Project Title: Characterizing Key Factors That Influence Recalcitrance of Lignocellulosic Biomass Deconstruction via *in-situ* liquid IR SNOM Approach

Call Topic: This proposal is in response to Functional and Systems Biology on EMSL's suite of surface science capabilities are ideal for chemical imaging organic matter mineral interactions. Capabilities include *in-situ* liquid IR SNOM.

Specific Aims The most expensive operations in biological processing of cellulosic biomass to fuels and chemicals are for releasing sugars from this naturally recalcitrant material. Unfortunately, such bioprocessing technologies have not yet been commercialized, at least partly because high enzyme doses are currently required to hydrolyze cellulose to glucose at the high yields vital to economic success and to compensate for the rapid fall-off in hydrolysis rate as conversion progresses [1-6].

Infrared scattering-scanning near-field optical microscopy (IR *s*-SNOM) can provide chemical information with nanoscale spatial resolution and high sensitivity by combining infrared spectroscopy with atomic force microscopy (AFM). Briefly, the apex of an AFM tip is illuminated with infrared radiation and probes the infrared absorption of a sample with nanoscale spatial resolution. This allows infrared spectroscopy with far higher spatial resolution than conventional techniques and can be used to create chemical images with ~10 nm spatial resolution.[7] This technology will be applied to reacted cellulose to uncover spatial variations of the chemical changes during enzymatic hydrolysis of cellulose correlated with surface morphology. The new capability extends IR *s*-SNOM to liquid environments, which will allow *in-situ* measurements of cellulose hydrolysis. Using a novel, bottom illumination geometry, in-liquid IR *s*-SNOM will probe chemical changes throughout the cellulose hydrolysis process *in-situ*.[8] Thus, this project seeks to quantitatively study enzymatic deconstruction of cellulosic biomass across multiple spatial and temporal scales by imaging the interaction of proteins with the cellulose surface in order to develop groundbreaking new mechanisms for enzymatic biofuel production.

Mission Relevance: The mechanisms of hydrolysis reactions and factors that limit hydrolysis effectiveness remain unclear, and consequently it limits many promising applications in the real world. In particular, the mechanism of rapid decline of cellulose hydrolysis rate during enzymatic hydrolysis, which contributes to high enzyme demands and high cost of biomass bioconversion, is not understood despite of decades of research efforts. This project strongly supports the aims of the Functional and Systems Biology Area by exploring the interaction of enzymes with biomass to "leading to improved strategies for designing plants, fungi, and microbes for biofuels and bio-based products, as well as unraveling the complexities of carbon, nutrient, and elemental cycles within cells and their immediate environment".

Work Plan: Overall, this research is to apply innovative techniques to develop a fundamental understanding of the relationship between dynamic change of substrate structure and the functionality of various cellulases components during enzymatic hydrolysis of cellulose. Effects of enzyme-substrate interactions on reaction rates will be studied using key cellulase components from wild type *Trichoderma reesei* and its variants with different levels of *N*-linked glycosylation (heterologously expressed in *Aspergillus niger*) with provided cellulose substrates.

SPECIFIC OBJECTIVES:

Objective 1: Image the chemical maps of cellulose surfaces with nanoscale spatial resolution with IR s-SNOM; **Objective 2:** Validate *in situ* hydrolysis processes via IR s-SNOM as cellulose degrades/transforms/reacts; **Objective 3:** Measure in-situ hydrolysis processes of cellulose degradation/transformation/reaction via a combination of solid-state NMR and SFG-VS imaging. **Objective 4:** Develop a hybrid automation model based on quantum mechanics/molecular mechanics (QM/MM) dynamics simulation models to assist the interpretation the IR s-SNOM signals in experiments.

