Chemical Engineering Science 325 (2026) 123422

Contents lists available at ScienceDirect

Chemical Engineering Science

journal homepage: www.elsevier.com/locate/ces

Multiscale modeling of area-selective atomic layer deposition of SiO_2 on Al_2O_3 and SiO_2 surfaces with intermediate etching steps

Feiyang Ou [a], Abdulrahman Alghamdi [a], Chun-Pei Lin [a], Gerassimos Orkoulas [b], Panagiotis D. Christofides [a,c,*]

[a] Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, CA, 90095-1592, USA
[b] Department of Chemical Engineering, Widener University, Chester, PA, 19013, USA
[c] Department of Electrical and Computer Engineering, University of California, Los Angeles, CA, 90095-1592, USA

ARTICLE INFO

Keywords:
Semiconductor manufacturing
Microscopic simulation
CFD simulation
Multiscale simulation
Area-selective atomic layer deposition
Atomic layer etching

ABSTRACT

Area-Selective Atomic Layer Deposition (ASALD) has attracted increasing attention as a bottom-up patterning technique capable of achieving self-aligned fabrication and circumventing the alignment challenges inherent in traditional top-down lithography. In this work, a microscopic Monte Carlo-based collision model was developed to describe SiO_2 ASALD on both SiO_2 (growth area (GA)) and Al_2O_3 (non-growth area (NGA)) surfaces. The model incorporates small-molecule inhibitors (SMIs) with precursor and co-reactant reactions of bis(diethylamino)silane (BDEAS) and O_3, with validation against reported experimental data, accurately reproducing 22% monodentate inhibitor coverage and 6.7% unwanted BDEAS nucleation on the NGA. In addition, an atomic layer etching (ALE) process employing trimethylaluminum (TMA) and HF was integrated as an additional, intermediate step to remove undesired nucleation on NGA and enhance overall selectivity, with simulations demonstrating that a 1.0 s etch step maintains selectivity over 40 batches under ideal microscopic conditions. The effects of etching time and batch number on film selectivity were systematically analyzed. To bridge microscopic surface kinetics with reactor-scale dynamics, a macroscopic computational fluid dynamics (CFD) model of the ALD reactor was developed, and was experimentally verified in collaboration with NIST. This CFD model was coupled with the microscopic model through PyFluent to form a multiscale CFD model for the entire process. This integrated model captures wafer-level pressure dynamics and precursor delivery delays to the surface, identifies the need for an extended etch time (1.6 to 1.8 s) to sustain selectivity under realistic reactor conditions and provides an effective digital twin for ASALD process optimization.

1. Introduction

For decades, nanoscale semiconductor manufacturing has relied predominantly on top-down fabrication processes comprising multiple lithography, etching, and deposition steps. Traditional top-down approaches employing Atomic Layer Deposition (ALD) and Atomic Layer Etching (ALE) enable atomic-level precision in film thickness, allowing the formation of highly conformal and ultrathin layers essential for complex three-dimensional architectures such as Fin Field-Effect Transistors (FinFETs) (Leskelä and Ritala, 2002; George, 2010) and Gate-All-Around (GAA) (Loubet et al., 2017) structures. These designs substantially increase transistor density, thereby enhancing computational performance while reducing power consumption. Consequently, significant research effort has focused on improving and optimizing atomic layer-based operations to achieve higher yields and tighter quality control. Both ALD and ALE are characterized by sequential, self-limiting surface reactions, typically involving two half-cycles separated by inert gas purging, that afford sub-nanometer control of material growth or removal. This cyclic mechanism has been applied to a wide range of materials, including metals and metal oxides commonly used as gate dielectrics, such as SiO_2, Al_2O_3, HfO_2, TiO_2, and Ta_2O_5, enabling the precise and uniform deposition of functional thin films critical to advanced semiconductor device fabrication.

With shrinking device dimensions, the semiconductor industry faces mounting challenges in achieving reliable pattern fidelity at the nanoscale. One of the most critical issues is edge placement error (EPE) (Mameli et al., 2017; Mackus et al., 2019), which is the cumulative misalignment that arises during the sequential deposition, lithography, and etching steps used in conventional top-down fabrication. These small positional deviations between the intended and actual

* Corresponding author.
E-mail addresses: feiyangou@g.ucla.edu (F. Ou), abdulghamdi1435@gmail.com (A. Alghamdi), andrealin0514@g.ucla.edu (C.-P. Lin), gorkoulas@widener.edu (G. Orkoulas), pdc@seas.ucla.edu (P.D. Christofides).

https://doi.org/10.1016/j.ces.2026.123422
Received 21 November 2025; Received in revised form 12 January 2026; Accepted 21 January 2026
Available online 27 January 2026

feature locations can lead to unwanted material growth or etching at undesired areas (Merkx et al., 2020b). Although such errors were tolerable in earlier technology nodes with larger gate widths, they become increasingly detrimental in sub-5 nm manufacturing, where even nanoscale misalignments can severely degrade device performance. EPE ultimately limits the achievable stacking complexity of three-dimensional architectures, such as multilayer GAA transistors.

To address these limitations, Area-Selective Atomic Layer Deposition (ASALD) has emerged as a promising bottom-up self-alignment technique for high-volume manufacturing (Fang and Ho, 2015; Parsons and Clark, 2020). In ASALD, the substrate is chemically divided into a growth area (GA) where deposition is desired, and a non-growth area (NGA) which is passivated to inhibit nucleation. This selectivity is achieved through surface chemical modification and the careful choice of small-molecule inhibitors (SMIs) that preferentially adsorb and deactivate the NGA without interfering with reactions on the GA. When the inhibitor effectively blocks precursor adsorption on the NGA, ASALD enables self-aligned, layer-by-layer film growth that reduces the need for conventional lithography and etching steps. Consequently, this approach not only improves process yield and lowers fabrication cost but it also facilitates the realization of more complex three-dimensional device architectures with higher stacking layers (Mackus et al., 2019).

Several strategies have been proposed to achieve selective inhibition on the NGA, among which the use of self-assembled monolayers (SAMs) has become particularly prevalent in academic studies. SAMs consist of long aliphatic tail chains that self-organize through van der Waals interactions, forming a densely packed organic layer in which the molecular heads are chemically anchored to the substrate surface. This configuration provides an effective diffusion barrier that suppresses precursor nucleation on the NGAs, allowing deposition to proceed primarily on the desired GA. The successful application of SAM-based inhibitors has been reported in multiple studies (Chen et al., 2005; Sinha et al., 2006; Huang et al., 2014; Minaye Hashemi et al., 2015; Haider et al., 2016). However, SAMs exhibit several intrinsic limitations. First, their coverage is rarely defect-free, restricting reliable inhibition to only a few nanometers of film thickness on GA. More critically, SAM formation typically relies on wet-chemical processing and requires extended reaction times on the order of tens of minutes to achieve sufficient surface passivation. These constraints hinder their integration with conventional vapor-phase ALD and ALE processes, where rapid cycling and compatibility with high aspect ratio features are essential. As a result, SAM-based inhibition is unsuitable for high-volume manufacturing of dielectric films in advanced gate structures that require both high selectivity and conformal coverage.

To overcome the limitations associated with SAM-based inhibition, small-molecule inhibitors (SMIs) have been proposed as a more versatile alternative. Unlike SAMs, which rely on long-chain organic molecules that block precursor adsorption primarily through hydrophobicity and physical steric hindrance, SMIs employ short, volatile molecules delivered in the vapor phase to passivate the NGA via a combination of chemical bonding and steric shielding (Merkx et al., 2022). The key advantage of SMIs lies in their compatibility with existing ALD process conditions: they can be rapidly introduced and reacted in the gas phase, enabling straightforward integration into standard reactor configurations. By incorporating an additional inhibition step, the conventional AB-type ALD cycle can be extended into an ABC-type ASALD cycle (Mameli et al., 2017), where:

- Step A involves selective adsorption of the inhibitor onto the NGA to prevent precursor nucleation.
- Step B introduces the precursor, which reacts exclusively on the unblocked GA.
- Step C supplies the co-reactant to complete the surface reaction on the GA.

This three-step ASALD structure not only ensures selective film growth but it also allows the use of reactive oxidant co-reactants in Step C. Such co-reactants can both finalize the desired deposition on the GA and simultaneously remove residual inhibitors from the NGA, thereby regenerating the surface for the next cycle. This vapor-phase inhibition strategy has been successfully demonstrated in SiO_2 ASALD systems employing SiO_2 (GA) and Al_2O_3 (NGA) substrates, with validation from both experimental studies (Mameli et al., 2017) and density functional theory (DFT) assisted kinetic Monte Carlo (kMC) simulations (Yun et al., 2022a). In these implementations, the inhibitor layer is cyclically removed during the ozone oxidation step (Step C) and re-adsorbed in the subsequent inhibition step (Step A) to maintain selectivity over multiple deposition cycles.

However, prior experimental work has shown that the inhibitor is not always perfect. For example, in area-selective ALD of SiO_2 on a SiO_2 growth surface with an Al_2O_3 non-growth surface, the acetylacetone (Hacac) inhibitor does not completely block precursor nucleation on the NGA (Merkx et al., 2020a). In fact, up to about 8% of the BDEAS precursor can still adsorb on the Al_2O_3 NGA despite the inhibitor (Merkx et al., 2020a), indicating that the blocking is imperfect and the selectivity can be easily lost. This unwanted deposition will accumulate with continued cycling, and eventually the growth rate in NGA becomes identical to that in GA after an initial nucleation delay of roughly 10-20 ALD cycles (Merkx et al., 2020b; Vos et al., 2019; Vallat et al., 2017). One possible solution to mitigate this loss of selectivity is to add an etching step, after the deposition steps, forming an ABCD-type ALD sequence with step D serving as a selective etch (Mackus et al., 2019). The 'step D' here denotes an arbatrily comprehensive etching procedure that may actually contain multiple internal steps to complete the etching process. The etching process can selectively remove the unwanted nuclei on the NGA while incurring only a minimal loss of material on the GA. Indeed, such ABCD-type (ALD-etch supercycle) processes have been demonstrated for a variety of material systems, for instance, area-selective deposition of Ruthenium (Vos et al., 2019), Titanium Nitride (Merkx et al., 2020b)and of Ta_2O_5 (Vallat et al., 2017) with an additional thermal ALE etching step whereby periodic removal of nuclei on the NGA maintains high selectivity without significantly compromising the deposition efficiency on GA. In this work, the term "batch" refers to one minimal repeatable process unit, consisting of a complete sequence of elementary steps. Depending on the process configuration, a batch may include different numbers of steps: two steps for conventional ALD, three steps for traditional AS-ALD, and five steps (ABCDE) for the AS-ALD with integrated etching considered here.

Despite the aforementioned merits, there remains a notable gap in both computational and experimental studies analyzing the multi-batch performance of SMI-based ASALD processes-particularly those employing an ABCD-type cycle with integrated etching steps to suppress unwanted nucleation on the NGA. Prior computational work by Yun et al. (2022a), assumes ideal inhibitor performance and neglects precursor nucleation on the NGA, a simplification that contradicts experimental observations indicating incomplete blocking and progressive selectivity loss. To address this, the present study introduces a Monte Carlo-based collision model that captures the stochastic surface interactions of inhibitor adsorption, precursor deposition, co-reactant oxidation, and exposure-limited etching. This microscopic model enables a detailed investigation into how inhibitors adsorb, how nucleation events initiate and accumulate on the NGA under imperfect blocking conditions, and how etching parameters such as reaction time and exposure conditions can be tuned to recover and sustain high selectivity during multi-batch, high-thickness ASALD. Building on this, a computational fluid dynamics (CFD) model of the ALD reactor that is developed in collaboration with NIST and validated by optical $MoCl_5$ flow experiments is coupled with the Monte Carlo simulation via PyFluent to construct a multiscale digital twin of the ASALD process. This integrated framework captures the real-time reactor-scale pressure dynamics and enables cycle-accurate simulation of surface coverage evolution, providing a predictive, application-ready platform for optimizing ASALD systems in realistic fabrication environments.

2. Microscopic Monte-Carlo based collision model

The objective of the ASALD process is to enable selective thin-film metal oxide deposition on the GA that is represented by the $\beta - SiO_2$ (101) surface, while suppressing deposition on the NGA which is typically $\alpha - Al_2O_3$ (0001). The full ASALD cycle in this work consists of five steps. Step A involves the adsorption of the small-molecule inhibitor acetylacetone (Hacac) as it has significant difference in activation energy barriers in GA and NGA (Yun et al., 2022a), which can be selectively chemisorbed onto the NGA surface to passivate it and prevent precursor nucleation. Step B introduces the silicon precursor bis(diethylamino)silane (BDEAS), which reacts with surface hydroxyl groups to form Si-H terminations on the GA, but it also leads to limited and undesired nucleation on the NGA due to imperfect inhibition. In Step C, the co-reactant ozone oxidizes the precursor-modified surface, converting Si-H terminations to Si-OH to complete both ideal and unwanted SiO_2 deposition on both surfaces. Simultaneously, the ozone step also removes other species, including residual inhibitors, from both surfaces, preparing them for the next cycle. To restore and maintain high selectivity over multiple cycles, the process incorporates a fully thermal, HF-based atomic layer etching (ALE) sub-sequence that is compatible with vapor-phase integration and is consistent with established TMA/HF conversion–etch mechanisms for oxide materials (Dumont et al., 2017; Rahman et al., 2018). In this scheme, Step D is a TMA exposure that serves two coupled roles at the microscopic level: (i) it removes any AlF_3-terminated surface segments generated during the HF step of the previous batch by forming volatile Al-containing products, and (ii) it converts exposed SiO_2 segments (including undesired nuclei on the NGA) into an Al-O-Si / Al_2O_3-like modified state via a conversion reaction, without immediate geometric thickness removal. Step E is a thermal HF exposure that fluorinated these Al-containing modified segments to an AlF_3-like state, which is then removed during the subsequent batch's Step D.

To simulate the microscopic surface reactions of the ASALD process, a Monte Carlo collision model was developed by discretizing the surface into a two-dimensional grid of reactive sites. This model allows for stochastic simulation of molecular adsorption, steric interactions, and nucleation events. Each surface representing either SiO_2 as the GA or Al_2O_3 as the NGA is initialized as a regular lattice, where each site may host an adsorbed species depending on the reaction conditions and spatial constraints. Molecules such as inhibitors (e.g., acetylacetone), precursors (e.g., BDEAS), and reaction intermediates are represented with geometric footprints defined by spatial structure and the van der Waals radii of the functional groups of molecules. This geometric representation ensures that new adsorbates cannot overlap with existing ones, thereby introducing steric exclusion effects that influence local surface coverage. The resulting grid-based simulation provides physically meaningful outputs, such as coverage fractions, inhibitor retention, and precursor breakthrough events, which are critical for quantifying selectivity loss and guiding etch process integration. An illustration of the initialized empty grids for both GA and NGA surfaces is shown below in Figs. 1 and 2 to demonstrate the layout before molecular adsorption.

The general Monte Carlo simulation scheme proceeds by repeatedly sampling adsorption and reaction events on the discretized surface. In each trial, a site is randomly selected from the surface grid, and a physisorption time increment is added to the cumulative simulation time. If the chosen site is already occupied, the algorithm immediately moves on to the next trial. If the site is unoccupied, the simulation evaluates whether the target molecule (e.g., inhibitor, precursor) can be adsorbed at the location without overlapping any nearby molecule, determined by checking steric compatibility using van der Waals radii.

To resolve steric interactions between adsorbed species and enforce non-overlapping placement in the surface Monte Carlo simulation, all molecular templates are constructed using a computational

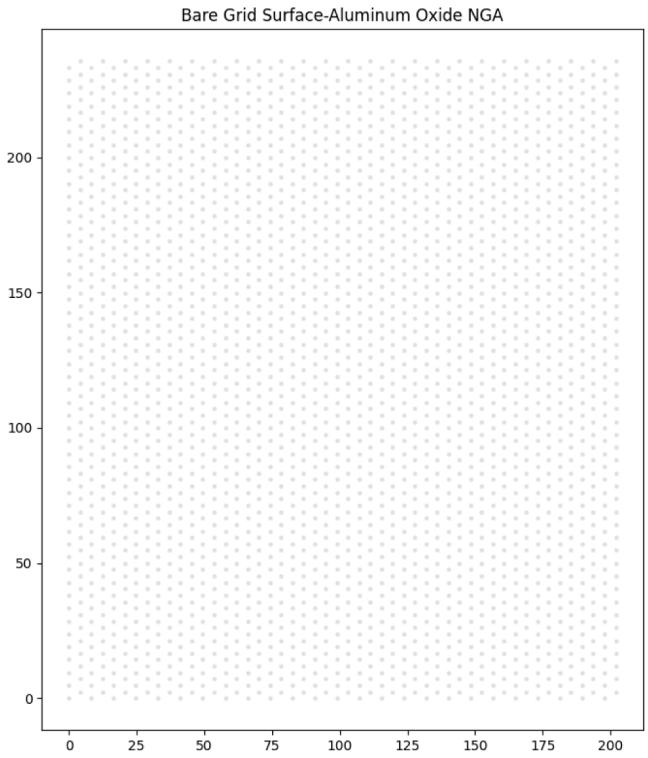


Fig. 1. Bare Al_2O_3 staggered arrangement grid surface with uniform 4.76Å site distance.

