Efficient Search for New Physics Using Active Learning in the ATLAS Experiment

Zubair Bhatti[1*], *Kyle* Cranmer[2], *Irina* Espejo[1], *Lukas* Heinrich[3], *Phillip* Gadow[4], *Patrick* Rieck[1], and *Janik* von Ahne[5] on behalf of the ATLAS Computing Activity

[1]Department of Physics, New York University, USA
[2]Department of Physics, University of Wisconsin-Madison, USA
[3]Department of Physics, Technische Universität München, Germany
[4]CERN, Geneva, Switzerland
[5]DESY, Hamburg and Zeuthen, Germany

Abstract. Searches for new physics set exclusion limits in parameter spaces of typically up to 2 dimensions. However, the relevant theory parameter space is usually of a higher dimension but only a subspace is covered due to the computing time requirements of signal process simulations. An Active Learning approach is presented to address this limitation. Compared to the usual grid sampling, it reduces the number of parameter space points for which exclusion limits need to be determined. Hence it allows to extend interpretations of searches to higher dimensional parameter spaces and therefore to raise their value, e.g. via the identification of barely excluded subspaces which motivate dedicated new searches. In an iterative procedure, a Gaussian Process is fit to excluded signal cross-sections. Within the region close to the exclusion contour predicted by the Gaussian Process, Poisson disc sampling is used to determine further parameter space points for which the cross-section limits are determined. The procedure is aided by a warm-start phase based on computationally inexpensive, approximate limit estimates such as total signal cross-sections. The procedure is applied to a dark matter search on data collected by the ATLAS detector at the LHC, extending its interpretation from a 2 to a 4-dimensional parameter space while keeping the computational effort at a low level. The result is published in two formats: on one hand there is a publication of the Gaussian Process model. On the other hand, a visualization of the full 4-dimensional contour is presented as a collection of 2-dimensional exclusion contours where the 2 remaining parameters are chosen by the user.

1 Introduction

Beyond Standard Model (BSM) searches for new physics at the Large Hadronic Collider (LHC) by the ATLAS Collaboration [1] look for an excess of events over Standard Model (SM) predictions and barring discovery, seek to exclude highly parameterized models governing BSM processes subject to constraint from experimental data. The computational demand for making constraints by estimating exclusion contours is often intractable even for models

*e-mail: zb609@nyu.edu

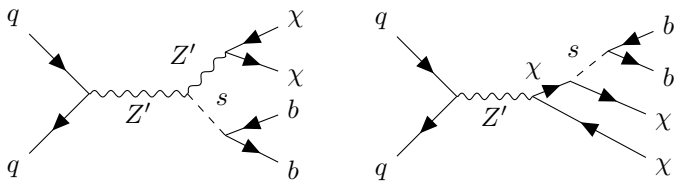

Figure 1. Feynman diagrams of the production of a dark Higgs boson s together with a pair of dark matter particles χ, mediated by a Z' particle which also interacts with the initial state quarks.

with reduced parameter sets, forcing searches which further reduce the dimensionality of the model being interpreted.

Introduced in Ref. [2] as a method for selecting data to add to the training set for a Machine Learning task, Active Learning (AL) is used here as an efficient strategy for querying exclusion limits in a high dimensional BSM search. Over a few iterations, small batches of points are evaluated to obtain upper limits on the cross-section of the BSM signal. A Gaussian Process (GP) is used to suggest which points should be evaluated for the following batch. This search, fully detailed in Ref. [3], utilizes a previously preserved analysis with the RECAST protocol and a computationally cheap exclusion limit estimation technique based on the SimpleAnalysis framework. Details about RECAST and SimpleAnalysis are available in Ref. [4] and Ref. [5].

AL is applied here in a search for dark matter produced in association with a SM Higgs boson that decays into b-quarks. This search covers four dimensions of the parameter space that was previously limited to two. This proceeding is organized as follows: a description of the dark matter signal process, the pipelines used for determining exclusion limits based on RECAST and SimpleAnalysis, how Active Learning is used for this search, and the results.

