Copenhagen Ultrathon on Precision Medicine

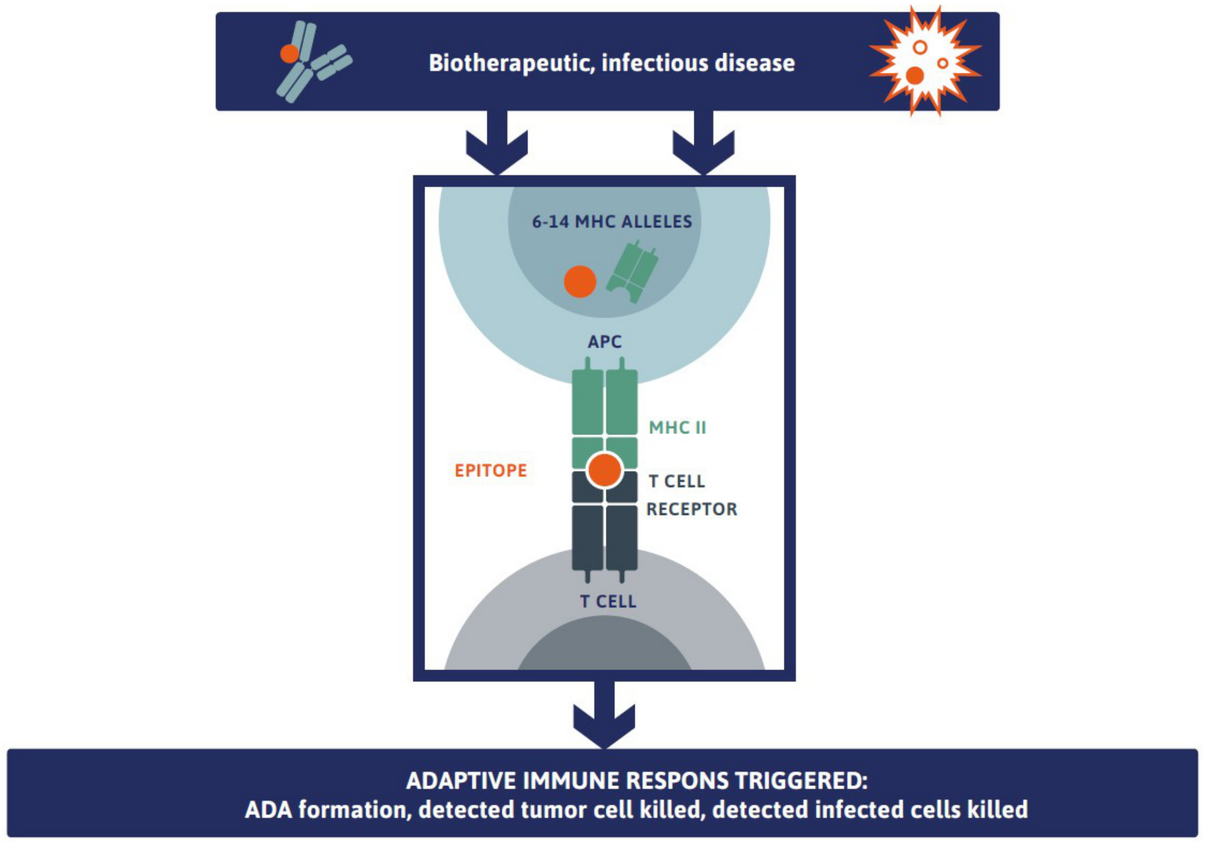
Marek Prachar
Sune Justesen
Daniel B. Steen-Jensen
Frederik O. Bagger

Prediction of peptide-epitope binding - the key to immune response, vaccine design and drug design

