arXiv:2601.06939v1 [cond-mat.dis-nn] 11 Jan 2026

独立费米子神经量子态极小值之间几乎不存在能垒

David D. Dai[1], * 和 Marin Soljačić[1], †
[1]麻省理工学院物理系，美国马萨诸塞州剑桥市 *02139*（日期：2026年1月13日）

神经量子态（NQS）已被证明能够高效表示量子多体波函数，但其损失景观仍未得到充分理解，且存在争议。本文证明，NQS 的损失景观比以往认为的更为良性，与传统深度学习更为相似，并呈现出模式连通性：独立训练的 NQS 可通过参数空间中的路径彼此连通，且几乎不存在能垒。为构造这些路径，我们开发了 GeoNEB，这是一种路径优化器，将高效的随机重构与弹性带方法相结合，用于构造最小能量路径。对于由一个含 1.6M 个参数的 Psiformer 建模的强相互作用六电子量子点，我们发现了两个独立的极小值，其预期能垒比系统的整体能量尺度小 $\sim 10^{-5}$ 倍，比线性路径的能垒小 $\sim 10^{-3}$ 倍。该路径在实现低能量的同时也遵循物理对称性，角动量在整个过程中保持良好的量子化。我们的工作首次构造出独立训练的 NQS 之间的优化路径，并表明 NQS 的损失景观或许并不像过去担心的那样病态。

引言

深度神经网络为解决量子多体问题带来了新的范式。神经量子态（NQS） [1–11] 将波函数参数化为深度神经网络 Ψ_θ，并通过最小化试探波函数的期望能量 [12–18]，在没有外部数据的情况下训练参数 θ。近期用于连续空间费米子的架构，如 FermiNet [6, 19] 和 Psiformer [9]，能够达到化学精度，并在原子和分子体系 [6, 9, 19]中超越传统方法。

目前尚不清楚，传统深度学习中的直觉和模式在多大程度上适用于神经量子态 [20]。在自然语言处理和计算机视觉 [23], 中，随机梯度下降（SGD） [21]或 Adam [22] 等一阶优化器已是标准且足够；而神经量子态通常需要复杂的二阶优化器 [24–28] 才能取得最佳结果。例如，与 KFAC（克罗内克分解近似曲率）相比，使用 Adam 训练 FermiNet 和 Psiformer 会得到显著更差的能量 [6, 9]。这种不确定性导致了人们对神经量子态损失景观结构的不同看法。一种观点认为，使用随机重构或 KFAC 等能力足够强的优化器，就能训练大型且灵活的模型并从中获益 [6, 9]。另一些人则认为，神经量子态优化受到“一般而言崎岖而复杂的损失景观” [29]的阻碍，因此主张改用受物理约束的试探波函数 [20]。

在本文中，我们从模式连通性的角度研究神经量子态的损失景观。在传统深度学习领域（例如图像分类）中，已有充分证据表明，独立训练得到的局部极小值可通过损失近乎恒定且较低的连续路径相连 [30, 31]。我们将这一分析拓展至量子情形。神经量子态优化需要从持久的马尔可夫链中采样，并使用量子几何张量进行预条件处理，因此许多传统深度学习中的路径搜索方法都无法适用。为应对神经量子态特有的挑战，我们提出 GeoNEB：量子几何张量引导的弹性带方法。该方法结合了弹性带方法 [31–33] 与利用核技巧 [34, 35], 高效实现的精确自然梯度下降，并明确遵循参数空间的量子几何。

对于强相互作用的六电子量子点，我们证明，两个独立训练的 1.6M 参数 Psiformer 模型可通过一条非线性路径相连，其能垒为 7.2×10^{-4} 哈特里；相比之下，基态能量高达 40.817 哈特里，而线性路径的能垒为 1.86 哈特里。我们进一步表明，这条路径符合物理对称性，且在整个路径上，角动量方差始终较小。所展示的结果来自训练得到的第一对极小值，且超参数调优极少。模式连通性的发现为反驳“崎岖景观”直觉提供了有力的经验依据，并表明神经量子态的损失景观是连通的，且通过随机重构即可较容易地探索。

神经量子态

我们研究的是 N 电子哈密顿量

$$H = -\frac{1}{2}\sum_i \nabla_i^2 + \sum_{i<j} V_{\text{int}}(\mathbf{r}_i - \mathbf{r}_j) + \sum_i V_{\text{ext}}(\mathbf{r}_i). \quad (1)$$

试探波函数 $\Psi_\theta(\mathbf{R})$ 必须满足粒子交换下的反对称性，在两个电子相互靠近时满足尖点条件，并在无穷远处满足正确的边界条件。我们采用了 Psi-

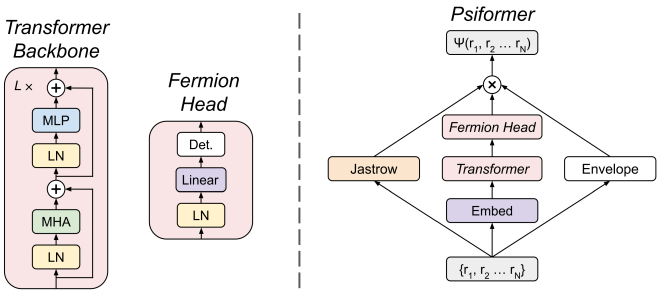


图 1. Psiformer 架构示意图。电子坐标经过嵌入后，再由 Transformer 层处理。行列式用于确保反对称性，而 Jastrow 因子和包络函数则用于处理尖点和边界条件。

former 架构（图 1） [9]，一种改编自 Transformer 模型的方法，这些模型在自然语言处理领域取得了巨大成功 [36] 并在计算机视觉领域也表现出色 [37]。