Tasks 1. Image the chemical maps of cellulose surfaces with nanoscale spatial resolution with IR s-SNOM.IR *s*-SNOM: By combining the high spatial resolution of atomic force microscopy with the chemical specificity of infrared spectroscopy, IR *s*-SNOM is a versatile tool for chemical nano-imaging. IR *s*-SNOM can be used to perform vibrational spectroscopy at the nanoscale to probe protein location and secondary structure with few protein sensitivity as well as to track chemical changes and heterogeneity therein correlated with sample nanoscale morphology. IR absorption images will locate protein by its amide I band (~1650 cm^{-1}). The cellulose surface will be chemically mapped before and after reaction to assess chemical and crystallinity changes through hydrolysis. Specifically, the lateral order index (LOI, 1427 cm^{-1}/895 cm^{-1}), the total crystallinity index (TCI, 1378 cm^{-1}/2900 cm^{-1}), and the hydrogen-bond intensity (HBI, 3350 cm^{-1}/1337 cm^{-1}) have been used to assess crystallinity during enzymatic hydrolysis of cellulose.[9] These values will be mapped by acquiring images at these respective wavelengths, to check for the hypothesized preferential dissolution of amorphous regions versus the crystalline regions.[10] Additionally, the chemical signatures of lignin and hemicellulose will also be mapped to understand their role in hydrolysis.

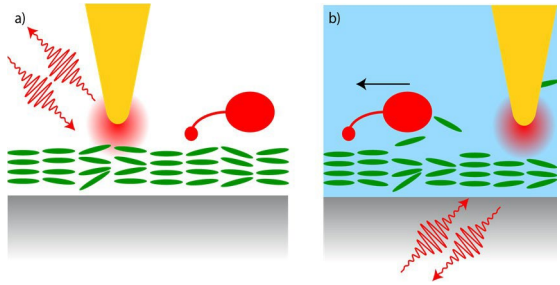

Figure 1. a) *s*-SNOM of cellulose surface before and after reaction will be used in task 1 to characterize changes in crystallinity and chemical identity due to enzymatic hydrolysis. b) In task 2, cellulose surface will be imaged in-liquid to obtain real-time chemical information of the hydrolysis process.

The new capability extends IR *s*-SNOM to liquid environments, which will allow *in-situ* measurements of cellulose hydrolysis. Using a novel, bottom illumination geometry, in-liquid IR *s*-SNOM will probe chemical changes throughout the cellulose hydrolysis process *in-situ*.[8] Thus, this project seeks to use IR s-SNOM imaging to quantitatively study enzymatic deconstruction of cellulosic biomass across multiple spatial and temporal scales in order to discover groundbreaking new mechanisms of the enzymatic hydrolysis.

Task 2: Validate dynamic characterization in situ hydrolysis as cellulose degrades/transforms/reacts via in-liquid IR *s*-SNOM. A newly developed capability enables IR *s*-SNOM measurement of samples in liquid environments. In-liquid IR *s*-SNOM uses a novel total internal reflection illumination geometry which minimizes IR absorption while maintaining high spatial resolution and specificity. This unique capability is now a versatile platform to investigate chemical and biological processes at the liquid-solid interface. This will allow real-time tracking of enzymatic hydrolysis of cellulose on image acquisition times on the few minute timescale. The amide IR signature of enzyme proteins can be used to track their location and conformation to uncover the role and function of cellulase during the enzymatic hydrolysis process of cellulose. The objective of this research is to apply innovative techniques to develop a fundamental understanding of the relationship between dynamic change of substrate structure and the functionality of

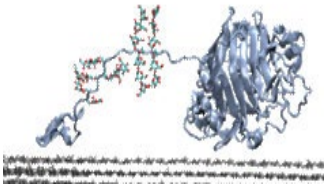
Figure 2. Numerical model of *T. reesei* Cel7A acting on a cellulose surface.

various cellulases components during enzymatic hydrolysis of cellulose (4, 5). Effects of enzyme-substrate interactions on reaction rates will be studied using key cellulase components from wild type *Trichoderma reesei* and its variants with different levels of *N*-linked glycosylation (heterologously expressed in *Aspergillus niger*) [10]. Using in-liquid IR s-SNOM, the dynamic hydrolysis behaviors of individual key cellulase components such as adsorption and desorption will be monitored. As described in task 1, heterogeneity in the chemical composition and crystallinity cellulose surface can be correlated with reactivity and protein location during the hydrolysis deconstruction, however, in real time with in-liquid *s*-SNOM. Cellulose reactivity with enzymes, monomeric/oligomeric sugars, and reducing ends groups will be characterized over the time course of cellulose hydrolysis. Effects of cellulose reactivity, accessibility to cellulases, and cellulase processivity on enzymatic hydrolysis of cellulose will be addressed in the study. The goal is to provide new data, models, insights,

and mechanisms that will facilitate the development and commercialization of low cost enzymatic hydrolysis technologies. Improving our knowledge of enzymatic hydrolysis of cellulosic biomass will also speed applications for both new and existing technologies by giving investors more confidence in their commercial use**.** To our knowledge, this is the first direct kinetic evidence that has used IR s-SNOM to monitor the enzymatic hydrolysis of cellulose by cellulases.