Challenge ID: U21-08

Main research question: Predict stability of binding from peptide sequence. Identifying and understanding receptor-ligand interactions is vital to comprehend biology on a cellular level. In the case of the immune systems' ability to recognize pathogens and elicit a cellular immune response the single most selective step is binding of peptides to the Major Histocompatibility Complex (MHC). The MHC is highly polymorphic and insights into the mechanisms governing the interactions in the binding cleft requires great investments in experimental works. Artificial Neural Networks have been applied as a high performing model to predict binding of peptides to the MHC class I, based on training from experimental measurements. However, the training data have been based on affinity (ability to bind) and not stability (staying bound). We have found the latter to be much more predictive for getting an immune reaction. Being able to predict immune response is critical for vaccine design (you want immune response), drug design (you don't want an immune response). The main technical challenge is the integration of data from several types of experiments, and the ability of coping with different lengths of peptide of input. Solving this problem will have large impact on understanding of the immune system and scientific and commercial efforts within vaccines (virus, personalized cancer vaccines) and production of safe drugs.

Secondary research question(s): Explore data enrichment from public sources, optimal peptide encoding and attention models for varying length peptides. Test model generalization on public data.



Biotherapeutic, infectious disease

ADAPTIVE IMMUNE RESPONS TRIGGERED:
ADA formation, detected tumor cell killed, detected infected cells killed

KEY

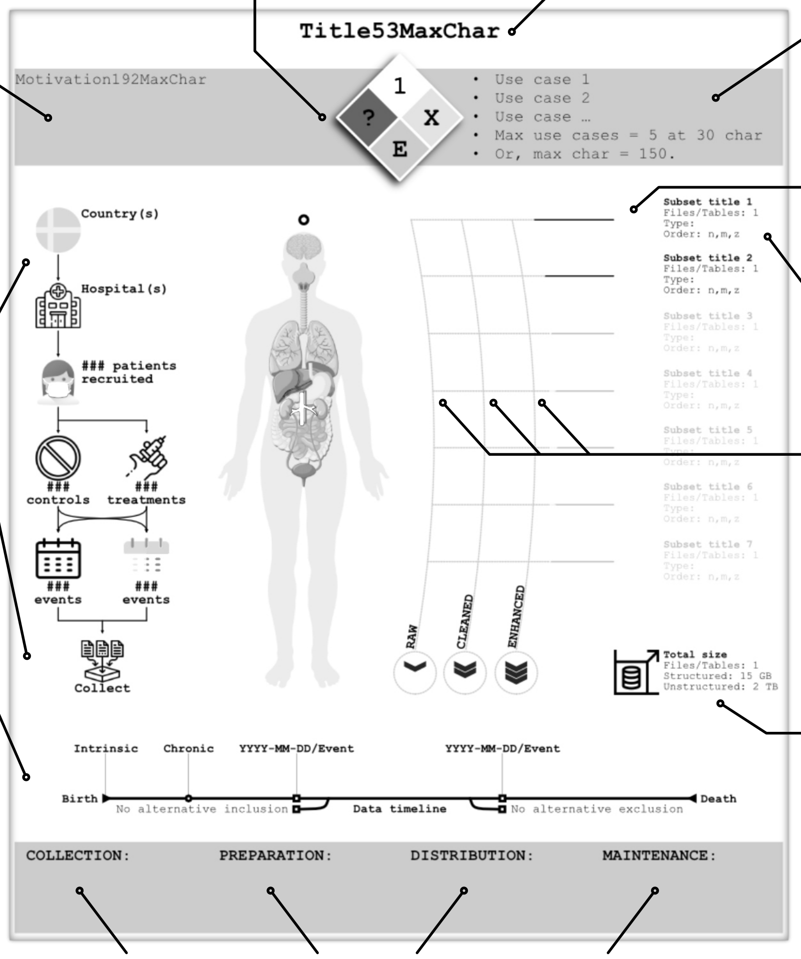
Document score from 1 to 3, 1 being basic and 3 being excellent.

Synthetic score from E to A, E having no synthetic data and A having synthetic data equivelant to the original.

Sample size in orders of magnitude {X^1, X^2, X^3, etc}.

Independent evaluation of bias and fairness in the dataset with respect to the population it was created to represent. Symbols mean:

- "?" No evaluation has taken place.
- "!" Extreme levels of bias.
- "✓" Almost no bias found.
- "~" A mix, be wary.



