The following is all the relevant information that has to be stored in the DB (includes an example and description):

* **AlphaMissense:** Annotates missense variants with the pre-computed [AlphaMissense](https://www.science.org/doi/10.1126/science.adg7492" \t "_blank) pathogenicity scores. It predicts the pathogenicity of single nucleotide missense variants. **Alphamissense** refers to a specific tool or database, developed to predict the impact of missense variants on protein function using deep learning models. It is different from simply identifying a missense variant; it involves predicting whether a missense variant is deleterious (likely to affect protein function) or benign.

"alphamissense": {

                    "am\_pathogenicity": 0.2711,

                    "am\_class": "likely\_benign"

                },

* **CADD** (Combined Annotation Dependent Depletion): includes deleteriousness scores for single nucleotide variants (also supports sus\_scrofa), indels and structural variants (only supported in GRCh38).
* **cadd\_phred**: The PHRED-like scaled CADD score, where higher scores indicate greater likelihood of the variant being deleterious.
* **cadd\_raw**: The raw CADD score, which is the raw output from the CADD model before scaling.

"cadd\_raw": -0.620477,

"cadd\_phred": 0.066,

* **spliceaAI:** Retrieves pre-calculated annotations from SpliceAI a deep neural network, developed by Illumina, Inc that predicts splice junctions from an arbitrary pre-mRNA transcript sequence (value **1** will be used: Ensembl/GENCODE v24 canonical transcripts (masked scores)).
  + **Delta scores (DS)** range from 0 to 1, where values close to 1 indicate a high likelihood of splicing disruption.
  + **Distance parameters (DP)** indicate how close the variant is to the site in question; negative values may indicate upstream locations, and positive values downstream locations relative to the variant.

In this SpliceAI annotation, "SYMBOL": "VCAM1" indicates that the sequence variant being analyzed is located in or near the **VCAM1** gene, and the potential impact of the variant on the splicing of this gene is being predicted.

"spliceai": {

                    "DS\_AG": 0,

                    "DP\_DG": -28,

                    "DP\_AL": -45,

                    "SYMBOL": "VCAM1",

                    "DS\_DL": 0,

                    "DP\_DL": 8,

                    "DS\_AL": 0,

                    "DS\_DG": 0,

                    "DP\_AG": 13

                },

* **gencode\_basic:** Limit your analysis to transcripts belonging to the GENCODE basic set. This set has fragmented or problematic transcripts removed. If Gencode Basic were not used, you might see more transcripts and possibly different annotations that are not part of this trusted subset, which could complicate the interpretation by including less reliable predictions. Each transcript\_consequence entry will now pertain to transcripts marked as GENCODE basic. Information for transcripts not included in this set will be omitted from the output.
* "transcript\_consequences": [
* {
* ...
* [
* "gene\_id": "ENSG00000168685",
* "gene\_symbol": "IL7R",
* "transcript\_id": "ENST00000511982",
* "biotype": "protein\_coding",
* "consequence\_terms": [
* "downstream\_gene\_variant"
* ],
* **hgvs**: Include HGVS nomenclature based on Ensembl stable identifiers. The output will include HGVS (Human Genome Variation Society) notation for the variants. HGVS notation provides a standardized way to describe variants in DNA, RNA, and protein sequences.
  + "hgvsc": "ENST00000648951.1:n.949-1839C>T"
  + "hgvsp": "ENSP00000421207.1:p.Val138Ile",

Each "hgvsc" entry is an HGVS-compliant description of the variant at the transcript level. Here’s how the HGVS notation is broken down:

