

Abschlussbericht

zum Teilvorhaben:	Materialsimulation
im Verbundprojekt:	Halbleiter-Quantenprozessor mit shuttlingbasierter skalierbarer Architektur
Akronym:	QUASAR
Zuwendungsempfänger:	HQS Quantum Simulations GmbH
Förderkennzeichen:	13N15653
Laufzeit des Teilvorhabens:	01.02.2021 – 31.01.2025

Kurzbericht

Wir, HQS Quantum Simulations, haben uns im Rahmen des Arbeitspaket 2, „Simulation und Modellierung“, am QUASAR Projekt beteiligt. Das Arbeitspaket 2 verfolgte das Ziel, zentrale Materialparameter von Silizium-Germanium-Quantentöpfen – vor allem das sogenannte Valley-Splitting – numerisch vorherzusagen. Diese Quantentöpfe bildeten das Fundament der im Projekt vorgesehenen Shuttling-Architektur für Spin-Qubits. Im Laufe der vier Projektjahre wurde ein durchgängiger Simulations-Workflow entwickelt, der von ersten DFT-Machbarkeitsstudien bis hin zu GPU-beschleunigten, Cloud-basierten Großrechnungen reichte. Die nachstehenden Abschnitte geben einen chronologisch gegliederten Überblick über die erzielten Ergebnisse.

Grundlagen und DFT-Machbarkeitsstudie

Im ersten Jahr stand die Charakterisierung von reinem und mechanisch verspanntem Silizium mit Hilfe Dichtefunktionaltheorie (DFT)-basierter Elektronenstrukturechnungen, ausgeführt beispielsweise mit dem [GPAW-Code](#) [1], im Mittelpunkt. Für die Berechnung von mechanischen Eigenschaften, wie der elastischen Konstanten, wurden zwei ergänzende Ansätze untersucht: einerseits die Ableitung des Spannungstensors aus Energie-Deformations-Kurven, andererseits das direkte (DFT-)Berechnung des Spannungstensors. Die Kombination beider Methoden lieferte robuste Vorhersagen, die mit Literaturwerten übereinstimmten. Parallel dazu wurden aus den DFT-Ergebnissen durch [Wannierisierung](#) [2] erste Tight-Binding-Modelle abgeleitet. Diese erwiesen sich jedoch als zu komplex, um realistische Si/Ge-Heterostrukturen in experimentellen Größenskalen effizient zu beschreiben. Der Austausch mit unseren Konsortialpartnern machte deutlich, dass die Nutzung eines etablierten Slater-Koster-Modells für Si/Ge-Legierungen deutlich erfolgversprechender war. Entsprechend wurde Ende 2021 mit der Implementierung eines [10-Band Slater-Koster-Tight-Binding-Modells](#) [3] begonnen.

Aufbau des Tight-Binding-Workflows

Das zweite Jahr war geprägt von der Fertigstellung und Validierung des 10-Band Slater-Koster-Modells für beliebig verspannte Si/Ge-Heterostrukturen. Darauf aufbauend entstand ein vollständiger Workflow: Eine auf der Zinkblende-Struktur basierende Einheitszelle wurde zunächst mechanisch verspannt, anschließend in allen drei Raumrichtungen zur Superzelle erweitert und schließlich mit stochastisch generierten Legierungsrealisationen gefüllt. Zusätzliche wurden laterale Lokalisierungspotentiale sowie elektrische Felder eingebettet, bevor ein, als Expertise ins Projekt

eingebachter, HQS-entwickelter Spektralfilter herangezogen wurde um die im Quantentopf lokalisierten Quantenpunktzustände zu extrahieren und das Valley-Splitting, die energetische Aufspaltung der untersten beiden Quantenpunktzustände, zu berechnen. In einer ersten Pilotstudie wurden zwei Legierungsprofile verglichen – ein abruptes Stufenprofil („well“) und ein linear abfallendes Keilprofil („wedge“). Die Simulationen zeigten, dass das Keilprofil das mittlere Valley-Splitting vergrößerte, allerdings auch eine stärkere Streuung induzierte. Darüber hinaus wurde festgestellt, dass ein externes elektrisches Feld die statistische Streuung unabhängig von seiner Richtung verkleinerte und damit die Qubit-Stabilität erhöhte. Noch offen blieb zu diesem Zeitpunkt die Frage bezüglich der laterale Konvergenz der Simulation, da der Quantenpunkt in der Modellierung mit fünf Nanometern Durchmesser erheblich kleiner war als im realen Chip.