Dataset title. May be shortened as compared to the actual title.

A shortened version of the *MOTIVATION* section in the *Statement of intent* meant to clarify why the dataset exists.

Use cases. A brief summary of the *USES* section found in the *Statement of intent* meant to inspire.

Subset complexity. A visual depiction of tensor order.

Titles and statistics of subsets found within the dataset.

Flow chart summarizing cohort selection, treatment, and sampling.

Levels of data cleaning and feature engineering. Subsets containing data derived from other populations such as clinical risk scores or polygenic risk scores will have entries in the enhanced column.

Timeline of sampling and research design with respect to critical events. Datatypes are represented as glyphs with their position(s) denoting time and frequency.

Total disk size and file count of structured vs. unstructured data in the dataset.

Brief keyword summaries of the *COLLECTIONv*, *PREPARATION*, *DISTRIBUTION*, and *MAINTENANCE* sections found in the *Statement of intent*.

MAIDS Provenance

PROTO
V0.1
use
MAIDS
V0.1
fork
Ultra thon

The MAIDS specification and its version number detailing what a MAIDS document needs to describe.

The MAIDS repo providing a code base from which to build MAIDS documents.

This version of the MAIDS document forked from the original repository and with unique content added for the Ultrathon.

MEDICAL AI DATA SHEET

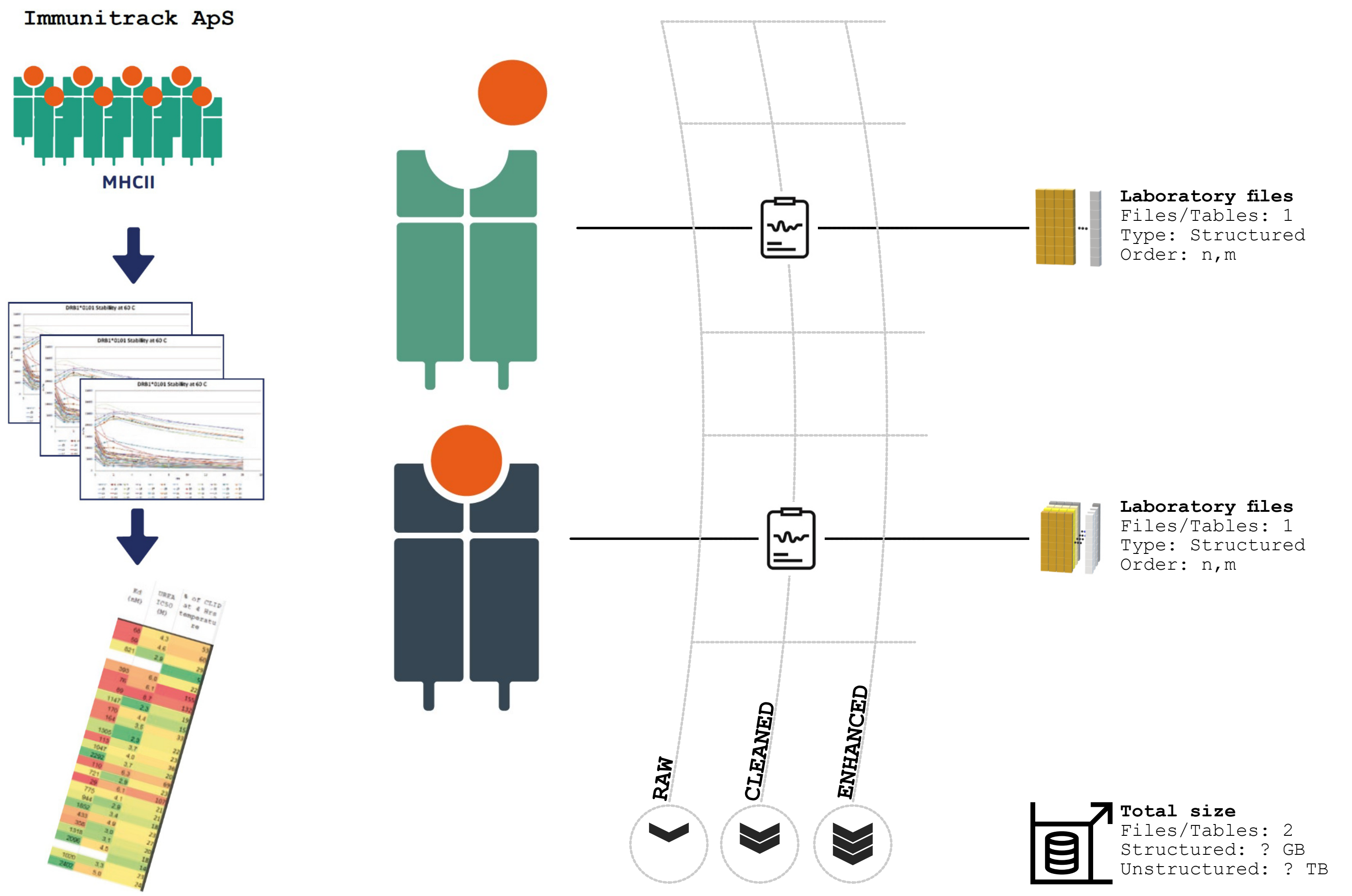
A principled standard for clinical data communication

