环境。

许多扰动会诱导共同的细胞响应，同时保留扰动特异性成分[5,8]。现有方法利用了这一结构的某些方面。例如，COMPASS[5] 将扰动效应分解为一个细胞系特异性方向和一个扰动特异性残差，从而能够迁移到

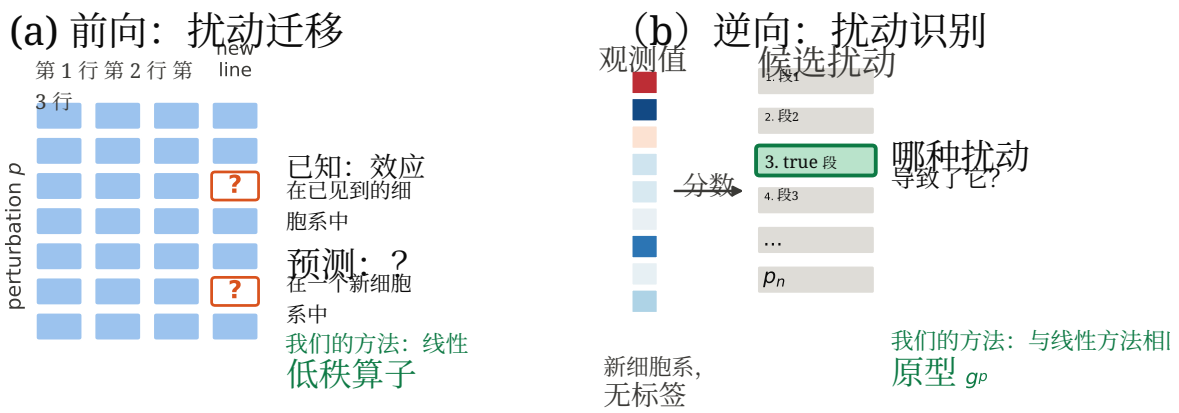

图1。一种线性分解，两个方向。两项任务都基于对扰动效应的一种分解：一部分是在细胞系之间共享的，另一部分则是扰动特有的。（a）前向：根据测量过该扰动的细胞系，预测其在新细胞系中的效应（第4.1节）。（b）逆向：在新细胞系中，根据一个观测到的 State 判断其背后的扰动，且不使用来自该细胞系的标签（第4.2节）。

未见细胞系。尽管 COMPASS 表明扰动效应包含可迁移的结构，但其分解假设细胞环境通过单一共享方向改变这些效应，而其余扰动特异性成分在不同细胞系间保持不变。这一假设限制较大：不同细胞环境可能以不同程度调节扰动响应的不同成分。因此，我们将这一分解推广为低秩算子，使每个细胞系都能对共享扰动表示应用自身的线性变换。

为前向任务设计的模型无法解决逆问题[7]。对于逆向识别[6,7]，我们利用相同的分解，从训练细胞系中构建扰动特异性原型。通过移除共享细胞响应，并将单个细胞与这些原型进行匹配，我们得到了一种简单的免训练线性方法，用于识别未见细胞系中的扰动。

重要的是，我们发现，跨细胞系可迁移性不仅与预测扰动效应有关，还连接了前向任务与逆任务。仅根据训练细胞系估算的迁移分数，就能有力地预测某种扰动能否在此前未见过的细胞系中被识别。反过来，在一项独立筛选中，我们表明，在某种细胞类型内的可识别性会限制其向其他细胞类型的迁移。综合来看，这些结果表明，共同的线性结构既能支持扰动效应的迁移，也能支持从细胞状态中识别扰动效应。

我们的贡献主要有三点：

1. **扰动迁移与识别之间的双向联系。**仅在训练细胞系上测得的可迁移性，就能预测 BiPert-ID 和深度网络在留出细胞系中的识别能力；反过来，在一项独立筛选中，某种细胞类型内的可识别性会限制向其他细胞类型的迁移。

2. **用于扰动迁移的低秩模型。**我们将跨细胞系迁移建模为结构化矩阵补全问题，并提出低秩算子 BiPert，以推广 COMPASS；同时提出其小样本扩展 BiPert-FS，用于在测量值有限的情况下适配新细胞系。

3. **免训练的扰动识别。** 我们将共享分解扩展至逆向识别，并提出免训练线性原型模型 BiPert-ID。该模型在未见细胞系上的表现优于当前最先进的方法。

2 相关工作

前向：用于扰动效应预测的深度模型。 大多数遗传扰动效应预测方法都采用深度网络。CPA 将细胞编码到潜在空间，并为每种扰动添加一个学习得到的嵌入[9]；GEARS 在已知基因关系图上传播扰动，以覆盖未见基因及其组合[10]；在数百万个细胞上预训练的 Transformer scGPT 会针对同一任务进行微调[11]；GenePert 根据语言模型生成的基因嵌入回归预测效应[12]；表格基础模型 TabICL 利用外部基因知识及六组先验基因特征（涵盖蛋白质序列、文本、形态、蛋白质相互作用和基因依赖性），在给定情境下预测效应[13,14]；TxPert 整合多个生物学知识图谱，以预测训练分布之外的效应[15]；State 是一种在多种细胞情境下训练的 Transformer，旨在实现这些情境之间的迁移[4]。其中，只有 State 和 TxPert 评估了向训练时未纳入的细胞系进行迁移的能力。

线性基线。 越来越多的研究开始质疑单细胞基础模型预测扰动响应的有效性。Ahlmann-Eltze、Huber 和 Anders 证明，在训练扰动上拟合的线性模型，对于留出的扰动效应，其表现可以媲美甚至超过深度模型和基础模型[16], ；更广泛的单细胞基础模型基准测试也得出了相同结论[17,18]。COMPASS 将这一线性视角扩展到跨细胞系迁移，通过将扰动 p 在细胞系 ℓ中的效应分解为细胞系特异性方向，以及跨细胞系共享的扰动表示（第 4节）。这一共享表示能够预测未见细胞系中的结果，在该设置下优于 GEARS、scGPT 和 State[5]。一项同期的评估研究将每个响应分解为跨环境共享的部分和环境特异性部分，并发现当前的预测器主要只能恢复前者[19]；该研究没有提出模型，也未研究逆任务。PerturbMap[20] 拟合单个筛选实验中不同处理条件之间的线性映射，以预测留出的扰动，但并未留出细胞系。我们从两个方向拓展 COMPASS 的分解方法：首先，将其推广为低秩算子形式[21]，该形式将 COMPASS 作为特例，并且仅凭最少 10 个测量值即可适应新细胞系；其次，将该框架扩展到逆问题，即根据产生的细胞状态识别扰动。我们的少样本模型建立在子空间对齐[22] 以及向源模型进行正则化的基础上[23,24]。其思想与几何矩阵补全相近，后者利用侧信息[25,26]。

因子模型。因子模型将多个数据集分解为共享的和数据集特有的潜在因子。JIVE[21] 将联合与个体变异分离，而 MOFA 和 MOFA+[27, 28] 则将因子分析扩展到多个组学层和样本组。引导式稀疏因子分析也通过单个 CRISPR 筛选中的潜在因子对扰动效应进行建模[29]。这些方法主要用于解释观测数据中的变异，通常需要为新数据集额外估计因子或载荷。BiPert-FS 解决的是另一项任务：仅凭少量观测值，预测新细胞系中未测量的扰动响应。它不对新细胞系进行分解，而是保持来源响应程序和平均效应不变，仅学习正则化的仿射校正（公式（7））。该校正在未用于拟合的基因敲低上进行评估。

逆向：从细胞推断扰动。与给定扰动预测细胞响应不同，识别导致观测到的转录状态的扰动这一逆问题，受到的关注相对较少。RNA 指纹分析[6] 通过将单个细胞与已知扰动相关的转录响应特征进行匹配，来解决这一任务。PROGENy 等足迹方法从大量扰动实验中提取一致的响应特征，并对新的表达谱进行评分，以推断通路活性[30]；BiPert-ID 则将同一原理应用于未见过的细胞系中单细胞的单个基因敲低。ExPert[7], 目前是最先进的方法，采用一种学习式方法：它在多个扰动筛选中训练神经编码器，并通过将细胞表示与学习得到的每基因原型进行比

较，对细胞进行分类。在 ExPert 的留出细胞系基准测试中[7]，为正向预测设计的模型（GEARS、scGPT 和 GenePert）表现接近随机水平，说明预测扰动效应并不一定能识别其成因。然而，这些模型都未利用上文介绍的线系不变扰动表示。

3 背景

扰动效应为何可以分解。 细胞组织的三个特性为第4节中引入的模型提供了依据。

基因在功能模块中发挥作用。 蛋白质在复合物和通路中运作，这意味着扰动同一模块的不同成员可能会破坏共享的细胞功能。与此一致，全基因组规模的扰动筛选表明，属于同一蛋白质复合物的基因的基因敲低会产生强相关的转录效应[1]。这表明扰动效应占据由共享细胞程序构成的低维空间。

不同扰动汇聚于共享响应方向。 不同的基因敲低可以调动重叠的细胞程序，从而产生相似的转录响应。例如，对核心生长机制的扰动会诱发广泛且刻板的响应，而对更特异调控因子的扰动则会产生额外的基因依赖性效应[5]。关键在于，某一扰动调动该共享响应方向的强度在各细胞系之间相对较为保守，其保守程度高于基因特异性剩余部分[5]。这促使我们将保守的扰动组分与细胞系依赖性效应分离开来。

细胞环境决定共享程序如何表达。 不同细胞系表达不同的基因子集，并具有各自独特的活跃调控网络。因此，同一底层细胞程序会因细胞环境不同而影响不同的下游基因，影响强度也有所不同。因此，相关细胞系之间的扰动响应比不相关细胞系之间更相似[2]。这促使我们对细胞系特异性载荷进行建模，以确定共享程序如何转化为基因表达变化。