Modellverfeinerung und GPU-Beschleunigung

Zu Beginn des dritten Jahres wurden im Tight-Binding-Modell zusätzliche Korrekturterme für einzelne Atome und abstandsabhängige Kopplungen zwischen den Atomen implementiert und anhand von Referenzdaten überprüft. Die Resultate legten nahe, dass diese Verfeinerungen das Valley-Splitting weniger beeinflussten als die intrinsische Grenzflächenrauigkeit, d.h. die Tatsache, dass Grenzflächen stochastisch erzeugt werden. Eine wesentliche Entscheidung betraf die Behandlung der Atompositionen: Eine vollständige strukturelle Relaxation der Millionen Atome umfassenden Heterostrukturen hätte den Rechenaufwand wesentlich gesteigert und eine GPU-Optimierung nahezu ausgeschlossen. Daher blieb es bei der Annahme paarweiser Gleichgewichtsabstände, was zugleich eine speicher- und recheneffiziente Gruppierung der Kopplungsmatrizen auf der GPU erlaubte. Die Portierung des Kern-Workflows auf Grafikprozessoren ermöglichte erstmals Simulationen von Strukturen mit Ausmaßen von $50 \times 50 \times 40 \text{ nm}^3$, also etwa fünf Millionen Atomen, in etwa einer Stunde. Vergleiche mit halb-analytischen Effektive-Masse-Modellen der Universität Konstanz bestätigten die wesentlichen Trends, beispielsweise den Einfluss von Quantentopfbreite, Quantenpunktdurchmesser, angelegten Feldern und Grenzschichtbreite.

Statistische Studien und Cloud-Skalierung

Beginnend mit dem fünften Konsortialtreffen in Freiburg konnten Simulationen mit den realen Quantenpunkt-Abmessungen von rund $80 \times 80 \times 10 \text{ nm}^3$ präsentiert werden. Die Diskussion mit den Partnern zeigte jedoch, dass Einzelsimulationen nicht genügten, um belastbare Aussagen für die Chip-Fabrikation zu treffen; erforderlich waren statistische Auswertungen über viele Tausend Quantenpunkte. Nach Genehmigung einer Mittelumwidmung durch den VDI konnten deshalb umfangreiche GPU-Ressourcen in Cloud-Rechenzentren (AWS) erschlossen werden. Dank der parallelen Verfügbarkeit von 100-1000 GPUs ließen sich in der Zeit, die früher eine Einzelsimulation benötigte, bis zu tausend Quantenpunkte simultan berechnen. Als Eingangsdaten dienten experimentell bestimmte Legierungsprofile entlang der Wachstumsrichtung; dabei wurden zwei, nicht-destruktiv, im Rahmen des QUASAR Projekts gemessene, Profile und ein mittels [Atom-Probe-Tomography \(APT\) gewonnenes Profil der Universität Regensburg](#) [4] ausgewählt. Seit August 2024 lief eine schrittweise Hochskalierung hin zu mindestens zehntausend Valley-Splitting-Datenpunkten als Zielvorgabe für die Simulation. Erste Ergebnisse mit verringertem Quantenpunktdurchmesser erlaubten es, die räumliche Auflösung der Scans zu optimieren und Abweichungen von vereinfachten Effektive-Masse-Schätzungen zu quantifizieren, die anschließend gemeinsam mit den Konsortialpartnern ausgewertet wurden. Zum Ende des Projektes wurden die Ergebnisse (insgesamt etwa 100 Tausend Datenpunkte) mit unseren Projektpartnern geteilt.

Detaillierte Darstellung

Slater-Koster-Tight-Binding-Modell für Si/Ge-Random-Legierungen

Einordnung und Zielsetzung. Für die numerische Beschreibung der elektronischen Struktur in den im QUASAR-Projekt untersuchten Silizium/Germanium-Quantentöpfen wurde ein 10-Band-Slater-Koster-Tight-Binding-Modell ($sp^3d^5s^*$) implementiert. Das Modell von [Niquet et al. \(2009\)](#) [3] gehört bis heute zum Stand der Technik, weil es mit wenigen Parametern gleichzeitig die Valenzbandstruktur, die tief liegenden Leitungsbänder sowie die Deformation unter beliebiger mechanischer Verspannung quantitativ erfasst.