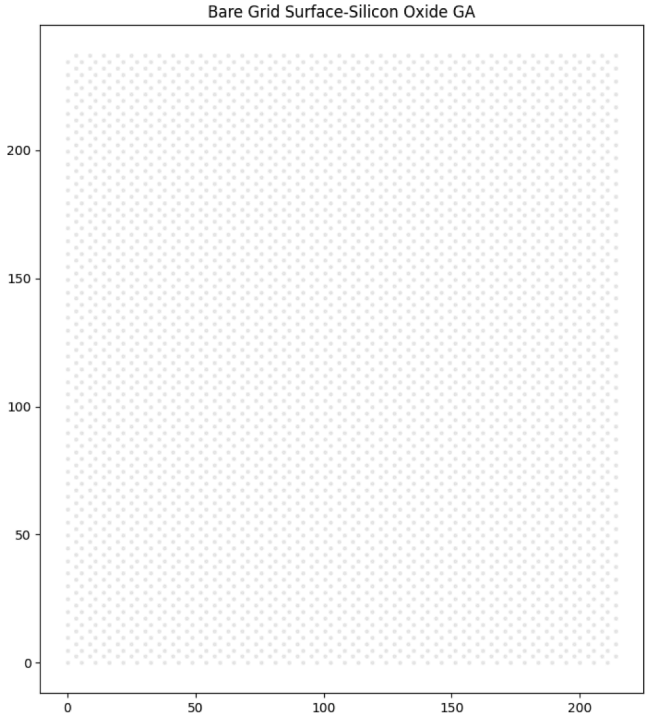


Fig. 2. Bare SiO_2 nonuniform staggered arrangement. Described in Yun et al. (2022a), surface grid separated into 2-column groups, within each group the vertical site distance 4.99Å, horizontal distance 3.13Å, between-group distance 5.40Å.

geometry framework based on convex polygonal approximations. Circular functional groups-such as methyl (CH_3) or silane (SiH_2) units-are represented as regular polygons constructed from uniformly spaced points along the circumference. Specifically, each circle is approxi-

mated by a 32-sided polygon, generated by discretizing the boundary using angular subdivision. These convex polygons are analytically tractable and allow robust geometric operations such as unions, intersections, and transformations. Composite molecular geometries are assembled by uniting multiple convex subcomponents into a single surface footprint.

Explicit electronic polarization of adsorbed molecules is not treated in the present model. Instead, molecular sizes and steric exclusion are represented using effective van der Waals-based geometries that capture the average spatial extent of adsorbed species on the surface. These effective geometries inherently incorporate, in a mean-field sense, the influence of electronic effects such as polarization and charge redistribution on intermolecular spacing. Importantly, the chosen molecular footprints are not arbitrary: they are validated by reproducing experimentally reported inhibitor coverages, precursor nucleation fractions, and growth behavior on both growth and non-growth areas. Within the scope of this work, the dominant factors governing selectivity and film evolution are steric hindrance, site availability, and exposure-limited surface kinetics, rather than subtle polarization-induced deformation of adsorbed molecules. Explicit treatment of polarization would require quantum-mechanical resolution of charge redistribution and geometry relaxation and is therefore incompatible with the lattice-based kinetic Monte Carlo framework and wafer-scale simulation objectives pursued in the present work. Neglecting explicit polarization thus represents a controlled approximation that preserves physical fidelity at the process-relevant scale while enabling tractable multiscale simulation.

Collision detection between any two molecular geometries A and B is based on evaluating the overlap of their occupied spatial regions. A steric conflict is defined mathematically as

$$A \cap B \neq \emptyset, \tag{1}$$

and the corresponding overlap area is computed as

$$A_{\text{overlap}} = \text{Area}(A \cap B). \tag{2}$$

If $A_{\text{overlap}} > 0$, the candidate placement is rejected due to geometric interference. Otherwise, the molecule can be safely placed on the surface. This check is applied between every candidate template and all currently placed geometries to maintain steric exclusion throughout the simulation.

To improve computational efficiency, especially in high-coverage surface regimes, a spatial indexing strategy is employed using a sort-tile-recursive tree (STRtree) structure. The STRtree is a spatial search tree that organizes the currently placed geometries by their bounding boxes, allowing rapid querying of nearby features. When a new molecule is proposed for placement, the tree is queried to retrieve only those existing geometries whose bounding boxes intersect a local search window around the candidate site. Exact polygonal overlap checks are then performed only within this reduced set, substantially reducing the number of pairwise intersection tests per Monte Carlo trial.

The overall surface is managed using a structured grid system initialized by a grid-construction routine. Each site contains spatial coordinates, state variables indicating its current chemical or physical condition, and (if occupied) the geometry associated with the adsorbed molecule. Geometry templates for each molecule type and orientation are precomputed to enable fast trial placement. At initialization, the grid includes empty containers for bidentate, monodentate, and trilate geometries; a list of all lattice site positions; and a uniform state vector tracking adsorption and transformation events. The spatial index tree is rebuilt as needed to reflect the evolving surface configuration. Together, this geometric and grid-based framework enables rigorous steric enforcement and scalable simulation of surface dynamics during the ASALD process.

If spatial placement is possible, the simulation then checks whether the molecule is eligible for further reaction steps. If subsequent surface reactions are defined for the molecule, each reaction is evaluated sequentially for feasibility and executed accordingly. Each successfully executed reaction contributes an additional reaction time increment to the total time counter.

The physisorption and surface reaction time steps are governed by kinetic models. Specifically, the reaction time increment $\frac{\Delta t}{N}$ (N is the number of sites in the simulation grid) for any event is computed as a stochastic sample from the uniform distribution using the general reaction rate k as shown in Eq. (3).

$$\Delta t = -\frac{\ln \gamma}{k} \tag{3}$$

where γ is a uniformly distributed random number in $(0, 1]$, and k is either k_{phys} for physisorption reactions or k_{surf} for surface reactions. The reaction rate k_{phys} for physisorption processes is computed using *collision theory* in Eq. (4).

$$k_{phys} = \frac{P A_{\text{site}} \sigma Z}{\sqrt{2\pi m k_B T}} \tag{4}$$

where P is the gas-phase pressure, A_{site} is the surface area per adsorption site, σ is the sticking coefficient, Z is the coordination number, k_B is the boltzmann constant, m is the molecular mass, and T is the temperature. The sticking coefficient used for Hacac is 1.0×10^{-4} (George, 2010), for BDEAS is 2.0×10^{-5} (Schwille et al., 2017), for O_3 is 4.5×10^{-5} (Lee et al., 2009), and for TMA and HF, because of a lack of experimental data on exact surfaces used in this work, the TMA sticking coefficient is set to be the same as the BDEAS precursor, and the HF applies the O_3 sticking coefficient. For thermally activated surface reactions (e.g., ligand exchange or bridge formation), the reaction rate

$$k_{surf} = \frac{k_B T}{h} \frac{Q^{\ddagger}}{Q} \exp\left(-\frac{E_a}{RT}\right) \tag{5}$$

Here, $Q^{\ddagger}$ and Q are the partition functions for the transition state and reactant, respectively, E_a is the activation energy, and h is Planck's constant. These kinetic formulations ensure that surface reactions are both spatially constrained and temporally resolved with physical accuracy. For simplicity, the $\frac{Q^{\ddagger}}{Q}$ is assumed to be 1 in this work. The activation energies of ASALD reactions calculated by Density Functional Theory (DFT) are demonstrated in Yun et al. (2022a).

In the present microscopic Monte Carlo framework, neither spontaneous thermal desorption nor lateral surface diffusion of chemisorbed species is explicitly modeled. Thermal desorption is neglected based on physical timescale considerations: experimental studies of small-molecule inhibitors and adsorbed precursor fragments in ASALD systems indicate that spontaneous desorption occurs on timescales of hours to days under typical ALD and ALE operating temperatures (Merkx et al., 2020a), whereas adsorption, surface reactions, and etching steps occur on timescales of seconds. As a result, natural desorption does not measurably affect surface coverage evolution within a single ALD or ALE half-cycle and can be safely neglected for the exposure-limited regimes investigated here. Surface diffusion of chemisorbed species is also not included explicitly. In the present model, adsorption events are stochastic in both spatial location and molecular orientation, and repeated adsorption across cycles already produces an effectively randomized surface configuration. From a statistical perspective, explicit surface diffusion would primarily serve to redistribute adsorbates without altering overall coverage or reaction probability, while substantially increasing computational complexity due to repeated geometric relocation and collision checks. Because the focus of this work is on multi-batch selectivity trends and exposure-limited growth and etching behavior rather than equilibrium surface rearrangement, neglecting surface diffusion represents a controlled and reasonable approximation consistent with prior lattice-based and kinetic Monte Carlo ALD modeling approaches.

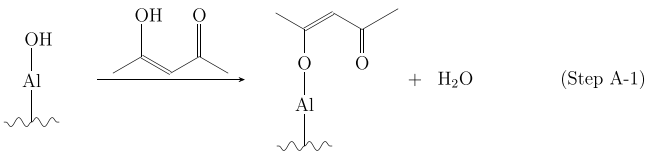
Fig. 3. Orange molecules are monodentates consisting of two circles, blue molecules are chelates consisting of a rectangle and semicircles.

2.1. Step A: inhibitor adsorption

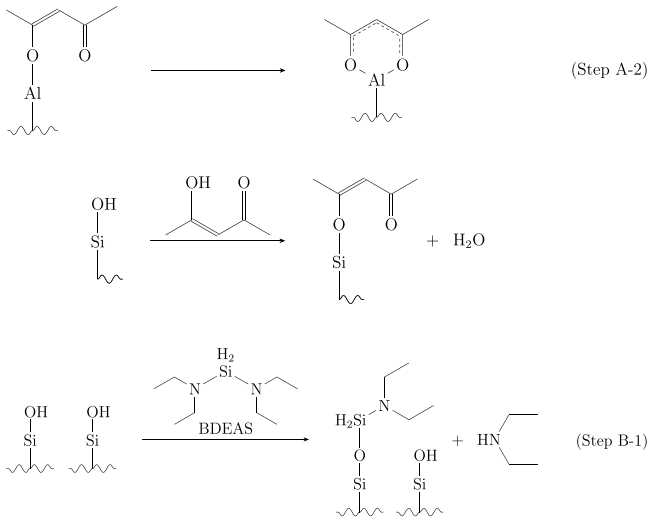
The inhibitor employed in this work is acetylacetone (Hacac), a molecule that exists in equilibrium between keto and enol tautomeric forms. In the gas phase, the enol configuration is known to be more thermodynamically stable than the keto form due to the stabilization provided by intramolecular hydrogen bonding (Folkendt et al., 1985). Upon introduction to the Al_2O_3 non-growth surface, which features surface hydroxyl groups in a vicinal diol arrangement, the acidic enol form of Hacac undergoes a condensation reaction with the basic -OH terminated surface. This results in monodentate adsorption, where the molecule binds via a single Al-O bond and releases H_2O as a byproduct.

Following monodentate adsorption, the Hacac molecule can undergo an energetically favorable transition to form a second Al-O bond, producing a chelate (or bidentate) configuration. This chelation results in a six-membered ring in which the π-electrons are delocalized, creating a stable conjugated system. To capture steric effects in simulation, the chelate geometry is represented as two circles (diameter: 4.0Å, corresponding to the van der Waals radius of CH_3 groups) spaced 4.8Å apart. The intermediate space, comprising the CH_2 bridge, is approximated as a convex shape-effectively modeled as a rectangle with semicircular ends. The sample shapes that are attached to the NGA is shown below in Fig. 3.

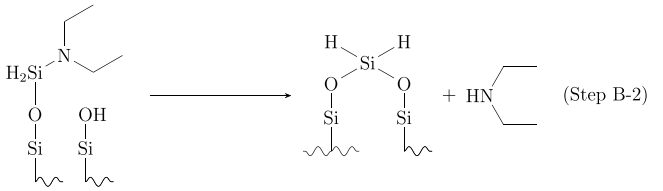
The monodentate configuration, which lacks the second Al-O bond, is modeled differently due to the out-of-plane displacement of one of the CH_3 groups (Yun et al., 2022a). It is represented by two circles: one with a 4.0Å diameter (CH_3) and the other with a projected effective diameter of 3.4Å, reflecting the horizontal projection of the non-planar methyl group. These circles are spaced 3.8Å apart to reflect the molecular geometry derived from quantum chemical calculations and steric mapping. This geometric abstraction enables accurate modeling of surface coverage and steric hindrance effects during inhibitor adsorption. The schematic diagrams of the modeling of chelates and monodentates are shown in Figs. 4 and 5, and the reaction mechanisms and molecule structures are demonstrated below as well to help understanding.



(6)



(7a)



(7b)

During the inhibitor adsorption process, the placement of each acetylacetone (Hacac) molecule is governed by a stochastic sampling approach. For each trial, a surface site is randomly selected, and a list of candidate rotation angles ranging from 0° to 345° in 15° increments is generated and randomly shuffled to ensure variability in orientation attempts. Each angle in the shuffled list is sequentially tested to determine whether a monodentate Hacac configuration can be accommodated at the selected site without overlapping adjacent molecules, based on van der Waals exclusion criteria. If none of the candidate orientations result in a sterically viable placement, the site remains unoccupied in that trial. The discretization with 15° step size is a trade-off between computational complexity and accuracy, as with thorough computational testing, it was found that step sizes below 15° do not lead to significant differences in the computed results. If a monodentate configuration fits successfully, the simulation then attempts to place a chelate (bidentate) configuration at the same site, starting from a reshuffled list of angles. This second angle list accounts for the fact that the single Al-O bond in monodentate species permits rotational flexibility, so the feasible orientations for chelate attachment are not constrained by the previously chosen monodentate angle. The process checks again for steric collisions in all attempted orientations and selects the first valid placement. If there are no available angles without collision with adjacent molecules, then it keeps the monodantate placed before. To reduce computational complexity, once either a monodentate or a chelate is successfully adsorbed, the molecule is considered geometrically fixed. That is, no further adjustment or rotation is permitted after placement. This assumption reflects a quasi-static treatment of adsorbed configurations and is consistent with other lattice-based surface Monte Carlo models that prioritize configurational diversity during trial selection rather than after placement (Pieck and Tonner-Zech, 2025).

2.2. Step B: precursor adsorption

After the adsorption of acetylacetone inhibitors in NGA, the bis(diethylamino)silane (BDEAS) precursor is introduced in Step B to initiate deposition on both surfaces. For simplicity, the model assumes that the

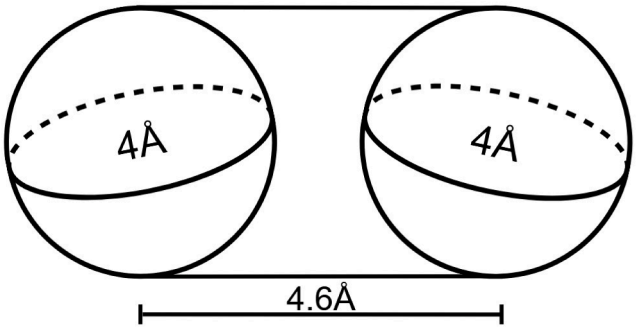


Fig. 4. Diagram of Chelate molecule. Each 4Å circle denotes a CH_3 group on the same horizon.

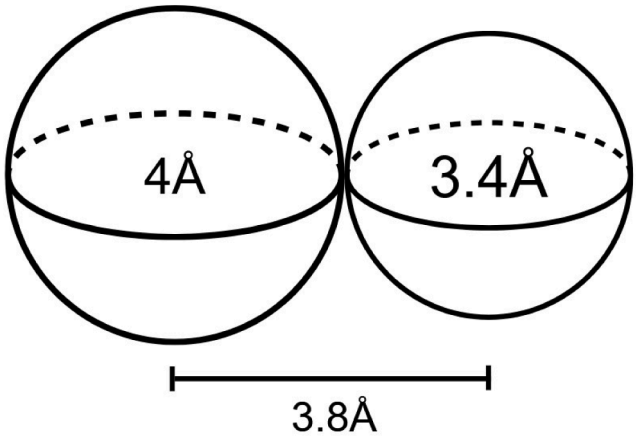


Fig. 5. Diagram of Monodentate molecule with two circles of 4Å diameter denoting a CH_3 group and 3.4Å diameter denoting projection of CH_3 on another side.

NGA is fully covered by inhibitor at the start of Step B with sufficiently long reaction time for Step A, and the GA remains completely available for precursor adsorption, so no need to consider the inhibitor-GA interactions. Full coverage means the available geometric space on the surface has been filled to the maximum extent permitted by steric constraints. BDEAS is a silicon-based precursor containing two diethylamine (DEA) ligands and two hydrogen atoms bonded to the silicon center. Upon exposure to hydroxyl-terminated SiO_2 surfaces, BDEAS undergoes a ligand-exchange reaction in which one DEA group is displaced, forming a covalent Si-O bond with the surface and releasing a DEA molecule into the vapor phase. DEA is a volatile byproduct and remains stable under the thermal ALD conditions employed.

Following this initial attachment, the adsorbed BDEAS species may seek out an adjacent hydroxyl site to undergo a second ligand-exchange reaction, releasing the remaining DEA group and forming a silicon bridge structure across the two sites. As described in Yun et al. (2022a), the geometric layout of the β-SiO_2 surface constrains bridge formation to occur only between specific neighboring columns. Due to the lateral spacing between surface sites, bridges can only form between pairs of vertical site columns, leading to incomplete surface coverage-a phenomenon also supported by experimental observations (Roh et al., 2022). Although the trilate DEA-silane configuration represents a necessary intermediate prior to silane bridge formation, such intermediates do not accumulate on the GA surface due to the distinct surface environment and reaction pathway. On the GA (SiO_2), the surface is free of inhibitor molecules and presents a high density of accessible hydroxyl sites. Under these conditions, once a BDEAS-derived trilate intermediate is formed, an adjacent hydroxyl site is almost always available, enabling rapid conversion into a bridged silane structure. This conversion step has a very low activation barrier (Yun et al., 2022a) and is effectively instantaneous on the timescale of the Monte Carlo simulation, preventing trilate intermediates from persisting. In contrast, on the NGA, inhibitor molecules introduce significant steric hindrance that frequently blocks access to neighboring hydroxyl sites even when they are chemically vacant, kinetically trapping the trilate configuration. On the GA, the only scenario in which a trilate could transiently exist is when both neighboring bridge-forming sites are already occupied. However, in this configuration, the surrounding bridged structures themselves impose sufficient steric exclusion to suppress additional precursor adsorption in the vicinity, preventing sustained trilate populations. Consequently, persistent trilate intermediates are not observed on the GA, and the resulting bridge-only growth mode naturally reproduces the experimentally reported SiO_2 surface coverage of approximately 86-94%.