这些观察共同促使我们采用以下模型分解：共享响应的强度取决于扰动和细胞系；一个保守的扰动特异性成分；以及对共享程序进行细胞系特异性加权。

4 BiPert：线性分解模型

我们考察与同一分解的前向和逆向相对应的两项互补预测任务。对于每项任务，我们先给出背景形式化描述，再介绍一种线性替代方案。实验、结果和消融实验见第5节。两项任务均采用跨细胞系泛化设定：模型在一组细胞系上训练，并在此前未见过的细胞系上评估。前向任务预测目标细胞系中尚未测量的扰动效应，而逆任务则根据由此产生的细胞表达谱识别扰动（图1）。对于每项任务，我们先说明问题和现有方法，再介绍我们的线性模型。

4.1 扰动迁移作为结构化矩阵补全

研究动机。 目前只有少数细胞系提供了扰动筛选数据，而在所有感兴趣的细胞类型中开展全基因组规模的筛选并不现实。因此，我们需要在不同细胞环境之间迁移知识：根据现有细胞系中测得的某种扰动效应，预测该扰动在新细胞系中的效应[5]。

任务。 令$x \in \mathbb{R}^G$表示单个细胞的表达向量，其各个分量量化G个基因的表达水平。对于细胞系ℓ，令$\mathscr{X}_\ell^0$表示对照细胞集合，$\mathscr{X}_{p,\ell}$表示基因$p \in \{1,\dots,P\}$被敲低的细胞集合。我们将细胞系ℓ中扰动p的效应 定义为相应平均表达谱之差：

$$y_{p,\ell} = \frac{1}{|\mathscr{X}_{p,\ell}|}\sum_{x\in\mathscr{X}_{p,\ell}} x - \frac{1}{|\mathscr{X}_\ell^0|}\sum_{x\in\mathscr{X}_\ell^0} x \in \mathbb{R}^G. \tag{1}$$

在源细胞系$y_{p,s}$中，所有扰动的效应均有观测值$s \in \mathscr{S}$。在目标细胞系$\ell^* \notin \mathscr{S}$中，只有部分扰动$P$}的效应可用。目标是预测每个未观测到的扰动$p \notin \mathscr{O}$对应的$y_{p,\ell^*}$。这可以视为具有向量值条目的矩阵补全问题：一个$P \times L$数组，每行对应一个扰动，每列对应$L = |\mathscr{S}| + 1$个细胞系之一，每个单元格中包含一个条目$y_{p,\ell} \in \mathbb{R}^G$，其中目标列仅有部分观测值。

COMPASS[5] 将每个扰动效应分解为沿着细胞系特异性单位向量$u_\ell \in \mathbb{R}^G$的分量，以及一个$g_p \in \mathbb{R}^G$在不同细胞系间共享的残差：

$$y_{p,\ell} \approx \beta_{p,\ell} u_\ell + g_p, \qquad \beta_{p,\ell} = \langle y_{p,\ell}, u_\ell \rangle. \tag{2}$$

这里，$\beta_{p,\ell} \in \mathbb{R}$是一个标量系数，而$u_\ell$表示某个细胞系在不同扰动下的主导共享响应，通过对扰动效应取均值并归一化后估算得出。从生物学角度看，这反映了与细胞生长减少和应激反应基因激活等过程相关的共同应激反应。由于该分量在不同扰动间共享，因此它本身无法区分受到扰动的是哪个基因；扰动特异性信息则由系数$\beta_{p,\ell}$和残差g_p体现。

对于此前未见过的细胞系ℓ^*，COMPASS-X 根据观测到的扰动效应$\{u_{\ell^*}\ y_{p,\ell^*}\}_{p\in\mathscr{O}}$进行估计。随后，它使用源细胞系的平均值预测未观测到的效应 $\hat{}y_{p,\ell^*} \in \mathbb{R}^G$：

$$\hat{y}_{p,\ell^*} = \bar{\beta}_p u_{\ell^*} + \bar{g}_p, \tag{3}$$

其中，$\bar{\beta}_p \in \mathbb{R}$ 和 $\bar{}g_p \in \mathbb{R}^G$分别表示相应量在源细胞系中的平均值。

我们将这一分解推广为：把每个扰动效应表示为低维编码$U_p \in \mathbb{R}^r$与特定于细胞系的载荷矩阵$V_\ell \in \mathbb{R}^{G\times r}$的乘积，其中低维编码在各细胞系间共享；这正是 JIVE[21] 的联合结构：

$$y_{p,\ell} \approx V_\ell U_p. \tag{4}$$

BiPert. 对于新的细胞系ℓ^*，我们通过对观测到的扰动进行最小二乘法拟合，估计其载荷矩阵：

$$V_{\ell^*} = \arg\min_{V\in\mathbb{R}^{G\times r}} \sum_{p\in\mathscr{O}} \|y_{p,\ell^*} - VU_p\|_2^2. \tag{5}$$

共享编码U_p取自秩为r的截断奇异值分解，其对象为$P \times |\mathscr{S}|G$矩阵，该矩阵的第p行汇集了各细胞系中的源细胞系效应$y_{p,s}$。拟合V_{ℓ^*}后，未观测到的扰动效应预测为 $\hat{}y_{p,\ell^*} = V_{\ell^*}U_p$。

COMPASS-X 作为一种特殊情况。 COMPASS-X 可视为该公式的一个受约束实例。假设共享残差具有低秩表示 $\bar{}g_p = Wh_p$，其中 $W \in \mathbb{R}^{G\times r'}$和 $h_p \in \mathbb{R}^{r'}$。那么，它的预测可写为

$$\bar{\beta}_p u_{\ell^*} + \bar{g}_p = \begin{bmatrix} u_{\ell^*} & W \end{bmatrix} \begin{bmatrix} \bar{\beta}_p \\ h_p \end{bmatrix}. \tag{6}$$

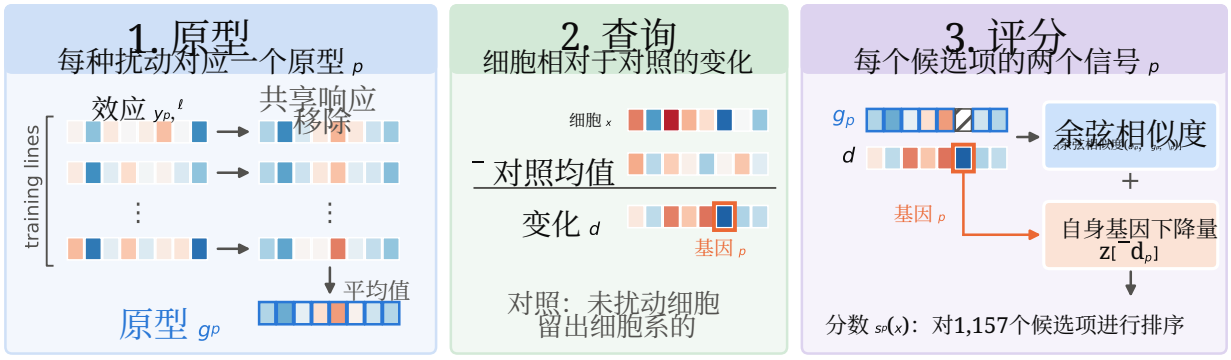

图2. BiPert-ID：我们的免训练线性原型模型。(1) 每种扰动的一个原型g_p 由训练细胞系构建。(2) 查询细胞由其相对于对照的变化d 表示。(3) 每个候选项p 通过余弦相似度加上自身基因下降量（公式8）计算分数。

堆叠载荷矩阵位于 $\mathbb{R}^{G\times(r'+1)}$，堆叠编码位于 $\mathbb{R}^{r'+1}$，因此这对应于秩为 4 的式（$r'+1$），但额外施加了以下约束：所有细胞系共享同一个载荷矩阵 W，且各细胞系的差异仅限于单一方向 u_ℓ。BiPert 不受此限制，允许每个细胞系为共享扰动编码分配自己的权重。

BiPert-FS：我们的小样本扩展。当目标细胞系中观测到的扰动数量较少时，拟合完整的载荷矩阵 V_{ℓ^*} 会变得不可靠。训练细胞系的平均响应为预测新细胞系提供了可靠的起点。因此，我们不从头学习其响应，而是将目标响应建模为对这一共享基线的低维修正。具体而言，我们引入一种受约束迁移模型，将源细胞系效应的均值，

$$\bar{y}_p = |\mathscr{S}|^{-1} \sum_{s\in\mathscr{S}} y_{p,s} \in \mathbb{R}^G,$$

通过低维基 $\Phi \in \mathbb{R}^{G\times k}$映射到目标细胞系，该基由源效应[22]中最大的奇异向量构成：

$$\hat{y}_{p,\ell^*} = \underbrace{\bar{y}_p}_{\text{source average}} + \underbrace{\Phi b}_{\text{shift}} + \underbrace{\Phi(C - I_k)\Phi^\top \bar{y}_p}_{\text{rescale and mix programs}} . \tag{7}$$

其中，Φ 具有正交归一列，I_k 表示 k ×k 单位矩阵。C 的对角元素用于缩放各个响应程序，而非对角元素则刻画程序间的相互作用；b 提供加性偏移。因此，令 C =I_k 且 b = 0，即可还原源平均值。我们通过对观测配对组 ($\bar{y}_p$, y_{p,ℓ^*}) 进行岭回归来估计 b $\in \mathbb{R}^k$和 C $\in \mathbb{R}^{k\times k}$，其中 p $\in \mathscr{O}$，并将 C 正则化至单位矩阵[23,24]（推导见附录 A）。这种参数化方法在低维子空间内学习仿射变换，同时保持正交分量不变。当 k = 30时，该模型仅有 k +k^2 = 930 个参数，因此适用于小样本迁移。