Transformer 主干网络。 输入构型 $\mathbf{R} = \{\mathbf{r}_1, \ldots, \mathbf{r}_N\}$ 使用原始笛卡尔坐标和“软半径”嵌入为特征向量 $\mathbf{h}_i^0$：

$$\mathbf{h}_i^0 = \mathbf{W}_{\text{embed}} \left(x_i, y_i, \sqrt{1 + r_i^2} \right). \tag{2}$$

对于量子点，我们发现使用软半径 $\sqrt{1+r_i^2}$ 至关重要。使用原始半径 r_i 会在原点引入人工尖点，导致局域能量发散；而完全去除径向信息，则会使网络失去一项重要特征。

将每个电子视为一个 token 后，这些特征会经过 L 个无掩码 Transformer 层。每层先应用多头自注意力（MHSA），随后通过 GELU 激活的多层感知机（MLP），并在这两个子层之前应用层归一化（LN） [38, 39]：

$$\tilde{\mathbf{h}}^\ell = \mathbf{h}^{\ell-1} + \text{MHSA}\left(\text{LN}\left(\mathbf{h}^{\ell-1}\right)\right), \tag{3}$$

$$\mathbf{h}^\ell = \tilde{\mathbf{h}}^\ell + \text{MLP}\left(\text{LN}\left(\tilde{\mathbf{h}}^\ell\right)\right). \tag{4}$$

Jastrow 和包络函数。Transformer 的输出经过一个 LN，再通过一个线性层，生成广义斯莱特行列式中使用的多体轨道。我们将行列式乘以一个 Jastrow 因子，以确保电子-电子尖点，并乘以一个包络函数，以确保 $r \to \infty$ 边界条件：

$$\begin{aligned} \Phi &= \mathbf{W}_{\text{orbital}} \text{LN}\left(\mathbf{h}^L\right) + \mathbf{b}_{\text{orbital}}, \\ \Psi_\theta(\mathbf{R}) &= \det\left[\Phi\right] \times e^{-\sum_{i<j} \mathcal{J}(\mathbf{r}_i - \mathbf{r}_j)} \times \prod_i \mathcal{K}(\mathbf{r}_i). \end{aligned} \tag{5}$$

量子点的 Jastrow 和包络函数形式为

$$\mathcal{J}(\mathbf{r}) = \frac{\beta \alpha^2}{\alpha + r}, \; \mathcal{K}(\mathbf{r}) = e^{-\frac{r^2}{2}}, \tag{6}$$

其中 α 是可训练的，而 β 是固定的，并与空间维度和耦合强度相匹配。

算法 1：GeoNEB（QGT引导的NEB）

要求： θ_L, θ_R (极小值), N (初始主元)
初始化： 在 θ_L, θ_R 之间线性布置 N 枢轴点
当尚未收敛**时执行**
  对于 $i = 1, \ldots, N$ **do**
    $\mathbf{R}_i \leftarrow$ 从 $|\Psi_{\theta_i}|^2$ 进行 MCMC 采样
    $\mathbf{g}_{\text{nat}} \leftarrow \text{SR}(\theta_i, \mathbf{R}_i)$ {方程。 8}
    $\hat{\tau} \leftarrow$ 归一化 $(\theta_{i+1} - \theta_{i-1})$
    $\mathbf{g}_\perp \leftarrow \mathbf{g}_{\text{nat}} - (\mathbf{g}_{\text{nat}} \cdot \hat{\tau})\hat{\tau}$
    $\mathbf{g}_\parallel \leftarrow k((2\theta_i - \theta_{i-1} - \theta_{i+1}) \cdot \hat{\tau})\hat{\tau}$
    $\theta_i \leftarrow \theta_i - \eta \, (\mathbf{g}_\perp + \mathbf{g}_\parallel)$
  结束循环
  如果iteration % $T_{\text{insert}} == 0$ **则**
    在损失较高的位置插入枢轴点
  结束条件判断
结束循环

图 2. GeoNEB 的伪代码。该曲线由相互耦合的神经量子态实例组成：每个实例都维护自己的马尔可夫链和局部量子几何张量，但会与相邻实例交互，以计算弹簧力和切线。

自然梯度优化。 训练过程通过最小化 $\langle\Psi_\theta| H |\Psi_\theta\rangle / \langle\Psi_\theta|\Psi_\theta\rangle = \mathbb{E}_{\mathbf{R}\sim|\Psi|^2}[E_L(\mathbf{R})]$，来进行，其中局域能量为 $E_L(\mathbf{R}) = H\Psi(\mathbf{R})/\Psi(\mathbf{R})$。为改善能量景观的病态曲率，我们采用随机重构（SR）。SR 最初被推导为将虚时间演化投影到变分流形上的方法，但实际上等价于使用量子几何张量（QGT）定义距离概念的自然梯度下降（NGD）。对于实波函数，QGT 等价于概率密度 $|\Psi_\theta|^2$ 的费希尔信息矩阵（FIM）。下文中，SR（QGT）和 NGD（FIM）将互换使用， $P \sim \mathcal{O}(10^6)$ 表示实参数的数量， $M \sim \mathcal{O}(10^3)$ 表示批次大小。