Orbitale Basis und Parametrisierung. Jedes Atom wird durch ein s-, drei p-, fünf d- und ein weiteres „angeregtes“ s-Orbital (s^*) beschrieben, woraus sich ein 10×10 -Hamiltonblock pro Gitterplatz ergibt. Die Matrixelemente bestehen aus (i) Atomorbital-Energien und (ii) zweizentrierten Kopplungsintegralen σ , π , δ , die – analog zu den Slater-Koster-Tabellen – allein vom Bindewinkel abhängen. Alle Parameter wurden in der Referenzpublikation so angepasst, dass die ab-initio-basierten Bandlücken, effektiven Massen und Deformationspotentiale von reinem Si und Ge sowie deren Mischung $Si_{1-x}Ge_x$ reproduziert werden.

Band-Offset als zentrales Eingangsparameter. Die Tiefen der Quantentöpfe werden über den Valenz- und Leitungsband-Versatz (band offset) zwischen Silizium und Germanium definiert. Das Tight-Binding-Modell liefert die Versätze als Funktion von Ge-Gehalt und Verspannung direkt integriert in den Hamiltonoperator. Für typisches Wachstumslayout (25 - 35 % Germanium außerhalb des Quantentopfs) ergibt sich aus einem band offset von 0.68 eV eine Potentialtiefe des Quantentopfs von etwa 0.2 eV.

Berücksichtigung von Verspannung. Für den Betrieb der Spin-Qubits sind die Heterostrukturen biaxial (senkrecht zur Wachstumsrichtung) auf die Gleichgewichtsstruktur der vorgegebenen Legierung (25 - 35 % Germanium) gedehnt. Das Modell enthält lineare Deformationsparameter, welche die Verschiebung aller Bandkanten bei hydrostatischer, uniaxialer oder Scherverspannung erfassen, ohne die Kristallsymmetrie zu brechen. Dadurch kann derselbe Parametersatz sowohl für entspannte als auch für stark verspannte Schichten verwendet werden.

„Random Alloy“ – Begriffsklärung. Der Quantentopf ist in einer zufällige Si/Ge Legierung („random alloy“) realisiert: An jeder Gitterposition sitzt mit Wahrscheinlichkeit $p(\text{Ge}) = x$ ein Germanium- und mit $p(\text{Si}) = 1 - x$ ein Silizium-Atom. Die Germaniumkonzentration x variiert entlang der Wachstumsrichtung und definiert somit den Quantentopf. Im Simulationsworkflow werden zwei Betriebsarten unterschieden:

1. Virtuelle Kristallnäherung (VCA) – der Hamiltonoperator wird über alle möglichen Realisierungen gemittelt. Der resultierende Hamiltonoperator liefert das sogenannte „deterministisches“ Valley-Splitting, d. h. den Mittelwert der Aufspaltung ohne statistische Streuung.
2. Realistische Probe (digital twin) – für jedes Realisierung des Legierungsprofils wird ein Hamiltonoperator generiert, wodurch die stochastische Inhomogenität explizit in das Modell eingeht. Aus den berechneten Valley-Splittings für mehrere Realisierungen des Legierungsprofils wird das aus der Chip-Fabrikation zu erwartende statistische Streuung des Valley-Splittings vorhergesagt.

Näherungen in der Strukturbeschreibung. Um die Wechselwirkung von bis zu fünf Millionen Atomen auf GPUs effizient behandeln zu können, wurden sämtliche Bindungen auf ihren jeweiligen Gleichgewichtsabstand fixiert. Eine vollständige Relaxation der Atompositionen verschiedene Bindungslängen erzwungen und damit eine effiziente Berechnung des Hamiltonoperators verhindert. Der gewählte Kompromiss ist geometrisch nicht exakt, führt jedoch zu deutlich kleineren Fehlern als wenn die Atome auf den „idealisierten“ Positionen der Basis der Zinkblende-Struktur festgehalten werden.