4.2 BiPert-ID：单细胞扰动识别

动机。逆任务旨在识别导致观测到的细胞状态的扰动[7]。在疾病分析等应用中，细胞以改变后的状态被观测到，但产生该状态的扰动尚不明确。差异表达能够捕捉下游响应，却无法直接确定其原因。扰动图谱提供了一个包含已知扰动及其细胞响应的参考，使未知状态能够与其最可能的原因相匹配。现有方法通常在单个筛选中构建这一参考

，并报告扰动效应在不同细胞环境之间的迁移能力有限[6,7]。我们的分解方法通过明确建模跨细胞系迁移来解决这一局限。

任务。 给定受扰动细胞的表达向量$x \in \mathbb{R}^G$，目标是识别被敲低的基因。我们使用 ExPert 基准[7]，，该基准会将整个细胞系留出。模型在其余细胞系的扰动筛选数据上训练，并在一个留出的细胞系上评估，因此这是一项零样本跨域识别任务。

我们的免训练线性原型模型（图 2）。 给定一个来自未见过细胞系的细胞，任务是识别被敲低的基因。我们的方法无需训练：它根据训练细胞系构建扰动特异性参考谱，并将查询细胞与这些参考谱进行匹配。

1. **为每种扰动构建一个原型。** 对于每种基因敲低p，测量其$y_{p,\ell}$在每个训练细胞系中的效应。减去该细胞系中所有基因敲低共有的响应——即扰动的一般响应——再对各细胞系的残差取平均值。所得结果$g_p \in \mathbb{R}^G$是基因特异性响应的原型，可迁移至其他细胞系。这与前向迁移模型中使用的残差相同。

2. **表示查询细胞。** 将查询细胞 x 与其所属细胞系中的未扰动对照细胞进行比较，得到其表达变化 $d \in \mathbb{R}^G$。与 ExPert 中的做法相同，从中减去所有受扰动查询细胞共有的响应，留下查询细胞特有的变化。

3. **使用两种信号为每个候选基因打分。** 首先，查询细胞的表达变化是否与该候选项的原型相似？其次，该候选基因本身是否下调？我们以相同权重结合这两种信号：原型相似度反映整体响应模式，而基因自身的表达则为其基因敲低提供直接证据。候选项按以下方式排名：

$$s_p(x) = z\left[\cos\left(d_{\setminus p}, g_{p,\setminus p}\right)\right] + z[-d_p], \tag{8}$$

其中 $s_p(x) \in \mathbb{R}$，$\setminus p$ 表示从两个向量中移除基因 p，$d_p \in \mathbb{R}$ 是基因 p 自身的变化，$z[\cdot]$ 则对1,157个候选项中的每个信号进行标准化。实现细节和分数消融实验见附录 B 和 C。

因此，这两项任务以相反的方向采用相同的分解方式。迁移从已知基因出发，推断其在新细胞系中的变化；识别则从新细胞系中的观测变化出发，找出与之匹配原型对应的基因。

5 实验与结果

除非另有标注，每项比较均采用已发表方法各自的基准（COMPASS[5] 或 ExPert[7]）。我们首先展示扰动迁移结果，然后展示扰动识别结果。

扰动迁移。我们遵循 COMPASS基准测试[5]，，该基准包含 CRISPR 干扰（CRISPRi）[31] 在四种细胞系中的筛选[1,2] 以及在 Orion对细胞系中的筛选[3]（附录A）。我们使用变异度最高的 2,000 个基因，并对每细胞归一化后的计数取对数$(1+x)$。10^4 每次留出一个细胞系；其 20% 的基因敲低构成测试集，共进行五次随机划分。去偏皮尔逊相关使用互不重叠的两组对照细胞来比较预测效应与真实效应，从而避免因共享对照细胞而导致相关性虚高。

实现。 BiPert使用秩50（附录C中的秩消融）。BiPert-FS使用$k = 30$，岭正则化仅依据源细胞系进行选择，依次将每一系视为伪目标。BiPert-ID没有自由参数。所有线性模型均可在CPU上于数秒至数分钟内完成拟合。

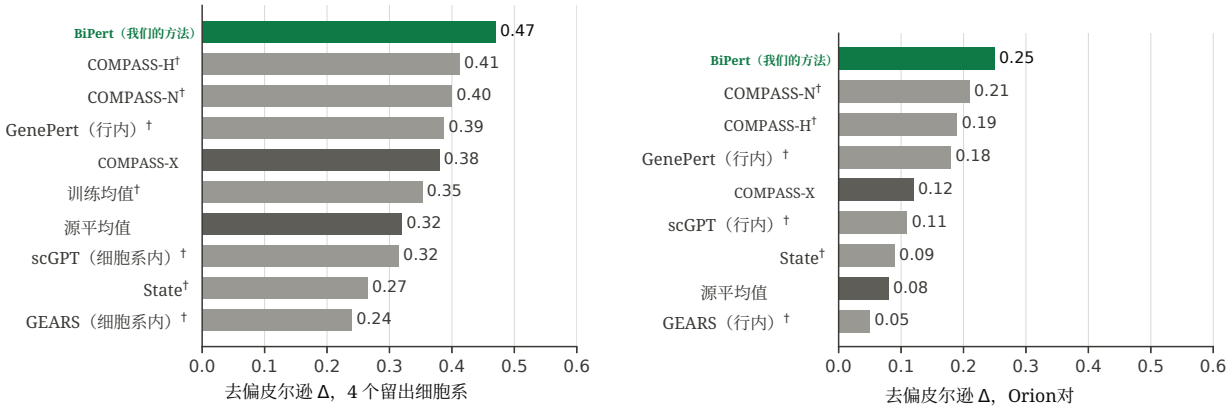

图3. COMPASS基准测试上的平均扰动迁移：四个留出细胞系（左）和 Orion 对（右）。†已发表的数值。

# 观测值	10	25	50	100
源平均值	0.320	0.320	0.320	0.320
COMPASS-X	0.322	0.355	**0.364**	0.370
BiPert（我们的模型）	0.332	0.351	**0.369**	**0.395**
BiPert-FS（我们的模型）	**0.355**	**0.361**	0.362	0.366

表 1. 新细胞系中的少样本预测：去偏皮尔逊 Δ 与该细胞系中观测到的基因敲低数量之间的关系，取四个留出细胞系和 5 个划分的平均值（划分间的标准差 ≈ 0.012）。在本表中，BiPert 使用秩 $r = 30$。

结果。 在COMPASS基准测试（图3）中，BiPert在四个细胞系上对留出扰动效应的预测准确率高于COMPASS-X（去偏皮尔逊：0.47对0.38），并在更难的Orion对上将准确率提高一倍以上（0.25对0.12）。它的表现也优于GenePert等基于嵌入的回归器（0.39和0.18）以及State等扰动条件深度模型[4]。作为COMPASS核心量的共享扰动编码也被更准确地恢复（0.34对0.22；Orion：0.20对0.07）。在观测扰动较少时（表1），BiPert-FS在10个观测时领先，在25–50个时与COMPASS-X持平，而BiPert在100个时超越二者。

扰动识别。我们采用 ExPert 的完整基准[7]：包含 12 折中的 22 个数据集，每折留出一个细胞系、一个 T 细胞供体或一种 T 细胞条件。在至少七种细胞情境中测量过的 1,157 个基因敲低作为候选项。我们使用 ExPert 的质量筛选条件，根据公开的细胞级数据重建每个数据集，并使其数据集统计量与 ExPert 发布的数据完全一致。在每一折中，BiPert-ID 使用与 ExPert 相同的数据集进行训练，并在相同的细胞上测试。按照 ExPert 的做法，将分数汇总（伪批量）：对每个基因敲低的测试细胞，先对每个细胞的对数 softmax 分数取平均，再进行排名。我们报告伪批量前10名召回率（真实扰动是否位于前 10 名）和平均倒数排名（MRR），并对基因敲低取平均值。

结果（表2）。ExPert[7], 是最先进水平的方法，是一种神经网络方法：在扰动筛选数据上训练的编码器会嵌入每个细胞，并通过与学习得到的每基因原型计算余弦相似度，为候选项评分。我们在两种设置下报告结果：迁移基准中四个留出细胞系，以及 ExPert 完整基准中的 34 个留出视图。我们使用这四个细胞系，是因为它们属于迁移基准，这使第6 节能够在相同的细胞系和扰动上比较两项任务。此外，在完整基准的全部十个留出细胞系中，BiPert-ID 在前10名召回率上领先 ExPert 的幅度与这四个细胞系上的幅度相同

	前10名召回率	MRR
GEARS†	0.01	0.01
scGPT†	0.04	0.02
GenePert†	0.06	0.03
仅自身基因下调†	0.18	0.16
RNA 指纹分析†	0.22	0.14
质心†	0.30	0.20
ExPert†	0.540	0.402
BiPert-ID（我们的）	**0.656**	**0.521**

表2. 四个留出细胞系（HepG2、Jurkat、K562、RPE1）上的扰动识别迁移基准中的各细胞系。

†ExPert 的作者随其基准发布的数值[7]；四个细胞系的均值。

		前10名召回率		MRR	
留出视图	*n*	ExPert	**BiPert-ID**（我们的方法）	ExPert	**BiPert-ID**（我们的方法）
新细胞系	10	0.426	**0.539**	0.309	**0.415**
新供体	12	0.258	**0.381**	0.182	**0.292**
新条件	12	**0.752**	0.442	**0.625**	0.340
所有视图	34	**0.482**	0.449	**0.376**	0.345

表3. 完整的 ExPert 基准（22 个数据集、12 折、34 个留出视图），在 ExPert 的确切测试细胞上；伪批量前10名召回率和平均倒数排名，按视图计算的均值。