值得注意的是，精确的 SR/NGD 更新

$$\Delta\theta = -\eta \left(\lambda \mathbf{I}_P + \mathbf{F}\right)^{-1} \nabla E, \tag{7}$$

其中 $\mathbf{F}$ 是 $P \times P$ 量子几何张量， λ 是正则化对角偏移，无需显式构造庞大的 $\mathbf{F}$ [34, 35] 即可执行。定义得分 $O_{i\mu} = \partial \log \Psi(\mathbf{R}_i)/\partial\theta_\mu$ 后，可以将更新重写为 [40]

$$\begin{aligned} \mathbf{X}_{i\mu} &= \frac{1}{\sqrt{M}} \left(O_{i\mu} - \frac{1}{M} \sum_j O_{j\mu} \right), \\ \mathbf{Y}_i &= \frac{2}{\sqrt{M}} \left(E_L(\mathbf{R}_i) - \frac{1}{M} \sum_j E_L(\mathbf{R}_j) \right), \\ \Delta\theta &= -\eta \mathbf{X}^T \left(\lambda \mathbf{I}_M + \mathbf{X}\mathbf{X}^T\right)^{-1} \mathbf{Y}. \end{aligned} \tag{8}$$

难以处理的 $P \times P$ 参数空间求逆已转变为成本相对较低的 $M \times M$ 样本空间求逆。与原始 Fer- miNet 和 Psiformer 实现所采用的 KFAC 相比 [6, 9]，精确

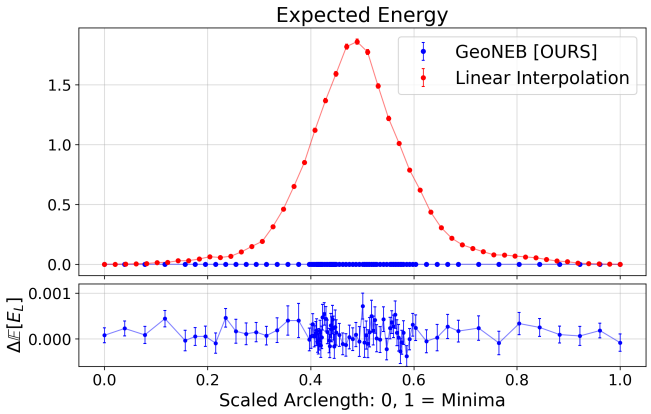


图 3. 能量势垒比较。GeoNEB 找到的路径最大势垒为 7.2×10^{-4} 哈特里，线性插值的势垒为 1.86 哈特里，基态能量为 40.817 哈特里。端点的欧几里得参数范数分别为 94.8 和 95.1，而它们之间的距离范数为 115.9，这证实端点相距甚远。每隔三个点取一个作为可训练枢轴点，其余点仅用于推理。

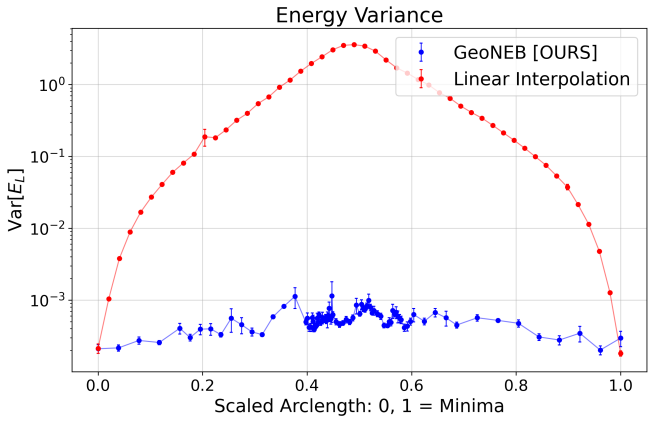


图 4. 能量方差比较。 为便于观察，图中使用对数刻度。GeoNEB 的峰值为 1.1×10^{-3} 哈特里^2，而线性插值的峰值为 3.6 哈特里^2。每隔一个点取一个可训练的枢轴点，其余点仅用于推理。

SR/NGD以一些额外开销为代价，避免了克罗内克因子结构近似 [27]；不过，这并未显著降低训练速度，因为拉普拉斯算子的计算无论如何都非常耗时。

路径优化算法

在两个极小值 θ_L 和θ_R 之间寻找低能量路径，是一个非平凡的优化问题。朴素的线性插值 $\theta(t) = (1-t)\theta_L + t\theta_R$ 通常会穿越高能垒。我们要寻找一条弯曲路径 $\theta(t)$，使沿途能量的最大值最小。

神经量子态中的挑战。为神经量子态寻找此类路径面临着独特的挑战。首先，神经量子态训练会使用持久的马尔可夫链，从 $|\Psi_\theta|^2$ 中采样。如果评估点 $\theta(t)$ 在每次迭代时都不断变化，例如，如果 t 是通过采样得到的 $\in [0,1]$ [30]，那么每次迭代都需要进行代价高昂的预热。其次，随机重配置（SR）需要在评估点处计算局部量子几何张量（QGT）。假设我们使用由中点 $\theta(0.5)$ [30],定义的贝塞尔曲线，就不清楚应如何对来自 $\theta(0.2)$ 的损失梯度进行预条件化，因为 $\theta(0.5)$ 是曲线的变分参数，而不是 $\theta(0.2)$；然而，这两个参数并不相同，且具有不同的局部几何。