Einbettung in den QUASAR-Workflow:

- **Superzellenaufbau:** Aus einer zinkblenden-basierten kubischen Einheitszelle (8 Atome) wurden $D \times D \times B$ Superzellen konstruiert. Wobei D den Durchmesser des Quantenpunkts umfasst und B etwa die (doppelte) Breite des Quantentopfs ist und entlang der Wachstumsrichtung ausgerichtet ist.
- **Legierungsprofil:** Vorgeschlagene oder experimentell gemessene Germaniumprofile wurden herangezogen um das Legierungsprofil entlang der Wachstumsrichtung implementiert, d.h., Germaniumatome anhand des Legierungsprofils zufällig ausgewählt.
- **Zusatzpotentiale:** Laterale (harmonische) Lokalisierungspotentiale (confinement) modellieren den elektrostatischen Quantenpunkt; des weiteren konnte ein konstantes elektrisches (und magnetisches) externe Feld angelegt werden.
- **Spektralfilter:** Ein in das Projekt als externe Expertise eingebrachter, von HQS entwickelter, Spektralfilter, isoliert die relevanten Quantenpunktzustände und liefert das Valley-Splitting als Eigenwertdifferenz. Ein wesentlicher Punkt für dessen Effizienz ist die effiziente Implementierung des Slater-Koster Modells.
- **GPU-Beschleunigung & Cloud:** Die effiziente Implementierung des Slater-Koster Modells für GPUs zusammen mit der Verwaltung des Simulationsworkflows mittels HQS Tasks, dem HQS-internen Werkzeug zur Organisation von wissenschaftlichen Simulationen, hat es uns ermöglicht die Simulation für Fabrikations-relevante Größen durchzuführen und gleichzeitig die benötigte Statistik für das Valley-Splitting zu sammeln.

Grenzen und Ausblick. Trotz der Transferierbarkeit des Modells wurden im Projekt Deformationseffekte nicht systematisch untersucht. Künftige Simulationen könnten hier größere Parameterbereiche von relevanten Deformationen und/oder Verspannungen berücksichtigen. Dies erfordert keine weitere Implementierungsarbeit, sondern lediglich die benötigten Hardwareressourcen (insbesondere GPUs). Die oben beschriebenen Näherungen in der Strukturbeschreibung, könnten verbessert werden indem die Simulation an eine Moleküldynamik (MD) Simulation des Silizium/Germanium-Quantentopfs geknüpft wird. Neben der Definition einer geeigneten Schnittstelle zwischen der Elektronenstruktursimulation und der MD Simulation, würde dies Anpassungen in der Implementierung der Elektronenstruktursimulation erfordern. Unsere bisherigen Studie haben keine Indizien dafür geliefert, dass dies eine genauere Spezifikation der mikroskopischen Eigenschaften des Spin Qubits liefern würde.

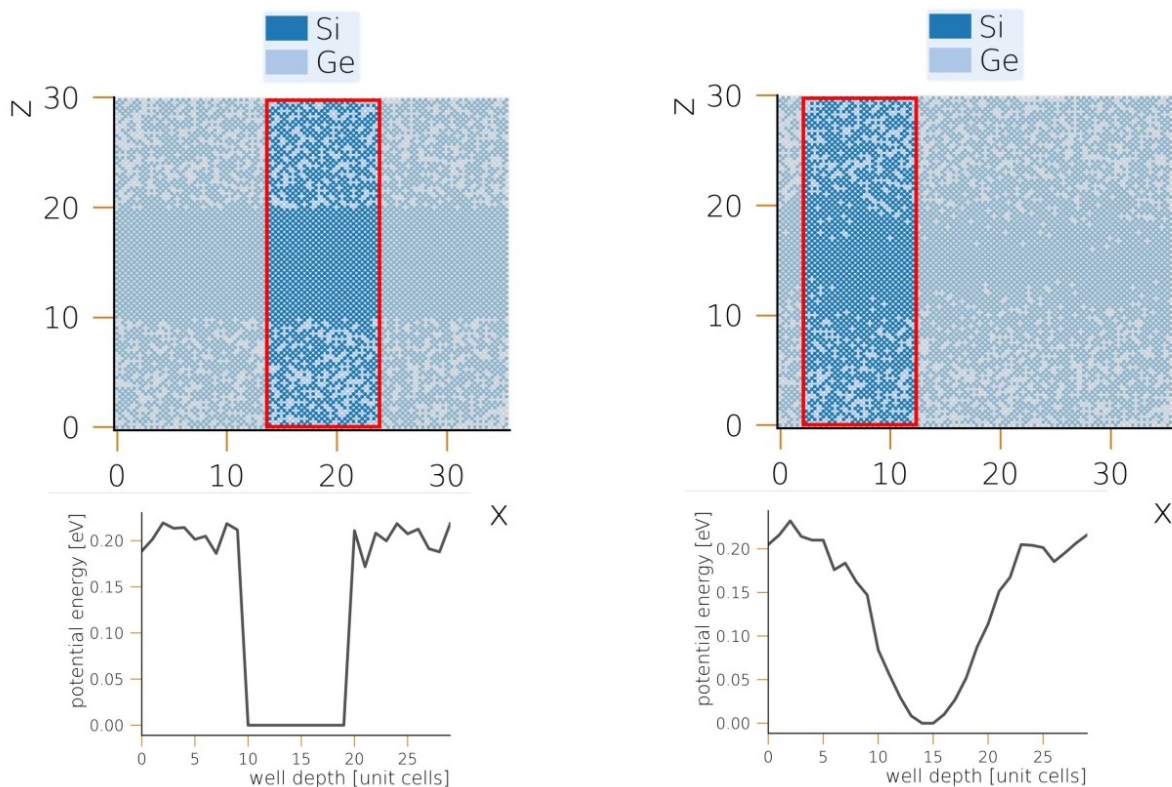
Zusammenfassend ermöglichte das 10-Band-Slater-Koster-Modell einen effizienten Workflow von der Materialparametrisierung (dem Legierungsprofil) bis zur industriell relevanten Statistik des Valley-Splittings. Es bietet damit das methodische Rückgrat aller von HQS im Arbeitspaket 2 durchgeführten Simulationen und lieferte eine wesentliche Kennzahlen, das Valley-Splitting, anhand derer die Chip-Architektur für Spin Qubits optimiert werden kann.