The key to immune response, vaccine, and drug design

The stability of peptide:MHC complex was found to better correlate with activation of the immune reaction. Ability to predict immune response is critical for vaccine design & drug design.

1 ? X^3 E

- No use case disclosed.



Laboratory files
Files/Tables: 1
Type: Structured
Order: n,m

Laboratory files
Files/Tables: 1
Type: Structured
Order: n,m

Total size
Files/Tables: 2
Structured: ? GB
Unstructured: ? TB

COLLECTION:

- Experimental
- Model
- No cohort

PREPARATION:

- Cleaned
- Internal preparation.

DISTRIBUTION:

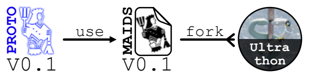
- Protected
- On request
- Collaboration
- Ultrathon 2021

MAINTENANCE:

- Undisclosed

Medical AI DataSheet (MAIDS) v0.1
Website: https://ultrathon.online
Project: github

MAIDS Provenance



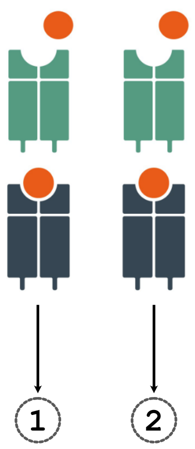
Connect
Twitter: https://twitter.com/UltrathonOnline
Email: ultrathon.rigshospitalet@regionh.dk
Author: Marek Prachar

Description of subsets

Table 1. Available Subsets

SID	Name	Format / Size	Purpose
1	Stability assay	CSV / 4919	Measurements of peptide:MHC stability. ELISA measurement, quantitative labels.
2	MS dataset 1	CSV / 2827	Measurements of peptide:MHC binding. Mass Spectroscopy, binary labels.

Subset relationships



Statement of intent

> MOTIVATION

Category 1-of-7 (4 questions)

The questions in this category are primarily intended to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.

M1: For what purpose was the dataset created? *Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.* Previous efforts of predicting peptide:MHC binding have been based on training data have been based on affinity (ability to bind) and not stability (staying bound). We have found the latter to be much more predictive for actually getting an immune reaction. Being able to predict immune response is critical for vaccine design (you want immune response), drug design (you don't want an immune response). The dataset consists of two types of data one is precise and expensive, the other less precise (binding/non-binding) and less expensive. Being able to make the full use the cheap data in a model would mean an explosion of available data for this type of problems. [By: Frederik O. Bagger]

M2: Who created the dataset (e.g. which team, research group) and on behalf of which entity (e.g. company, institution, organization)?. The stability dataset (D1) was created by Immunitrack ApS and the MS dataset (D2) was collected from IEDB. [By: Marek Prachar]