2 Mono-H($b\bar{b}$) Reinterpretation

Motivated by the need to generate mass for dark matter, this search targets a BSM model with a scalar dark Higgs boson s, a heavy spin-1 neutral Z' boson and heavy Majorana dark matter fermions χ, detailed in Ref. [6]. The production of a dark Higgs, which decays into b-quarks, and χ mediated by the Z' is the signal of interest here, shown in Figure 1. Sufficiently large mixing between the SM Higgs boson and the dark Higgs boson leads to the prompt decay of the dark Higgs, resulting in the coupling of the SM Higgs boson to SM fermions to be the same for the dark Higgs boson. The BSM parameter space is spanned by the masses of the dark Higgs, the Z' and χ and the Z'-χ and Z'-quark couplings: m_s, $m_{Z'}$, m_χ, g_χ, and g_q. The dominant background considered is the SM production of Higgs bosons decaying to b-jets with Z-bosons that decay into neutrinos.

The Mono-H($b\bar{b}$) analysis, detailed in Ref. [7], reinterpreted for this search selects events with the magnitude of missing transverse momentum $E_{\mathrm{T}}^{\mathrm{miss}}$ and at least two b-tagged jets originating from the decay of B-hadrons. Unlike those originating from the SM Higgs boson, the selected invariant mass of the b-jet pair originating from the dark Higgs can range from 50–280 GeV. Events belonging in the signal region must have $E_{\mathrm{T}}^{\mathrm{miss}} > 150$ GeV and not have charged leptons. The control region is defined orthogonally to the signal region by differing $E_{\mathrm{T}}^{\mathrm{miss}}$ requirements and requiring one or two leptons in the event. A simultaneous profile-likelihood fit to the control region and signal region is performed to set a limit on the signal cross-section.

3 Evaluating Parameter Points

A BSM search takes as input Monte Carlo (MC) samples of background processes, ATLAS experimental data, and MC samples of the signal at a single parameter setting and outputs an exclusion limit on the signal cross-section. For this dark Higgs model search, the space spanned by $\boldsymbol{\theta}$,

$$\boldsymbol{\theta} = \left(m_{Z'}, m_s, m_\chi, g_\chi\right) \tag{1}$$

is the input and the logarithm of the signal cross-section upper limit,

$$y \coloneqq \log\left(\frac{\sigma^{\text{Upper Limit}}_{\text{experiment}}}{\sigma_{\text{theory}}}\right), \tag{2}$$

is used to find the exclusion contour output at $y = 0$. To make fitting the cross-section contours easier, log is used for linearizing the limits. The g_q parameter factors out of the cross-section σ calculation and is fixed to 0.25 for this search. The map from parameter space to exclusion limits is approximated using a RECAST based pipeline and SimpleAnalysis based one, discussed below.

3.1 RECAST Pipeline

An analysis preserved using the RECAST protocol uses archived experimental data and background estimates and requires a new MC signal sample for its pipeline which follows these steps:

Job Options Requesting new signal samples begins with creating, validating and registering python-based configuration files. The files define which MC sample and parameter setting to use for event generation jobs on the Worldwide LHC Computing Grid (WLCG). The requester validates the configuration files locally by generating a small fraction of the total events that will be requested on the WLCG to estimate compute resources needed. Then registration of the configuration files occurs through an internal review process by system experts.

MC Production The requester then creates, submits and requests approval for WLCG jobs to begin running. The WLCG jobs step through event generation, detector simulation and reconstruct analysis-specific files.

Analysis The final step of the RECAST pipeline is the relatively straightforward running of analysis jobs on the REANA platform provided by CERN from which the exclusion limits are extracted. See Ref. [8] for details about REANA.

3.2 SimpleAnalysis Pipeline

The high accuracy of results from querying the RECAST based pipeline comes at a high computational cost and can involve multiple teams at ATLAS. To address this, an alternate method called SimpleAnalysis sacrifices some accuracy for a computationally cheaper analysis that can be run autonomously. SimpleAnalysis substitutes the most demanding steps in the RECAST pipeline for simplified detector responses, reconstruction and event selections. Discrepancies in the $E_{\text{T}}^{\text{miss}}$ requirement and matching variable-radius track jets with the large-radius jets were found between the SimpleAnalysis approach and the full analysis resulting

in differences of up to 20% in the signal selection efficiency. By first estimating the exclusion contour with SimpleAnalysis, this search was able to build a strategy to query the RECAST pipeline for even fewer parameter points.