（+0.113 对比 +0.116）。在四个留出细胞系上，BiPert-ID 在每个细胞系上的表现都优于 ExPert（表2）。GEARS、scGPT 和 GenePert 是前向模型，经 ExPert 的作者调整后用于识别；它们的表现接近随机水平（前10名召回率为 0.01–0.06，而随机排名为 0.009）。为了解其中原因，我们重新运行了 GenePert，并加入 State[4]（附录D）。两者都能复现训练过的基因敲低和细胞系，却无法复现新的基因敲低和细胞系；对于这两种模型，在相同评分方式下，训练数据的实测效应比模型的预测能够识别出更多基因敲低。

完整基准（表*3*）。我们还评估了 ExPert 的完整基准：包含 22 个数据集、分为 12 折，每折留出一个细胞系、一位 T 细胞供体或一种 T 细胞条件，并在 ExPert 指定的测试细胞上评分。在全部 34 种视图中取平均后，ExPert 略微领先（伪批量top-10召回率为 0.482，对比 0.449）。差距完全来自同一供体曾在训练中以另一种条件出现的视图。在每一种新供体视图以及 10 个新细胞系中的 9 个上，BiPert-ID 都明显领先：当测试细胞系或供体的任何数据都未在训练中出现时，线性分解的迁移效果优于深度网络[7]。

识别出哪些基因？（表4和5）平均召回率无法说明某种方法找到了哪些基因敲低。因此，我们探究 BiPert-ID 的领先优势是否仅限于效应较强的基因敲低，以及它是否只是通过读取 ExPert 未检测到的基因而取得优势。我们衡量每种基因敲低的效应大小，即它对训练细胞系中的细胞造成的变化程度，并将每个留出细胞系中的基因敲低按效应从弱到强分为四组。随后，我们衡量每种方法识别各组基因敲低的频率（表4）。两种方法识别效应较强的基因敲低的频率都远高于识别效应细微的基因敲低，而 BiPert-ID 在每组中均领先。

尽管 BiPert-ID 的简洁性使其能够读取查询细胞系中测得的每一个基因，ExPert 仅使用包含 3,096 个基因的固定基因面板。即便对于靶标位于该基因面板内的基因敲低，BiPert-ID领先

	ExPert	**BiPert-ID（我们的方法）**
效应最弱的四分位	0.21	**0.31**
第二四分位	0.49	**0.66**
第三四分位	0.66	**0.77**
效应最强的四分位	0.79	**0.87**

表 4. 按效应大小进行识别。在四个留出的细胞系中，伪批量排名前 10 的召回率。

靶基因	ExPert	**BiPert-ID（我们的方法）**
在 ExPert 的基因面板中	0.74	**0.80**
在基因面板之外	0.48	**0.61**

表5. ExPert 基因面板的识别。在四个留出细胞系中的伪批量前10名召回率。

迁移分数（四分位数）	ExPert	**BiPert-ID（我们的方法）**
最低	0.19	0.26
第二	0.39	0.59
第三	0.65	0.82
最高	0.90	0.95

表6. 正向迁移可预测逆向识别。在留出细胞系中正确识别的扰动比例，按仅使用训练细胞系计算的迁移分数的细胞系内四分位数分层。

识别范围	所有转录因子	最强的 10%
两种细胞类型	25 / 33	25 / 33
一种细胞类型	24 / 80	21 / 60
两者都不是	8 / 1,720	5 / 91

表7. 同一细胞类型内的可识别性限制迁移。成纤维细胞和RPE-1细胞中识别出的转录因子，占各组转录因子总数的比例。

（表 5），因此其优势并不能仅用可获取额外基因来解释。BiPert-ID 几乎识别出 ExPert 所识别的每一个基因敲低（仅有 6% 由 ExPert 单独发现），并进一步识别出 ExPert 遗漏的 18%。

6 迁移与识别之间的双向联系

迁移可预测识别（表 6）。 我们的分解表明，两项任务都取决于同一个量：扰动特异性效应gp 跨细胞系迁移的能力。我们仅使用训练细胞系对此进行检验。对于至少在两个训练细胞系中测量的每种扰动，我们计算其迁移分数：依次留出每个细胞系，计算该细胞系中的特异性效应与其他细胞系的平均效应之间的平均余弦相似度。然后，我们评估该扰动是否能在留出的细胞系中被识别，定义为使用合并查询细胞时，其排名进入1,157个候选项的前10名。

在3,340个扰动-细胞系对中，迁移能力较强的扰动更常被识别：对于两种方法，随着迁移分数从最低四分位升至最高四分位，被识别的比例稳步上升；四分位数是在每个留出细胞系内划分的。效应较强的扰动也更容易迁移（秩相关为0.54），但这种联系并非仅由效应大小所解释：在效应大小相同的情况下，两种方法仍然都更常识别迁移能力较强的扰动（附录 E）。这种联系对ExPert同样成立，尽管它并不对迁移进行建模；这表明，可迁移性通常是识别模型的一种有用归纳偏置。这种关联具有预测性，而非因果性。

细胞类型内的可识别性限制了迁移（表7）。我们在一项独立筛选中测试跨细胞类型迁移，该筛选使用CRISPR激活，在成纤维细胞和 RPE-1细胞中靶向 1,833 种转录因子（转录因子）[32]。对于每种转录因子，我们将其向导RNA分成两个互不重叠的集合，并测试根据一个集合估计的效应能否从另一个集合中识别出该转录因子（在 1,833 种转录因子中排名第一）。我们在每种细胞类型内以及跨细胞类型进行这项测试；只要任一方向（成纤维细胞到 RPE-1细胞，或反之）成功，就视为迁移成功。跨细胞类型识别与细胞类型内可重复性密切相关：在两种细胞类型中均被独立识别的 33 种转录因子中，有 25 种也能在两者之间迁移；而在两种细胞类型中均未被识别的 1,720 种转录因子中，只有 8 种能做到这一点。这种关联并非单纯由效应强度驱动。即使只看效应最强的 10% 转录因子，在两种细胞类型中均被识别的转录因子，其跨细胞类型识别频率也远高于两种细胞类型中均未被识别的转录因子（分别为 25/33 和 5/91）。在两种细胞类型内均可重复、但无法跨细胞类型重复的 8 种转录因子，是细胞类型特异性效应的候选项。比较时必须在两侧使用不同的向导；若使用相同的向导，会使跨细胞类型识别率被高估约四倍（附录F）。

7 局限性与结论

局限性。当训练数据中包含与测试数据同类型的细胞时，BiPert-ID 的表现不如深度网络 ExPert；例如，在同一供体以另一种条件出现在训练数据中的视图上（表3）。利用外部基因知识的方法也可能略胜 BiPert；BiPert 仅使用基因表达：在给定六组先验基因特征时，TabICL[14], 的去偏皮尔逊相关为 0.48，略高于 BiPert 的 0.47，尽管 BiPert 在 Orion对 上领先（0.25 对 0.24）[5]。

BiPert 是一种闭式线性分解，在留出细胞系和供体上评估时，于两项单细胞扰动任务上达到或超越了最先进的方法。所学参数具有直接的生物学解释：扰动编码U_P 揭示共享响应模式，而细胞系载荷 V_ℓ 刻画每种细胞环境如何对其重新加权。逆向识别得分可进一步分解为原型相似度以及敲低基因自身的表达下降。当测试细胞环境与其训练数据相似时，深度模型表现更好，正如 ExPert 在同供体视图上的表现。因此，基准测试应将在真正新的细胞环境上的性能与在密切相关的细胞环境上的性能分开报告。对于向新细胞环境的泛化，一个简单的线性模型是需要超越的基线。

最后，在控制扰动效应大小之后，两项任务之间的关系仍然存在。这表明，将扰动特异性响应跨细胞环境进行迁移的能力，与其从观测细胞中的可识别性密切相关。由于迁移分数仅使用训练细胞系计算，它还可以预先估计哪些扰动有望在新细胞系中被可靠识别，且与所使用的识别方法无关。

数据和代码可用性

所有数据集均已公开（Replogle 等[1], Nadig 等[2], X-Atlas/Orion[3], ExPert 基准[7] 以及 Southard等[32]）。用于复现结果的源代码可在以下网址获取：`https://github.com/bach-zouk/BiPert`。

致谢

我们感谢 Leandro Ventimiglia 提供的有益讨论。本研究获得生物技术与生物科学研究委员会（BBSRC）资助，项目编号为 BB/Y000730/1。计算资源由 Isambard-AI、伦敦国王学院国王学院计算研究、工程与技术环境（CREATE），以及瑞典国家超级计算学术基础设施（NAISS）提供；NAISS 由瑞典研究委员会资助。

AI 工具的使用

研究方向、方法和构想均由作者确定。作者使用 Claude Code 和 ChatGPT 协助进行科学分析、文献检索、语言编辑、语法检查，并提升稿件的清晰度和简洁性。所有科学构想、方法、实验、分析、结果和结论均经作者核实，作者对提交的内容承担全部责任。