GeoNEB 优化器。 我们发现，AutoNEB（自动 nudged 弹性带）算法 [31]最适合用于改造。它不使用连续曲线，而是将曲线表示为由“枢轴点”$\{p_0, \ldots, p_{K+1}\}$组成的离散链，其中 $p_0 = \theta_L$和 $p_{K+1} = \theta_R$。枢轴点通过弹簧力相互作用，以保持间距相等；除此之外，它们彼此独立，分别是各自拥有持久马尔可夫链和 SR 机制的神经量子态。从概念上讲，我们可以将寻找曲线视为一种集成训练，只是采用了一种不寻常的正则化形式。

力推动。 为防止枢轴点滑离势垒并导致采样不足，AutoNEB 会移除损失梯度中平行于链的分量。为防止弹簧力使链过度变直，AutoNEB 会移除弹簧力的垂直分量。令 τ_i 为枢轴点 i 处链的切线， k 为弹簧强度，则枢轴点 i的更新分量为

$$\mathbf{g}_i^{\perp} = \nabla_{\mathrm{nat}} E(\theta_i) - (\nabla_{\mathrm{nat}} E(\theta_i) \cdot \hat{\tau}_i)\,\hat{\tau}_i \tag{9}$$

$$\mathbf{g}_i^{\parallel} = k\,((2\theta_i - \theta_{i-1} - \theta_{i+1}) \cdot \hat{\tau}_i)\,\hat{\tau}_i \tag{10}$$

我们的贡献是将最初使用的原始损失梯度替换为自然能量梯度（经量子几何张量预处理）。我们不对弹簧力进行预处理，因为直观而言，它应直接作用于参数空间，以维持其中的等间距。

动态插入。 我们从少量枢轴点开始初始化。我们会定期在枢轴点之间的中点处评估插入判据，以检查路径分辨率。如果某个中点的判据劣于现有枢轴点中最差的判据，就将该中点作为额外的枢轴点插入。我们使用能量方差作为判据，因为它通常随能量误差线性变化 [41]，但信噪比更好。我们没有像 [31] 中那样重新分配枢轴点，因为这需要在每次迭代时对所有马尔可夫链进行预热。

结果

系统。 我们选择处于强相互作用区域、含有六个电子的自旋极化二维量子点作为测试平台

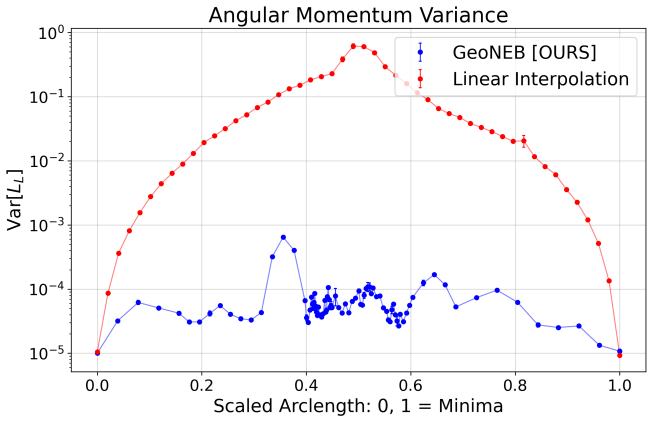


图 5. 角动量方差比较。图中采用对数刻度以便观察。GeoNEB 的峰值为 6.5×10^{-4} 原子单位^2，而线性插值的峰值为 0.61 原子单位^2。每隔一个点便是可训练枢轴点，其余点仅用于推理。GeoNEB 的势垒不可忽视，但从绝对意义上看，

$6.5\times$ 原子单位^2 10^{-4}

仍属良好表现；两者的差异反映的是端点的出色表现，而非路径的不足。

（耦合 $C=4$）作为测试平台。哈密顿量为

$$H=-\frac{1}{2}\sum_i\nabla_i^2+\sum_{i<j}\frac{C}{|\mathbf{r}_i-\mathbf{r}_j|}+\frac{1}{2}\sum_i\mathbf{r}_i^2. \qquad (11)$$

耦合 $C=4$ 不可忽视，因为电子之间的排斥作用很强，但又没有强到使电子高度局域化、从而易于用斯莱特行列式描述的程度。系统尺寸适中，因此能够对连接路径进行非常彻底的推理，以确认没有遗漏任何损失势垒。量子点的圆对称性使角动量成为衡量波函数质量的一项独立指标。在强耦合下，要获得良好量子化的角动量并非易事，因为对称性破缺带来的能量惩罚相较于整体能量尺度趋于消失。

设置。我们从随机初始化出发，训练了两个彼此独立的 Psiformer 模型，每个模型包含 1,586,182个实数参数。传统深度学习模型的规模可扩展至数十亿个参数，但 NQS 特有的拉普拉斯算子评估和蒙特卡洛采样成本高昂，因此实际可用的模型规模限制在约 $\mathcal{O}(10^6)$ 个参数。模型配置和训练超参数见补充材料。如表 I所示，两个模型均收敛至几乎相同的能量 ~40.817 哈特里，且能量方差极低$\sim3\times10^{-4}$ 哈特里^2。两个极小值在权重空间中相距甚远：其欧几里得参数范数分别为 94.8和 95.1，而二者之间的欧几里得距离为 115.9。

