Analytical High-Throughput Assays for Generation of Validated Datasets to Train ML Models for Cellulase Optimization

Simon Straß[1], Johanna Brückner[1], Sandra Maier[1], Evgenia Herbold[1], Tobias Hokamp[1], Philipp Schellenberger[2], Lenz Lorenz[2], Sven Benson[2], Philipp Kaiser[1] and Anne Zeck[1]

[1]NMI Natural and Medical Sciences Institute at the University of Tübingen
72770 Reutlingen, Germany

[2]Candidum GmbH
70569 Stuttgart, Germany

NMI Natural and Medical Sciences Institute
at the University of Tübingen
Markwiesenstrasse 55
72770 Reutlingen, Germany
Contact: simon.strass@nmi.de

Introduction

The efficient enzymatic hydrolysis of cellulose is essential for numerous biotechnological applications (e.g. recycling & biomass conversion). However, predicting cellulase activity on different substrates remains challenging.
Here, we present an experimental workflow for generating high-quality datasets that can serve as the foundation for ML-based cellulase activity predictions. While our dataset is still in development, preliminary data have been collected.
Our approach starts with a diverse library of cellulases which are tested in high-throughput assays to study the of the carbohydrate binding modules (CBM) and catalytic domains (CD) with papain digestion.

In general, all cellulases were tested following the scheme:

1. Proteolysis of cellulase domains with papain
2. Testing of enzymatic activity with different substrates
3. Confirmation & identification of proteolysis products by intact MS
4. Processivity assay & analysis via SEC & nanoDSF

By combining these structural and functional datasets, we provide a robust experimental framework that facilitates the correlation of biophysical properties with enzymatic performance.

Cellulose Degradation

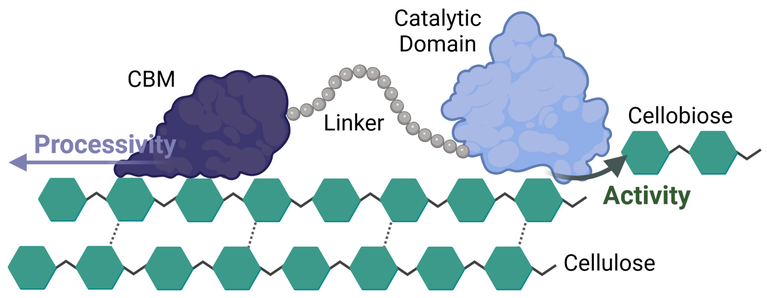


Figure 1: *Graphical abstract of cellulase structure with a carbohydrate binding module (CBM) and a catalytic domain (CD). Processivity refers to the ability of an enzyme to catalyze consecutive reactions without releasing its substrate.**

Structural Confirmation & Identification

Structural confirmation and identification was performed using **RPLC-MS** analysis of intact enzymes and papain-digested enzymes. Hereby, the cleavage sites of papain in the linker region were identified confirming separation of catalytic domain and CBM[1]. As alternative cleavage sites, clipping of terminal His-tag and TEV recognition site were observed. In general, **intact MS** can be used as high-throughput method for fast validation of protein primary amino acid sequence.

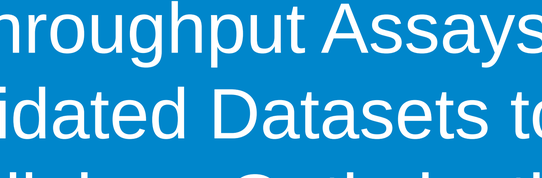


Figure 2: *Papain proteolysis cleaves native cellulases in linker domain to separate domains.**