4 Active Learning Strategy

Active Learning can be used to iteratively determine how to query a system for training data. In the case of BSM searches, the system to query is the computationally expensive full accuracy pipeline, preserved here with RECAST. For high dimensional searches, the number of queries to make scales exponentially so being conservative in the number of exclusion limits to request is necessary. A computationally efficient querying strategy should then only seek to evaluate points that provide the most information about the exclusion hyper surface. Each iteration of AL begins with training a GP to estimate the exclusion contour from the exclusion limits obtained so far. Then the trained GP is used to determine which additional queries to make. The AL querying strategy seeks to minimize uncertainties of a GP fit to exclusion limits over parameter space so points will be computed only if they sharpen the hyper surface. AL terminates when the exclusion contour estimated by the GP is satisfactorily sharp.

4.1 Gaussian Process

The final output of the AL loop is a trained GP that provides an estimate of the exclusion contour over parameter space. The GP is trained on independent datasets $\mathcal{D}_{SA} = \{(\boldsymbol{\theta}^i, y^i_{SA})\}_{i=1..p}$ and $\mathcal{D}_R = \{(\boldsymbol{\theta}^i, y^i_R)\}_{i=1..q}$ with p being the number of SimpleAnalysis evaluations and q the number of RECAST evaluations.

The 2-task GP is then defined as

$$\begin{pmatrix} y_{SA}(\boldsymbol{\theta}) \\ y_R(\boldsymbol{\theta}') \end{pmatrix} \sim \mathcal{GP}\left(\begin{pmatrix} m(\boldsymbol{\theta}) \\ m(\boldsymbol{\theta}') \end{pmatrix}, \Sigma(\boldsymbol{\theta}, \boldsymbol{\theta}')\right), \quad \Sigma(\boldsymbol{\theta}, \boldsymbol{\theta}') = \begin{pmatrix} k_{11}(\boldsymbol{\theta}, \boldsymbol{\theta}) & k_{12}(\boldsymbol{\theta}, \boldsymbol{\theta}') \\ k_{21}(\boldsymbol{\theta}, \boldsymbol{\theta}') & k_{22}(\boldsymbol{\theta}', \boldsymbol{\theta}') \end{pmatrix}, \tag{3}$$

where the mean and kernel with hyperparameters are given by

$$m(\boldsymbol{\theta}) = \boldsymbol{w}^T \boldsymbol{\theta} + b \ , \tag{4}$$

$$k_{ij}(\boldsymbol{\theta}, \boldsymbol{\theta}') = k(\boldsymbol{\theta}, \boldsymbol{\theta}')\kappa_{ij} + \epsilon^2 \delta(\boldsymbol{\theta}, \boldsymbol{\theta}') \ , \quad \delta(\boldsymbol{\theta}, \boldsymbol{\theta}') = \begin{cases} 1, & \text{if } \boldsymbol{\theta} = \boldsymbol{\theta}' \\ 0, & \text{else} \end{cases} \ , \tag{5}$$

$$k(\boldsymbol{\theta}, \boldsymbol{\theta}') = \exp\left(-\frac{\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2}{2\, l^2}\right) \ , \text{ and } \quad \kappa_{ij} = \begin{cases} \sigma_{SA} \text{ if } i = j = 1 \\ \sigma_R \text{ if } i = j = 2 \\ \sigma_{SA-R} \text{ if } i \neq j \end{cases} \ . \tag{6}$$

The hyperparameters $\boldsymbol{w}$, b, l, ε, σ_{SA}, σ_R and σ_{SA-R} are determined by maximum likelihood estimation given the observed $\mathcal{D}_{SA}$ and $\mathcal{D}_R$. The shared mean function between both tasks and kernel are used by the 2-task GP to learn the correlation between $\mathcal{D}_{SA}$ and $\mathcal{D}_R$ regression tasks from both datasets.