预期能量势垒。我们使用 GeoNEB 寻找连接这两个极小值的低能路径。图 3比较了 GeoNEB 优化路径与朴素线性插值路径上的能量分布。每

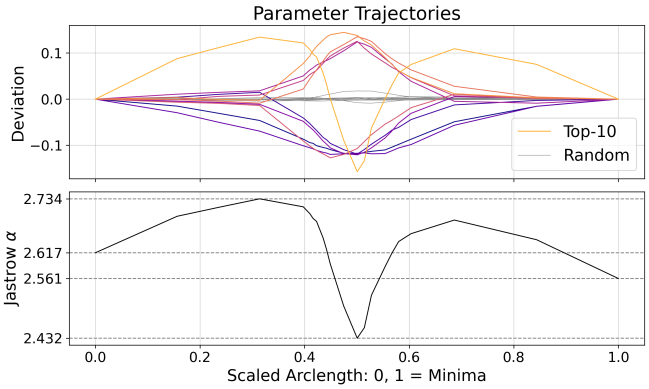


图 6. *GeoNEB* 参数轨迹。上图：线性插值偏差绝对值最大的前 10 个参数。下图：沿路径变化的 Jastrow α 参数。插值参数为 0.4 至 0.6 之间的 α 快速变化对应于最高的枢轴点集中度。

指标	单位	最小值 1	最小值 2
$\langle H\rangle$	Ha	40.81703(16)	40.81687(19)
var(H) $=\langle H^2\rangle-\langle H\rangle^2$	Ha^2	0.000210(28)	0.000296(72)
var(L) $=\langle L^2\rangle-\langle L\rangle^2$	原子单位^2	1.001(30)e-5	1.074(78)e-5

表 I.训练得到的极小值的结果。所有结果均使用训练期间从未见过的新测试游走者推断得出。这些极小值的欧几里得参数范数分别为 94。8 和 95。1，而它们之间的距离范数为 115。9，证实端点在参数空间中相距甚远。

第四个点是一个优化枢轴点，其余点则在推理时添加，用于检查是否遗漏了能量势垒。GeoNEB 路径上的最大能量升幅为 7.2×10^{-4} 哈特里，而线性路径上的最大能量升幅为 ≈1.86 哈特里，高出三个数量级以上。

算符方差势垒。为验证物理有效性，我们还考察了能量方差 Var$[E_L]$ 和角动量方差 Var$[L]$。图 4 显示，沿 GeoNEB 路径的能量方差始终较低，峰值为 1.1×10^{-3} 哈特里^2，而线性路径上的方差则飙升至 3。6哈特里^2。图 5 显示，角动量方差峰值为 6.5×10^{-4} 原子单位^2, ，也比线性插值低约三个数量级。尽管角动量从未直接出现在损失函数中，且架构并未显式强制圆对称性，我们仍取得了这些结果。虽然这一数值明显大于端点处的值，6.5×10^{-4} 原子单位^2 仍是非常好的表现；这种差异体现的是端点的出色表现，而非路径的不足。

参数轨迹。沿着优化后的路径，大多数参数都呈现平滑的拱形变化，如图 6所示。然而，Jastrow 因子中的 α参数控制着 Jastrow 因子的作用半径，在路径中心出现了一个尖锐的峰值

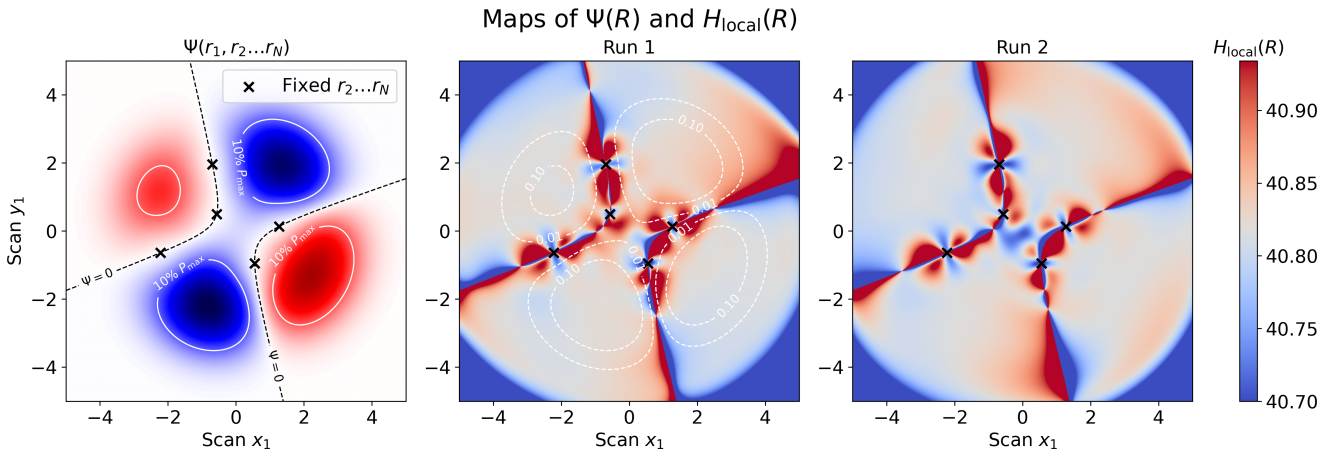


图 7. 训练极小值的空间可视化。 在所有子图中，一个电子在平面内移动扫描，其余电子固定不动。左：图像显示波函数振幅，颜色表示正负，饱和度表示大小；虚线标示节点面。两个模型的 $\Psi(\mathbf{R})$ 没有明显差异。中/右：两个极小值的局域能量图。尽管两个极小值在节点面附近都有误差，但其波动呈现不同的模式，尤其是在中心附近的 (0,0) 以及上方节点臂附近的 $(-1, 3)$。较窄的绘图范围在视觉上放大了偏差。