参考文献

[1] Joseph M. Replogle, Reuben A. Saunders, AngelaN. Pogson, Jeffrey A. Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J. Wagner, Karen Adelman, Gila Lithwick-Yanai, Nika Ire-

madze, Florian Oberstrass, Doron Lipson, Jessica L. Bonnar, Marco Jost, Thomas M. Norman, and Jonathan S. Weissman。《利用全基因组 Perturb-seq 绘制信息丰富的基因型-表型景观》。*Cell*，185(14):2559–2575.e28，2022年。doi: 10.1016/j.cell.2022.05.013.[2]Ajay Nadig, Joseph M. Replogle, Angela N. Pogson, Mukundh Murthy, Steven A. McCarroll, Jonathan S. Weissman, Elise B. Robinson, and Luke J. O'Connor。《扰动图谱中差异表达的转录组范围分析》。*Nature Genetics*，57(5):1228–1237，2025年。doi: 10.1038/s41588-025-02169-3.[3] Ann C. Huang, Tsung-Han S. Hsieh, Jiang Zhu, Jackson Michuda, Ashton Teng, Soohong Kim, Eliza- beth M. Rumsey, Sharon K. Lam, Ikenna Anigbogu, Philip Wright, Mohamed Ameen, Kwontae You, Christopher J. Graves, Hyunsung John Kim, Adam J. Litterman, Rene V. Sit, Alex Blocker, and Ci Chu。X-Atlas/Orion：通过可扩展的固定-冷冻保存平台构建全基因组 Perturb-seq 数据集，用于训练剂量依赖型生物学基础模型。*bioRxiv*，2025年。doi: 10.1101/2025.06.11.659105.[4] Abhinav K. Adduri, Dhruv Gautam, et al.《利用 State 预测不同背景下的扰动细胞响应》。*bioRxiv*，2025年。[5] Huan Liang and Rohit Singh。COMPASS：对共享和基因特异性扰动响应进行分量级推断。*bioRxiv*，2026年。doi 10.64898/2026.08.03.742643.[6] Isabella N. Grabski, Junsuk Lee, John D. Blair, Carol Dalgarno, Isabella Mascio, Alexandra Bradu, David A. Knowles, and Rahul Satija。《利用 RNA 指纹分析绘制细胞扰动的转录响应》。*Cell*，2026年。doi: 10.1016/j.cell.2026.09.017.[7] Leon Schwartz, Jacob L. Schlamowitz, Benjamin Kuznets-Speck, Conor W. Johnson, Diptatanu Das, Emanuelle I. Grody, Tao Bai, Hanxiao Sun, Irish Senthilkumar, Gillian Primavera, Jaekwon Jung, Britta Kuusik, Rogan A. Grant, Enrique Rojas, Abdalla M. Abdrabou, Brandon Peiffer, Zhaoli Sun, Anna Mae Diehl, Carsten Marr, Navdeep S. Chandel, Divakar S. Mithal, Samuel E. Weinberg, Mazhar Adli, Auinash Kalsotra, Ekta Prashnani, and Yogesh Goyal。ExPert 从生物学和临床背景下的细胞状态推断因果扰动。*bioRxiv*，2026年。doi: 10.64898/2026.11.02.762549.[8] Ramon Viñas Torné, Maciej Wiatrak, Zoe Piran, Shuyang Fan, Liangze Jiang, Sarah A. Teichmann, Mor Nitzan, and Maria Brbi´c。Systema：一种用于评估遗传扰动响应预测的方法，可超越系统性变异。*Nature Biotechnology*，44(6):1050–1059，2025年。doi: 10.1038/s41587-025-02777-8.[9] Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L. Ibarra, Sanjay R. Srivatsan, Mohsen Naghipourfar, Riza M. Daza, Beth Martin, Jay Shendure, Jose L. McFaline-Figueroa, Pierre Boyeau, F. Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J. Theis。《在高通量筛选中预测复杂扰动下的细胞响应》。*Molecular Systems Biology*，19(6):e11517，2023年。doi: 10.15252/msb. 202211517.[10] Yusuf Roohani, Kexin Huang, and Jure Leskovec。利用 GEARS 预测新型多基因扰动的转录结果。*Nat. Biotechnol.*，42:927–935，2024年。[11] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang。scGPT：利用生成式人工智能构建单细胞多组学基础模型。*NatureMethods*，21(8):1470–1480，2024年。doi: 10.1038/s41592-024-02201-0。

[12] Yiqun T. Chen和James Zou。GenePert：利用GenePT嵌入进行基因扰动预测。*bioRxiv*，2024年。doi: 10.1101/2024.10.27.620513.[13] Jingang Qu、David Holzmüller、Gaël Varoquaux和Marine Le Morvan。TabICLv2：更优、更快、可扩展且开放的表格基础模型。*arXiv*预印本 *arXiv:2602.11139*，2026年。[14] Giovanni Palla、Alexander Hillsley、Yang-Joon Kim和Loic A. Royer。表格基础模型是跨生物尺度具有竞争力的细胞扰动预测器。*bioRxiv*，2026年。doi: 10.64898/2026.06.28.735106.[15] Frederik Wenkel、Wilson Tu、Cassandra Masschelein、Hamed Shirzad、Liam Hodgson、Ihab Bendidi、Cian Eastwood、Shawn T. Whitfield、Craig Russell、Yassir El Mesbahi、Jiarui Ding、Marta M. Fay、Berton Earnshaw、Emmanuel Noutahi和Alisandra K. Denton。TxPert：利用多个知识图谱预测转录组扰动效应。《自然·生物技术》，2026年。doi: 10.1038/s41587-026-03113-4.[16] Constantin Ahlmann-Eltze、Wolfgang Huber和Simon Anders。基于深度学习的基因扰动效应预测尚未超越简单的线性基线。《自然·方法》，22(8):1657–1661，2025年。doi: 10.1038/s41592-025-02772-6.[17] Gerold Csendes、Gema Sanz、Kristóf Z. Szalay和Bence Szalai。面向扰动后RNA测序预测的细胞基础模型基准评测。《*BMC*基因组学》，26(1):393，2025年。doi: 10.1186/s12864-025-11600-2.[18] Aaron Wenteler、Martina Occhetta、Nikhil Branson、Victor Curean、Magdalena Huebner、William Dee、William Connell、Siu Pui Chung、Alex Hawkins-Hooker、Yasha Ektefaie、César Miguel Valdez Córdova和Amaya Gallagher-Syed。PertEval-scFM：面向扰动效应预测的单细胞基础模型基准评测。*bioRxiv*，2025年。doi: 10.1101/2024.10.02.616248.[19] Yongqi Huang、Hanzhi Wang、Chenyu Li和Parker Wilson。细胞扰动响应预测的可恢复分辨率。*bioRxiv*，2026年。doi: 10.64898/2026.09.30.755756.[20] Panpan Cui、Yiqi Liu和Wenhao Sun。PerturbMap：单细胞扰动响应的跨情境迁移。*arXiv*预印本 *arXiv:2607.28090*，2026年。doi: 10.48550/arXiv.2607.28090.[21] Eric F. Lock、Katherine A. Hoadley、J. S. Marron和Andrew B. Nobel。面向多种数据类型整合分析的联合与个体变异解释（JIVE）。《应用统计年刊》，7(1):523–542，2013年。doi: 10.1214/12-AOAS597.[22] Basura Fernando、Amaury Habrard、Marc Sebban和Tinne Tuytelaars。利用子空间对齐的无监督视觉领域自适应。载于国际计算机视觉大会，第2960–2967页，2013年。[23] Theodoros Evgeniou和Massimiliano Pontil。正则化多任务学习。载于知识发现与数据挖掘大会，第109–117页，2004年。[24] Ilja Kuzborskij和Francesco Orabona。稳定性与假设迁移学习。载于国际机器学习大会，第942–950页，2013年。[25] Amit Boyarski、Sanketh Vedula和Alex Bronstein。谱几何矩阵补全。载于数学与科学机器学习（*MSML*），2021年。arXiv:1911.07255.[26] Abhishek Sharma和Maks Ovsjanikov。图上的矩阵分解：一种函数视角。载于*IEEE*国际声学、语音与信号处理会议（*ICASSP*），2022年。

[27] Ricard Argelaguet、Britta Velten、Damien Arnol、Sascha Dietrich、Thorsten Zenz、John C. Marioni、Florian Buettner、Wolfgang Huber和Oliver Stegle。多组学因子分析——一种用于多组学数据集无监督整合的框架。*Molecular Systems Biology*，14(6):e8124，2018年。doi: 10.15252/msb.20178124.[28] Ricard Argelaguet、Damien Arnol、Danila Bredikhin、Yonatan Deloro、Britta Velten、John C. Marioni和Oliver Stegle。MOFA+：一种用于多模态单细胞数据全面整合的统计框架。*Genome Biology*，21(1):111，2020年。doi: 10.1186/s13059-020-02015-1.[29] Yifan Zhou、Kaixuan Luo、Lifan Liang、Mengjie Chen和Xin He。一种新的贝叶斯因子分析方法提升了单细胞 CRISPR 筛选中受扰动影响的基因与生物过程的检测能力。《自然·方法》，20(11):1693–1703，2023年。doi: 10.1038/s41592-023-02017-4.[30] Michael Schubert、Bertram Klinger、Martina Klünemann、Anja Sieber、Florian Uhlitz、Sascha Sauer、Mathew J. Garnett、Nils Blüthgen和Julio Saez-Rodriguez。扰动响应基因揭示癌症基因表达中的信号传导特征。*NatureCommunications*，9:20，2018年。doi: 10.1038/s41467-017-02391-6.[31] Lei S. Qi、Matthew H. Larson、Luke A. Gilbert、Jennifer A. Doudna、Jonathan S. Weissman、Adam P. Arkin和Wendell A. Lim。将CRISPR重新用作基因表达序列特异性调控的RNA 引导平台。*Cell*，152(5):1173–1183，2013年。doi: 10.1016/j.cell.2013.02.022.[32] Kaden M. Southard、Rico C. Ardy、Anran Tang、Deirdre D. O'Sullivan、Eli Metzner、Karthik Guruvayurappan和Thomas M. Norman。全面的转录因子扰动重现成纤维细胞的转录状态。*Nature Genetics*，57:2323–2334，2025年。doi: 10.1038/s41588-025-02284-1.