4.2 Acquisition Function

In the context of querying strategies, an efficient approach should seek the most information about the exclusion contour from minimal queries to the RECAST pipeline. The benefit of using a GP over another regression method is that the GP has an intrinsic uncertainty

encoded in the kernel k. To leverage this feature of the GP to suggest optimal points to query, the exclusion probability is introduced:

$$p_{\text{excl}}(\boldsymbol{\theta}) = \int_{-\infty}^{0} g\left(y \,|\, \mu(\boldsymbol{\theta}),\, \sigma(\boldsymbol{\theta})\right) dy \,. \tag{7}$$

The probability for a point in parameter space to be excluded is modeled from Gaussians g of exclusion limits y conditional on means $\mu(\boldsymbol{\theta})$ and uncertainties $\sigma(\boldsymbol{\theta})$ in their limit estimates. To get a direct measure of the uncertainty about the exclusion of a point $\boldsymbol{\theta}$, it is useful to introduce the exclusion entropy over the exclusion contour:

$$H_{\text{excl}}(\boldsymbol{\theta}) = -p_{\text{excl}}(\boldsymbol{\theta}) \log p_{\text{excl}}(\boldsymbol{\theta}) - (1 - p_{\text{excl}}(\boldsymbol{\theta})) \log (1 - p_{\text{excl}}(\boldsymbol{\theta})) \,. \tag{8}$$

This is a direct measure of the uncertainty around the exclusion of a point $\boldsymbol{\theta}$ with a minimal value at $p_{\text{excl}} = 0$ and $p_{\text{excl}} = 1$ and maximum when $p_{\text{excl}} = 0.5$. Restated then, an efficient querying strategy will minimize exclusion entropy across the parameter space. This Maximum Entropy Search (MES) is an extension of the single point search introduced in Ref. [9]. One drawback of relying on MES for the search is in its strict focus which can lead to close bunching of suggestions. To address this, a mix of Poisson disc sampling,

$$P = \{\boldsymbol{\theta} \text{ s.t. } \bar{y}_{exp}(\boldsymbol{\theta}) - 2\bar{y}_{-1\sigma}(\boldsymbol{\theta}) < \bar{y}_{exp}(\boldsymbol{\theta}) < \bar{y}_{exp}(\boldsymbol{\theta}) + 2\bar{y}_{+1\sigma}(\boldsymbol{\theta})\} \tag{9}$$

implemented in Ref. [10], with MES was used as the acquisition function for this search:

$$\{\boldsymbol{\theta}^{*1}, ..., \boldsymbol{\theta}^{*q}\} = \text{argmax}_{\text{top-}\frac{q}{2}}\, H_{\text{excl}}(\boldsymbol{\theta}) \cup P \,. \tag{10}$$

The Poisson disc sampling is bounded by twice the standard deviation $\bar{y}_{\pm 1\sigma}(\boldsymbol{\theta})$ in the expected limit contour $\bar{y}_{exp}(\boldsymbol{\theta})$. Since RECAST pipeline queries can be executed in parallel, batches of points are requested at a time rather than serially. Equation 10, selects a batch with q points where 50% ($\frac{q}{2}$) are MES points and 50% are from Poisson disc sampling.

4.3 Warm Start

In order for the initial step of the AL loop to start, a batch of queries must already have been processed which clearly introduces an issue. To address this, computationally inexpensive SimpleAnalysis derived limits on a regular and fine grid are used to train a separate single task GP. By calculating the exclusion entropy from this GP, an initial batch of points to evaluate using RECAST are determined. This warm start step could be informed by even simpler cross-section based estimates as well. The full algorithm is given in Algorithm 1.

Algorithm 1 Active Learning approach to efficiently obtain 4D exclusion contours

Warm start input: SimpleAnalysis dataset $\mathcal{D}_{SA}$, Gaussian Process f, acquisition function $a(\boldsymbol{\theta})$ Eq. 10
Train f on $\mathcal{D}_{SA} = \{(\boldsymbol{\theta}^i, y^i_{SA})\}_{i=1..q}$
Select initial batch $\{\boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^q\}$ from $a(\boldsymbol{\theta})$
Evaluate $\{\boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^q\}$ with RECAST to get dataset $\mathcal{D}_{R,1} = \{(\boldsymbol{\theta}^i, y^i_R)\}_{i=1..q}$
Warm start output: Datasets $\mathcal{D}_{SA}$, $\mathcal{D}_{R,1}$
Active Learning input: Initial datasets $\mathcal{D}_{SA}$, $\mathcal{D}_{R,1}$, 2-task Gaussian Process g, acquisition function $a(\boldsymbol{\theta})$
$\mathcal{D}_R \leftarrow \mathcal{D}_{R,1}$
$j \leftarrow 1$
repeat
    Train g on $\mathcal{D}_{SA}$ and $\mathcal{D}_R$
    Determine H_{excl} and stopping condition
    $j \leftarrow j + 1$
    Select next batch $\{\boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^q\}_j$ from $a(\boldsymbol{\theta})$
    Evaluate $\{\boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^q\}_j$ with RECAST
    Add batch $\{\boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^q\}_j$ and its evaluations to $\mathcal{D}_R$
until H_{excl} is low
return $\mathcal{D}_{SA}$, $\mathcal{D}_R$, g
Active Learning output: Final datasets $\mathcal{D}_{SA}$, $\mathcal{D}_R$, g