Pilotstudie – Vergleich der Legierungsprofile „Well“ vs. „Wedge“

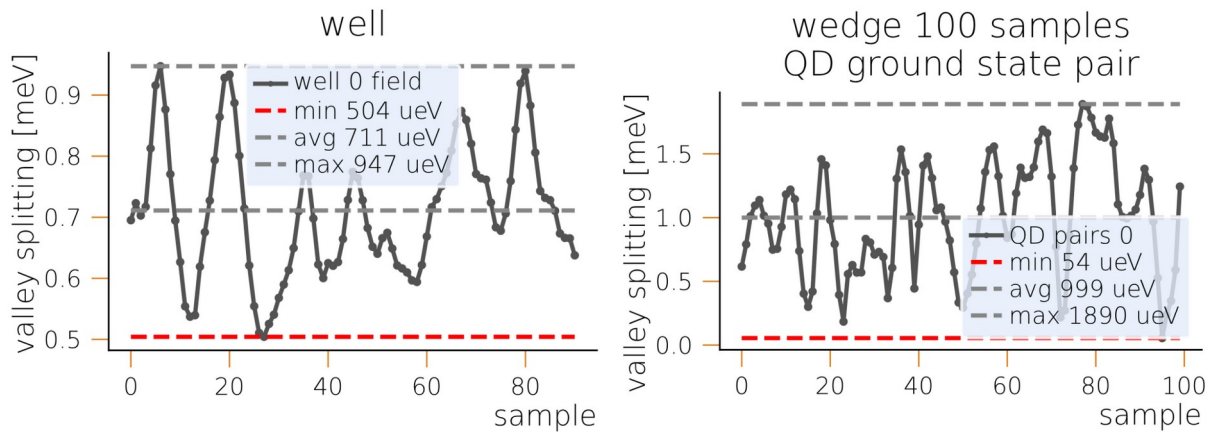
Motivation und Zielsetzung. Nach der Implementierung des 10-Band Slater-Koster Tight-Binding-Modells diente eine erste Pilotstudie als durchgängiger Test des neuen Simulations-Workflows. Vorrangige Ziele waren die Debugging-Phase abzuschließen, den Konsortialpartnern ein einsatzbereites Werkzeug für spätere Großrechnungen zu präsentieren und erste physikalische Trends des Valley-Splittings in Abhängigkeit vom Germaniumprofil und von externen elektrischen Feldern abzuleiten.

Simulationsaufbau. Der Quantenpunkt wurde als rechteckige Box mit 5 nm Seitenlänge in einem 10 nm tiefen Si/Ge-Quantentopf modelliert. Zwei vertikal gerichtete Germaniumkonzentrationsprofile wurden untersucht: ein abruptes Stufenprofil („well“) mit verschwindendem Germanium-Anteilen im Inneren des Quantenpunkts und scharf definierten Grenzflächen sowie ein linear abfallendes Keilprofil („wedge“) mit sanftem „trailing edge“. Zusätzlich wurde ein homogenes elektrisches Feld entlang der Wachstumsrichtung in positiver und negativer Polarität angelegt, um den Quantenpunkt wahlweise zur oberen oder unteren Grenzfläche zu verschieben.