补充材料

补充材料的组织结构如下。附录A 列出迁移基准的细胞系，推导 BiPert-FS 的少样本拟合，并分析其各部分在真实数据上的贡献。附录 B 给出 BiPert-ID 的实现细节。附录C 报告消融实验：BiPert 的秩以及 BiPert-ID 的分数。附录 D 探讨用于识别的前向模型为何分数较低。附录 E 表明在控制效应大小后，迁移与识别之间的关联仍然成立。附录 F 更详细地给出正文中的识别结果：在 ExPert 完整基准上比较的不确定性、各留出细胞系的结果，以及跨细胞类型分析为何比较不同的向导。

A BiPert-FS

概述。图S1展示了 BiPert-FS 与传统因子模型的区别。JIVE 等方法[21] 和 MOFA[27,28] 会将多个数据集联合分解为共享因子和数据集特异因子。它们的主要目标是解释观测数据，而纳入新数据集则需要估计其因子载荷（图S1，下图）。

BiPert-FS 处理的是另一项任务：仅根据少量观测到的基因敲低，预测新细胞系中未测量的扰动响应。它不会分解新细胞系，而是保留从源细胞系中学习到的基底 Φ，并估计一个较小的校正项，该校正项由b 和C参数化。矩阵C 通过正则化趋近于单位矩阵，并在未用于适应的基因敲低上评估该校正项。

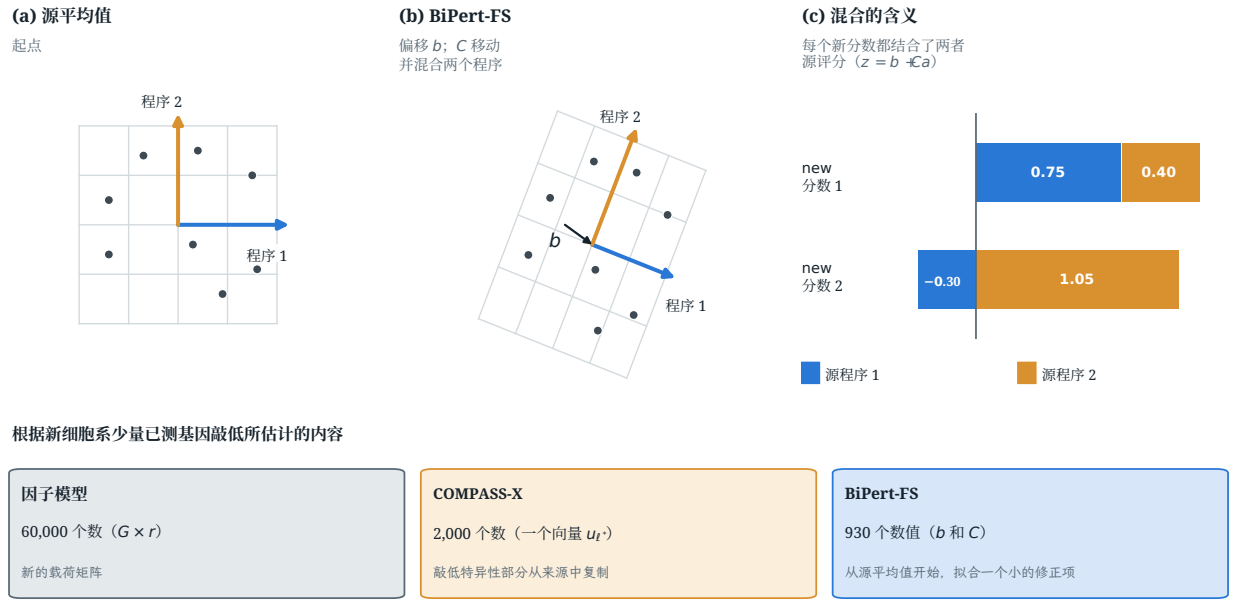

图 S1。BiPert-FS 如何适应新细胞系。二维示意图。(a) 源平均值中的基因敲低（圆点）。(b) BiPert-FS 根据源平均得分a_P 拟合仿射映射$z_p \approx b + Ca_P$，将其映射到新细胞系得分z_P：b 使所有基因敲低整体偏移，而C 则移动两个程序。(c) 混合：每个新得分都结合两个源得分。同一程序的权重（0.75、1.05）会重新缩放该程序；另一个程序的权重（0.40，−0.30）则会混合这两个程序。底部：对于新细胞系，每种方法所估计的内容，其中有$G = 2,000$ 个基因、$r = 30$ 个因子和$k = 30$ 个程序。

基准。 COMPASS基准测试包含四个细胞系：K562、RPE1、HepG2 和 Jurkat[1,2], , 四者均测量了 2,052 个基因敲低。Orion对由 HCT116 和 HEK293T[3], 组成，两者均测量了 2,270 个基因敲低。

输入。令 $\mathscr{S}$ 表示源细胞系，不包括目标细胞系 ℓ^*。对于每个源细胞系 s，令Y_s 表示其基因敲低效应矩阵。我们利用源矩阵的前k 个右奇异向量构造 $\Phi \in \mathbb{R}^{G\times k}$，其中$k = 30$；在将每个矩阵按基因中心化并堆叠后，取这些向量。$\Phi$ 的列是正交归一的：

$$\Phi^\top \Phi = I_k.$$

对于基因敲低 p，源细胞系的平均响应为

$$\bar{y}_p = \frac{1}{|\mathscr{S}|}\sum_{s\in\mathscr{S}} y_{p,s}.$$

设 $\mathscr{O}$ 为目标细胞系中观测到的n 个基因敲低，其响应为$y_p = y_{p,\ell^*}$。

低维表示。将源平均和目标响应投影到 Φ：

$$a_p = \Phi^\top \bar{y}_p, \qquad z_p = \Phi^\top y_p.$$

a_p 和 z_p 都只有k 个坐标。正文中的方程（7）给出的预测为

$$\hat{y}_p = \bar{y}_p + \Phi\big(b + (C - I_k)\, a_p\big).$$

等价地，

$$\hat{y}_p = \Phi(b + Ca_p) + (I - \Phi\Phi^\top)\,\bar{y}_p,$$

其中I 是$G \times G$ 单位矩阵。因此，BiPert-FS 仅在学习基内调整源预测，保持其正交分量不变。

降维至k 维。预测误差可分解为位于 Φ 张成空间内和空间外的分量：

$$y_p - \hat{y}_p = \Phi(z_p - b - Ca_p) + (I - \Phi\Phi^\top)(y_p - \bar{y}_p). \tag{S1}$$

由于这些分量相互正交，

$$\begin{aligned}\|y_p - \hat{y}_p\|_2^2 = {} & \|z_p - b - Ca_p\|_2^2 \\ & + \|(I - \Phi\Phi^\top)(y_p - \bar{y}_p)\|_2^2.\end{aligned} \tag{S2}$$

第二项与b 或C无关。因此，最小化基因上的平方预测误差，可归结为在k 维空间中拟合从a_p 到z_p 的回归。

正则化目标函数。我们通过最小化下式来估计b 和C：

$$\mathscr{L}(b, C) = \sum_{p\in\mathscr{O}} \|z_p - b - Ca_p\|_2^2 + \mu\,\|C - I_k\|_F^2. \tag{S3}$$

惩罚项会促使C 保持接近单位矩阵，而偏移量b 不受惩罚。让C 趋近单位矩阵而非零矩阵，意味着在数据较少时，模型会回退到源程序，而不是将其舍弃。这是必要的，因为C 包含$k^2 = 900$ 个条目，远多于通过 10 次基因敲低观测到的$nk = 300$个坐标。偏移量只有k 个条目，其解$b = \bar{z} - C\bar{a}$使用了z_p 和a_p 在所有观测到的基因敲低中的均值，因此即使n 较小，也能得到充分确定，无需惩罚项。令$C = I_k + D$，目标函数便成为以a_p为自变量、以$z_p - a_p$ 为因变量且包含无惩罚截距的岭回归。换言之，模型学习目标细胞系与源平均响应之间的差异。当观测到的基因敲低少于$k + 1 = 31$ 个时，惩罚项也能确保拟合问题定义良好。

	去偏皮尔逊				PDS			
	源平均值	仅平移	偏移 + 对角.C	完整（b，C）	源平均值	仅平移	偏移 + diag.C	完整（b，C）
Mean	0.320	0.353	0.359	0.366	0.266	0.258	0.241	0.261

表 S1. 少样本拟合的组成部分（100 个观测到的基因敲低，在四个留出细胞系和五种划分上取均值）。仅平移：$C = I$. 偏移 + *diag*. C：每个程序分别重新缩放，不进行混合。 完整：式（7），少样本拟合表中的 BiPert-FS 行。

闭式解。定义

$$X = \begin{bmatrix} 1 & a_1^\top \\ \vdots & \vdots \\ 1 & a_n^\top \end{bmatrix}, \quad Z = \begin{bmatrix} z_1^\top \\ \vdots \\ z_n^\top \end{bmatrix}, \quad \Theta = \begin{bmatrix} b^\top \\ C^\top \end{bmatrix}.$$

这里，$X \in \mathbb{R}^{n\times(k+1)}$、$Z \in \mathbb{R}^{n\times k}$和 $\Theta \in \mathbb{R}^{(k+1)\times k}$。令

$$\Lambda = \mathrm{diag}(0, \mu I_k), \qquad \Theta_0 = \begin{bmatrix} 0 \\ I_k \end{bmatrix}.$$

因此，目标函数可写为

$$\mathscr{L}(\Theta) = \|Z - X\Theta\|_F^2 + \mathrm{tr}\big((\Theta - \Theta_0)^\top \Lambda (\Theta - \Theta_0)\big).$$

令其梯度为零，得到

$$(X^\top X + \Lambda)\,\Theta = X^\top Z + \Lambda\Theta_0,$$

因此

$$\boxed{\Theta = (X^\top X + \Lambda)^{-1}\,(X^\top Z + \Lambda\Theta_0).} \tag{S4}$$