5 Results

The region of parameter space this search investigates is bounded by $m_{Z'} \in [500, 5000]$ GeV, $m_s \in [50, 150]$ GeV, $m_\chi \in [100, 1200]$ GeV, $g_\chi \in [0.5, 2.0]$ and g_q set to 0.25. While Active Learning does not require a warm start in general, the multitask GP was informed with 5000 queries of the low-fidelity SimpleAnalyis pipeline to help reduce the number of RECAST evaluations needed. In each iteration, small batches of 200 queries of the RECAST pipeline were requested for a total of about 800 queries over 4 iterations until convergence. The full Run 2 ATLAS dataset was also used for the analysis. The GPyTorch library, introduced in Ref. [11], was used to fit the multitask GP on a single V100 NVIDIA GPU with 32 GB of RAM. To estimate the excluded region, the Bayesian Optimization procedure implemented in the python package *excursion* was used. See Ref. [12] for details about *excursion*.

Physics analyses at ATLAS inherently involve multiple steps due to their complex nature and some searches from the collaboration are implementing automation where feasible. See Ref. [13] for notable progress towards end-to-end automation. Our search benefited from improvements in the Job Options and MC Production steps and iteration of the Active Learning loop only took approximately 1 week to complete. After the final update of the GP, the contour is sharp enough and the exclusion entropy is low farther away from the contour. The GP posterior can be sampled at any point in the investigated 4D parameter space to provide expected and observed limits on the signal cross-section under various parameter settings. Some 2D slices are provided for visualization in Figure 2.

