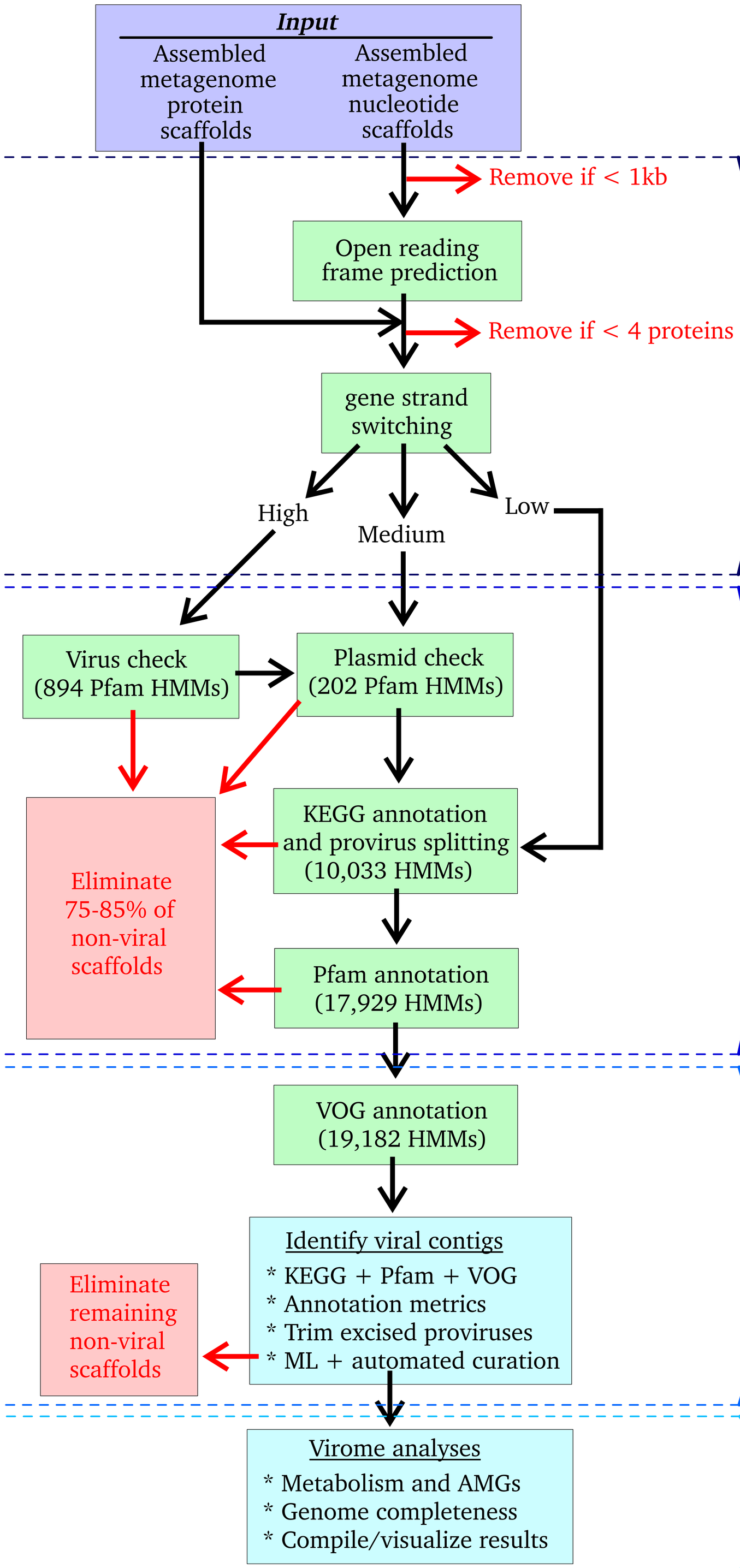

Step 1

Nucleotide scaffolds input: remove all scaffolds below 1kb or user-set minimum limit, then predict open reading frames (encoded proteins) using Prodigal.

Nucleotide or protein input: remove all scaffolds that encode less than 4 proteins or user-set minimum limit.

Bin scaffolds according to program-set cutoffs of high, medium and low level of open reading frame strand switching.

Step 2

High level: annotate with 894 Pfam HMM profiles representative of common viral proteins. Remove if zero proteins are annotated.

Medium level and passed high level: annotate with 202 Pfam HMM profiles representative of common plasmid proteins. Remove if 3 or more proteins are annotated.

Low level and passed high/medium level: annotate with 10,033 KEGG HHM profiles. Split scaffold if a putative provirus region is identified based on annotation v-scores of 4-protein sliding windows. Remove scaffolds that display non-vial signatures of KEGG annotation. Annotate with 17,929 Pfam HMM profiles and remove scaffolds that display non-vial signatures of Pfam annotation

Step 3

Annotate remaining scaffolds with 19,182 VOG HMM profiles. Compile 27 annotation metrics for each scaffold using KEGG/Pfam/VOG abundances, v-scores and discrete annotation signatures. Trim putative provirus regions to eliminate non-viral ends.

Apply the 27 annotation metrics to identify viruses using a neural network machine learning (ML) model. Implement automated curation steps to review and verify the ML results. Remove remaining non-viral scaffolds and compile identified viruses.

Step 4

For all identified viruses: assess virome metabolic capacity using auxiliary metabolic genes (AMGs), estimate genome completeness, identify circular genomes, and compile all results of annotations and metrics.