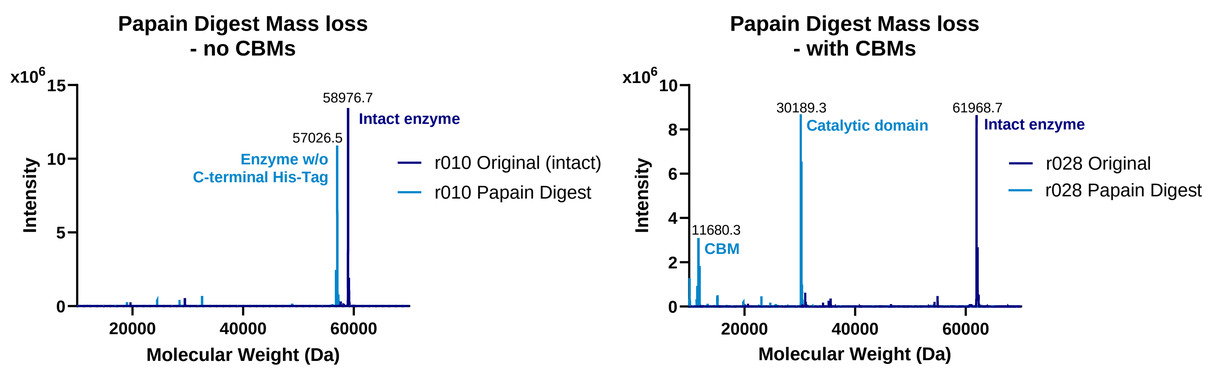


Figure 3: *Charge-deconvoluted mass spectrum of the main chromatographic peaks of cellulases r010 & r028 before and after papain proteolysis.*

Enzyme Activity Determination

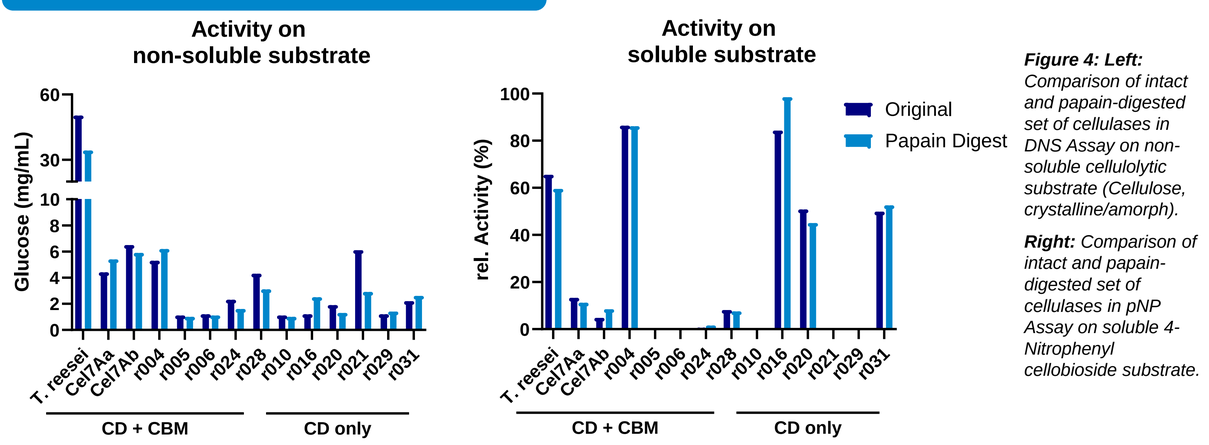
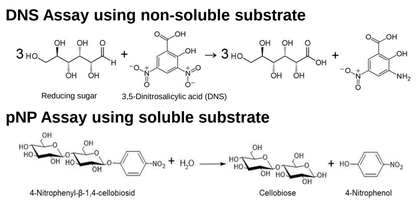


Figure 4: Left: *Comparison of intact and papain-digested set of cellulases in DNS Assay on non-soluble cellulolytic substrate (Cellulose, crystalline/amorph).*

Right: *Comparison of intact and papain-digested set of cellulases in pNP Assay on soluble 4-Nitrophenyl cellobioside substrate.*

Different assays are used to determine the ability of cellulases to hydrolyze specific substrates. Regardless of the substrate, most of the tested, individual enzymes generally showed a low overall activity. Using the soluble substrate **4-Nitrophenyl (pNP) cellobioside**, however, some enzymes showed high activity. Enzymes with CBM showed a slight trend for higher activity with insoluble substrates (with **DNS assay** on crystalline/amorph cellulose). Papain digestion had little to no effect on enzyme activity.