由此可得少样本适应参数的闭式解，无需迭代训练。

选择正则化强度。我们将惩罚项参数化为

$$\mu = \alpha\, n\, \bar{v},$$

其中 $\bar{v}$ 是源坐标ap 在不同基因敲低条件下的平均方差。乘以n 进行缩放，可使惩罚项与数据拟合项保持成比例，而 $\bar{v}$则考虑了扰动效应的尺度。

我们仅使用源细胞系，针对每种观测预算分别选择 $\alpha \in \{0.01, 0.1, 1, 10\}$。依次将每个源细胞系视为伪目标，并使用其余细胞系构建源表示。我们保留使皮尔逊相关系数与 PDS 之和最大的 α 值。留出的目标细胞系绝不会用于超参数选择。

各组成部分的贡献。为了确定式（7）中的哪些部分能够提升性能，我们在相同的少样本划分上，并采用相同的正则化选择方法，评估了四种变体：源平均值（$b = 0$，$C = I_k$）、仅平移（$C = I_k$）、平移加逐程序缩放（对角 C），以及完整的仿射拟合（表 S1）。

在四种细胞系上取平均后，偏移解释了相较于源平均值的大部分提升。在有 10 个观测到的基因敲低时，仅靠偏移就能使去偏皮尔逊相关从 0.320 提升至 0.347，而完整拟合的结果为 0.355。在 10 个基因敲低时，仅靠偏移就解释了 0.025 的

BiPert-FS 相较于 COMPASS-X 的 0.033 优势。然而，在 25 个基因敲低时，仅靠偏移的表现低于 COMPASS-X（0.351 对 0.355），优势则完全来自C。

偏移带来的平均提升由一个细胞系 RPE1 驱动：在 100 个基因敲低时，皮尔逊相关系数从 0.360 提高到 0.517，但 PDS 从 0.312 降至 0.141。RPE1 是四个细胞系中唯一的非癌细胞系，这或许是造成这一差异的原因；但我们并未对此进行检验。在其他三个细胞系中，仅使用偏移并未优于源平均值，而相较于偏移所带来的提升来自C（在 100 个基因敲低时为 0.005–0.026）。混合还改善了扰动判别：相较于对角重缩放，它在每个细胞系中都提高了 PDS，在 100 个基因敲低时，Jurkat 中的提升最高可达 0.055。

只有在较大的预算下，C 的非对角线元素才会在不同随机抽取的观测到的基因敲低样本之间呈现可重复性。在 100 个基因敲低时，不同抽样结果之间的相关性为 0.50–0.68；相比之下，在 10 个时为 0.03–0.22。因此，在观测样本极少时，C 主要起到接近单位矩阵的正则化校正作用，而非作为经过可靠估计的程序混合变换。随着观测到的基因敲低样本增多，C 的效果会更好：相较于仅使用偏移，其在皮尔逊相关系数上的提升平均从 10 个基因敲低时的 0.007 增至 100 个时的 0.012；在 100 个基因敲低时提升最大的细胞系 Jurkat 中，则从 0.012 增至 0.026。

B BiPert-ID：实现细节

本节给出BiPert-ID的实现细节，如第4.2节所述 .

原型。对于每个训练细胞系，扰动效应$y_{p,\ell}$ 按基因进行标准化（除以各扰动上稳健的逐基因离散度）。通过投影去除各基因敲低的平均效应方向来移除共享响应，并将每个残差归一化为单位长度。原型g^p 是在测量p的训练细胞系上这些残差的平均值。

相似度加权平均。训练细胞系与留出细胞系的相关性并不相同：一个留出的T细胞样本与其他T细胞情境的相似程度，高于其与例如K562的相似程度。因此，我们根据每个训练细胞系 ℓ 的未扰动状态与留出细胞系 ℓ^*的相似程度对其进行加权。设c_ℓ、$\mathrm{c}_{\ell^*} \in \mathbb{R}^G$为其对照细胞的平均表达（计数归一化至每细胞 10^4 后的$\log(1+\mathrm{x})$）。在跨 $\{\mathrm{c}_{\ell^*}\} \cup \{\mathrm{c}_\ell\}$对每个基因进行中心化后，我们计算 $\rho_\ell = \cos(\mathrm{c}_\ell, \mathrm{c}_{\ell^*})$，并在训练˜ ˜细胞系上对这些值进行标准化，$\tilde{\rho}_\ell = (\rho_\ell - \bar{\rho})/\mathrm{sd}(\rho)$，并设$\mathrm{w}_\ell = \mathrm{e}^{\rho_\ell} / \sum_{\ell'} \mathrm{e}^{\rho_{\ell'}}$。原型变为加权平均值$\mathrm{g}_p = \sum_{\ell \ni p} \mathrm{w}_\ell \, \tilde{}\mathrm{g}_{p,\ell} / \sum_{\ell \ni p} \mathrm{w}_\ell$，其中 $\tilde{}\mathrm{g}_{p,\ell}$为上述残差，求和范围是测量p的训练细胞系。仅使用留出细胞系的无标签对照细胞，即已作为查询参考的同一批细胞，因此该方法保持零样本。该规则没有需要调优的参数（softmax温度在标准化单位下固定为1）。与对训练细胞系的未加权平均值相比，该加权使四个留出细胞系上的伪批量top-10召回率提高 0.014，并在全部34个视图上提高0.025。该规则，包括仅当其在全部34个视图上增益的95%置信区间完全位于零以上时才予以采用的标准，均在评估之前预先规定。

查询。 查询细胞的变化 d 是相对于留出细胞系中对照细胞的均值计算的，并除以这些对照细胞在各基因上的离散程度。遵循 ExPert 的方法，通过投影去除该细胞系所有查询细胞均值变化的方向，从而移除共享响应。

评分。 计算余弦相似度前，会从两个向量中移除候选项自身的基因。正文公式（8）中的两项会在1,157个候选项之间进行标准化，并以相同权重相加；没有经过调优的参数。对于来自同一基因敲低的一组细胞，会先对每个细胞的对数软最大值分数取平均，再进行排序（ExPert的伪批量评估）。

C 消融实验

BiPert：秩。 准确率从秩为10时的0.43提升至秩为50–100时平台期的0.47，随后略有下降，在秩为300时达到0.44。扰动判别分数（PDS）随秩稳步提升，从秩为10时的0.15增至秩为300时的0.25，仅在秩为200–300时与COMPASS-X相当。因此，秩控制着预测准确率与扰动判别之间的权衡：在秩为50时，BiPert的PDS低于COMPASS-X（0.21对0.25）。我们还测试了BiPert是否能像JIVE一样，先将所有细胞系共有的成分与细胞系特异性成分分离开来[21]。结果并非如此：在不同秩下，用JIVE替代普通奇异值分解后，准确率最多变化0.004，PDS最多变化0.010。因此，BiPert不需要进行这种拆分；性能提升来自低秩结构本身。

BiPert-ID：评分。 我们对评分规则（公式（8））进行消融，以评估余弦项和自身基因下降项各自的重要性。两者缺一不可：单独使用其中任何一项，识别出的扰动都远少于二者结合使用时（图 S2）。从查询细胞系中每个实测基因读取自身基因下降量，而非仅从所有训练细胞系共有的基因中读取，还能进一步提升效果。

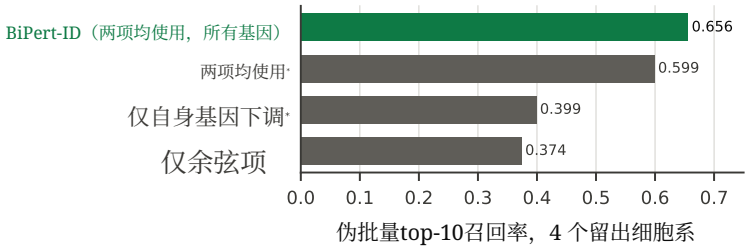

图 S2. BiPert-ID 消融实验。 在 ExPert 的测试细胞上，对表 2 中的四个留出细胞系取均值。 *自身基因下降量仅从每个训练细胞系中均有实测数据的基因读取。

D 前向模型识别效果不佳的原因

介绍了 BiPert-ID 如何识别基因敲低后，我们现在探究，为何针对前向任务构建的模型在同一任务上的表现如此不佳（表 2）。前向模型可以通过为每个候选基因敲低预测一个表达谱，并将每个测试细胞与这些表达谱进行匹配来用于识别；该表中的前向模型评分由 ExPert 的作者发布[7]。我们在四个留出细胞系上重新运行了 GenePert[12]，以探究其评分较低的原因，并加入了 State[4],，这是一个旨在预测未见细胞系中效应的前向模型。对于每个模型，我们提出两个问题：移除所有基因敲低的平均效应后，基因敲低的预测表达谱是否与其实测效应一致？采用相同的评分方式对实测效应进行评分，能否比对预测进行评分识别出更多基因敲低？图S3 给出了具体数据。

GenePert 通过岭回归，将基因的文本嵌入映射到基因敲低的效应。对于纳入拟合的基因敲低，其效应谱是实测效应的平滑副本（一致性为 0.82）；将这些效应谱用于匹配，可识别出留出细胞系中 36% 的基因敲低，而直接使用实测效应则为 42%。对于未纳入拟合的基因敲低（五折交叉验证），一致性降至 0.10，识别率降至 8%。ExPert 的作者报告 GenePert 的识别率为 6%；据我们理解，他们的适配器使用的是拟合得到的效应谱，因此我们无法通过自己的重新实现复现这一数值（每个训练上下文对应一个模型）。

State 在三个细胞系上训练，并根据第四个细胞系的对照细胞预测该细胞系中的每种基因敲低。在训练过的细胞系中，其预测结果与实测效应一致（0.62）。在留出的细胞系中，