On the NGA (Al_2O_3), the situation differs. Experimental studies have shown that Hacac chelate structures are chemically robust and effectively block BDEAS substitution (Merkx et al., 2020a). However, the bonding between the monodentate inhibitor and the surface is weaker and can be displaced by BDEAS. In the simulation logic, precursor adsorption begins by randomly selecting a candidate site. If the site is occupied by a chelate inhibitor, the trial is aborted, as chelates are considered non-substitutable. Otherwise, those surrounding monodentates are temporarily removed, and the algorithm attempts to place a trilate structure, which is a geometric abstraction of the single DEA silane structure after first chemisorption of BDEAS molecule.

The trilate is modeled using three circles: a central circle (diameter 5.0 Å) representing the silane group, and two side circles (diameter 4.0 Å) representing the middle CH_2 of ethyl groups. The terminal CH_3 groups are ignored as they are located on the upper horizon. The two wings are placed at a 45° angle from the central axis, with each wing offset 3.44 Å from the center. This structure is visualized in Fig. 6. If none of the tested orientations fit without overlapping neighboring adsorbates, the trial is discarded and the monodentates are restored. If the trilate is successfully placed, any previously removed monodentates that still geometrically fit are reintroduced around the trilate to preserve partial inhibitor coverage.

After successful trilate placement, the simulation checks whether an adjacent site is available for bridge formation. The bridge structure is a simplified remnant of the trilate, retaining only the central circle, as both DEA groups are eliminated, and extending across to an adjacent surface site, provided no steric collisions occur. If such a connection is feasible, the bridging structure is formed, representing the completion of Si-O-Si bond formation between two hydroxyl sites on both surfaces.

2.3. Step C: ozone oxidation

Following the precursor adsorption stage in Step B, the process proceeds to Step C, in which the surface is exposed to ozone to initiate oxidation and cleaning. The ozone exposure performs two primary functions: (1) it oxidizes the terminal -H groups on surface-bound silicon

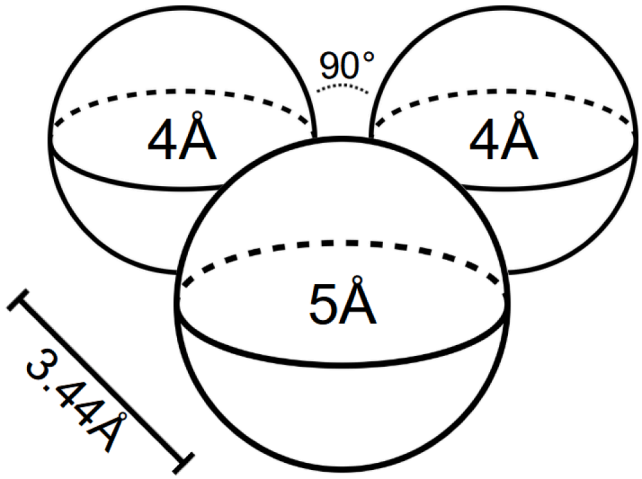


Fig. 6. Diagram of Trilate DEA-Silane structure. Center 5Å diameter circle denotes the Si-H_2 silane group, and each 4Å diameter circle denotes the ethyl group.

atoms into -OH groups, thereby completing the SiO_2 deposition process by creating the hydroxyl-terminated SiO_2 surface and has volatile O_2 as byproduct, and (2) it removes remaining surface-bound species, including monodentate and chelate inhibitors on the NGA and unreacted trilate precursor structures on both surfaces. It is important to distinguish between the chemical fate of fully incorporated silane bridge species and that of unreacted precursor-derived intermediates during the ozone step. Bridged silane structures are doubly anchored to adjacent surface hydroxyl sites and are therefore structurally integrated into the surface network. Experimental studies of SiO_2 ALD consistently show that such bridged species undergo oxidation of residual Si-H bonds during ozone exposure (Merkx et al., 2020a), yielding stable, hydroxyl-terminated SiO_2 surface species. In contrast, unreacted trilate BDEAS intermediates are only singly bound to the surface and retain a largely molecular character. Under ozone conditions, these weakly incorporated species do not follow the film-forming oxidation pathway, instead, they undergo ligand combustion and oxidative fragmentation, producing volatile byproducts that desorb from the surface rather than contributing to oxide growth.

In the simulation, ozone reactions are executed by randomly sampling surface sites. If a selected site is empty, the algorithm proceeds to the next trial. If the selected site corresponds to an endpoint of a silane bridge structure (i.e., a site occupied by a BDEAS-derived Si-H species), the simulation changes the state of the site to represent a newly formed Si-OH group. It is important to note a modeling simplification adopted here: although the two -H groups on the silicon atom are physically located between the two bridged surface sites (perpendicular to the bonding axis), the model assigns the oxidation product to the original endpoint sites. That is, ozone oxidation updates the state of the selected endpoint site to reflect SiO_2 formation, even though the actual -OH group lies spatially between the sites. This abstraction enables a clean and consistent update of grid-based surface state variables while avoiding the need for explicit sub-grid spatial resolution.

Furthermore, once both ends of a bridge are oxidized, the corresponding pair of sites is fixed and preserved as an intact deposited SiO_2 segment. In subsequent ALD cycles, when BDEAS adsorbs onto one of these previously deposited SiO_2 sites, the simulation enforces an immediate bridge formation with the adjacent paired site from the prior batch. This assumption ensures that deposited SiO_2 layers remain laterally uniform in height (i.e., all bridged sites belong to the same thickness layer), which simplifies thickness tracking and avoids complex three-dimensional structure resolution or adjacency-based bridging logic.

Lastly, if ozone selects a site occupied by residual adsorbates such as monodentate or chelate inhibitors, or unreacted trilate BDEAS structures, the associated molecular geometries and corresponding entries in the simulation grid are removed. This cleanup operation resets those sites to an unoccupied state, making them available for inhibitor adsorption or precursor attachment in the next cycle.

2.4. Step D: TMA exposure for AlF_3 removal and conversion modification

Following the main deposition steps, Step D introduces a trimethylaluminum (TMA) exposure that implements the reactive half-cycle of a thermal TMA/HF ALE scheme. In the mechanism, Step D performs two distinct but physically linked functions that depend on the chemical state of the local surface. First, TMA removes AlF_3-terminated segments that were created during the HF step of the previous batch by converting them into volatile Al-containing products, thereby producing true material removal in the model. Second, on hydroxyl-terminated SiO_2 segments, TMA drives a conversion reaction that forms Al-O-Si / Al_2O_3-like modified surface species. This conversion does not immediately reduce the geometric film thickness; rather, it prepares the affected segments for fluorination during Step E, after which they become AlF_3 and are removed during the subsequent batch's Step D. This state-dependent, cycle-decoupled treatment reflects the well-established thermal ALE principle that modification (conversion/fluorination) and removal can occur in different half-cycles (DuMont et al., 2017; Rahman et al., 2018).

From a simulation perspective, TMA molecules are modeled as stochastically sampling lattice sites on the surface. If the selected site is empty, the trial is skipped. If the selected site belongs to a deposited bridge segment, the model applies a state-dependent rule as follows:

(i) **Removal of AlF_3 (previous-batch product).** If the sampled bridge is in an AlF_3-terminated state (generated during Step E of the previous batch), the segment is eligible for removal during Step D. In this case, the entire bridge (both endpoints) is removed as a single correlated event, representing volatilization of the fluorinated layer during TMA exposure. If the removed segment resides on the first deposited layer (layer index = 1), the corresponding geometry is deleted and the sites return to the underlying substrate state. If the segment resides on a higher layer (layer index > 1), its layer index is decremented by one to represent net vertical removal while retaining the underlying stack.

(ii) **Conversion of SiO_2 (current-batch modification).** If the sampled bridge is in a hydroxylated SiO_2 state, TMA does not directly remove material in the same batch. Instead, the bridge is converted into an Al-containing modified state (Al-O-Si / Al_2O_3-like) while preserving its geometric representation and layer index. This chemical-state update captures conversion modification that enables subsequent fluorination during Step E. As in the deposition steps, if TMA samples one endpoint of a bridge, both connected sites are updated simultaneously to ensure consistent segment-level chemistry.

To more accurately represent the reactivity contrast between NGA and GA, Step D includes an exposure-based accessibility model. Since not all surface sites are equally accessible to incoming TMA molecules due to geometric crowding and steric shielding from surrounding features, a local exposure score is computed for each site before TMA attachment is accepted. Exposure is modeled as the fraction of a hemisphere above the surface site that remains unblocked by nearby deposited bridges.

This is computed in two stages:

1. **Horizontal (2D) shielding:** As shown in Fig. 7, for each adjacent bridge within a two-step hexagonal neighborhood, the tangent lines to its circular footprint define the angular sector it obstructs. For a single nearby bridge, this horizontal angular block is computed as

$$\theta_{\text{block}}^{(j)} = 2 \cdot \arcsin\left(\frac{r}{d_s^{(j)}}\right) \tag{8}$$

where r is the effective blocking radius (typically the midpoint of the bridge), and $d_s^{(j)}$ is the center-to-center surface distance between the current site and adjacent bridge j.

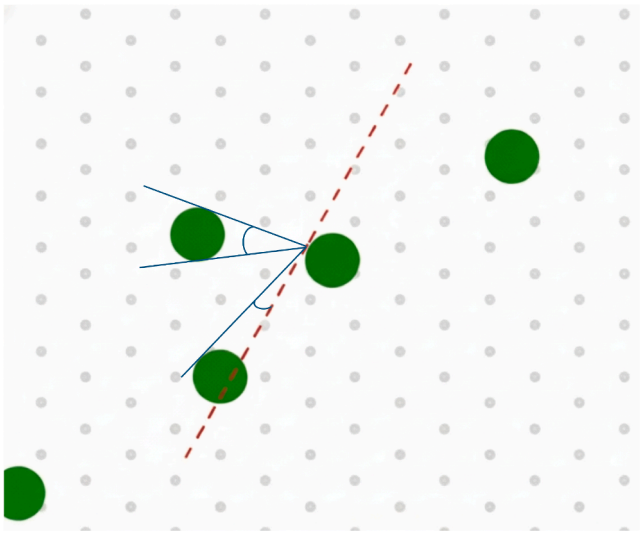


Fig. 7. The diagram for 2D exposure mechanism. The selected bridge would block half of the hemisphere, and the adjacent bridges would block by angles between tangent lines.

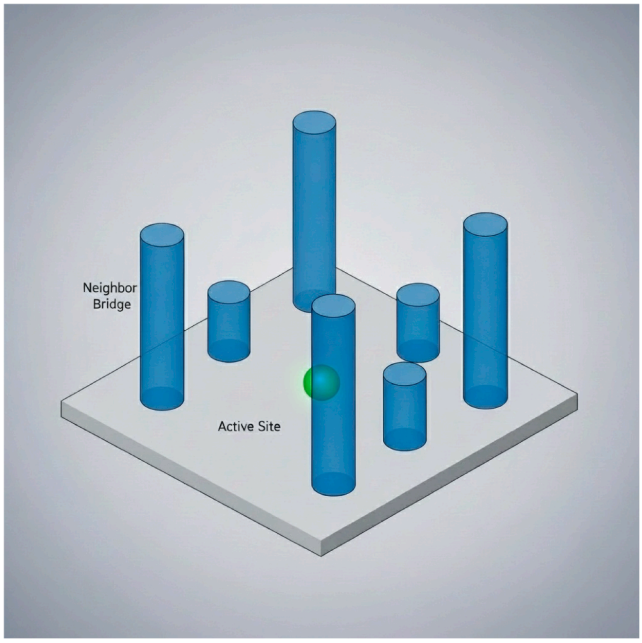


Fig. 8. The diagram for 3D exposure mechanism. The height difference between the selected site and surrounding deposited sites can results in different obstructions.

2. **Vertical (elevation) shielding:** As shown in Fig. 8, the elevation angle blocked by each neighbor is calculated as

$$\phi_{\text{block}}^{(j)} = \arctan\left(\frac{(L_j - L_i)\cdot H_{\text{layer}}}{d_s^{(j)}}\right) \tag{9}$$

where L_j and L_i are the discrete layer indices (heights) of the adjacent and current sites, respectively, H_{layer} is the height per monolayer in Å, and $d_s^{(j)}$ is the lateral distance between these two sites.

Assuming hemispherical access, the total blocked surface area fraction on the hemisphere is estimated from the combined solid angle of horizontal and vertical obstructions. However, since TMA is assumed to attack the root of a surface bridge from one side only, the molecule can only access half the hemisphere as the other half is geometrically obstructed by the bridge itself.

The hemisphere coverage blocked by a single adjacent bridge j is approximated as

$$f_{\text{block}}^{(j)} = \frac{\theta_{\text{block}}^{(j)} \cdot \sin\left(\phi_{\text{block}}^{(j)}\right)}{2\pi} \tag{10}$$

and the total unblocked exposure at a site is computed as

$$\text{Exposure} = 0.5 - \sum_j f_{\text{block}}^{(j)} \tag{11}$$

For computational efficiency, only bridges whose midpoint lies within a hexagonal neighborhood of radius 2 (i.e., two hex steps from the central site) are considered in the blocking calculation. The hex-distance of 1 is defined by the six closest neighbors, and hex-distance of 2 includes the twelve next-closest sites forming a second shell.

Finally, the NGA exhibits higher effective exposure in early batches due to its lower nucleation density. This causes the initially sparse islands on the NGA to experience greater etch reactivity, while the denser GA remains more shielded. Similar phenomena have also been shown in the higher reaction rates of the protruding area in other surface reactions (Mokhtarzadeh et al., 2022), and plasma etching can smooth the rough surface (Martin and Cunge, 2008). This exposure-driven etch contrast serves as the foundation for selective ALE that can suppress unwanted nucleation on the NGA while preserving most of the growth on the GA, thereby maintaining long-term process selectivity. In this work, exposure-based accessibility is explicitly considered only for the TMA modification step (Step D). While exposure limitations and steric hindrance are present for all surface reactions, Steps A-C are treated as self-limiting ALD processes and are simulated with sufficiently long exposure times such that detailed exposure effects have a limited impact on the final surface state. In contrast, Step D is intentionally exposure-limited and directly controls the subsequent etching behavior; allowing this step to proceed to completion would eliminate net deposition. This simplifying assumption reduces computational cost while focusing on the dominant selectivity-controlling mechanism. Extending exposure-based modeling to all reaction steps will be considered in future work.

It is important to note that the thermal TMA/HF conversion–etch chemistry employed in this work is not intrinsically perfectly selective to SiO_2 over Al_2O_3. Under sufficiently long or repeated exposures, both oxide materials can undergo fluorination and subsequent removal through Al-containing intermediate formation (Rahman et al., 2018). Accordingly, the ALE step described here does not imply absolute chemical preservation of the Al_2O_3 non-growth area (NGA). The focus of this study is instead on exposure-limited, kinetically driven ALE, in which effective selectivity emerges from differences in local exposure, nucleation density, and roughness evolution between the growth area (GA) and the NGA. In practical integration, prolonged ALE would gradually consume the underlying NGA oxide; this limitation can be mitigated by employing a sufficiently thick blocking layer prior to ASALD so that gradual consumption during corrective ALE does not compromise device functionality over the targeted number of cycles. Future extensions of this work could explicitly incorporate NGA thickness consumption into the multiscale framework and investigate run-to-run control strategies that jointly optimize ALE exposure time and initial blocking-layer thickness to ensure long-term selectivity and process robustness.

Step E: HF fluorination of the TMA-modified surface

After the TMA exposure in Step D, Step E applies a thermal hydrogen fluoride (HF) exposure that fluorinates the Al-containing modified surface segments formed during the current batch. In this step, Al-O-Si / Al_2O_3-like modified segments are converted into an AlF_3-terminated state. Consistent with the thermal HF-based ALE mechanism, this fluorination step is treated as a self-limiting chemical transformation that

prepares the modified layer for removal during the next batch's TMA exposure, rather than producing immediate thickness loss within the same batch (DuMont et al., 2017; Rahman et al., 2018).

In the microscopic simulation, HF exposure updates the chemical state of any TMA-modified bridge segment to an AlF_3 state while preserving its geometric representation and layer index. HF is not allowed to remove material directly in the model; instead, removal is executed in Step D of the subsequent batch when TMA encounters an AlF_3-terminated segment and volatilizes it. This explicit decoupling ensures that the model captures the cycle-dependent nature of thermal ALE, in which fluorination and removal occur in different half-cycles.

2.5. Definition of film thickness and selectivity

In the present work, film thickness is computed from the microscopic Monte Carlo simulation by tracking the average number of deposited surface layers on the growth area (GA) and non-growth area (NGA). Each completed silane bridge formed during Step B and oxidized during Step C is treated as one deposited SiO_2 layer element in the geometric stack. The normalized average layer number on each surface, denoted as L_{GA} and L_{NGA}, is obtained by averaging the discrete layer indices over all surface sites. The physical film thickness is then calculated as

$$T_{GA} = L_{GA} \cdot h_{SiO_2}, \qquad T_{NGA} = L_{NGA} \cdot h_{SiO_2}, \tag{12}$$

where h_{SiO_2} is the effective monolayer thickness of SiO_2 per ALD cycle. In this study, h_{SiO_2} is taken as constant and chosen to be consistent with experimentally reported growth-per-cycle values for SiO_2 ALD using BDEAS and ozone (typically 0.9–1.1 Å per cycle).