* **ENST00000648951.1**: This part represents the transcript ID (ENST00000648951) and its version (1).
* **n.949-1839C>T**: This represents the nucleotide change. The n. prefix indicates that it’s a description at the cDNA (nucleotide) level. The C>T indicates that at this position, a cytosine (C) has been replaced with thymine (T).
* **mane**: Include MANE Select annotations (GRCh38 only). MANE provides a high-quality set of matched annotations for the human genome.
  + "mane\_plus\_clinical": "NM\_001145661.2",
  + "mane\_select": "NM\_032638.5",
    - The mane\_select annotation represents the principal transcript for a gene, chosen as the most biologically relevant isoform. It is consistent across NCBI's RefSeq and Ensembl's GENCODE annotations.
    - **Example**: "mane\_select": "NM\_032638.5". This indicates that for the gene in question, the transcript identified by RefSeq as NM\_032638.5 is the MANE Select transcript. This transcript is the main reference used for clinical and functional studies.
    - The mane\_plus\_clinical annotation includes additional transcripts beyond the MANE Select. These transcripts are included because they are important for clinical interpretation, even though they might not be the primary isoform.
* **Example**: "mane\_plus\_clinical": "NM\_001145661.2". This indicates that the transcript NM\_001145661.2 is part of the MANE Plus Clinical set. This set is curated to include transcripts that are clinically relevant but may differ from the MANE Select transcript.

**• most\_severe\_consequence**: "downstream\_gene\_variant" (This means that the variant is located downstream of the CDKN2B-AS1 gene and may affect its expression or regulation without altering the actual coding sequence of the gene.)  
• **gene\_id:** "ENSG00000240498" (unique identifier given by Ensembl)  
• **gene\_symbol**: "CDKN2B-AS1" (name of gene)  
• **gene\_symbol\_source**: "HGNC" (source/database where gene symbol is standardized)  
• **hgnc\_id:** "HGNC:34341" (unique identifier for the gene provided by the HGNC database. It allows for precise reference and retrieval of gene-related information.)  
• **transcript\_id:** List all related transcript IDs ("ENST00000428597", "ENST00000580576", "ENST00000585267") (These are unique identifiers for the different transcripts (versions of the gene's mRNA) produced by the gene. Each transcript might have different exons and splicing patterns.)  
• **consequence\_terms:** [ "downstream\_gene\_variant" ] (lists the types of consequences or impacts that the genetic variant may have.  
**• impact**: "MODIFIER" (The impact level of the variant on gene function. "MODIFIER" indicates that the variant is expected to have a minor effect on the gene's function, often related to gene regulation or expression rather than coding sequence changes.)  
• **clin\_sig**: [ "risk\_factor" ] (Clinical significance of the variant. "risk\_factor" means that this variant is associated with an increased risk of developing a particular disease or condition.)

"clin\_sig": [

                    "benign"

                ],

"clin\_sig": [

                    "benign",

                    "pathogenic"

                ],

**• clin\_sig\_allele:** "C:risk\_factor" (Specifies which allele (in this case, "C") is associated with the clinical significance mentioned (e.g., increased risk).)

"clin\_sig\_allele": "G:benign",

**• phenotype\_or\_disease**: 1 (This indicates the presence of a phenotype or disease association. The value 1 typically denotes that the variant is associated with a particular condition or phenotype.)

"phenotype\_or\_disease": 1,

• frequencies: Include detailed frequencies for different populations (the following will be listed: gnomadg\_nfe, gnomadg, gnomadg\_amr, sas, etc.)

"frequencies": {

                    "G": {

                        "gnomadg\_amr": 0.4817,

                        "gnomadg\_mid": 0.6234,

                        "eas": 0.4256,

                        "gnomadg\_eas": 0.426,

                        "eur": 0.6322,

                        "sas": 0.7127,

                        "gnomadg\_sas": 0.7019,

                        "gnomadg": 0.5742,

                        "gnomadg\_fin": 0.6432,

                        "gnomadg\_afr": 0.5086,

                        "af": 0.5377,

                        "afr": 0.4614,

                        "gnomadg\_ami": 0.532,

                        "gnomadg\_oth": 0.5684,

                        "gnomadg\_asj": 0.631,

                        "amr": 0.4625,

                        "gnomadg\_nfe": 0.6239

                    }

                },

• **colocated\_variants:** Field that includes relevant colocated variant IDs and their clinical significance, frequencies, and synonyms. (information about other variants that are located at the same genomic position as the primary variant). But before there is already listed a few points that are in this field: “phenotype\_or\_disease”, “clin\_sig”, “clin\_sig\_allele”, frequencies (gnomadg\_amr...).