Size Analysis

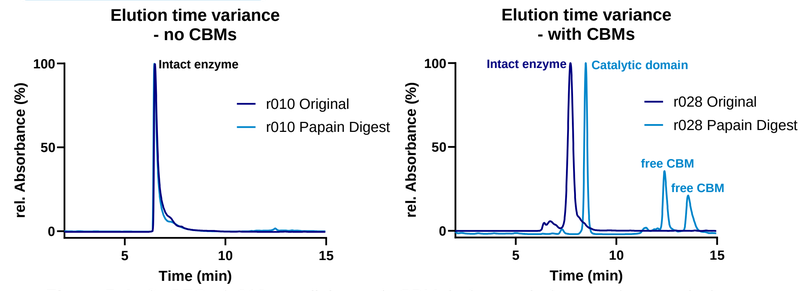


Figure 5: Left: *SEC of r010, a cellulose w/o CBM, before and after papain proteolysis.*
Right: *SEC of r028, a cellulose with CBMs, before and after papain proteolysis.*

Size exclusion chromatography (**SEC**) is used to separate and quantify the individual domains following papain digestion. This allows simple and high-throughput predictions to be made about the domain structure of cellulases. With the aid of **processivity assays**[2], statements can be made about the binding capacity of CBMs to defined substrates and be quantified via SEC.

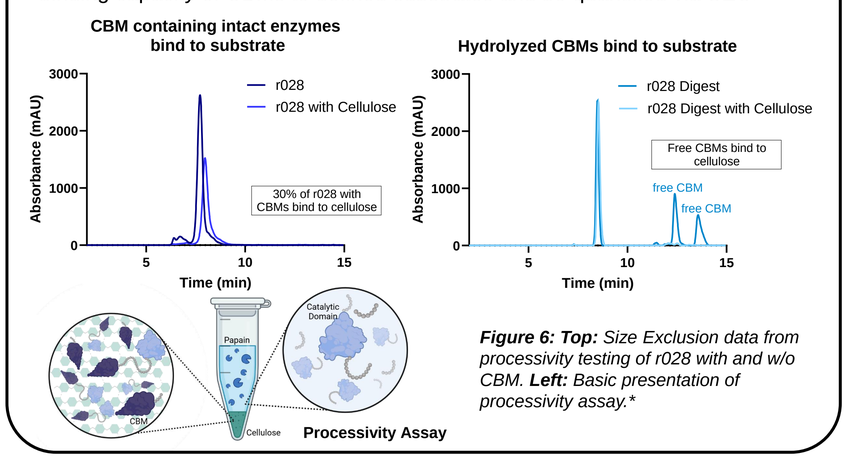


Figure 6: Top: *Size Exclusion data from processivity testing of r028 with and w/o CBM.* ***Left:*** *Basic presentation of processivity assay.**

Thermal Stability & Molecular Radius

Nano Differential Scanning Fluorimetry (**nanoDSF**) is a label-free, high-precision method for analyzing protein stability, folding, and unfolding by monitoring changes in intrinsic tryptophan and tyrosine fluorescence as the protein is heated.

The influence of CBMs on enzyme stability was determined by measuring the intrinsic shift in tryptophan fluorescence during **thermal unfolding**. Inflection points revealed that hydrolyzed enzyme domains are more stable towards thermal unfolding.

Proteolysis with papain of cellulases correlate with reduction in **hydro-dynamic radius**, determined with nanoDSF.

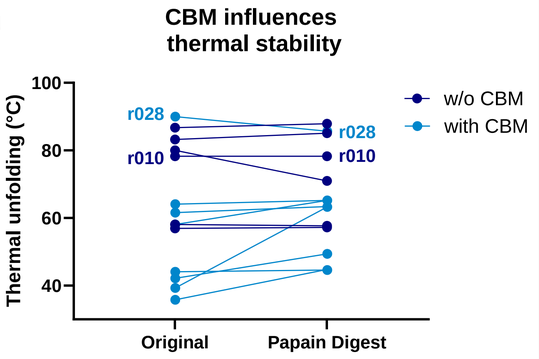


Figure 7: *Inflection point of thermal unfolding of cellulases determined before and after hydrolysis of CBM via papain digestion.*

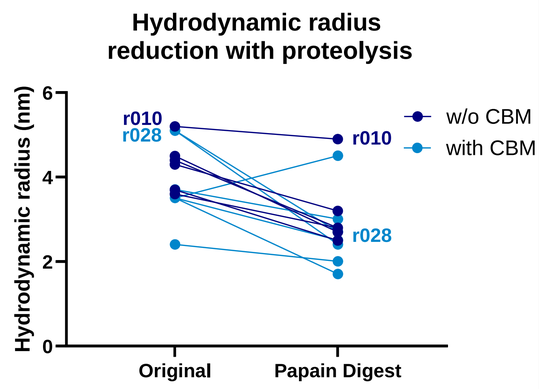


Figure 8: *Hydrodynamic radius of cellulases determined before and after hydrolysis of CBM via papain digestion.*

Dataset

The data set is based on 14 cellulases derived from different organisms. Of these, 13 enzymes were produced recombinantly (eleven from *E. coli*; two from corn). The exception is *T. reesei*, which was used as a cellulase mix. Seven cellulases have one CBM, one cellulase has two CBM, six cellulases have no CBM. The data set contains eight endoglucanases, three exoglucanases and three multivariants. The molecular weight of the enzymes are between 39 and 102 kDa.