A key feature of the revised thermal TMA/HF ALE mechanism is that chemical conversion and physical removal are temporally decoupled across batches: Step E (HF) converts Al-containing modified segments into AlF_3, while physical removal of AlF_3 occurs during the subsequent batch's Step D (TMA). Therefore, the instantaneous end-of-batch surface state after Step E may contain AlF_3-terminated segments that are chemically prepared for removal but not yet physically removed. To enable consistent batch-wise reporting of net thickness change and selectivity, we define the effective post-batch thickness as the thickness obtained after removing all AlF_3-terminated segments in a bookkeeping operation that mirrors the removal action that will occur at the start of the next batch's TMA exposure. This bookkeeping operation is used only for metric evaluation and does not represent an additional physical process step in multi-batch simulation.

Selectivity is defined as a normalized thickness contrast between the growth area and non-growth area:

$$S = \frac{T_{GA} - T_{NGA}}{T_{GA} + T_{NGA}}. \tag{13}$$

This definition yields $S = 1$ for ideal area-selective deposition with no net growth on the NGA and $S = 0$ when the GA and NGA exhibit identical thicknesses. This selectivity metric is used consistently throughout the manuscript to quantify the effectiveness of inhibition and the TMA/HF-based ALE correction across multiple deposition batches.

3. Macroscopic computational fluid dynamics model

To accelerate the commercialization and optimization of area-selective atomic layer deposition (ASALD) technology, it is essential to connect the surface-level phenomena of ASALD with the conditions encountered in full-scale industrial reactor systems. While experimental approaches remain invaluable for understanding process behavior, they are often costly, time-consuming, and limited in their ability to provide spatially and temporally resolved in situ data. In contrast, multiscale in-silico modeling offers a cost-effective solution to these limitations by simultaneously analyzing surface kinetics and fluid transport phenomena within the reactor environment.

In this work, we expand upon the previously developed microscopic Monte Carlo surface model by integrating it with a computational fluid dynamics (CFD) framework, thereby enabling simulation of real-time ASALD behavior under dynamic reactor conditions. The objective is to develop a fully functional digital twin that captures both local reaction kinetics and global transport characteristics, with particular emphasis on reproducing the spatial pressure and temperature profiles on the wafer surface. Such a model facilitates accurate prediction of film growth behavior under realistic operating conditions and provides a tool for process optimization and control.

To ensure industrial relevance and physical accuracy, we build upon an experimentally established ALD reactor developed by the National Institute of Standards and Technology (NIST). This reactor system includes a well-characterized precursor delivery mechanism, reaction chamber, and outlet piping. Importantly, it is supported by comprehensive experimental measurements, which allow us to verify the accuracy of our CFD modeling approach. The reactor geometry and flow architecture are representative of industrial-scale ALD systems, enabling direct comparison between simulated and experimental results and thereby enhancing the credibility of the macroscopic simulation framework.

The following Section 3.1 will describe the reactor design, computational domain, and mesh generation for the CFD simulation. We then present detailed comparisons between simulation results and experimental measurements to validate the CFD setup, boundary conditions, and model parameters in Section 3.2.

3.1. Reactor design and CFD model development

The simulations were performed using a reactor geometry based on the optically accessible perpendicular-flow single-wafer ALD chamber developed at NIST (Kimes et al., 2012). The chamber consists of an expansion cone leading into a cylindrical reactor body with three optical access ports and a centrally located heated wafer chuck. Precursor and carrier gases are introduced through four stainless-steel delivery lines (~4.8 mm ID) positioned symmetrically at the top of the chamber.

Computational domain and mesh: The computational domain included the full chamber volume, inlet manifolds, and exhaust lines (Fig. 9). The mesh was an unstructured tetrahedral grid with prism layers along solid surfaces to resolve near-wall gradients. Local refinement was applied in three critical regions: (i) the inlet jet zone, (ii) the optical imaging plane, and (iii) the wafer surface. Mesh quality was maintained with a minimum orthogonal quality of ≈0.3 and an aspect ratio less than 4. A mesh-independence study confirmed that both peak $MoCl_5$ arrival time and integrated wafer-plane concentration varied by less than 3%, and the intermediate mesh was adopted for all cases.

Governing equations and species model: Argon served as the carrier gas, and $MoCl_5$ was treated as a non-reactive tracer to isolate transport effects (reaction kinetics will be introduced in subsequent multiscale coupling). A multicomponent diffusion model was used with a user-defined binary diffusivity D_{MoCl_5-Ar} in the range of 2×10^{-4} to 1×10^{-3} $m^2\ s^{-1}$. Flow symmetry and plume structure were insensitive to variations within this range, indicating convective dominance under the examined flow conditions. The reactor temperature was fixed at 393 K to match the experiments.

Turbulence modeling: At relatively low flow rates-such as the 0.1 L min^{-1} (100 SCCM) case with approximately 3 mol% $MoCl_5$ injection-the flow remained in a low-Reynolds-number transitional regime. Under these conditions, a laminar model successfully reproduced the experimentally observed symmetric plume. At higher flow (0.5 L min^{-1}, ≈ 2.4 kPa), however, transitional vortices developed near the wafer surface. Standard RANS models (k-ω SST, Transition SST) excessively damped this unsteady vortex motion, resulting in artificial asymmetry in the $MoCl_5$ distribution. To resolve large-scale unsteadiness while maintaining RANS stability near walls, the Scale-Adaptive Simulation (SAS) model was employed. The SAS approach restored the experimentally observed symmetric flow structure and accurately captured the natural vortex dynamics under the higher-flow regime.

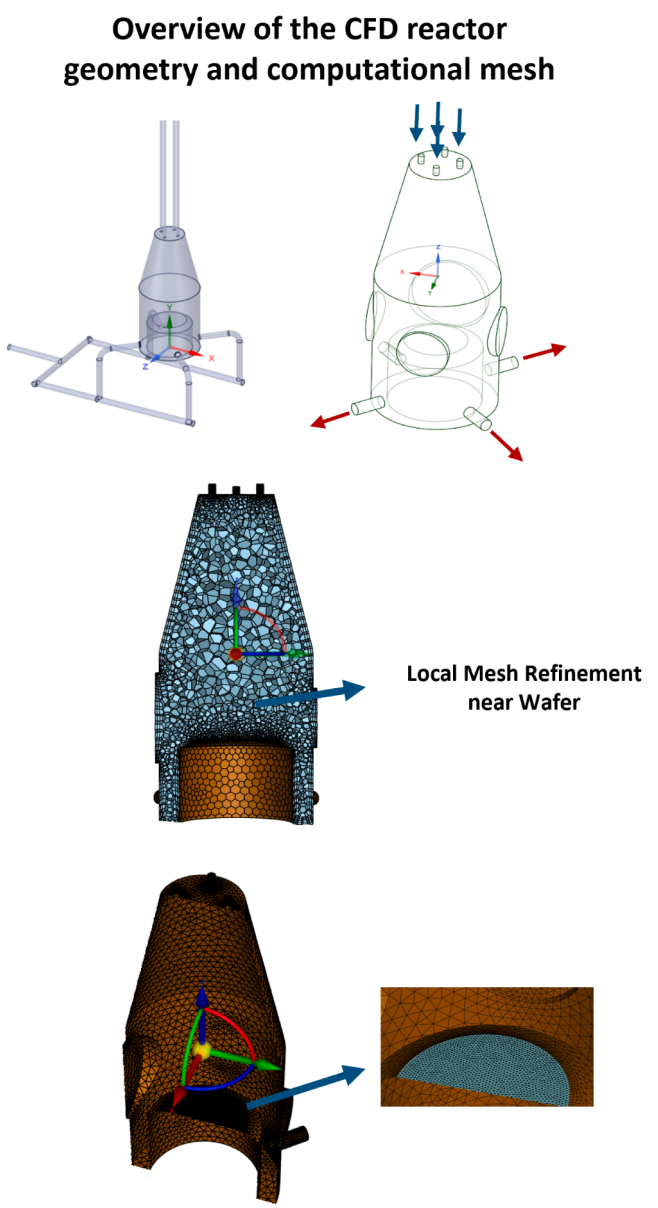


Fig. 9. Overview of the CFD reactor geometry and computational mesh. (Top) Optically accessible ALD reactor with inlet and outlet configurations. (Middle) Unstructured tetrahedral volume mesh showing refined regions near the optical window and wafer. (Bottom) Detailed view of the prism boundary layers and local refinement applied at the wafer surface.

3.2. *Verification of CFD model*

(a) Feed-line and outlet flow validation.

To ensure accurate boundary conditions, separate inlet- and outlet-line simulations were conducted prior to full reactor modeling. For the inlet line, velocity and pressure were benchmarked against experimental data Kimes et al. (2012) to confirm realistic flow delivery. The simulated upstream inlet pressure (1044 Pa) agreed with the measured 998 Pa within 4.6%, and the outlet velocity, i.e., inlet reactor velocity (40.9 m s^{-1}), matched the measured 41 m s^{-1}. A separate outlet simulation verified the expected pressure drop through the exhaust network, yielding 308 Pa inlet pressure and 58.5 m s^{-1} outlet velocity, consistent with the full reactor's ≈310 Pa prediction. These results confirm that both inlet and outlet configurations accurately reproduce the measured pressure-velocity relationships, providing reliable boundary inputs for subsequent transient $MoCl_5$ transport simulations.

(b) Reactor flow structure and quantitative validation.

For validation, the 100 SCCM case was used to assess the CFD model's ability to reproduce key flow and transport parameters observed experimentally. The simulated chamber pressure (318 Pa) closely matched the NIST-measured value of 317.6 Pa at the chamber location (P2) in Fig. 10, demonstrating excellent agreement. In addition, the inlet pressure was consistent with the experimental range of 300-340 Pa after accounting for the expected pressure drop through the inlet delivery lines.

The transient $MoCl_5$ concentration buildup and decay also reproduced the measured absorbance profiles at multiple points (inlet, mid-chamber, and outlet), confirming correct temporal evolution and flow symmetry as shown in Fig. 11.

The velocity distribution at the wafer surface was predominantly between 0.10 and 0.12 m s^{-1} (blue region), with only localized areas reaching up to 0.33 m s^{-1}, aligning with the experimentally derived value of 0.1 m s^{-1} shown in Fig. 12.

The residence time, obtained from the simulated $MoCl_5$ mass-fraction evolution as shown in Fig. 13, was approximately 0.8 s (from 0.7 to 1.5 s), matching the experimentally observed gas residence time under the same flow and pressure regime.

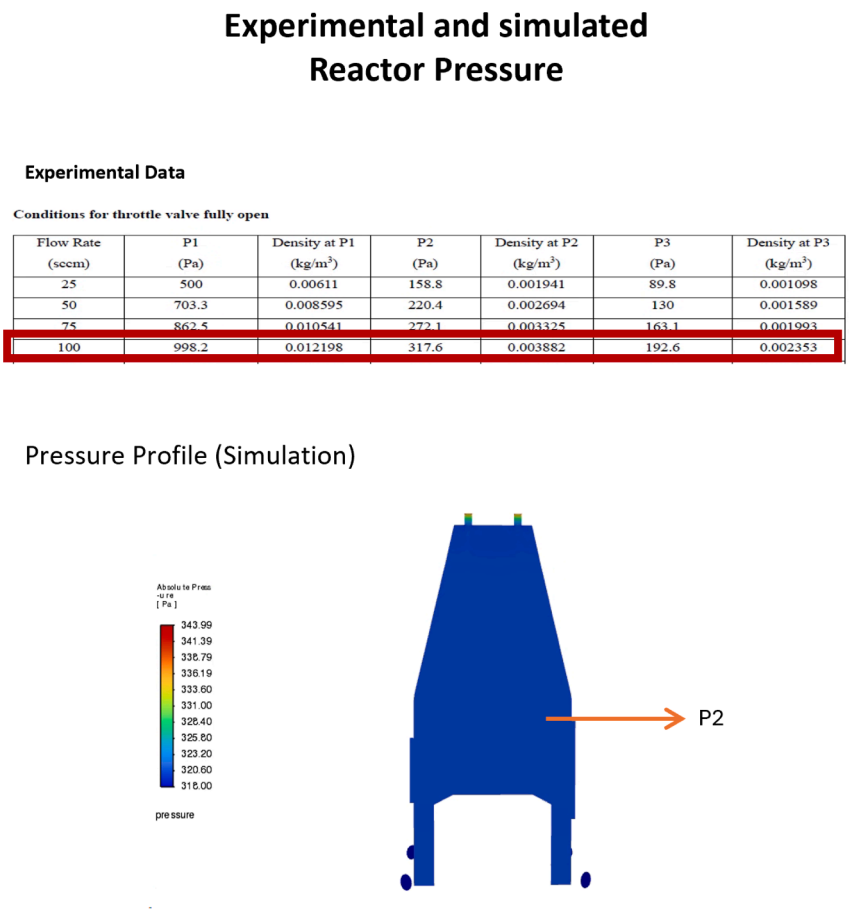
(c) Image-based validation via circular-masked planes.

To directly compare CFD predictions with the optical measurements from the NIST setup, simulated $MoCl_5$ concentration fields were post-processed into circular-masked planes corresponding to the reactor's viewing window, through which the telecentric lens and CMOS camera captured time-resolved absorbance images (Maslar and Kalanyan, 2025). The resulting sequence of simulated images reproduced the buildup and decay of $MoCl_5$ during pulsed injection with high spatial and temporal fidelity as shown in Fig. 14. Both the symmetry of the spreading front and the decay dynamics matched the experimental observations, with consistent transition timing between frame intervals (~1.12-1.45 s).

This image-based validation confirms that the CFD model accurately captures the spatial uniformity, transient $MoCl_5$ transport, and gas residence-time characteristics observed experimentally. Combined with the quantitative agreement in chamber pressure, wafer-surface velocity, residence time, and $MoCl_5$ buildup/decay profiles, these results validate the CFD model and establish its predictive reliability for subsequent reaction-coupled ALD simulations.

Although the CFD model uses Ar and $MoCl_5$ as a heavy, non-reactive tracer system for validation, this choice does not affect the fidelity of the multiscale integration because the CFD framework is designed to provide reactor-scale pressure and transport dynamics that are independent of specific molecular chemistry. The purpose of the validation step is to ensure that the turbulence model, boundary conditions, residence time, and transient precursor delivery are accurately captured. Once these macroscopic flow-field characteristics are verified, Fluent applies its built-in multicomponent diffusion model to simulate the transport of the ALD precursors such as BDEAS, O_3, TMA, and HF, using their individually defined diffusivities and material properties, making the model fully transferable and scalable to different species. Moreover, species transport in this reactor operates in a convection dominated regime, so differences in molecular diffusivity have only a minor influence on macroscopic flow and delivery patterns. Additionally, ALD is an atomic-layer process in which only a single molecular layer is deposited per cycle; the associated reactant consumption and byproduct formation occur within an extremely thin region above the wafer and do not significantly perturb chamber-scale hydrodynamics. Consequently, validating the CFD flow field with a heavy, non-reactive tracer reliably ensures accurate prediction of precursor transport and near-surface delivery for the ALD species used in this study.

The coupled multiscale simulation assumes the absence of gas-phase side reactions among precursors and co-reactants. This assumption is consistent with the fundamental operating principle of atomic layer deposition and atomic layer etching processes, in which reactive species



Flow Rate (sccm)	P1 (Pa)	Density at P1 (kg/m^3)	P2 (Pa)	Density at P2 (kg/m^3)	P3 (Pa)	Density at P3 (kg/m^3)
25	500	0.00611	158.8	0.001941	89.8	0.001098
50	703.3	0.008595	220.4	0.002694	130	0.001589
75	862.5	0.010541	272.1	0.003325	163.1	0.001993
100	998.2	0.012198	317.6	0.003882	192.6	0.002353

Fig. 10. (Top) Experimental data from NIST showing pressure at multiple locations. (Bottom) Simulated absolute pressure contour with highlighted chamber pressure region (P2), showing excellent agreement (317 Pa experimental vs. 318 Pa simulated).

are introduced sequentially rather than simultaneously, with inert gas purge steps separating each exposure. As a result, only a single reactive species is present in the reactor at any given time, and gas-phase reactions between precursors are intentionally suppressed by process design. Under these exposure-limited conditions, film growth and etching are governed by surface-mediated reactions, while gas-phase chemistry plays a negligible role. Potential gas-phase decomposition of individual precursors is minimal under the operating temperatures and residence times considered and does not contribute appreciably to material deposition or removal. Neglecting gas-phase side reactions therefore does not compromise the accuracy of the model for the surface-controlled ALD and ALE regimes investigated in this work and is consistent with established experimental interpretation and modeling practice in the ALD/ALE literature.

4. Multiscale real-time simulation

Multiscale integration of microscopic and macroscopic models

After independently developing the microscopic Monte Carlo model and the macroscopic computational fluid dynamics (CFD) model, as well as verifying the accuracy of the CFD-based fluid flow predictions, both components are combined into a real-time, coupled multiscale simulation framework. This integration forms the basis for a digital twin capable of resolving both surface kinetics and reactor-scale transport phenomena simultaneously. The motivation for such multiscale coupling lies in two key aspects: (1) the need to use real-time local pressure data to drive the microscopic surface reactions, and (2) the necessity of accounting for the effect of surface reactions as dynamic source terms within the CFD domain.

First, it is well established that the precursor partial pressure at the wafer surface does not instantaneously reach its target value upon dosing. Instead, there exists a finite transport delay governed by the reactor geometry, inlet configuration, and flow resistance. Many industrial ALD and ALE systems use showerhead or diffusive inlet structures to promote lateral gas uniformity (Yun et al., 2022b; Ding et al., 2019), but these structures inherently slow down the rate of precursor buildup. As a result, the local pressure on the wafer surface evolves dynamically, and assuming a constant or pre-defined pressure, as is commonly done in purely microscopic simulations, can lead to unrealistic kinetics. By feeding real-time pressure data from the CFD model into the Monte Carlo simulation, the multiscale model captures the spatial and temporal evolution of the precursor environment more accurately.