M3: Who funded the creation of the dataset? *If there is an associated grant, please provide the name of the grantor and the grant name and number.* Immunitrack ApS Innovation Foundation and Rigshospitalet. [By: Marek Prachar]

M4: Any other comments? The project is well suited for attention models or recurrent neural networks. The attention span is known (9 positions). The labels are well-defined, and it is possible to test on experimentally validated data. Recent outbreak of COVID-19 has shown how important this research is, and how much room there is for an improvement. Previous publications showing that current methods are not working well [Prachar et al. 2019], which could potentially have aided vaccine design, both in terms of speed-up and mutation resistance. [By: Frederik O. Bagger]

> COMPOSITION

Category 2-of-7 (17 questions).

Most of these questions are intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for specific tasks. The answers to some of these questions reveal information about compliance with the EU's General Data Protection Regulation (GDPR) or comparable regulations in other jurisdictions.

C1: What do the instances that comprise the dataset represent (e.g., samples, images, people)? *Are there multiple types of instances (e.g., samples, images, and people), interactions (e.g., nodes and edges), resolutions (e.g., genetic data, single cell expression vs. tissue expression, cell counts, different image technologies, etc.)? Please provide a description.* The whole dataset represents the event or lack thereof a step leading to activation of the immune response. D1: Peptides measured in a stability assay, investigating whether a there is a complex formed with the MHC molecule and how stable that complex is. Two columns: one with peptide (9 amino acids), one with measurement. ELISA measurement. The MS dataset represents peptide ligands investigated to be binding to MHC. D2: Each instance is a peptide found to be bound to MHC. Peptides vary in length. Only positive (binding) instances are present. Peptides with modifications (e.g. + OX(M14)) can be disregarded. [By: Marek Prachar]

C2: How many instances are there in total? *Provide an exact integer value for each type mentioned in question C1.* D1: 4919 (each instance represents a unique peptide:MHC) D2: 2827 (each instance represents a unique peptide:MHC). [By: Marek Prachar]

C3: Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representative-ness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., an active decision to cover a more diverse range of instances, because instances were withheld or unavailable).* The datasets represent a small sample from the whole peptide space. If peptides are tested randomly only around 1% of the measured peptides binds to MHC. The provided datasets are thus enriched and allow building a model that can contain features describing the studied problem. [By: Marek Prachar]

C4: What data does each instance consist of? *"Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.* Each data instance is a measurement of stability of a unique peptide:MHC complex, in case of the Stability dataset. For the MS datasets each instance is representing whether the peptide was/was not present in the eluted ligands. [By: Marek Prachar]

C5: Is there a label, target, or outcome (e.g., mortality) associated with each instance? *If so, please provide a description and indicate its actual presence within the dataset or whether it is represented by a* *proxy or compounded (e.g., a multi-cause event).* Yes, each peptide sequence is associated with an outcome. D1: normalized stability, measured with ELISA. Normalisation (% stability) is per batch, to a reference peptide (100%). D2: detection of the peptide via Mass Spec (binding/non-binding) binary value. [By: Marek Prachar]

C6: Is any information missing from individual instances? *If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.* No. [By: Marek Prachar]

C7: Are relationships between individual instances made explicit (e.g., familial links, or samples derived from the same patient or same exposure)? *If so, please describe how these relationships are made explicit.* No. [By: Marek Prachar]

C8: Are there recommended data splits (e.g., training, development/validation, testing)? *If so, please provide a description of these splits, explaining the rationale behind them.* For iterative models it is common to use 5-fold cross validation. Sometimes nested, such that there is also one rotating bin left for early stopping. [By: Frederik O.Bagger]

C9: Are there any errors, sources of noise, or redundancies in the dataset? *If so, please provide a description.* Yes. There are several duplicates in D1. Some of which are the control measurements, those can be disregarded when using the data. The control peptides are longer than 9 residues. D1 is known to have less noise than D2. [By: Marek Prachar]