In situ Characterization of the cellulase structure and conformation changes: Because the cellulase bonded to the solid cellulose substrate is usually separated from the aqueous phase, monitoring the conformation changes of cellulases during the conversion reaction requires surface specific techniques. We propose to use in-liquid IR s-SNOM to study and monitor the structural and conformation changes of cellulose during hydrolysis deconstruction. With single wavelength images across a broad spectral range, we will obtain maps of the LOI, TCI, and HBI, as well as the protein amide I signatures at regular time intervals in the deconstruction process. This, correlated with the nanoscale sample morphology, will give valuable information on the rate and key parameters properties, which drive or inhibit cellulose deconstruction. The conformation with the binding and catalytic reaction that matches the level of mechanical behavior determined from IR s-SNOM experiments will be the most likely conformations of the cellulase catalytic domain and binding modules. To validate the changed conformation, we will conduct mutagenesis studies. For example, the PI proposes to use enzymes such as cellobiohydrolase I (CBH I) from a catalytically-inactive mutant E212Q, a *T. reesei* exoglucanase that has an inactive catalytic domain but other properties similar to those of active cellulase [11], [12] and evaluate the impact of altering amino acid residue(s) at the interface between cellulase and cellulose on the mechanical energy.

Task 3: Measure in-situ hydrolysis processes of cellulose degradation/transformation/reaction via combining Solid-state NMR and SFG-VS imaging.

Utilizing the sealed rotor system developed at EMSL, we can monitor the CP/MAS spectrum of the solid cellulose while simultaneously observing the high resolution solution-state NMR of solvated intermediates and products. In this way, we can measure the kinetics of the hydrolysis reaction as well as observe the chemical processes as they occur. Previous results found kinetic evidence of using broadband stimulated Raman spectroscopy (BBSRS) to monitor the enzymatic hydrolysis of cellulose, and confirmed that surface sum-frequency generation vibrational spectroscopy (SFG-VS) is capable of detecting the structure of cellulose from various sources for the first time [13](Figure 3). In addition, by using novel nonlinear vibrational spectroscopic techniques such as the Total Internal Reflection Sum Frequency Generation Vibrational Spectroscopy (TIR-SFG-VS) combined with the conventional SFG-VS (non-TIR) to selectively characterize the molecular structures of surface layers and the crystalline core of cellulose. From the SFG spectra in the C-H and O-H regions, it was found for the first time that the surface layers of Avicel are essentially amorphous while the surface layers of Iβ cellulose are crystalline but with different structural and spectroscopic signatures than that of its crystalline core[14] (Figure 4). As the second order nonlinear spectroscopy, SFG-VS is surface/interface selective and have submonolayer sensitivity. In addition, recent SFG-VS studies have shown that the chiral SFG-VS is uniquely sensitive to protein secondary structure at interfaces. SFG-VS measures the vibrational spectra of interfacial molecular groups and the polarization dependent SFG-VS data can determine the orientation and conformation information

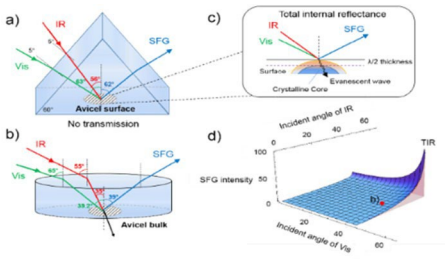

Figure 3. Experimental HR-BB-SFG-VS spectra of Avicel, cellulose Iβ from *Halocynthiaroretzi* tunicate, cellulose Iβ from red tunicate, and cellulose Iα from alga *Valonia ventricosa*

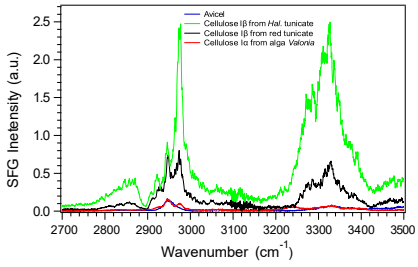

Figure 4. (Non) TIR SFG-VS mechanisms and applications in cellulose characterization.