Second, the influence of surface reactions on fluid dynamics must be considered. Prior works have demonstrated coupled frameworks where surface kinetics are modeled using microscopic techniques and reactor flow is resolved using CFD (Yun et al., 2022b). However, in many such studies, the coupling is one-way: the CFD model is first solved independently, and the resulting pressure and temperature profiles are used as inputs to the microscopic simulation via interpolation. This neglects the feedback of surface consumption on precursor concentration near the wafer, which becomes especially important when reactions are spatially heterogeneous or kinetically fast. In practice, surface reactions locally deplete precursor species, generate volatile products, and can alter the boundary layer flow, meaning that their effect should be reflected back in the CFD simulation via a surface-based source term.

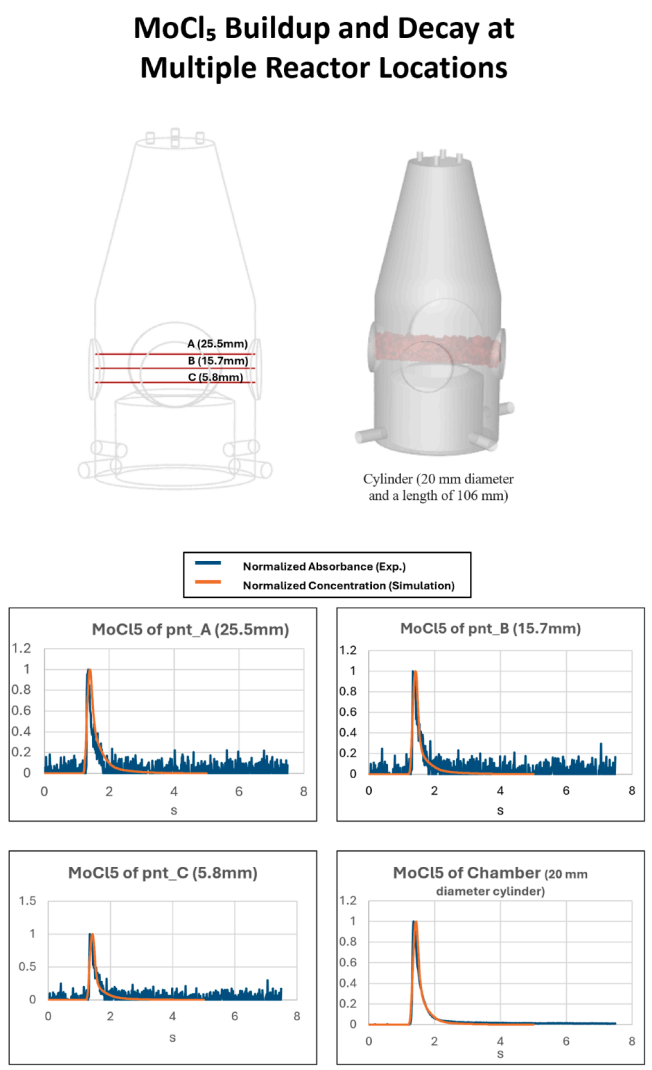


Fig. 11. Comparison of simulated $MoCl_5$ concentration and experimentally measured absorbance at multiple chamber locations (Maslar and Kalanyan, 2025). Top: Reactor geometry showing monitoring points A (25.5 mm), B (15.7 mm), C (5.8 mm), and the wafer-plane cylinder (20 mm diameter, 106 mm length). Bottom: Normalized time-series data illustrating $MoCl_5$ buildup and decay, showing close agreement in transient behavior and peak timing between simulation and experiment.

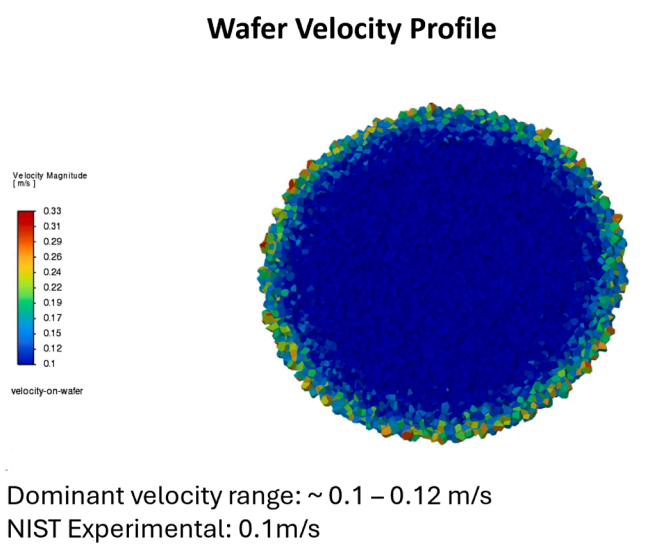


Fig. 12. Wafer surface velocity distribution showing dominant velocity range of 0.10-0.12 m s^{-1} (blue region), consistent with the NIST experimental measurement of 0.1 m s^{-1}.

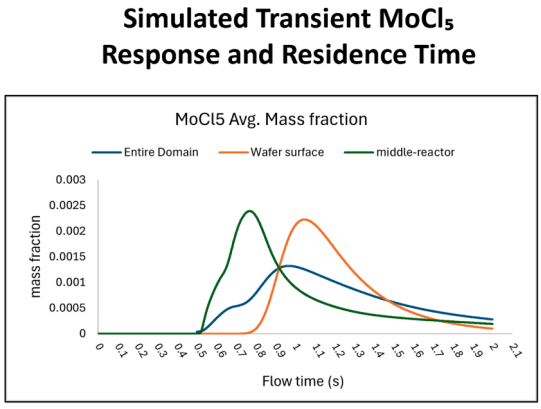


Fig. 13. Transient $MoCl_5$ concentration response showing buildup during precursor pulse (0.5-0.7 s) and subsequent decay during purge. The average residence time of ≈0.8 s closely matches the experimentally reported value (Maslar and Kalanyan, 2025).

The full multiscale simulation loop is illustrated schematically in Fig. 15. Each cycle begins with a steady-state CFD simulation under N_2 purge flow, which is saved and used to simulate post-purge initial conditions. The simulation proceeds as follows:

1. Read the instantaneous facet-average pressure across the wafer surface from the CFD field.
2. Use the local partial pressure of reactants as input to the microscopic simulation to compute real-time reaction rates.
3. Run the Monte Carlo surface simulation over a short time interval (0.1 s in this work) to calculate the normalized coverage increment for each surface region.
4. Convert the coverage increment into a surface reaction source term using:

$$S = \frac{\Delta\theta \cdot Y_{\max}}{\Delta t} \tag{14}$$

where $\Delta\theta$ is the change in normalized coverage, $Y_{\max}$ is the maximum molecular yield per unit area, and $\Delta t = 0.1$ s.

5. Inject the species source term back into the CFD model via a user-defined function (UDF), assigning it to mesh cells directly above the wafer using a user-defined memory (UDM) scheme to identify the corresponding surface locations. The thermal source from surface reactions is ignored, and the consistent isothermal operating condition is assumed.

This cycle is repeated iteratively to produce a time-resolved coupling between surface chemistry and reactor-scale fluid dynamics. The choice of 0.1 s for the coupling interval reflects a balance between numerical stability and physical fidelity. The selection of a 0.1 s coupling interval represents a practical compromise between temporal resolution and the intrinsic stochastic resolution of the microscopic Monte Carlo surface model. Although this interval is smaller than the characteristic reactor residence time (approximately 0.8 s), further reduction of the coupling step does not proportionally improve accuracy in the present framework. The microscopic model operates on a discrete surface lattice and advances time through stochastic reaction events, with typical

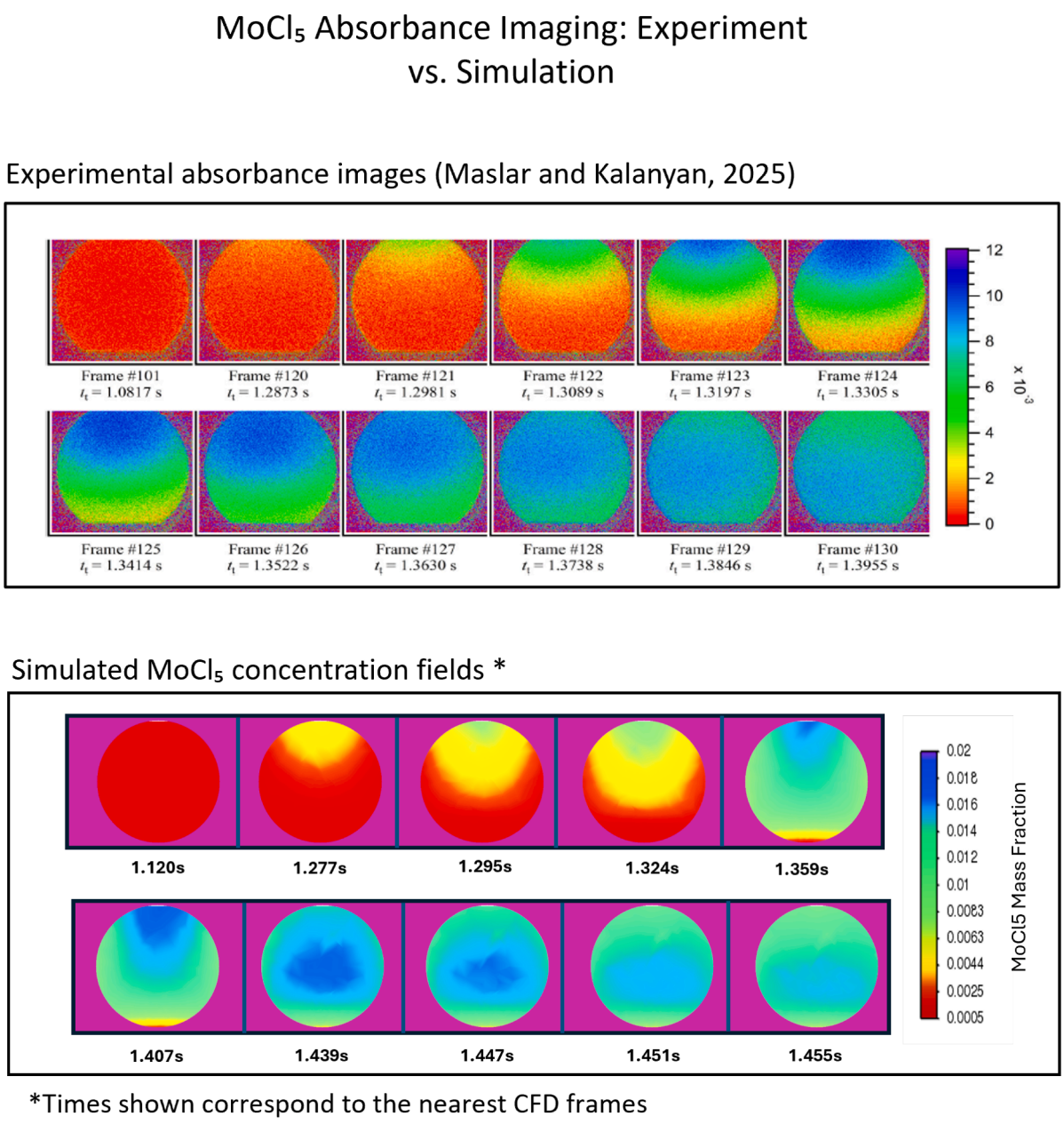


Fig. 14. $MoCl_5$ image comparison between experiment and simulation during a pulsed injection sequence in the NIST optically accessible ALD reactor. **Top:** Experimental absorbance images acquired through the optical viewing window using a telecentric lens and CMOS camera (Maslar and Kalanyan, 2025). **Bottom:** CFD-predicted $MoCl_5$ concentration fields extracted from circular-masked mid-plane slices, aligned to match the camera field of view.

individual reaction time increments already on the order of several milliseconds. Under these conditions, the dominant source of temporal uncertainty arises from the finite spatial resolution and statistical variance of the Monte Carlo simulation rather than from the CFD-MC coupling frequency. Over-refinement of the coupling interval would therefore primarily increase computational cost and stochastic noise without yielding commensurate gains in physical fidelity. The chosen interval is sufficient to resolve the gradual precursor delivery transient at the reactor scale while remaining consistent with the effective temporal resolution of the surface kinetics model. Achieving stricter time-step independence would require substantially increasing the microscopic model resolution (e.g., larger surface grids or ensemble averaging), which is computationally prohibitive at present. Given that the NGA region contains approximately 2500 reactive sites and the GA contains around 3840 CFD cells, individual reaction time increments are typically on the order of 10^{-3} s, so the 0.1 s interval allows for sufficient reaction resolution while maintaining acceptable statistical variance inherent to the stochastic Monte Carlo simulation.

In the present multiscale framework, heat generation from surface reactions is neglected and the reactor is treated as isothermal. This assumption is justified by both the process chemistry and the operating conditions considered in this work. Following the revision of the etching mechanism, the ALE step is formulated as a fully thermal, HF-based process without plasma excitation. Under these conditions, only a single molecular layer reacts per cycle, and the total amount of reacted material is extremely small relative to the thermal mass of the wafer, chuck, and reactor hardware. In industrial ALD and thermal ALE systems, the wafer temperature is actively controlled and maintained under near-isothermal conditions, such that any heat released by surface reactions is rapidly dissipated and does not result in measurable local

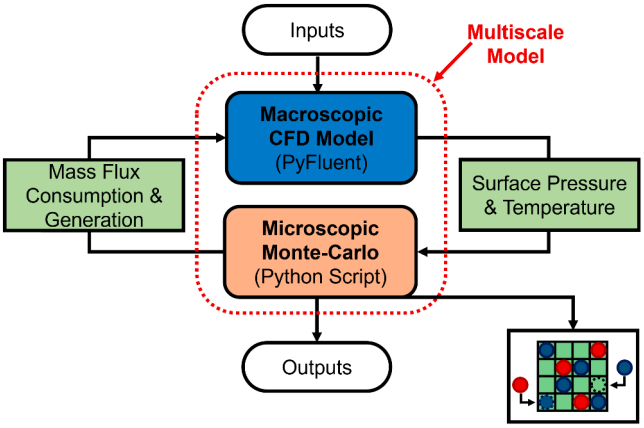


Fig. 15. The schematic diagram of the multiscale simulation. The CFD program is started and controlled in PyFluent environment, sending temperature and pressure to microscopic python scripts, and receive the outputs of microscopic model as sources of species.

temperature excursions. As a result, reaction heat does not significantly perturb surface kinetics, precursor transport, or selectivity in the present process window. Neglecting reaction heat is therefore a reasonable and widely adopted approximation in ALD and thermal ALE modeling, and does not affect the qualitative or quantitative conclusions of this study.

Through this multiscale framework, we realize a fully coupled, time-resolved digital twin of the ASALD process. The model faithfully captures both stochastic reaction dynamics and reactor-scale transport effects, enabling predictive analysis and optimization of self-aligned deposition and etching cycles under realistic industrial conditions.

5. Results and analysis

This section presents the simulation results obtained from both the microscopic and multiscale ASALD models, focusing on the evolution of surface coverage, film thickness, and selectivity across multiple deposition-etching cycles. The results demonstrate the capability of the predictive model to capture key physical trends and process sensitivities observed in ASALD systems.

The analysis begins with single-batch microscopic simulation results in Section 5.1, including both the first and second batches with sufficient reaction time for saturated reaction. While the first batch illustrates the initial surface adsorption and reaction dynamics, the second batch is critical for revealing the onset of nucleation accumulation on NGA. This comparison highlights how partially formed nuclei from earlier batches serve as reactive sites for continued growth, and clarifies the mechanism by which selectivity is gradually lost. Including the second batch also serves to establish how the multi-batch Monte Carlo scheme captures surface history and accumulation effects over time. Following this, multi-batch microscopic simulation results without any etching steps are provided to show how selectivity degrades as the number of cycles increases. The model predicts a characteristic nucleation delay period on the NGA, followed by accelerated growth that ultimately compromises selective deposition. Subsequent simulations introduce etching into the cycle and analyze the effect of different etch times on long-term selectivity. The results would reveal the dependence of batch-wise selectivity on etch duration.

The final part of the results examines the behavior of the multiscale model in Section 5.2, which integrates real-time precursor pressure evolution with microscopic surface kinetics. Comparisons between multiscale and pure microscopic simulations at fixed etch durations show that the multiscale model predicts a different effective etch rate under otherwise identical conditions, attributed to the finite pressure delivery and development time within the reactor.

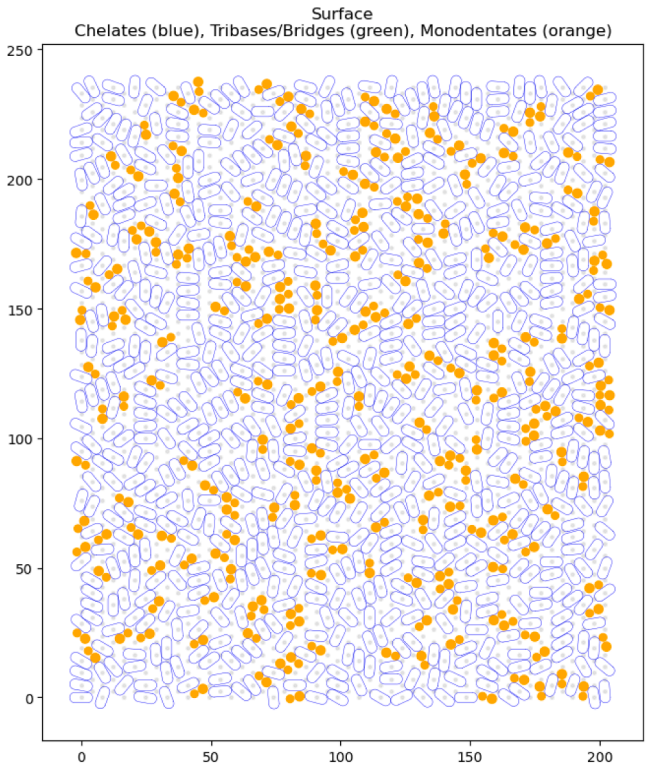


Fig. 16. Inhibitor adsorption on NGA on bare grid. Orange molecules are monodentates, blue molecules are chelates.

