We care more about the **shape of the derivative** of an activation function than the function itself because the derivative directly affects **gradient flow** during backpropagation. The activation function determines how signals pass through the network, but its derivative controls how efficiently weights are updated. If the derivative is too small (vanishing gradient, like in sigmoid), learning slows down, while if it is too large (exploding gradient), updates become unstable. A well-shaped derivative, like in ReLU or Swish, ensures stable and efficient learning, making the network train effectively without getting stuck or diverging.