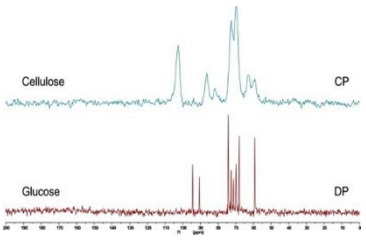

Figure 5. Demonstration of mixed-phase NMR with cellulose/glucose (50/50) in water.

of these molecular groups, thus providing *in situ* information on the cellulase and cellulose surfaces. In doing such SFG-VS measurements, the laser can be directed to the surface of an optical window in contact with the cellulose substrate bounded with cellulase. The conformation with the binding and catalytic reaction that matches the level of mechanical behavior determined from SFG-VS experiments will be the most likely conformations of the cellulase catalytic domain and binding modules on cellulose surface. In addition, CP/MAS NMR can characterize the solid cellulose while simultaneously observing the high resolution solution-state NMR of solvated intermediates and products utilizing the CLASSIC NMR approach strategy.[23] This experiment observes both phases nearly simultaneously *in situ*, thus we can measure the kinetics of the hydrolysis reaction as well as observe the chemical processes as they occur. As a demonstration, a sample prepared with a mixture of cellulose and glucose in H_2O and optimized NMR conditions such that CP only shows the insoluble cellulose while the DP shows only the solution phase (Figure 5).

Task 4. Calculating IR spectra based on QM/MM simulation of *Trichoderma reesei* (and its variants) on the hydrolysis of cellulose with or without a solid cellulose surface. The calculated IR spectra will assist us on interpreting the experimental signals from IR s-SNOM. Hence, we will properly analyz the enzymatic catalysis by combing the experimental and simulation results. Similar to the literature [15], we will apply molecular mechanics force fields to simulate the entire enzymatic reaction and apply the mixed QM methods such as density functional theory (DFT) and semi-empirical methods (e.g., AM1) to treat the catalytic sites and the reacting ligand moiety. The molecular dynamics trajectory will be used to calculate the IR spectra (i.e., IR intensity I) using the autocorrelation of molecular dipole derivatives $\mu\cdot$ in terms of time t, according to the formula $I(\omega) \propto \mathrm{f}\langle\mu\cdot(\tau)\mu\cdot(t+\tau)\rangle_{\mathrm{r}}\, e^{-\mathrm{i}\omega t}dt$, where ω is the vibrational frequency[16]. We will compare this method to the static IR calculation based on the formula of $I(\omega) \propto \left(\frac{\mathrm{a}\mu}{\mathrm{aq(w)}}\right)^2$. Here, we will compute molecular dipole μ or molecular dipole derivatives $\frac{\mathrm{a}\mu}{\mathrm{aq(w)}}$ in terms of normal mode q using a 'divide-and-conquer' *ab intio* quantum chemistry method (see Figure 6), which will give us accurate results of these values, leading to highly accurate IR spectra prediction. Such a divide-and-conquer method has been used for computing sum frequency generation (SFG) spectra for beta-sheet proteins. The dynamic IR spectra calculation can provide accurate spectra by capturing the anharmonic effect of molecular vibrations, while the cost of calculation can be higher due to the integration. For the static IR sepectra calculation, the results can be less accurate due to the harmonic approximation, while the computational cost can be lower without an integration over time. Here, the QM/MM simulation at the ONIOM(B3PW91/6-31+G(d,p):AM1) level using the Gaussian1 16 program, with the Amber force-fields for the MD region. After that, the Gaussian 16 Package Program will be used to calculate μ or $\frac{\mathrm{a}\mu}{\mathrm{aq(w)}}$ for each fragment of the protein or large molecules at the level of B3LYP/6-31G*, and in-house Python scripts will be used to compute μ or $\frac{\mathrm{a}\mu}{\mathrm{aq(w)}}$ for a whole protein or complex structure. The μ and $\frac{\mathrm{a}\mu}{\mathrm{aq(w)}}$ data for the complex structures will be used to compute IR spectra using TRAVIS (http://www.travis-analyzer.de/) or in-house Python codes. In summary, we will provide a highly accurate *ab initio* quantum chemistry methods to simulate the IR spectra obtained from IR s-SNOM, enabling us to reveal the catalytic process by *Trichoderma reesei* (or it variants) at the molecular level by combined experimental and simulation evidences.