5.1. Multi-batch microscopic simulation

The NGA after 3.0s of inhibitor adsorption at $P_{Hacac} = 300$ Pa is shown in Fig. 16 The orange-filled geometries represent monodentate inhibitors, while the blue-outlined geometries correspond to chelate structures. The simulation shows that monodentates comprise approximately $22 \pm 2\%$ of the total adsorbed molecules, which is in good agreement with the experimentally reported fraction of $20 \pm 1.5\%$ (Merkx et al., 2020a). The overall inhibitor density is found to be 1.92 molecules per square nanometer, which is slightly higher than the experimentally observed value of 1.7 molecules per square nanometer (Merkx et al., 2020a). This deviation can be attributed to minor differences in site spacing used during surface grid generation. Nonetheless, both the fractional and total coverage values fall within a reasonable range of the experimental data, supporting the physical validity of the surface adsorption model.

The precursor adsorbed on the NGA after 3.0s reaction at $P_{BDEAS} = 300$ Pa is shown below in Fig. 17 The green circle is the final silane bridge structure, and the trilate structure (three circles) is the intermediate silane-DEA structure where only one DEA molecule is eliminated and cannot form the bridge structure either because the adjacent sites are not available or the formed bridge would collide with other molecules. To calculate the normalized coverage, a simulation of direct BDEAS adsorption on NGA surface without inhibitor is conducted and shown in Fig. 18. The average normalized coverage obtained from multiple simulation runs with different random seeds is approximately $6.7 \pm 1.2\%$, whether considering the total normalized coverage or only the coverage at bridge endpoints. This value is in good agreement with the experimentally reported normalized coverage of 8% (Merkx et al., 2020a), further validating the accuracy of the microscopic model in capturing unwanted precursor nucleation on NGA. For the BDEAS adsorption on GA, the result is demonstrated in Fig. 19.

As illustrated in Fig. 19, the GA surface exhibits a geometric limitation due to its two-column arrangement. Specifically, if a given site

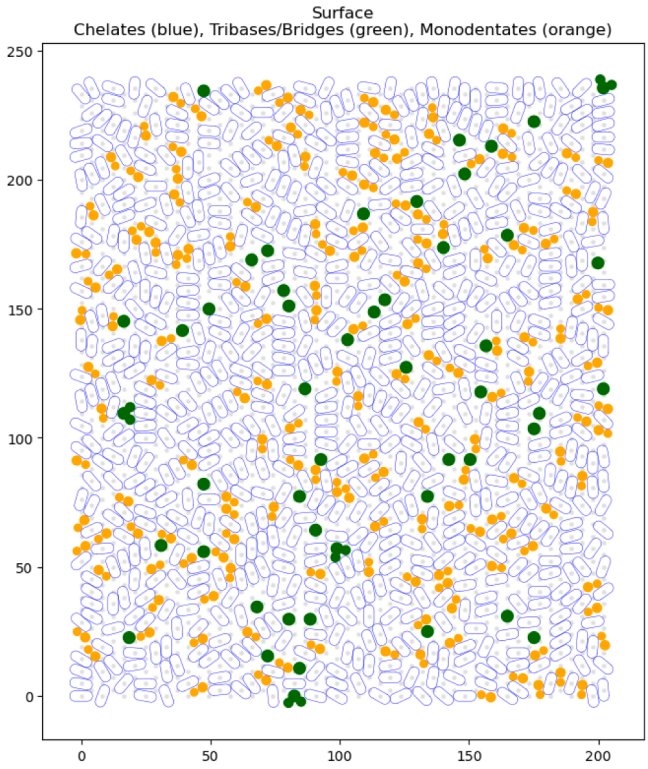


Fig. 17. Precursor substitution on inhibitor-adsorbed NGA surface. The green circle is the final Silane bridge structure, the green trilate is the intermediate Silane-DEA structure.

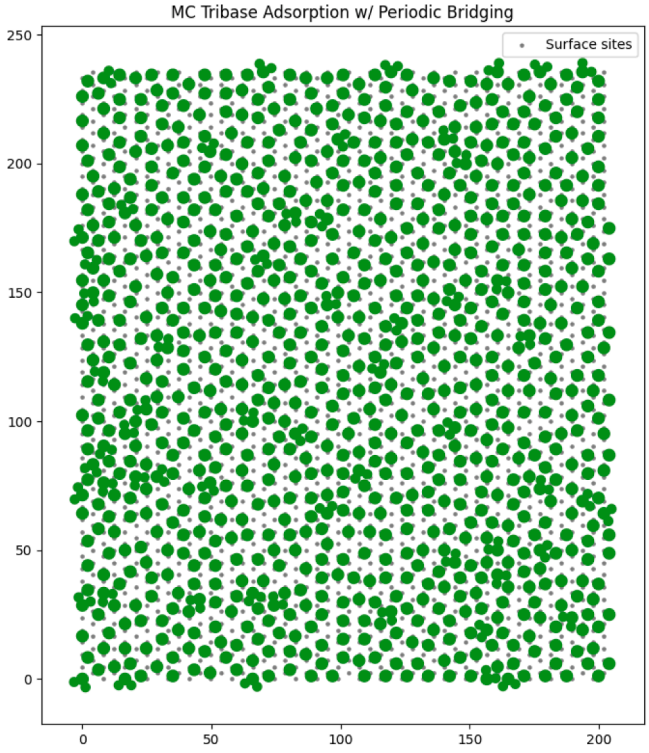


Fig. 18. Normalized benchmark of BDEAS direct adsorption on NGA surface, with 898 total sites adsorbed and 838 of them are end points of bridge structure.

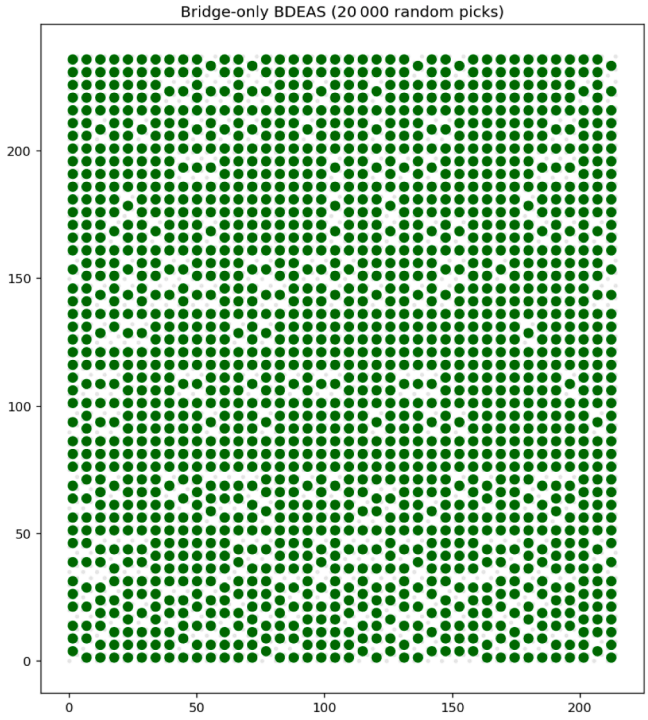


Fig. 19. BDEAS precursor adsorption on GA with the microscopic mechanism described in Yun et al. (2022a). The surface grid are separated into 2-column groups and the bridge can only formed between the two columns in each group.

is between an existing bridge structure formed by the two sites directly above and another formed by the two sites directly below, it becomes isolated and incapable of forming a bridge with any neighboring site. This constraint results in so-called dead points where no further precursor attachment can occur. Experimental studies have reported that the actual surface coverage on the GA ranges from 86% to 94%. The simulation presented in Fig. 19 yields a final coverage of 90%, which lies well within the experimentally measured range, further supporting the accuracy and physical relevance of the microscopic model.

The grid after Step C, the ozone oxidation for 1.0s at $P_{O_3} = 60$ Pa is shown in Fig. 20 The ozone oxidation step does not alter the overall deposited layer structure. Its primary function is to oxidize the surface-bound silane species, converting them into hydroxyl-terminated SiO_2 structures. Simultaneously, the ozone removes residual species such as inhibitors on the NGA and unreacted trilate precursors, thereby preparing the surface for the subsequent ALD batch.

If no etching step is applied, the process proceeds directly to a second inhibitor adsorption cycle. To illustrate the multi-batch simulation mechanism under this condition, the second round of Step A is shown in Fig. 21. The second inhibitor adsorption cycle illustrates how inhibition proceeds in the presence of pre-existing SiO_2 nucleation sites on the NGA surface. Due to the reduced available surface area caused by these existing deposits, the total number of adsorbed inhibitor molecules decreases from 936 in the first cycle to 901 in the second cycle. Moreover, as monodentate molecules occupy less surface area than chelate structures, their relative proportion increases under increasingly constrained spatial conditions. This shift in molecular distribution leads to a sustained rate of contamination and nucleation accumulation on the NGA, as the absolute number of monodentate molecules remains relatively stable during the early batches. Although the total number of inhibitors declines with each cycle, the rising fraction of monodentates compensates by maintaining available nucleation sites. The evolution of monodentate and chelate counts over the first 20 batches is shown in Fig. 22. As a reference, the surface grid status after the second round of Step B and Step C is shown in Figs. 23 and 24.

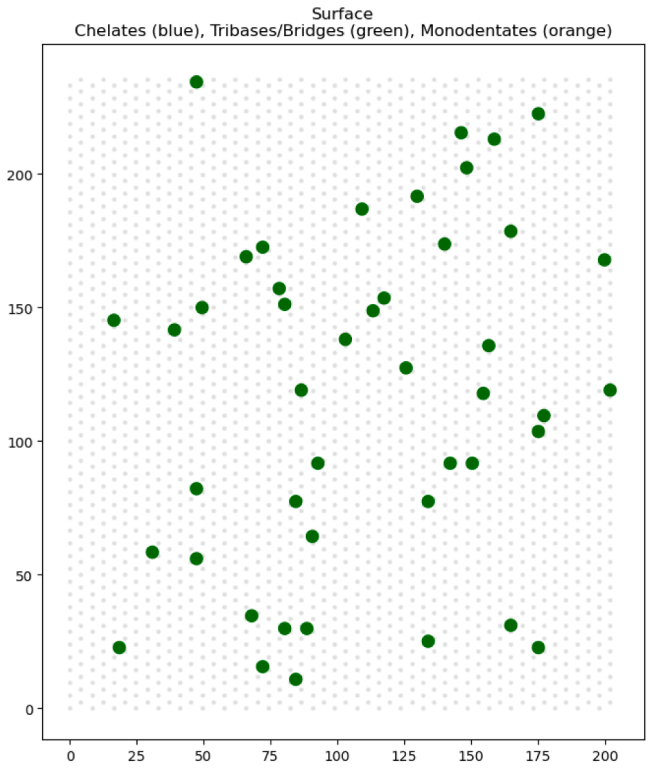


Fig. 20. The ozone would complete the deposition of the deposited silane structure by converting the -H to -OH structure, at the same time strip off all other adsorbed inhibitor molecules and trilate molecules to prepare the surface for next batch.

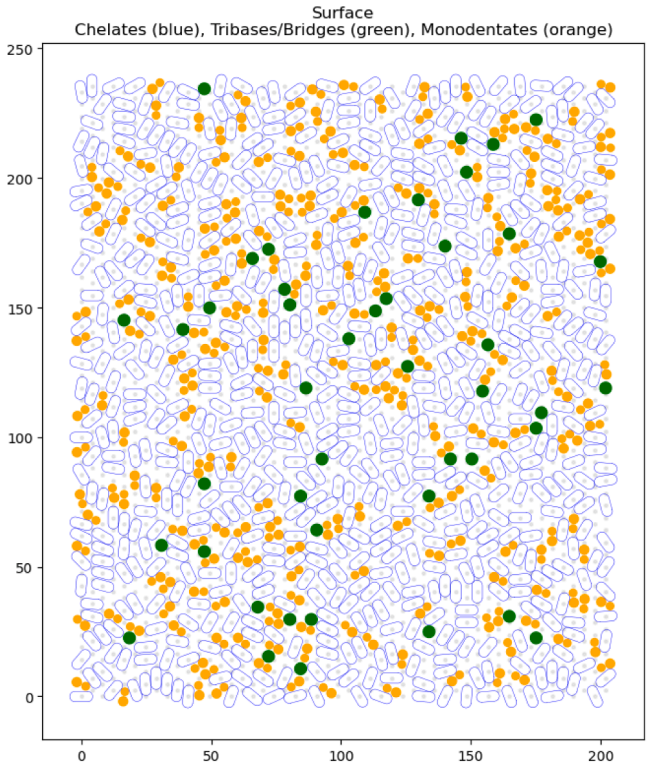


Fig. 21. Second round of inhibitor adsorption on NGA surface. The pre-existing deposited SiO_2 takes some space that reduces the total number of inhibitors on surface and increases the proportion of monodentate inhibitor as the monodentate takes less spaces on the surface.

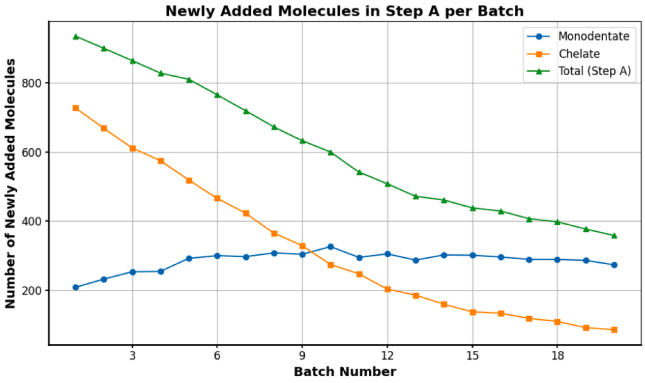


Fig. 22. Number of monodentates and Chelates added in first 20 batches. Monodentate number doesn't change too much because of reduction in spaces. Chelates and total adsorbed molecule number decrease quickly with batches.

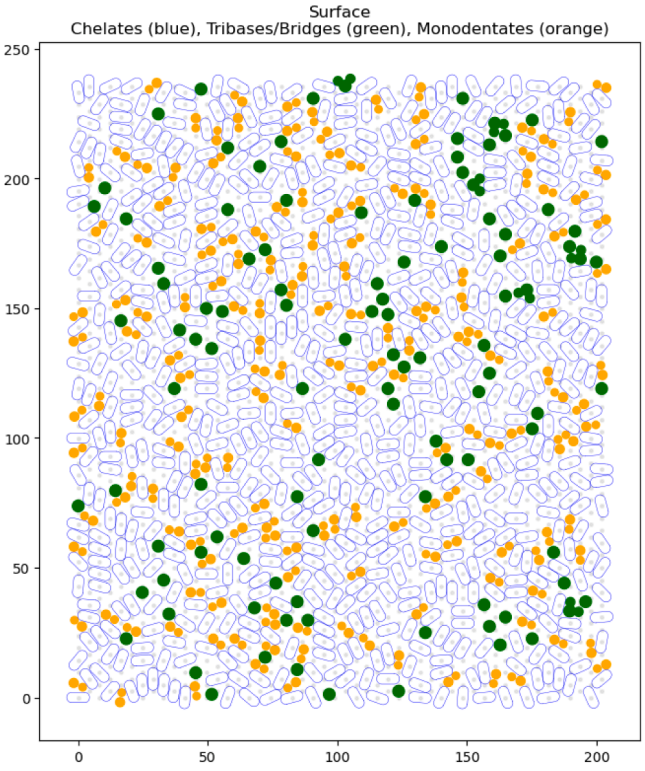


Fig. 23. Second round of Step B, more contamination is formed on NGA surface.

Under the precursor accumulation without etching, the NGA thickness, GA thickness and selectivity for the first 40 batches are shown below in Fig. 25. The multi-batch simulation results demonstrate that, in the absence of an etching step, the selectivity cannot reach the threshold of 0.99 typically required for semiconductor manufacturing. Selectivity declines rapidly over successive batches; by the end of 40 batches, it approaches zero. This loss in selectivity is attributed to the progressive accumulation of unwanted nucleation on the NGA. After an initial nucleation delay of approximately 10–15 batches, the growth rate on the NGA becomes comparable to that on the GA, effectively resulting in the complete transformation of the Al_2O_3 NGA into SiO_2, indistinguishable from the GA. This convergence underscores the necessity of introducing an etching step after each batch to remove unwanted deposition from the NGA while preserving most of the material grown on the GA. The test of different random seed proves the low variance and robustness of the simulation method. The following simulations use random seed 42 as default value.

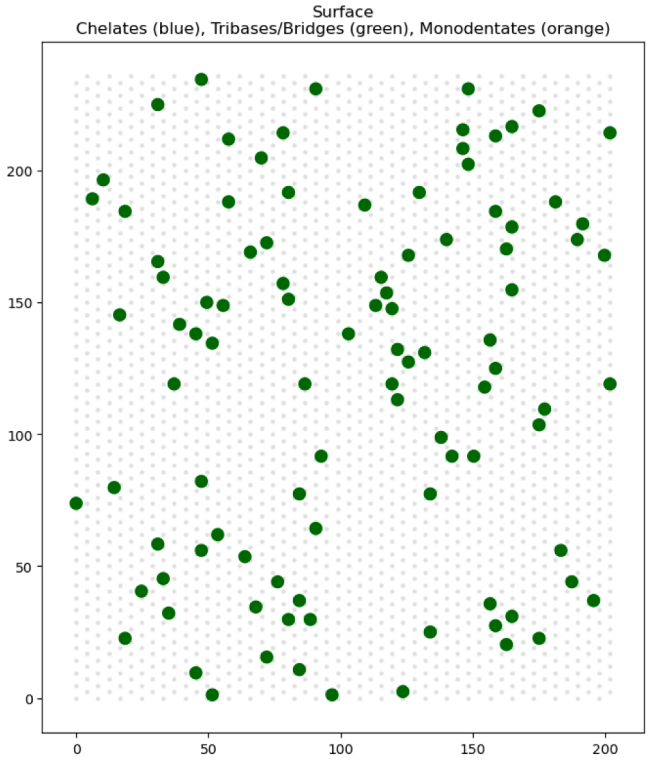


Fig. 24. Second round of Step C, more unwanted nucleation is completed on NGA surface.

