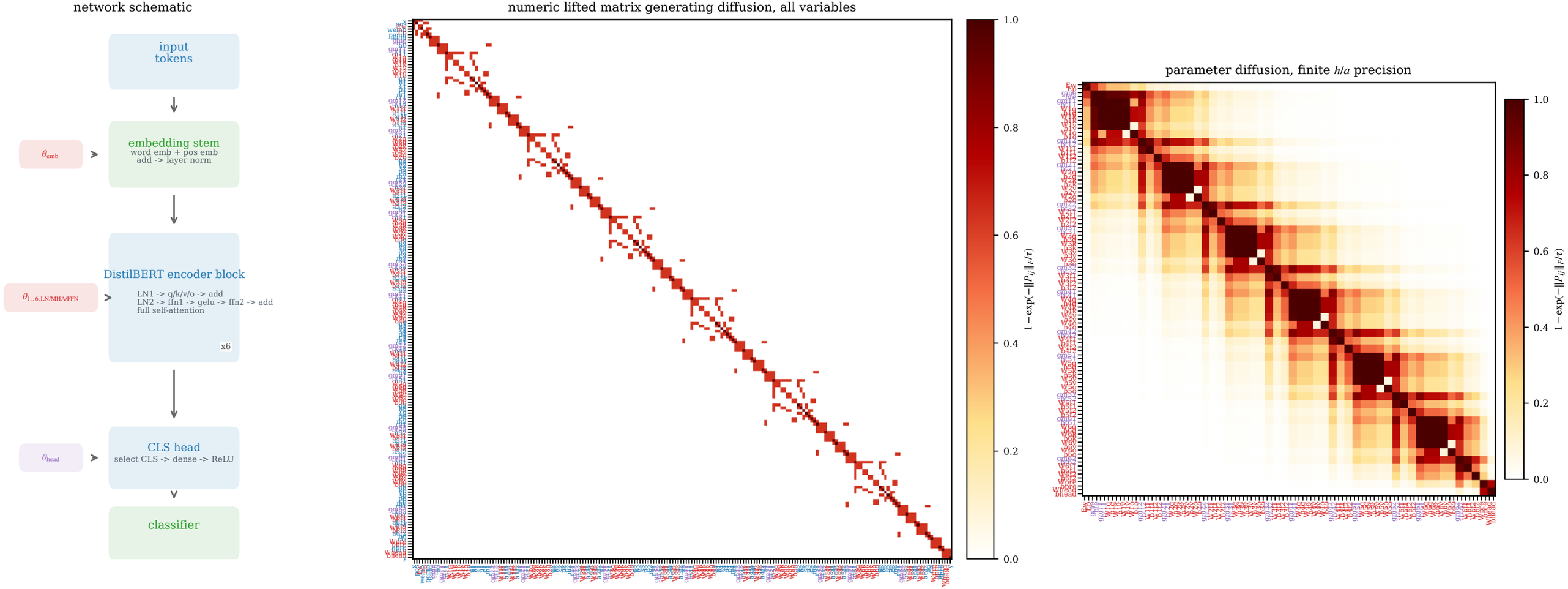
network schematic
input tokens
embedding stem
word emb + pos emb
add -> layer norm
θ_{emb}
DistilBERT encoder block
LN1 -> q/k/v/o -> add
LN2 -> ffn1 -> gelu -> ffn2 -> add
full self-attention
x6
$\theta_{1..6,\text{LN/MHA/FFN}}$
CLS head
select CLS -> dense -> ReLU
θ_{head}
classifier
numeric lifted matrix generating diffusion, all variables
$1-\exp(-\|P_{ij}\|_F/\tau)$
parameter diffusion, finite h/a precision
$1-\exp(-\|P_{ij}\|_F/\tau)$