。由于 α 是一个对模型输出影响显著的单一标量，它在插值参数 0.4 到 0.6 之间的快速变化要求在该区间进行密集采样。与此同时，沿路径在合理范围内调整 α 仍能取得良好结果，这表明模型的其余部分可以补偿 α 的变化。

局域能量图。 图 7展示了实空间中的局域能量$E_L(\mathbf{R})$，并表明这些模型确实对应不同的波函数。两者在节点面附近都出现偏差，其中 $\Psi = 0$，但这些误差的具体局域结构存在一些差异。节点面附近的性能下降既可以理解，也难以修正，因为节点面对期望值的贡献很小，同时接收到的训练信号也很少。

讨论

我们首次通过优化路径观测到神经量子态中的模式连通性。由于我们的路径在实现低能量的同时，还实现了量子化角动量，因此损失景观不仅从纯粹能量最小化的角度看是连通的，从整体物理有效性的角度看也是如此。我们在尝试的第一对极小值之间就取得了这一结果，而且只需对曲线优化超参数进行最少调整，这表明模式连通性具有稳健性。尽管神经量子态的损失景观与传统深度学习的损失景观确实存在差异，但我们的工作表明，随机重构消除了大多数病态问题，并产生了相当良性的景观，展现出模式连通性等类似优点。

我们的工作还揭示了 Jastrow 在神经量子态优化中发挥的关键作用。Jastrow α参数在路径中心附近的快速变化表明，这类“关键参数”可能会使景观几何变得复杂，并损害优化效果。据我们所知，这种参数在神经量子态架构中独有。由于低能流形上存在一段非平凡的 α 值范围，而且直接训练得到的端点甚至具有不同的 α，将 Jastrow 固定为超参数或许能在几乎不损失精度的情况下改善优化。目前，我们也不清楚这种快速变化究竟是优化伪影，还是损失景观中真实且不可避免的特征。根据我们训练神经量子态的经验，我们怀疑这可能是由曲线优化期间采用的更激进的学习率调度所导致的伪影，但具体研究仍留待未来工作。

将模式连通性从传统深度学习推广到神经量子态，远非理所当然，尽管两者在高层次直觉上相似。现有关于模式连通性的理论，很大程度上依赖于监督分类和回归中特有的结构性假设。例如，噪声稳定性理论假设数据集是静态的，且逐样本损失仅取决于网络输出，并且关于网络输出是凸的 [42]。神经量子态的无监督能量最小化违反了这两项假设：数据分布由模型自身定义，而逐样本局域能量损失取决于网络的拉普拉斯算子。对于深度学习，我们的工作通过在一种与典型监督分类任务特征迥异的任务中观察到模式连通性，拓展了我们对模式连通性的理解。对于物理学，我们的工作表明，神经量子态的损失景观并不像人们曾担心的那样病态，而且过参数化并不会妨碍优化，甚至可能有益。

关于并行工作的说明。一篇并行预印本 [43]研究了连接极小值的线性插值，而非优化曲线。因此，他们发现独立模型之间存在较高的势垒，仅在通过微调或迁移学习相关联的模型之间找到了低势垒路径。我们的工作首次在严格意义上观察到模式连通性——即连接独立初始化并训练的模型的优化路径——这一定义源自传统深度学习文献。我们感谢他们从互补角度对迁移学习所做的探讨，并建议读者参阅他们的工作，以深入了解这一领域。

致谢

DDD 获得了美国国家科学基金会研究生研究奖学金的资助，资助编号为 2141064。作者感谢 MIT SuperCloud 和林肯实验室超级计算中心提供高性能计算资源。

* dddai@mit.edu
† soljacic@mit.edu

[1] G . Carleo 和 M.Troyer，利用人工神经网络解决量子多体问题，《科学》**355**，602（2017），https://www.science.org/doi/pdf/10.1126/science.aag2302 .