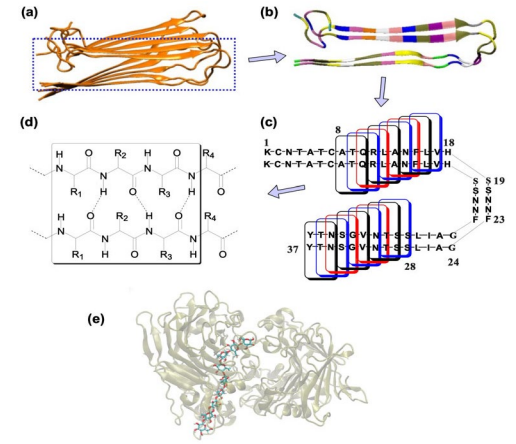

Figure 6. (a)-(d) A divide-and-conquer *ab initio* quantum chemistry method to compute Sum Frequency Generation (SFG) spectrum for beta-sheet proteins, see *[16]*. Such a method will be used to compute the IR spectra of Trichoderma reesei (e) and its variants.

Appendix 1_ List of References:

[1] T. Eriksson, J. Karlsson, F. Tjerneld, A model explaining declining rate in hydrolysis of lignocellulose substrates with cellobiohydrolase I (Cel7A) and endoglucanase I (Cel7B) of Trichoderma reesei, Applied Biochemistry Biotechnology 101 (2002) 41-60.
[2] S.G. Desai, A.O. Converse, Substrate reactivity as a function of the extent of reaction in the enzymatic hydrolysis of lignocellulose, Biotechnology Bioengineering 56(6) (1997) 650-655.
[3] A.J. Ragauskas, C.K. Williams, B.H. Davison, G. Britovsek, J. Cairney, C.A. Eckert, W.J. Frederick, J.P. Hallett, D.J. Leak, C.L. Liotta, J.R. Mielenz, R. Murphy, R. Templer, T. Tschaplinski, The path forward for biofuels and biomaterials, Science 311(5760) (2006) 484-489.
[4] B. Yang, D.M. Willies, C.E. Wyman, Changes in the enzymatic hydrolysis rate of Avicel cellulose with conversion, Biotechnology and Bioengineering 94(6) (2006) 1122-1128.
[5] B. Yang, C.E. Wyman, Effect of xylan and lignin removal by batch and flowthrough pretreatment on the enzymatic digestibility of corn stover cellulose, Biotechnol Bioeng 86 (2004).
[6] B. Yang, C.E. Wyman, BSA treatment to enhance enzymatic hydrolysis of cellulose in lignin containing substrates, Biotechnology and Bioengineering 94(4) (2006) 611-617.
[7] H.A. Bechtel, S.C. Johnson, O. Khatib, E.A. Muller, M.B. Raschke, Synchrotron infrared nano-spectroscopy and -imaging, Surface Science Reports 75(3) (2020) 100493.
[8] B.T. O'Callahan, K.-D. Park, I.V. Novikova, T. Jian, C.-L. Chen, E.A. Muller, P.Z. El-Khoury, M.B. Raschke, A.S. Lea, In Liquid Infrared Scattering Scanning Near-Field Optical Microscopy for Chemical and Biological Nanoimaging, Nano Letters 20(6) (2020) 4497-4504.
[9] N. Kruer-Zerhusen, B. Cantero-Tubilla, D.B. Wilson, Characterization of cellulose crystallinity after enzymatic treatment using Fourier transform infrared spectroscopy (FTIR), Cellulose 25(1) (2018) 37-48.
[10] B. Yang, Z. Dai, S.-Y. Ding, C.E. Wyman, Enzymatic hydrolysis of cellulosic biomass, Biofuels 2(4) (2011) 421-450.
[11] G. Beldman, A.G.J. Voragen, F.M. Rombouts, M.F. Searlevanleeuwen, W. Pilnik, Adsorption and kinetic behavior of purified endoglucanases and exoglucanases from Trichoderma viride, Biotechnology Bioengineering 30(2) (1987) 251-257.
[12] J. Stahlberg, C. Divne, A. Koivula, K. Piens, M. Claeyssens, T.T. Teeri, T.A. Jones, Activity studies and crystal structures of catalytically deficient mutants of cellobiohydrolase I from Trichoderma reesei, JOURNAL OF MOLECULAR BIOLOGY FIELD Publication Date:1996 Nov 29 264(2) 337-49. FIELD Reference Number: FIELD Journal Code:2985088R FIELD Call Number:.
[13] L. Zhang, Z. Lu, L. Velarde, L. Fu, Y. Pu, S.-Y. Ding, A.J. Ragauskas, H.-F. Wang, B. Yang, Vibrational spectral signatures of crystalline cellulose using high resolution broadband sum frequency generation vibrational spectroscopy (HR-BB-SFG-VS), Cellulose (Dordrecht, Neth.) 22(3) (2015) 1469-1484.
[14] L. Zhang, L. Fu, H.-f. Wang, B. Yang, Discovery of Cellulose Surface Layer Conformation by Nonlinear Vibrational Spectroscopy, Sci. Rep. 7 (2017) 44319.
[15] C.S. Pereira, R.L. Silveira, M.S. Skaf, QM/MM Simulations of Enzymatic Hydrolysis of Cellulose: Probing the Viability of an Endocyclic Mechanism for an Inverting Cellulase, Journal of Chemical Information and Modeling 61(4) (2021) 1902-1912.
[16] D. Xiao, L. Fu, J. Liu, V.S. Batista, E.C.Y. Yan, Amphiphilic Adsorption of Human Islet Amyloid Polypeptide Aggregates to Lipid/Aqueous Interfaces, Journal of Molecular Biology 421(4) (2012) 537-547.