C10: Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, public databases, other datasets and/or private silos)? *If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.* It is self-contained. Validation data (known immunogenic peptides) can be found at IEDB (www.iedb.org). One-hot encoding can be done without external data, but it is also possible to use evolutionary information (BLOSUM) or other distance matrix to encode each of the amino acids. [By: Frederik O.Bagger]

C11: Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? *If so, please provide a description.* Yes, D1 is confidential, unpublished and with commercial interest. Anyone needs to sign an NDA. [By: Marek Prachar]

C12: Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? *If so, please describe why.* The data is related to a mouse genotype (allele). When expanding this type of models it is important to consider that genotypes are covered in an unbiased manner (some alleles are more prevalent in some ethnicities). [By: Frederik O.Bagger]

C13: Does the dataset not relate to people (e.g., animals, cell lines, environment)? *A short answer is sufficient. If no relation to people, you may skip the remaining questions in this section.* Yes, it does not relate to people. The dataset relates to mice. [By: Marek Prachar]

C14: Does the dataset identify any subpopulations (e.g., by age, gender, etc.)? *If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.* Answer. [By: Name]

C15: Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? *If so, please describe how.* Answer. [By: Name]

C16: Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? *If so, please provide a description.* Answer. [By: Name]

C17: Any other comments? No. [By: Marek Prachar]

> COLLECTION PROCESS

Category 3-of-7 (13 questions).

If possible, dataset creators should read through these questions prior to any data collection to flag potential issues and then provide answers once collection is complete. In addition to the goals of the prior category, the answers to questions here may provide information that allow others to reconstruct the dataset without access to it.

L1: How was the data associated with each instance acquired? *Was the data directly observable (e.g., raw text, instrument measurements), reported by subjects/physicians (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses, scores, etc.)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.* Data was directly observed using an instrument. In the case of stability measurements the data is expressed as a binding percentage to a known stable binder. [By: Marek Prachar]

L2: What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? *How were these mechanisms or procedures validated?* D1: ELISA; D2: Mass Spectroscopy. [By: Frederik O.Bagger]

L3: If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? *Please describe.* Sampling strategies include selection with MS and known T cell epitopes from databases. [By: Marek Prachar]

L4: Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., salaried, immaterial through prizes / authorship / etc) and how much (e.g., according to competitive scales mandated by [insert body or institution])? D1: Employees at Immunitrack payed by Immunitrack ApS and Innovation Foundation. D2: Sofron et al. 2016 (doi: 10.1002/eji.201545930). [By: Marek Prachar]

L5: Over what timeframe was the data collected? *Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent data from old biobanked samples, or recent data dump from a 5-year-old registry)? If not, please describe the time frame in which the data associated with the instances was created.* Not relevant. [By: Marek Prachar]

L6: Were any ethical review processes conducted (e.g., by an institutional review board)? *If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.* No, not necessary. [By: Marek Prachar]

L7: Does the dataset not relate to people (e.g., animals, cell lines, environment)? *A short answer is sufficient. If no relation to people, you may skip the remaining questions in this section.* Yes, it relates to mice. [By: Marek Prachar]

L8: Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? *Please explain.* Answer. [By: Surname, name]

L9: Were the individuals in question notified about the data collection? *If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.* Answer. [By: Surname, name]

L10: Did the individuals in question consent to the collection and use of their data? *If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.* Answer. [By: Surname, name]

L11: If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? *If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).* Answer. [By: Surname, name]

L12: Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? *If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.* Answer. [By: Surname, name]

L13: Any other comments? No. [By: Marek Prachar]

> PREPROCESSING / CLEANING / LABELING

Category 4-of-7 (4 questions).

If possible, dataset creators should read through these questions prior to any preprocessing, cleaning, or labeling and then provide answers once these tasks are complete. The questions in this category are intended to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks.