[2] Y. Nomura、A. S. Darmawan、Y. Yamaji 和 M. I mada，利用受限玻尔兹曼机学习求解强关联量子系统，Phys. Rev. B 96, 205152 2017( )。[3] G. Carleo、K. Choo、D. Hofmann、J. E. Smith、T. Wester- hout、F. Alet、E. J. Davis、S. Efthymiou、I. Glasser、S.-H. Lin、M. Mauri、G. Mazzola、C. B. Mendl、E. van Nieuwen- burg、O. O' Reilly、H. Théveniaut、G. Torlai、F. Vicen- tini 和 A. Wietek，NetKet：用于多体量子系统的机器学习工具包，SoftwareX 10, 100311 (2019)。[4] K. Choo、T. Neupert 和 G. Carleo，利用神经网络量子态研究二维受挫 J_1-J_2 模型，Phys. Rev. B 100, 125124 (2019)。[5] O. Sharir、Y. Levine、N. Wies、G. Carleo 和 A. Shashua，用于高效变分模拟多体量子系统的深度自回归模型，Phys. Rev. Lett. 124, 020503 (2020)。[6] D. Pfau、J. S. Spencer、A. G. D. G. Matthews 和 W. M. C. Foulkes，利用深度神经网络从第一性原理求解多电子薛定谔方程，Phys. Rev. Res. 2, 033429 (2020)。[7] J. Hermann、Z. Schätzle 和 F. Noé，电子薛定谔方程的深度神经网络求解，《自然·化学》12, 891–897 (2020)。[8] L. L. Viteritti、R. Rende 和 F. Becca，用于受挫量子自旋系统的 Transformer 变分波函数，Phys. Rev. Lett. 130, 236401 (2023)。[9] I. von Glehn、J. S. Spencer 和 D. Pfau，用于第一性原理量子化学的自注意力试探函数，载于《第十一届国际学习表征会议（2023）。[10] J. Hermann、J. Spencer、K. Choo、A. Mezzacapo、W. M. Foulkes、D. Pfau、G. Carleo 和 F. Noé，《利用神经网络波函数进行从头算量子化学》，《自然评论：化学》7，692–709（2023）。[11] H. Lange、A. Van de Walle、A. Abedinnia 和 A. Bohrdt，《从架构到应用：神经量子态综述》，《量子科学与技术》9，040501 2024( )。[12] R. P. Feynman，《液氦双流体模型的原子理论》，Phys. Rev. 94，262（1954）。[13] W. L. McMillan，《液氦的基态》[4], Phys. Rev. 138，A442（1965）。[14] D. Ceperley、G. V. Chester 和 M. H. Kalos，《多费米子研究的蒙特卡罗模拟》，Phys. Rev. B 16，3081（1977）。[15] D. M. Ceperley 和 B. J. Alder，《利用随机方法求电子气基态》，Phys. Rev. Lett. 45，566（1980）。[16] B. Tanatar 和 D. M. Ceperley，《二维电子气的基态》，Phys. Rev. B 39，5005（1989）。[17] W. M. C. Foulkes、L. Mitas、R. J. Needs 和 G. Rajagopal，《固体的量子蒙特卡罗模拟》，Rev. Mod. Phys. 73，33（2001）。[18] R. J. Needs、M. D. Towler、N. D. Drummond 和 P. López Ríos，《连续介质变分与扩散量子蒙特卡罗计算》，Journal of Physics: Condensed Matter 22，023201（2009）。[19] J. S. Spencer、D. Pfau、A. Botev 和 W. M. C. Foulkes，《更优、更快的费米子神经网络》（2020），arXiv:2011.07125 [physics.comp-ph]。[20] M. S. Moss、A. Orfi、C. Roth、A. M. Sengupta、A. Georges、D. Sels、A. Dawid 和 A. Valenti，《双重下降：神经量子态何时能够泛化？》（2025），arXiv:2508.00068 [cond-mat.dis-nn]。[21] I. S utskever、J. Martens、G. Dahl 和 G. Hinton，《深度学习中初始化和动量的重要性》，载于《第30届国际机器学习会议论文集》，《机器学习研究论文集》第28卷，S. Dasgupta 和 D. McAllester 编（PMLR，美国佐治亚州亚特兰大，2013），第1139–1147页。[22] D. P. Kingma 和 J. Ba，《Adam：一种随机优化方法》，载于国际学习表征会议（2015）。[23] Y. LeCun、Y. Bengio 和 G. Hinton，《深度学习》，《自然》521，436（2015）。[24] S.-i. Amari，《自然梯度在学习中高效奏效》，Neural Computation 10，251 1998( )，https://direct.mit.edu/neco/article-pdf/10/2/251/813415/089976698300017746.pdf。[25] S. Sorella，《使用随机重构的格林函数蒙特卡罗》，Phys. Rev. Lett. 80，4558（1998）。[26] S. Sorella，《用于变分量子蒙特卡罗的广义兰佐斯算法》，Phys. Rev. B 64，024512（2001）。[27] J. Martens 和 R. Grosse，《利用克罗内克分解近似曲率优化神经网络》，载于《第32届国际机器学习会议论文集》，《机器学习研究论文集》第37卷，F. Bach 和 D. Blei 编（PMLR，法国里尔，2015），第2408–2417页。[28] J. Stokes、J. Izaac、N. Killoran 和 G. Carleo，《量子自然梯度》，Quantum 4，269（2020）。