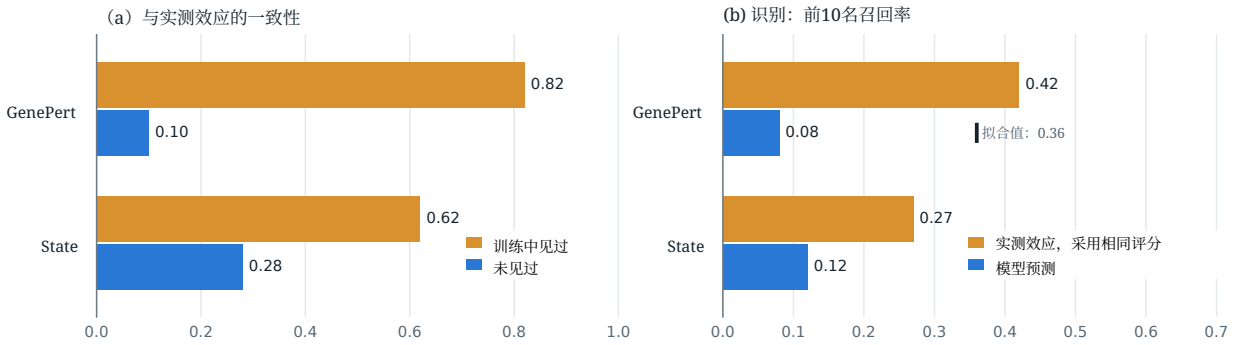

图 S3. 前向模型的预测内容，以及其识别基因敲低的表现。(a) 对 GenePert 和 State 而言，移除所有基因敲低的平均效应后，基因敲低的预测效应与实测效应之间的相关性。“已见”和“未见”分别指 GenePert 中的基因敲低，以及 State 中的细胞系。(b) 在四个留出细胞系上，从1,157个候选项中进行伪批量top-10召回率评估，比较模型的预测谱与其训练数据中的实测效应，评分方式相同。对于 GenePert，柱形表示未纳入拟合的基因敲低，刻度线表示纳入拟合的基因敲低。State 是我们自行改编的版本，使用作者公开的零样本模型（嵌入输入），不属于 ExPert 的基准。模型使用不同的训练数据和基因集；仅比较同一模型内的柱形。

一致性为 0.28，不高于三个训练细胞系中实测效应的简单平均值（0.28）；而且，预测谱识别出的基因敲低不到该平均值的一半（0.12 对 0.27）。State 将基因敲低编码为标签，因此仅能为 1,157个候选项中的约 920 个生成谱。

摘要。 GenePert 和 State 对其训练过的基因敲低和细胞系保留了敲低特异性信号，但对新的基因敲低和细胞系则会失去这种信号。对于这两个模型，在采用相同评分方法时，实测效应识别出的基因敲低多于模型预测。因而，在新细胞系中的识别取决于候选基因敲低的实测效应，而 BiPert-ID 直接将这些效应用作原型。

E　Tr控制效应大小后，迁移可预测识别

研究动机。 第 6 节和表 6 表明，迁移分数较高的扰动更有可能在留出细胞系中被识别。然而，效应较大的扰动可能更容易被识别，也可能具有更稳定的迁移表现。事实上，迁移分数与效应大小的秩相关为 0.54。因此，我们检验迁移能否在独立于效应大小的情况下预测识别。

实验设置。 我们分析了 3,340 个扰动-细胞系对。对于每一对，如果该扰动在留出细胞系中（对细胞进行汇总后）被识别为 1,157个候选项中的前10名之一，则结果为$y = 1$；否则为$y = 0$。我们使用两个预测因子，二者均仅根据训练细胞系计算：迁移分数t（定义见正文）和效应大小e，后者以扰动平均效应范数的对数衡量。两个预测因子均经过标准化，使其均值为零、方差为一。

我们拟合了一个逻辑回归模型，

$$\log \frac{P(y=1)}{P(y=0)} = \beta_t t + \beta_e e + \gamma_\ell,$$

其中，γ_ℓ 是每个留出细胞系单独的截距，用于解释识别难度的差异。系数 β_t 衡量在控制效应大小和细胞系后，迁移与识别之间的关联，β_e 则是效应大小对应的系数。一个标准差的增加

配对组	方法	每 +1 个标准差的识别优势比		模型拟合度（伪-R^2）
		迁移分数 [95% 置信区间]	效应大小	不使用 → 使用迁移
全部（3,340）	ExPert	×4.4 [3.8, 5.1]	×1.7	0.16 → 0.29
	BiPert-ID	×5.2 [4.4, 6.2]	×1.9	0.17 → 0.30
效应最大的半数（1,671）	ExPert	×4.5 [3.7, 5.4]	×1.2	0.03 → 0.21
	BiPert-ID	×4.1 [3.3, 5.1]	×1.4	0.04 → 0.19

表S2。在控制效应大小后，迁移可预测识别。采用逻辑回归分析某个基因敲低是否在留出细胞系中被识别（在汇总所有细胞后，位列1,157个中的前10名），并为每个细胞系设置一个截距。每个乘数表示在其他预测变量固定时，预测变量每增加一个标准差，识别优势比的变化；×1表示无效应。两个预测变量都仅使用训练细胞系。

会使迁移分数对应的识别优势比乘以 $\exp(\beta_t)$。如果观测到的关联完全由效应大小解释，那么 β_t 应接近于零。

结果。表S2 报告了拟合得到的优势乘数，即 $\exp(\beta)$。迁移仍然是两种方法的强预测因子，其优势乘数的95%置信区间均远高于1。迁移分数每增加一个标准差，识别优势比对于ExPert约乘以4.4，对于BiPert-ID约乘以5.2；相比之下，效应大小同样增加时，优势比分别乘以1.7和1.9。

在仅包含效应大小和细胞系截距的模型中加入迁移分数后，ExPert的McFadden伪R^2 从0.16增至0.29，BiPert-ID则从0.17增至0.30。

作为额外检验，我们仅使用每个细胞系中效应最大的半数扰动-细胞系对重复分析。迁移仍与识别显著相关（ExPert 和 BiPert-ID 的优势分别扩大至 4.5 倍和 4.1 倍），而效应大小的影响要小得多（×1.2和 ×1.4）。在这半数配对组中，效应大小和细胞系单独只能解释很少的差异（伪R^2 0为 0.03 和 0.04），加入迁移后，这一数值升至 0.21 和 0.19。

总之，这些结果表明，**迁移对识别的预测作用，无法仅凭扰动效应大小来解释。**

F 识别结果详述

完整 ExPert 基准：结果的不确定性。在 ExPert 完整基准的全部 34 个留出视图中（表 3 和第 5节），ExPert 的伪批量top-10召回率均值为 0.482，而 BiPert-ID 为 0.449，差值为 -0.033 (BiPert-ID 减去 ExPert)。

为量化不确定性，我们对基因敲低进行配对自助法。在每种视图中，我们有放回地重抽样基因敲低，并在同一组重抽样数据上计算两种方法的前10名召回率。然后，我们对34种视图中的差值取平均值，赋予每种视图相同的权重，与 ExPert 的标题指标一致。

在 2,000 次自助重抽样中，均值差的 95% 置信区间为 $[-0.040, -0.026]$。该区间不包含零，表明在完整基准上，ExPert 整体具有虽小但确实存在的优势。

然而，这一优势完全来自这样的视图：同一供体在另一种条件下出现在训练集中。如正文所述，BiPert-ID 在所有新供体视图以及10种新细胞系视图中的9种上表现更好。新细胞系和新供体对应的区间均高于零，而同一供体视图对应的区间则低于零。

每个留出细胞系的识别结果。表S3 报告了表2 中正文（第5节）中各细胞系的结果，这些结果构成了平均值，并在 ExPert 的测试细胞上进行评估。BiPert-ID 在每个

	前10名召回率				MRR			
	HepG2	Jurkat	K562	RPE1	HepG2	Jurkat	K562	RPE1
GenePert†	0.049	0.063	0.059	0.084	0.030	0.031	0.034	0.040
仅自身基因下调†	0.189	0.177	0.189	0.171	0.166	0.155	0.169	0.143
RNA 指纹分析†	0.249	0.203	0.156	0.254	0.166	0.135	0.109	0.164
质心†	0.329	0.314	0.311	0.255	0.221	0.210	0.216	0.166
ExPert†	0.597	0.503	0.471	0.589	0.457	0.376	0.340	0.434
BiPert-ID（本文方法）	**0.688**	**0.652**	**0.647**	**0.637**	**0.561**	**0.515**	**0.520**	**0.485**

表 S3. 各留出细胞系的扰动识别。†ExPert 的作者发布的数值及其基准[7]。

留出细胞系] 细胞系上均在两项评估指标中优于 ExPert。对于 K562，我们采用 ExPert 的必需基因筛选结果。

可识别性与迁移：为何需要不同的向导。在 Southard等人[32], 开展的 CRISPR 激活筛选中，每个转录因子（TF）都由多个向导RNA靶向。单个向导也可能产生与其预期靶向的 TF 无关的效应。因此，在不同细胞类型间比较相同的向导，可能会高估 TF 响应的迁移效果。

在 1,833 个转录因子中，成纤维细胞与 RPE-1细胞对相同向导的响应之间的平均相关性为 0.060–0.065。该数值高于同一细胞类型内靶向同一转录因子的不同向导之间的相关性（0.035–0.039）。

这种差异也会影响识别。跨细胞类型使用相同向导可正确识别 153–160 个转录因子（在 1,833 个候选项中排名第一），而使用不同向导时仅能正确识别 36–46 个。因此，重复使用向导会使表面上的跨细胞类型识别率虚高约四倍。

为避免这一混杂因素，第 6 节和表7中的分析将每个转录因子的向导分成两个互不重叠的集合，并始终比较跨细胞类型使用不同集合所获得的响应。