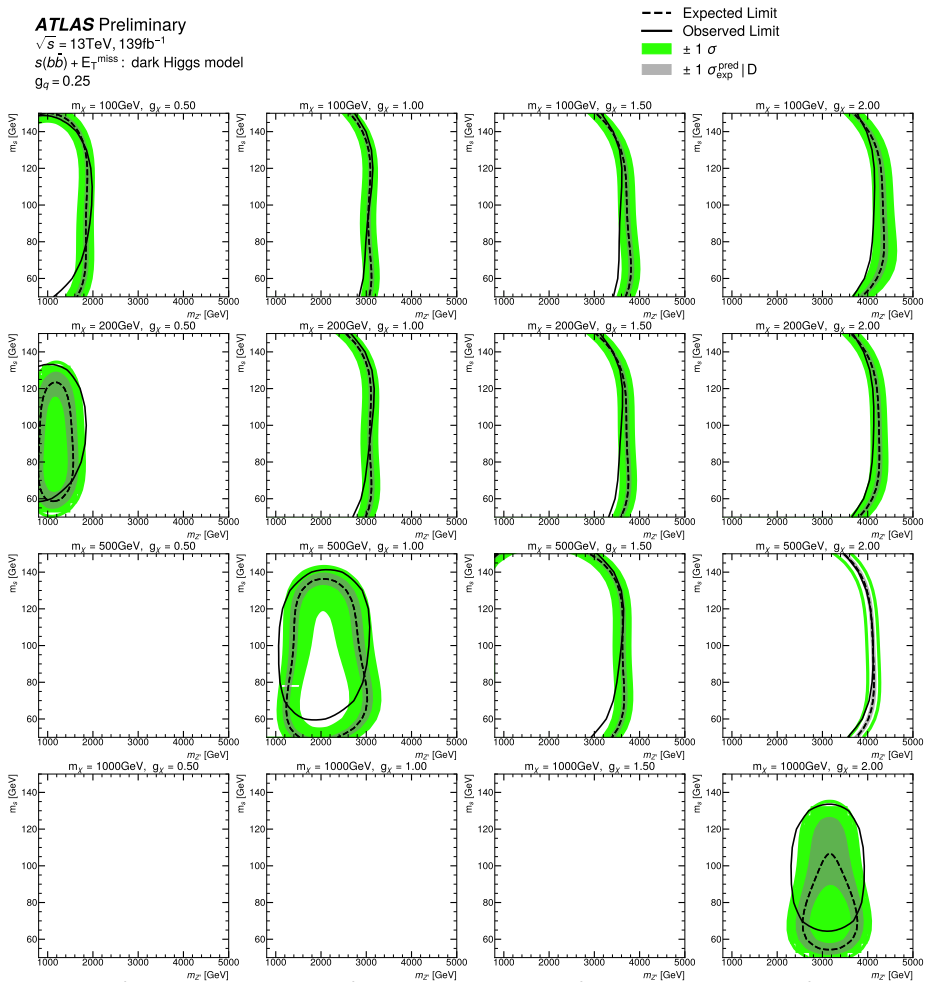

Figure 2. Exclusion limits on the mediator mass $m_{Z'}$ and dark Higgs mass m_s where the coupling g_q=0.25 and each slice corresponds to a fixed value for dark matter mass m_χ and the coupling g_χ detailed in Ref. [3]. The regions to the left of the contours are excluded. The exclusion limits are determined by evaluating the multitask GP in a four-dimensional grid. Blank plots are slices where no points were excluded.

6 Conclusion

In conclusion, this proceeding details how AL was used to set exclusion limits for a dark Higgs boson model using the full Run 2 dataset collected by the ATLAS detector at the LHC. This search reinterprets a previous Mono-H($b\bar{b}$) dark matter search with a pair of b-jets and $E_{\mathrm{T}}^{\mathrm{miss}}$ in the final state. As a querying strategy, AL determines which parameter space points should be evaluated through a pipeline that sets signal strength upper limits. By exploiting the intrinsic uncertainty of GPs, only evaluations that provide the most information about the exclusion contour are executed. The inherent latency in iteratively querying the pipeline was also improved on during this search by taking advantage of the automated production of simulation samples and utilizing RECAST and REANA for executing the analysis steps.

Exploring whether using total cross-sections may be sufficient for the first task of the GP is worth investigating, potentially dropping the need to use the SimpleAnalysis framework.

References

[1] ATLAS Collaboration, JINST **3**, S08003 (2008)

[2] D.A. Cohn, Z. Ghahramani, M.I. Jordan, JAIR **4**, 129 (1996)

[3] ATLAS Collaboration (2022), ATL-PHYS-PUB-2022-045, https://cds.cern.ch/record/2839789

[4] K. Cranmer, I. Yavin, JHEP **2011**, 38 (2011)

[5] ATLAS Collaboration (2022), ATL-PHYS-PUB-2022-017, https://cds.cern.ch/record/2805991

[6] M. Duerr, A. Grohsjean, F. Kahlhoefer, B. Penning, K. Schmidt-Hoberg, C. Schwanenberger, JHEP **2017**, 143 (2017)

[7] ATLAS Collaboration (2018), ATLAS-CONF-2018-039, https://cds.cern.ch/record/2632344

[8] T. Šimko, L. Heinrich, H. Hirvonsalo, D. Kousidis, D. Rodríguez, EPJ Web Conf. **214**, 06034 (2019)

[9] Z. Wang, S. Jegelka (2018), 1703.01968

[10] M. Vanga, *Poisson Disk Sampling in Processing* (2018), https://sighack.com/post/poisson-disk-sampling-bridsons-algorithm

[11] J.R. Gardner, G. Pleiss, D. Bindel, K.Q. Weinberger, A.G. Wilson (2021), 1809.11165

[12] L. Heinrich, G. Louppe, K. Cranmer, *diana-hep/excursion: v0.0.1-alpha* (2018), https://doi.org/10.5281/zenodo.1634428

[13] ATLAS Collaboration (2023), ATL-PHYS-PUB-2023-010, https://cds.cern.ch/record/2857975