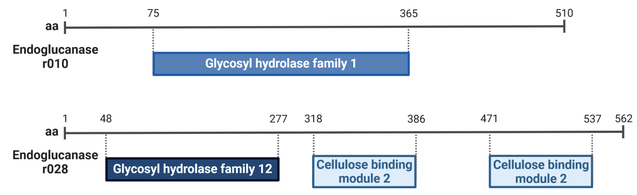


Figure 9: *Domain structure of highlighted cellulases r010 & r028.**

ML-training data

We can provide a broad spectrum of analytical assays with the ability for high-throughput as a basis for solid datasets for ML-based training data.
E.g., Random Forest would be well-suited for this task due to its ability to handle small, heterogeneous datasets, model complex nonlinear interactions, and provide feature importance metrics for biological interpretation.
As we continue to expand our dataset, we aim to establish a comprehensive resource for the scientific community to develop ML models for cellulase activity prediction and enzyme engineering.
Our work supports data-driven advancements in cellulase research and industrial biocatalyst development.

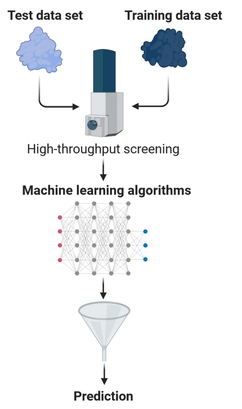


Figure 10: *Workflow for dataset generation with high-throughput assays for ML-based prediction.**

Summary & Outlook

By integrating structural and functional datasets shown here, we provide a robust experimental framework that facilitates the correlation of biophysical properties with enzymatic performance. As we continue to expand our dataset, we aim to establish a comprehensive resource for the scientific community to develop ML models for cellulase activity prediction and enzyme engineering. Our work supports data-driven advancements in cellulase research and industrial biocatalyst development. This dataset can be extended by integrating data gathered from HDX-MS[3] or Biolayer Interferometry to study CBM/catalytic domain binding to substrate in detail.
The analytical methods presented can be used to characterize enzyme preparations and/or novel enzyme candidates. They are useful to identify structural differences and to screen enzyme candidates for stability, activity and special properties. Furthermore, they can assist enzyme engineering to optimize the enzyme structure according to the desired properties using low sample amount and within a reasonable time frame.

References

1 Strobel KL, Pfeiffer KA, Blanch HW, Clark DS (2015) Structural Insights into the Affinity of Cel7A Carbohydrate-binding Module for Lignin. *The Journal of Biological Chemistry*, Vol. 290, No. 37, pp. 22818 –22826. doi 10.1074/jbc.M115.673467.

2 Wilson DB, Kostylev M (2012) Cellulase Processivity. *Methods Mol Biol*, 908: 93–99. doi 10.1007/978-1-61779-956-3_9.

3 Gramlich M, Hays HCW, Crichton S, Kaiser PD *et al*. (2021) HDX-MS for Epitope Characterization of a Therapeutic ANTIBODY Candidate on the Calcium-Binding Protein Annexin-A1. *Antibodies*, *10*, 11. doi 10.3390/antib10010011.

*Figures created with BioRender.

Grant No BW1_5021/02

Investition in Ihre Zukunft.

Baden-Württemberg Ministerium für Wirtschaft, Arbeit und Tourismus

Grant No RegioInn_2449401

The work was performed as part of the project “Encycling”, funded by investBW BW1_5021/02.
The RSLC U3000 HPLC system and the maXis HD UHR-TOF mass spectrometer used for intact mass analysis were funded by the State Ministry of Baden-Wuerttemberg for Economic Affairs, Labor and Tourism (#7-4332.62-NMI/55 and WM-4332-3/6)).
The Agilent 1260 HPLC system for SEC and the Prometheus Panta nanoDSF system were partially funded by RegioInn_2449401 EFRE project.