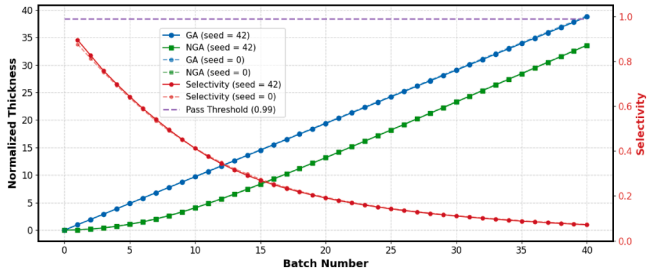


Fig. 25. Pure microscopic simulation multi-batch result without etching. After a nucleation delay of about 10–15 batches, the NGA has identical deposition rate as the GA. The selectivity never achieves pass criterion (0.99) and decreases very quickly. The comparison between random seed of 42 and 0 shows the robustness and low variance of this simulation method under 50×50 scale.

Etching results for durations of 0.6 s, 0.8 s, and 1.0 s under $P_{TMA} =$ 300 Pa, $P_{HF} = 60$ Pa and $T = 523\ K$ are presented in Figs. 26–28, respectively. A comprehensive overview of the GA thickness, NGA thickness, and resulting selectivity as functions of batch number and etch time is shown in Fig. 29. At an etch time of 0.6 s, the selectivity cannot maintain above the pass criterion at any time, which cannot fulfil the requirements of any application. At an etch time of 0.8 s, the selectivity remains above the 0.99 threshold for approximately 5 to 10 batches before declining rapidly, which can fulfil the requirements for very thin film on GA. In contrast, the 1.0 s etch time maintains selectivity above the pass criterion throughout all 40 simulated batches, indicating that this duration is saturated with respect to effective removal of unwanted NGA growth. Increasing the etch time beyond 1.0 s yields diminishing returns, as it further reduces deposition efficiency on the GA without improving selectivity. This saturation point represents a critical trade-off in the deposition-etching sequence. Etching on the growth area is reduced with increasing batch number, which is observed on all three cases, are attributed to the roughness development on surface that decreases the site exposure.

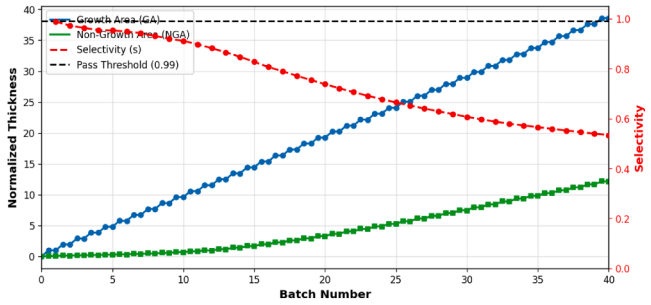


Fig. 26. Under fixed 0.6s etching time, the selectivity cannot be maintained above pass criterion (0.99) at anytime, which cannot fulfil any thickness requirement.

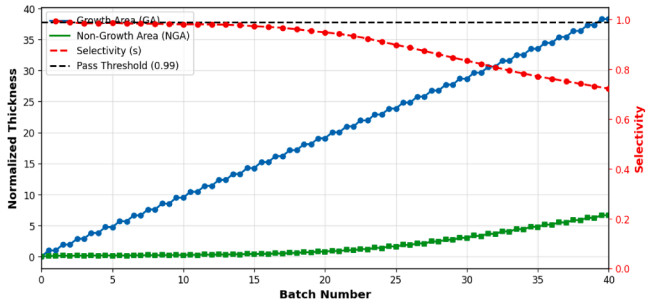


Fig. 27. Under fixed 0.8s etching time, the selectivity can be maintained between 5–10 batches, which is suitable for very thin film requirement on GA, but for thicker film that requires over 10 layers the selectivity cannot fulfil the requirements.

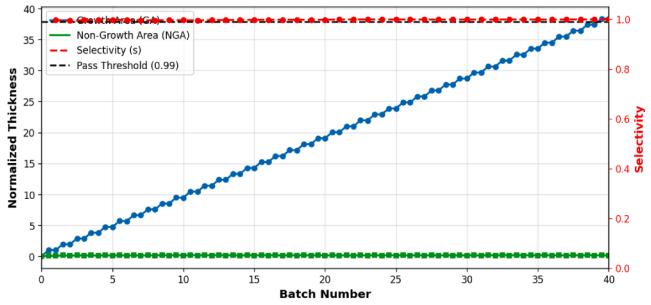


Fig. 28. Under fixed 1.0s etching time, the selectivity is maintained for the whole 40 batches at perfect level, which indicates the etching on NGA is almost saturated.

It should be noted that this critical etch time is not universal and may vary depending on the specific reaction chemistry and operating conditions. The optimal etch duration must therefore be determined on a case-by-case basis. To emphasize the impact placed by exposure mechanisms introduced in this work, the results of 0.6s and 1.0s etching without exposure impact is shown below in Fig. 30. The final average layer number on GA is about 6.4% lower than the case exposure is actively involved (38.37 v.s. 35.90), the final average layer number on NGA is about 17% lower (12.11 v.s. 10.03), which results in apparent increase in selectivity that have the risk of misguiding the implementation.

5.2. *Multiscale simulation*

The pure microscopic simulation framework provides insight into surface dynamics under idealized conditions, assuming that the partial pressures of all precursor species at the wafer surface remain constant throughout the reaction cycle. This simplification neglects the transient nature of gas-phase transport within practical atomic layer deposition (ALD) systems, where precursor delivery from the inlet to the wafer

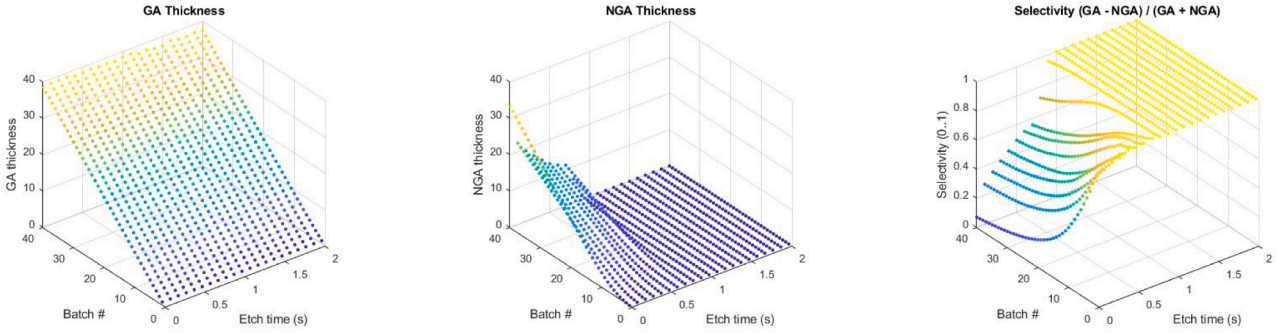


Fig. 29. Holistic view of the GA/NGA thickness and selectivity at different batch numbers and etch times. The Selectivity improves very quickly with increased etch time and saturates at about 1.0s. The GA thickness decreases with higher etch time and shows nucleation delay-like behavior at long etch times because of surface roughness developed by etching.

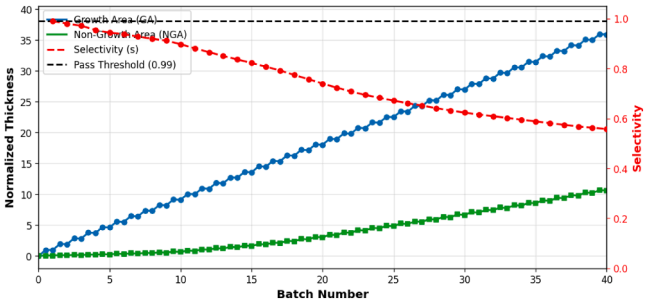


Fig. 30. Without the involvement of exposure mechanism, the etching is more aggressive on both surfaces, which can be inferred by comparing with Fig. 26 whose final average layers are higher on both GA and NGA than in this figure. Without exposure involvement, the selectivity is also apparently higher, which can misguide the implementation.

surface occurs over finite time and is influenced by reactor geometry, flow resistance, and purge cycles. As a result, purely microscopic models cannot capture the dynamic coupling between gas transport and surface reactions observed in real reactors.

Although atomic layer deposition and atomic layer etching are governed by self-limiting surface reactions, self-limitation does not imply instantaneous, spatially uniform, or cost-free reactant delivery. In practical reactor environments, the rate at which self-limited reactions approach saturation is determined by the local precursor partial pressure at the wafer surface, which evolves dynamically due to finite transport, mixing, and purge processes. Industrial ALD and ALE systems operate under strict constraints on throughput, precursor utilization efficiency, and cost, such that exposure times cannot be arbitrarily extended to guarantee saturation. Instead, the objective is to identify the minimum injection and dwell times required to achieve complete surface reaction across the entire wafer. Under these conditions, delays in precursor arrival, pressure non-uniformity, and transport-limited regions can significantly increase the effective saturation time, even for nominally self-limited reactions. The CFD-based gas-flow simulations in this work explicitly resolve the transient and spatially resolved delivery of reactants to the wafer surface, enabling quantitative prediction of the time-dependent surface pressure that governs reaction kinetics. When coupled with the microscopic Monte Carlo model, this multiscale framework reveals how reactor geometry, flow configuration, and operating conditions translate into required exposure times for full coverage, rather than assuming idealized, instantaneous pressure equilibration. As demonstrated in the results that follow, neglecting gas transport leads to systematic underestimation of the required reaction time for time-sensitive steps such as ALE surface modification, whereas incorporating realistic gas delivery dynamics enables accurate correction of exposure times needed to maintain selectivity under industrially relevant conditions.

To address this limitation, the multiscale simulation integrates the real-time pressure and temperature profiles obtained from the macroscopic CFD model as inputs to the microscopic Monte Carlo surface simulation. In turn, the microscopic model provides time-dependent reaction source terms-such as precursor consumption or byproduct generation-that are fed back into the CFD domain. This bidirectional coupling enables the model to capture how surface reactions influence, and are influenced by, the evolving fluid flow field within the reactor. The following results highlight the differences between the idealized microscopic model and the fully coupled multiscale framework and demonstrate how pressure transients and reaction feedback alter the evolution of selectivity and film growth over multiple ALD cycles. The influence of multiscale simulation is not only on Step D, the major etching step, but on all steps. The multiscale simulation inlet mole fraction is set to be:

$$f_{in} = \frac{P_{micro}}{P_{OP}} \tag{15}$$

Where f_{in} is the inlet precursor mole fraction, P_{micro} is the microscopic simulation pressure, and P_{OP} is the operating pressure set in CFD simulation. The P_{OP} applied in this work is 600 Pa, near vacuum, the typical working condition for ALD/ALE processes, and the operating temperature is 523 K. This setup can ensure a steady-state precursor partial pressure in multiscale simulation identical with the pressure applied in microscopic simulation, but with a development and delivery process. The comparison of coverage progress on the first batch between microscopic and multiscale is shown below in Fig. 31 with $f_{in,Hacac} = f_{in,BDEAS} = f_{in,TMA} = 0.5, f_{in,O_3} = f_{in,HF} = 0.1$.

Due to the time update algorithm defined in Eq. (3), the reaction time increment is inversely proportional to the reaction rate. Since the reaction rate itself is a function of local precursor partial pressure in Eq. (4), extremely low pressures at the start of the reaction can result in artificially large time increments, thereby underestimating reaction activity in the initial phase. To address this issue, a threshold pressure is introduced to prevent premature time advancement under negligible reaction conditions. In this work, the threshold is set to 20% of the final target partial pressure. This approach is commonly adopted in multiscale Monte Carlo simulations to suppress the effect of initial pressure near zero (Yun et al., 2022b; Jansen, 2012). The pressure threshold employed in the multiscale simulation is introduced as a numerical stabilization mechanism rather than a physically intrinsic parameter. During the initial precursor delivery transient, the local partial pressure at the wafer surface can be several orders of magnitude lower than its steady-state value. Directly applying the stochastic time-advancement scheme under such conditions can lead to unphysically large time increments that do not correspond to meaningful surface reaction progress. To avoid this artifact, a lower-bound threshold on the local precursor partial pressure

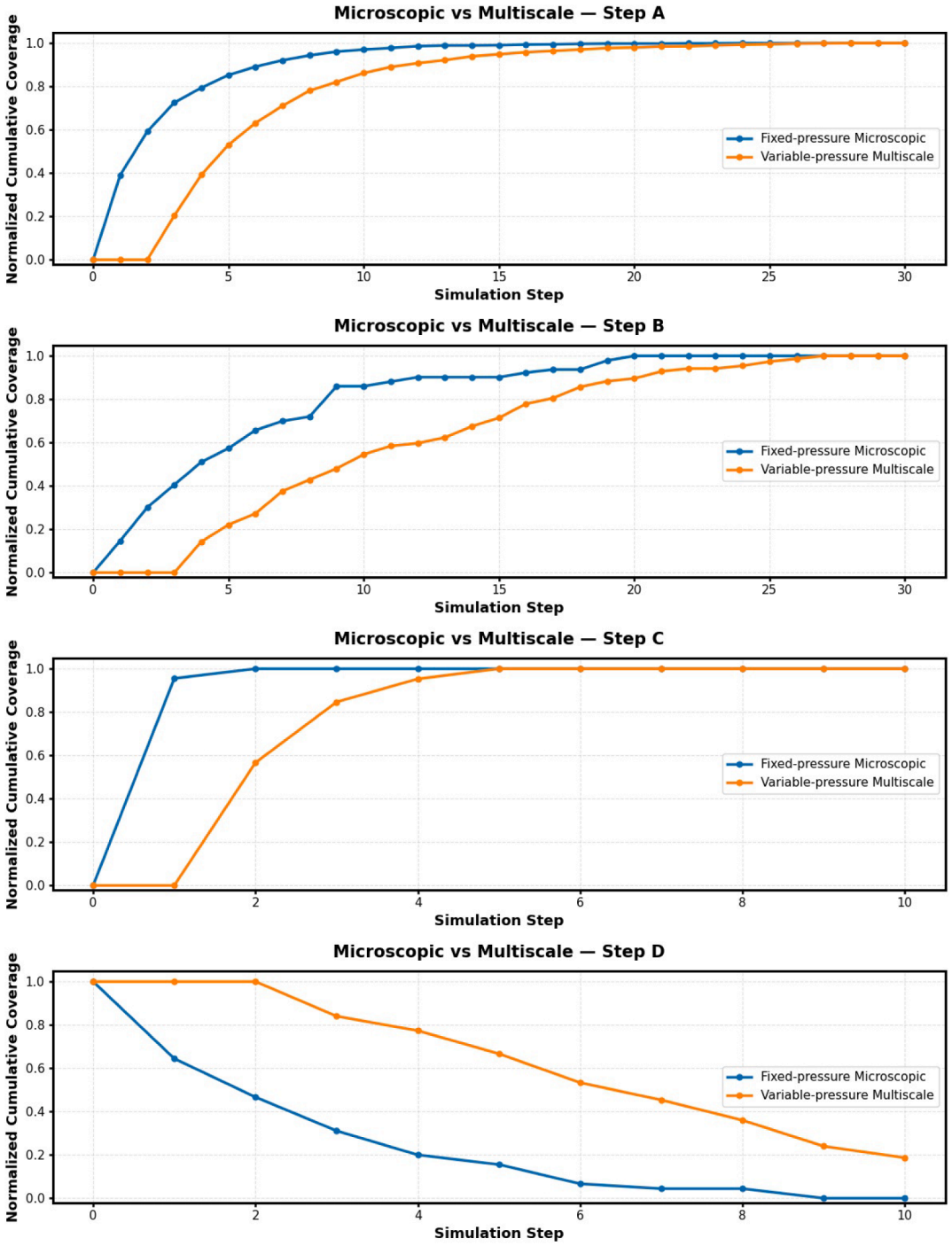


Fig. 31. The comparison on first batch between microscopic and multiscale simulations. Because of the pressure threshold, the reaction starts with a lag and because of lower partial pressure of precursors, the initial slope of all steps is smaller for multiscale simulations.

is imposed, below which surface reactions are temporarily suppressed. In this work, the threshold is set to 20% of the final target partial pressure, consistent with prior validated multiscale ALD (Yun et al., 2022a) and ALE studies where this value yielded reaction completion times and saturation behavior in good agreement with experimental observations. While the precise numerical value of the threshold influences the absolute onset time of surface reactions, it does not alter the qualitative trends or conclusions regarding selectivity evolution or the relative correction required for the etching time. As is common in industrial process modeling, such calibrated numerical parameters are used to ensure numerical robustness and physical consistency when experimental reactor-scale transients are partially characterized or unavailable, and are subsequently refined as additional experimental validation becomes available.

As a consequence of this pressure-based thresholding, the multiscale simulation captures a delayed onset of surface reactions relative to the pure microscopic model, which assumes ideal and instantaneous pressure delivery. This effect manifests as a lag in reaction initiation and a smaller initial slope in surface coverage or thickness evolution plots for the multiscale case because of the lower pressure. To ensure complete conversion in Steps A, B, C, and E under this delayed regime, sufficiently long reaction durations are applied to allow full surface transformation.

However, Step D-the primary etching stage-requires precise control over reaction time to achieve selective removal without over-etching.