P1: Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? *If so, please provide a description. If not, you may skip the remainder of the questions in this section.* Yes, the stability data is normalized as a percentage in relation to a known stable binder (control). [By: Marek Prachar]

P2: Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? *If so, is it available and needs to be done to gain access? If open without restriction then please describe a means to access this "raw" data.* Yes. [By: Marek Prachar]

P3: Is the software used to preprocess/clean/label the instances available? *If so, please provide a link or other access point and describe with enough detail so that others* *might reproduce it. If a custom script was used will you include it within the MAIDS repository or otherwise make it available.* Processing of the stability data was done using Microsoft Excel. [By: Marek Prachar]

P4: Any other comments? Answer. [By: Marek Prachar]

> USES

Category 5-of-7 (6 questions).

These questions are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harm.

U1: Has the dataset been used for any tasks already? *If so, please provide a description. A detailed response will help others determine the value of this dataset by example.* D1: data is novel. D2: Sofron et al. 2016 (doi: 10.1002/eji.201545930). [By: Marek Prachar]

U2: Is there a repository that links to any or all papers or systems that use the dataset? *If so, please provide a link or other access point. Will you compile such a list and make it available in the MAIDS repository.* No. [By: Marek Prachar]

U3: What (other) tasks could the dataset be used for? *Please provide as much inspiration as you can. Distinguish between tasks the dataset is ideal for versus those tasks where the dataset is not entirely suited. Describe why the dataset might not be suitable.* Vaccine design for mice. Understanding the biology and binding preference of IAB. [By: Frederik O.Bagger]

U4: Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? *For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?* Data has commercial interest, is confidential and is under NDA. Findings can be published in agreement with Immunitrack ApS. The raw data can also be published with the paper, if needed, and in agreement with Immunitrack ApS (Who are eager to publish a nice paper on this). [By: Frederik O.Bagger]

U5: Are there tasks for which the dataset should not be used? *If so, please provide a description.* Data and any use and application is confidential, and should be discussed with Immunitrack ApS. [By: Frederik O.Bagger]

U6: Any other comments? Answer. [By: Marek Prachar]

> DISTRIBUTION

Category 6-of-7 (7 questions).

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

D1: Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? *If so, please provide a description. If not, then disregard the rest of the questions.* No, not allowed. [By: Frederik O.Bagger]

D2: How will the dataset be distributed (e.g., tarball on website, API, GitHub)? *Does the dataset have a digital object identifier (DOI).* Should stay on server. [By: rederik O.Bagger]

D3: When will the dataset be distributed? *A cautious response is more useful than an optimistic one.* For developers under NDA: When needed. For public: At time of publication. [By: rederik O.Bagger]

D4: Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.* Yes. Attached NDA. [By: rederik O.Bagger]

D5: Have any third-parties imposed IP-based or other restrictions on the data associated with the instances? *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.* Yes. Patent pending. [By: rederik O.Bagger]

D6: Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.* No. [By: rederik O.Bagger]

D7: Any other comments? No. [By: rederik O.Bagger]

> MAINTENANCE

Category 7-of-7 (8 questions).

As with the previous category, dataset creators should provide answers to these questions prior to distributing the dataset.

These questions are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers.

T1: Who is supporting/hosting/maintaining the dataset? *Please be as thorough as possible.* Immunitrack ApS. [By: Marek Prachar]

T2: How can the owner/curator/manager of the dataset be contacted (e.g., email address)? By email, at: mprachar @immunitrack.com. [By: Marek Prachar]

T3: Is there an erratum? *If so, please provide a link or other access point.* No. [By: Marek Prachar]

T4: Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? *If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub).* There is a possibility that more instances will be added from the public sources (MS data). [By: Marek Prachar]

T5: If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? *If so, please describe these limits and explain how they will be enforced.* The dataset does not relate to people. [By: Marek Prachar]

T6: Will older versions of the dataset continue to be supported/hosted/maintained? *If so, please describe how. If not, please describe how its obsolescence will be communicated to users.* No. Not expected to be a problem. [By: Marek Prachar]

T7: If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? *If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/ distributing these contributions to other users? If so, please provide a description.* It is possible to extend the MS data, IEDB is a great source (www.iedb.org). [By: Surname, name]

T8: Any other comments? No. [By: Marek Prachar]