Data-mining. A lot of effort has been made to collect as many samples for each of the 6 datasets as we could find online. This is to assure that the results are a good estimate of the real performance of the pipeline.

* Antiviral – Number of samples after removing similar entries:
  + Positive: taken from
    - SATPdb: <http://crdd.osdd.net/raghava/satpdb/down.php> [27] and
    - UniProt: [www.uniprot.org](http://www.uniprot.org), looking for reviewed sequences of length between 4 and 42 with a keyword of antiviral protein
  + Negative: taken from
    - UniProt: [www.uniprot.org](http://www.uniprot.org), looking for reviewed sequences of length between 4 and 42 without a keyword of antiviral protein
* Cell penetrating – Number of samples after removing similar entries:
  + Positive: taken from
    - cellppd: <http://crdd.osdd.net/raghava/cellppd/dataset.php> by merging the sequences found in CPPsite1, CPPsite2 [28], Independent, Sanders2011-b [11], Dobchev-2010 [29], Hansen-2008 [12], Hallbrink-2005 [16]
  + Negative: taken from
    - UniProt: [www.uniprot.org](http://www.uniprot.org), looking for reviewed sequences of length between 4 and 34
* Antibacterial – Number of samples after removing similar entries:
  + Positive: taken from
    - SATPdb: <http://crdd.osdd.net/raghava/satpdb/down.php>
    - UniProt: [www.uniprot.org](http://www.uniprot.org), looking for reviewed sequences of length between 4 and 100 and annotated as having antibacterial function
    - APD database [27] latest version with discretion of Dr Wang and his team
  + Negative: taken from
    - UniProt: [www.uniprot.org](http://www.uniprot.org), looking for reviewed sequences of length between 4 and 100 and not annotated as having antibacterial function
* Anticancer – Number of samples after removing similar entries:
  + Positive: taken from
    - SATPdb: <http://crdd.osdd.net/raghava/satpdb/down.php> ,
    - cancerPPD: <http://crdd.osdd.net/raghava/cancerppd/downseq.php> [10] (under sequences having natural amino acids),
    - sequences from the supplementary information of Chen et al. [7],
    - anticp: <http://crdd.osdd.net/raghava/anticp/datasets.php> [30] by merging the sequences found in main and Independent:1 dataset from and
    - UniProt: [www.uniprot.org](http://www.uniprot.org), looking for reviewed sequences of length between 4 and 50 and annotated as having anticancer function
  + Negative: taken from
    - UniProt: [www.uniprot.org](http://www.uniprot.org), looking for reviewed sequences of length between 4 and 50 and not annotated as having anticancer function
* Toxic – Number of samples after removing similar entries:
  + Positive: merged from
    - SATPdb: <http://crdd.osdd.net/raghava/satpdb/down.php> and
    - UniProt: [www.uniprot.org](http://www.uniprot.org), looking for reviewed sequences of length between 4 and 50 and annotated as having toxic dose or as venom.
  + Negative: taken from
    - UniProt: [www.uniprot.org](http://www.uniprot.org), looking for reviewed sequences of length between 4 and 50 and not annotated as having toxic dose or as venom
* Tumour homing - – Number of samples after removing similar entries:
  + Positive: taken from
    - TumorHPD [6] by merging the main and small datasets: <http://crdd.osdd.net/raghava/tumorhpd/down.php>
  + Negative: taken from

UniProt: [www.uniprot.org](http://www.uniprot.org), looking for reviewed sequences of length between 4 and 24