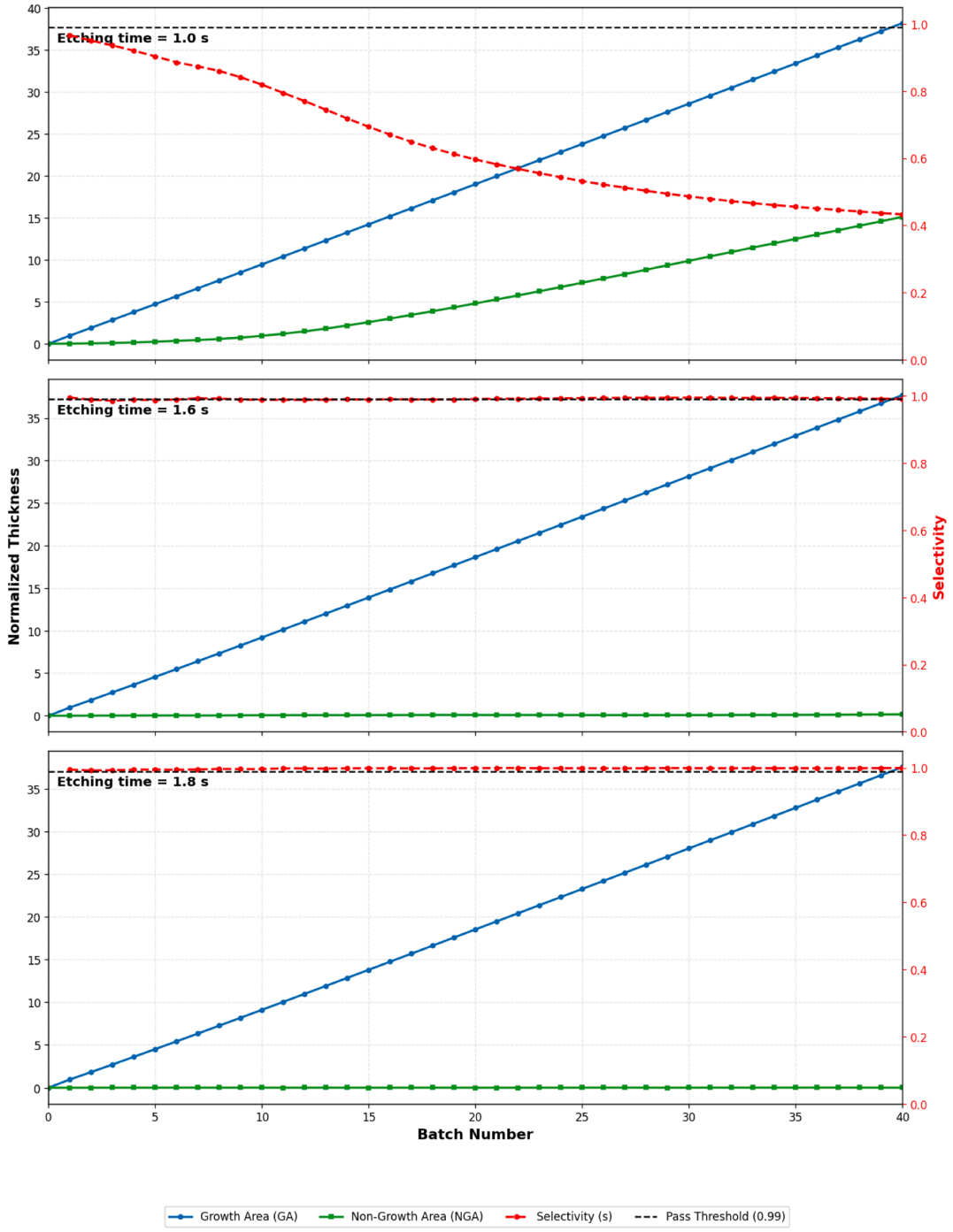


Fig. 32. Selectivity evolution over 40 batches under different etching times predicted by multiscale simulation. At an etching time of 1.0 s, the microscopic-only model predicts near-perfect selectivity across all batches; however, incorporation of precursor transport and delivery effects in the full ALD reactor model leads to a pronounced degradation in selectivity to an unacceptable level. Increasing the etching time to 1.6 s marginally maintains selectivity above the pass criterion, but the response remains unsaturated and exhibits a gradual downward trend, suggesting limited process margin for thicker depositions. Further increasing the etching time to 1.8 s drives the system close to saturation in the multiscale simulation, indicating an effective 0.8 s delay relative to predictions from the purely microscopic model.

For example, at an etch duration of 1.0 s, the microscopic simulation predicts complete removal of unwanted SiO_2 on the NGA after the first batch. In contrast, the multiscale simulation under the same etch time shows remnant material on the NGA surface due to the delayed reaction onset. These remnants can persist and accumulate in subsequent batches, ultimately degrading selectivity. The multiscale simulation result under the 1.0 s etch condition is illustrated in Fig. 32, highlighting the need for time-adjusted etch strategies in multiscale modeling as when considering the precursor delivery in the industrial reactor, the selectivity can just be maintained for 18 batches then decreases significantly with increasing batch, compared to perfect selectivity of 1.0s etching in the microscopic simulation.

To deal with this problem, it is necessary to properly extend the etching time in multiscale simulation. The result of a 1.6s etching mul-

tiscale simulation and the result of 1.8s etching multiscale simulation, also shown in Fig. 32, which shows that extend etching time to 1.6s can mitigate the phenomenon and 1.8s to completely solve the problem by reach saturation. Please note that the 0.8s extension is specific to the reaction and reactor applied in this work; this extension time is highly dependent on the reaction, reactor geometry, and the flow conditions.

The etching-time delay observed between the microscopic and multiscale simulations does not arise from differences in the underlying surface reaction mechanisms, which are identical in both models, but rather from the inclusion of precursor pressure delivery and transport dynamics in the multiscale framework. Finite pressure buildup and decay reduce the effective etching reaction rate, resulting in a small but systematic decrease in etching efficiency per batch. While this reduction has a small impact during the initial batches, the deposition process is inherently history dependent that any residual growth remaining on NGA persists and serves as an active surface for subsequent cycles. As deposition proceeds, these early-stage residuals are progressively amplified, leading to an accelerating accumulation of NGA thickness across batches. Consequently, even a minor difference in etching effectiveness during the early stages that is introduced by transport-limited precursor delivery can translate into a substantial divergence in selectivity after many batches. This cumulative amplification explains why the multiscale simulation requires a longer etching time to achieve the same long-term selectivity predicted by the microscopic model.

The approximately 0.8 s delay observed between precursor injection and effective surface reaction onset in the multiscale simulations is primarily governed by reactor geometry and gas-distribution design rather than by surface reaction kinetics. In the NIST reactor employed in this study, precursor delivery follows an inlet-to-wafer pathway that includes an expansion cone structure designed to promote lateral diffusion and improve spatial uniformity of precursor partial pressure across the wafer surface. While this diffuser geometry enhances uniformity, it also increases the effective transport distance and mixing volume between the inlet and the wafer, leading to a finite pressure buildup time at the surface. As a result, the observed delay is dominated by geometric characteristics such as inlet-wafer spacing, diffuser volume, and flow expansion design. Although precursor flow rate and total injected volume influence the overall gas residence time, ALD and ALE processes are self-limiting and require only a small fraction of the delivered precursor to react; increasing flow rate to reduce delivery delay would therefore lead to inefficient precursor utilization and is generally undesirable in industrial practice. More compact or spatially efficient gas-distribution schemes, such as showerhead-type injectors, can reduce delivery delay while maintaining uniformity, as demonstrated in prior reactor design studies. Consequently, the magnitude of the observed delay is reactor-specific and should be interpreted as a geometry-dependent transport effect rather than a universal kinetic parameter.

6. Conclusion

This work addresses the critical challenge of alignment errors in advanced semiconductor fabrication by modeling area-selective atomic layer deposition (ASALD), a bottom-up, self-aligned method capable of eliminating edge placement error. To capture the imperfect nature of current ASALD chemistry, a microscopic Monte Carlo-based collision model was developed to simulate the adsorption of small-molecule inhibitors (SMIs), precursor nucleation, and surface oxidation reactions on SiO_2 and Al_2O_3 substrates. The model quantitatively reproduces known experimental behaviors, including ~22% monodentate inhibitor coverage and ~6.7% unwanted precursor nucleation on NGA, both closely matching reported values (Merkx et al., 2020a).

To recover selectivity in the presence of imperfect inhibition, the ASALD cycle was extended to include an exposure-dependent thermal atomic layer etching (ALE) step using TMA and HF. Parametric studies revealed that etch durations of 0.6 s and 0.8 s could preserve selectivity above 0.99 for 10 and 20 deposition cycles, respectively, while 1.0 s etching maintained perfect selectivity across 40 batches. Beyond this point, additional etching yielded diminishing returns due to increased material loss from the GA.

To simulate industrially relevant conditions, the microscopic model was coupled to a CFD-based reactor model developed and validated with NIST, forming a multiscale digital twin of the ASALD process. The multiscale framework revealed that delayed precursor delivery in the reactor leads to slower initial reaction rates, requiring adjusted etch durations. For example, a 1.0 s etch time in multiscale simulation has much worse performance and becomes unacceptable, whereas extending the etch to 1.6 s restored selectivity across all 40 cycles, while extending to 1.8 s reaches saturation.

These findings demonstrate that integrating surface-level reaction dynamics with reactor-scale transport effects is essential for accurate prediction and optimization of ASALD processes. The modeling framework presented here can be readily adapted to other inhibitor chemistries, precursor systems, and reactor configurations, and serves as a foundation for future work in feedback control, surface defect engineering, and AI-guided reactor optimization.

CRediT authorship contribution statement

Feiyang Ou: Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization; **Abdulrahman Alghamdi:** Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization; **Chun-Pei Lin:** Writing – original draft, Methodology, Investigation, Conceptualization; **Gerassimos Orkoulas:** Writing – review & editing, Methodology, Investigation, Conceptualization; **Panagiotis D. Christofides:** Conceptualization, Investigation, Methodology, Writing – review & editing.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Financial support from the National Science Foundation and the Department of Energy is gratefully acknowledged. This work used computational and storage services associated with the Hoffman2 Shared Cluster provided by the UCLA Office of Advanced Research Computing's Research Technology Group.

References

Chen, R., Kim, H., McIntyre, P.C., Porter, D.W., Bent, S.F., 2005. Achieving area-selective atomic layer deposition on patterned substrates by selective surface modification. Appl. Phys. Lett. 86, 191910.

Ding, Y., Zhang, Y., Ren, Y.M., Orkoulas, G., Christofides, P.D., 2019. Machine learning-based modeling and operation for ALD of sio2 thin-films using data from a multiscale CFD simulation. Chem. Eng. Res. Des. 151, 131–145.

DuMont, J.W., Marquardt, A.E., Cano, A.M., George, S.M., 2017. Thermal atomic layer etching of sio2 by a "conversion-etch" mechanism using sequential reactions of trimethylaluminum and hydrogen fluoride. ACS Appl. Mater. Interf. 9, 10296–10307.

Fang, M., Ho, J.C., 2015. Area-selective atomic layer deposition: conformal coating, subnanometer thickness control, and smart positioning. ACS Nano 9, 8651–8654.

Folkendt, M.M., Weiss-Lopez, B.E., Chauvel, Jr, J.P., True, N.S., 1985. Gas-phase proton NMR studies of keto-enol tautomerism of acetylacetone, methyl acetoacetate, and ethyl acetoacetate. J. Phys. Chem. 89, 3347–3352.

George, S.M., 2010. Atomic layer deposition: an overview. Chem. Rev. 110, 111–131.

Haider, A., Yilmaz, M., Deminskyi, P., Eren, H., Biyikli, N., 2016. Nanoscale selective area atomic layer deposition of TiO_2 using e-beam patterned polymers. RSC Adv. 6, 106109–106119.

Huang, J., Lee, M., Lucero, A., Cheng, L., Kim, J., 2014. Area-selective ALD of TiO_2 nanolines with electron-beam lithography. J. Phys. Chem. C 118, 23306–23312.

Jansen, A.P.J. (Ed.), 2012. An Introduction to Kinetic Monte Carlo Simulations of Surface Reactions. Vol. 1. Academic Press.
Kimes, W.A., Moore, E.F., Maslar, J.E., 2012. Design and operation of an optically-accessible modular reactor for diagnostics of thermal thin film deposition processes. Rev. Sci. Instrum. 83.
Lee, G., Lee, B., Kim, J., Cho, K., 2009. Ozone adsorption on graphene: ab initio study and experimental validation. J. Phys. Chem. C 113, 14225-14229.
Leskelä, M., Ritala, M., 2002. Atomic layer deposition (ALD): from precursors to thin film structures. Thin Solid Films 409, 138–146.
Loubet, N., Hook, T., Montanini, P., Yeung, C.-W., Kanakasabapathy, S., Guillom, M., Yamashita, T., Zhang, J., Miao, X., Wang, J., et al., 2017. Stacked nanosheet gate-all-around transistor to enable scaling beyond finFET. In: 2017 Symposium on VLSI Technology. IEEE, pp. T230–T231.
Mackus, A. J.M., Merkx, M. J.M., Kessels, W. M.M., 2019. From the bottom-up: toward area-selective atomic layer deposition with high selectivity. Chem. Mater. 31, 2–12.
Mameli, A., Merkx, M. J.M., Karasulu, B., Roozeboom, F., Kessels, W. E. M.M., Mackus, A. J.M., 2017. Area-selective atomic layer deposition of SiO_2 using acetylacetone as a chemoselective inhibitor in an ABC-type cycle. ACS Nano 11, 9303–9311.
Martin, M., Cunge, G., 2008. Surface roughness generated by plasma etching processes of silicon. J. Vacuum Sci. Technol. B Microelectron. Nanometer Struct. Process. Measur. Phenom. 26, 1281–1288.
Maslar, J.E., Kalanyan, B., 2025. Visualizing molybdenum pentachloride flow during vapor deposition processes using absorption imaging. Appl. Spectrosc. 79, 1487–1496.
Merkx, M. J.M., Angelidis, A., Mameli, A., Li, J., Lemaire, P.C., Sharma, K., Hausmann, D.M., Kessels, W. M.M., Sandoval, T.E., Mackus, A. J.M., 2022. Relation between reactive surface sites and precursor choice for area-selective atomic layer deposition using small molecule inhibitors. J. Phys. Chem. C 126, 4845–4853.
Merkx, M. J.M., Sandoval, T.E., Hausmann, D.M., Kessels, W. M.M., Mackus, A. J.M., 2020a. Mechanism of precursor blocking by acetylacetone inhibitor molecules during area-selective atomic layer deposition of SiO_2. Chem. Mater. 32, 3335–3345.
Merkx, M. J.M., Vlaanderen, S., Faraz, T., Verheijen, M.A., Kessels, W. M.M., Mackus, A. J.M., 2020b. Area-selective atomic layer deposition of TiN using aromatic inhibitor molecules for metal/dielectric selectivity. Chem. Mater. 32, 7788–7795.
Minaye Hashemi, F.S., Prasittichai, C., Bent, S.F., 2015. Self-correcting process for high quality patterning by atomic layer deposition. ACS Nano 9, 8710–8717.
Mokhtarzadeh, M., Carulla, M., Kozak, R., David, C., 2022. Optimization of etching processes for the fabrication of smooth silicon carbide membranes for applications in quantum technology. Micro Nano Eng. 16, 100155.
Parsons, G.N., Clark, R.D., 2020. Area-selective deposition: fundamentals, applications, and future outlook. Chem. Mater. 32, 4920–4953.
Pieck, F., Tonner-Zech, R., 2025. Computational ab initio approaches for area-selective atomic layer deposition: methods, status, and perspectives. Chem. Mater. 37, 2979–3021.
Rahman, R., Mattson, E.C., Klesko, J.P., Dangerfield, A., Rivillon-Amy, S., Smith, D.C., Hausmann, D., Chabal, Y.J., 2018. Thermal atomic layer etching of silica and alumina thin films using trimethylaluminum with hydrogen fluoride or fluoroform. ACS Appl. Mater. Interf. 10, 31784–31794.
Roh, H., Kim, H.-L., Khumaini, K., Son, H., Shin, D., Lee, W.-J., 2022. Effect of deposition temperature and surface reactions in atomic layer deposition of silicon oxide using bis(diethylamino)silane and ozone. Appl. Surf. Sci. 571, 151231.
Schwille, M.C., Schössler, T., Schön, F., Oettel, M., Bartha, J.W., 2017. Temperature dependence of the sticking coefficients of bis-diethyl aminosilane and trimethylaluminum in atomic layer deposition. J. Vacuum Sci. Technol. A 35, 01B119.
Sinha, A., Hess, D.W., Henderson, C.L., 2006. Area selective atomic layer deposition of titanium dioxide: effect of precursor chemistry. J. Vacuum Sci. Technol. B Microelectron. Nanometer Struct. Process. Measur. Phenom. 24, 2523–2532.
Vallat, R., Gassilloud, R., Eychenne, B., Vallée, C., 2017. Selective deposition of ta2o5 by adding plasma etching super-cycles in plasma enhanced atomic layer deposition steps. J. Vacuum Sci. Technol. A 35.
Vos, M. F.J., Chopra, S.N., Verheijen, M.A., Ekerdt, J.G., Agarwal, S., Kessels, W. M.M., Mackus, A. J.M., 2019. Area-selective deposition of ruthenium by combining atomic layer deposition and selective etching. Chem. Mater. 31, 3878–3882.
Yun, S., Ou, F., Wang, H., Tom, M., Orkoulas, G., Christofides, P.D., 2022a. Atomistic-mesoscopic modeling of area-selective thermal atomic layer deposition. Chem. Eng. Res. Des. 188, 271–286.
Yun, S., Tom, M., Ou, F., Orkoulas, G., Christofides, P.D., 2022b. Multiscale computational fluid dynamics modeling of thermal atomic layer etching: application to chamber configuration design. Comput. Chem. Eng. 161, 107757.