Appendix 3: Active Collaborators List

CONFLICT OF INTEREST LIST				
Name	**Key Co-Author**	**Collaborator**	**Advisee / Advisor** *(specify)*	**Other** *(specify nature)*

Computing Resources

IMPORTANT: The EMSL computing systems available to users are not approved for use with sensitive data. The processing, storage, or transmittal of sensitive data (e.g. Personally Identifiable Information, Official Use Only, etc.) is thus prohibited on Tahoma, Cascade and Aurora. Due diligence must be used to prevent inadvertent disclosure of invention, patent, or other sensitive information. It is your responsibility to protect access to the information.

☒ **By checking this box, I am confirming that participants on this proposal will NOT process, store, or transmit sensitive data (e.g. Personally Identifiable Information, Official Use Only, etc.) on Tahoma, Cascade or Aurora.**

Total CPU Hours Request for first year of proposal: 100,000					
Total GPGPU Hours Request for first year of proposal:					
Total Data Archive Request for first year of proposal:					
Software Details	**Node Request (CPUs or GPGPUs)**	**Estimated # of jobs**	**Estimated Node Hours**	**Expertise of your investigators for these requests**	**EMSL Support Requested** *Specific Needs (e.g., compiling code, libraries needed, help running jobs, etc.)*
Gaussian 16	CPU	50	100,000	Expert User	Installing Gaussian 16 in Tahoma

Notes:
Tahoma allocations are awarded in units of wall-clock time expressed in node-hours. Tahoma's 160 CPU nodes each have 36 (3.1 GHz) Intel Xeon processor cores with 384 GB of memory and 2 TB of flash storage. Consequently, 10,000 Tahoma CPU node-hours are equal to 360,000 processor core-hours. Tahoma's 24 GPGPU nodes each have 36 processor cores and 2 Nvidia v100 GPGPUs, 1536 GB of memory, and 7 TB of flash storage. Tahoma's 10 PB global file system is capable of 100 Gigabyte/sec bandwidth. Tahoma can deliver a total of 1,500,000 node-hours per year.

Upon successful review and approval of a proposal, computing resources will be allocated for analysis and archiving of experimental data generated at EMSL.