[29] H. Lange、G. Bornet、G. Emperauger、C. Chen、T. Lahaye、S. Kienle、A. Browaeys 和 A. Bohrdt，《Transformer 神经网络与量子模拟器：模拟强关联系统的混合方法》，Quantum 9，1675（2025）。[30] T. Garipov、P. Izmailov、D. Podoprikhin、D. P. Vetrov 和 A. G. Wilson，《损失曲面、模式连通性与深度神经网络的快速集成》，载于《神经信息处理系统进展》第31卷，S. Bengio、H. Wallach、H. Larochelle、K. Grauman、N. Cesa-Bianchi 和 R. Garnett 编（Curran Associates, Inc.，2018）。[31] F. Draxler、K. Veschgini、M. Salmhofer 和 F. Hamprecht，《神经网络能量景观中几乎不存在势垒》，载于《第35届国际机器学习会议论文集》，《机器学习研究论文集》第80卷，J. Dy 和 A. Krause 编（PMLR，2018），第1309–1318页。[32] H. Jónsson、G. Mills 和 K. W. Jacobsen，《寻找跃迁最小能量路径的弹性带方法》，载于《凝聚相模拟中的经典与量子动力学》，B. J. Berne、G. Ciccotti 和 D. F. Coker 编（世界科学出版社，新加坡，1998），第385–404页。[33] E. L. Kolsbjerg、M. N. Groves 和 B. Hammer，《自动化弹性带方法》，《化学物理学杂志》145，094107（2016）。[34] A. Chen 和 M. Heyl，《通过高效优化增强深度神经量子态》，《自然·物理》20，1476 2024( )。[35] R. Rende、L. L. Viteritti、L. Bardone、F. Becca 和 S. Goldt，《一种用于优化大规模神经网络量子态的简单线性代数恒等式》，Communications Physics 7，260（2024）。[36] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. J ones、A. N. Gomez、L. u. Kaiser 和 I. Polosukhin，《注意力就是你所需要的一切》，载于《神经信息处理系统进展》第30卷，I. Guyon、U. V. Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett 编（Curran Associates, Inc.，2017）。[37] A. Dosovitskiy、L. Beyer、A. Kolesnikov、D. Weissenborn、X. Zhai、T. Unterthiner、M. Dehghani、M. Minderer、G. Heigold、S. Gelly、J. Uszkoreit 和 N. Houlsby，《一张图像胜过16×16个词：面向大规模图像识别的 Transformer》，载于国际学习表征会议（2021）。[38] R. Xiong、Y. Yang、D. He、K. Zheng、S. Zheng、C. Xing、H. Zhang、Y. Lan、L. Wang 和 T. Liu，《论 Transformer 架构中的层归一化》，载于《第37届国际机器学习会议论文集》，《机器学习研究论文集》第119卷，H. D. III 和 A. Singh 编（PMLR，2020），第10524–10533页。[39]K. He、X. Zhang、S. Ren 和 J. Sun，《深度残差网络中的恒等映射》，载于《计算机视觉——ECCV 2016》，B. Leibe、J. Matas、N. Sebe 和 M. Welling 编（Springer International Publishing，瑞士 Cham，2016），第630–645页。[40] 这里我们专门讨论实波函数，但按照 [34, 35]中的方法，所有内容都可以轻松推广到复波函数。[41] W. Fu、W. Ren 和 J. Chen，《神经网络变分蒙特卡罗的方差外推方法》，《机器学习：科学与技术》5，015016 (2024)。[42] R. Kuditipudi、X. Wang、H. Lee、Y. Zhang、Z. Li、W. Hu、R. Ge 和 S. Arora，《解释多层网络低成本解的损失景观连通性》，载于神经信息处理系统进展，第 32 卷，H. Wallach、H. Larochelle、A. Beygelzimer、F. d'Alché- Buc、E. Fox 和 R. Garnett 编（Curran Associates, Inc., 2019）。[43] V. Hernandes、T. Spriggs 和 E. Greplova，《神经量子态中的模式连通性》，载于北极光深度学习会议摘要 *2026*（2025）。

补充材料

S1. 架构细节

与 [9], 相比 我们使用前归一化而非后归一化，启用 QKV 偏置，并启用轨道偏置 s.

参数	值
Transformer 层	4
注意力头	4
模型维度	256
MLP 隐藏维度	256
广义行列式	1
参数总量	1,586,182

表 S1. 架构配置。

S2. 训练细节

我们对学习率采用线性预热调度，之后将其保持恒定。我们不使用动量或权重衰减。在我们的参数化机制下，由于形成矩阵 $\mathbf{X}\mathbf{X}^T$ 的成本，SR 的成本在经验上随批次大小的平方增长；形式上的 $\mathcal{O}(M^3)$ 求逆开销较小，并非主要成本。为降低这项开销，我们采用子采样/小批量处理：将包含 1024 个游走者的完整批次分成 4 个、每个包含 256 个游走者的分块，在各分块内独立计算 SR 更新，然后对分块更新取平均。该方法在保持训练稳定性的同时，将主要开销降低了 4 倍。

进行蒙特卡洛采样时，我们使用 Metropolis-Hastings 算法，并采用各向同性高斯提议分布移动所有坐标。我们会自适应调整提议宽度，使接受率维持在约 23.4%。每 10 次迭代进行一次调整：如果接受概率过高（过低），则将宽度乘以（除以）1.05。估计梯度时，使用稳健的基于中位数的估计量进行局部能量裁剪： $E_L^{\text{clip}} = \text{median}(E_L) \pm 5 \cdot \text{MAD}(E_L)$，其中 MAD 是中位数绝对偏差。

参数	值
总迭代次数	50,000
预热迭代次数	5,000
峰值学习率	0.25
SR 对角偏移	0.001
批次大小	1,024
SR 子采样	4
每次迭代的 MCMC 步数	50
预热迭代次数	1000

表 S2. 训练超参数。

参数	值
总迭代次数	5,000
预热迭代次数	250
初始主元	5
插入间隔	500次迭代
弹簧常数（k）	0.0
批次大小	512
SR 子采样	2
预热迭代次数	500

表S3. GeoNEB超参数。