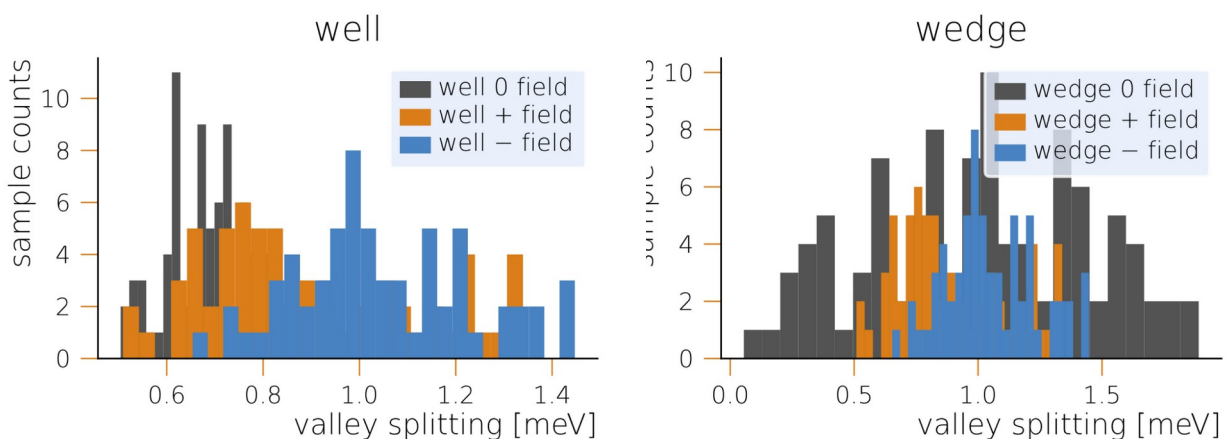
Für jede Parameterkombination wurden 100 stochastisch generierte Legierungsrealisationen erzeugt, um die intrinsische Grenzflächenrauigkeit abzubilden. Mithilfe eines spektralen Filters wurden anschließend die beiden energetisch niedrigsten Quantenpunkt(Dot)-Zustände extrahiert; ihre Energiedifferenz definiert das Valley-Splitting. Querschnittsdarstellungen der räumlich variierenden Germanium-Konzentration (siehe Abbildung, links „well“, rechts „wedge“) veranschaulichen die untersuchten Profile. Im unteren Teil der Abbildung wird das (teilweise gemittelte) Quantentopfpotential der Heterostruktur in Abhängigkeit der Wachstumsrichtung dargestellt.



Ergebnisse. Das Stufenprofil („well“) führt im Mittel zu einem kleineren Valley-Splitting von etwa 700 μeV , mit einer maximalen Varianz (für 100 Realisierungen) von rund $\pm 200 \mu\text{eV}$ auf und ist damit prognostisch sehr stabil. Das Keilprofil („wedge“) erhöht das mittlere Valley-Splitting auf etwa 1000 μeV , geht jedoch mit nahezu doppelt so großen Fluktuationen einher, was zu einer geringeren Vorhersagbarkeit führt. Dies ist in den folgenden zwei Abbildungen zu sehen, welche Valley-Splitting Scans für beide Profile zeigen.



Ein externes elektrisches Feld von einigen mV / nm steigert, unabhängig von seiner Ausrichtung entlang oder entgegen der Wachstumsrichtung, das Valley Splitting für das Stufenprofil und reduzieren die Streuung für das Keilprofil die Varianz bei gleichbleibendem mittleren Valley-Splitting. Der Symmetrietest, Austausch von „oben“ und „unten“ durch Umkehr des elektrischen Feldes, ergab ähnliche Valley-Splitting Verteilungen für das Keilprofil. Für das Stufenprofil, ergibt der Symmetrietest, dass es eine „Schokoladenseite“ des Quantentopfes gibt, trotz der Tatsache, dass das Germaniumprofil symmetrisch um den Quantentopfmittelpunkt ist.



Diskussion und Limitierungen. Die rechteckige 5 nm-Box ist deutlich kleiner als die 30-80 nm großen Quantenpunkte von tatsächlich realisierten Strukturen. Durch die künstlich kleine Struktur des Simulierten Quantenpunkts ist die statistische Varianz erhöht, ein kleiner konkreter Dot „sieht“ (mittelt über) weniger des ihn beherbergten Materials und ist damit anfälliger für statistische Fluktuationen. Darüber hinaus liefert eine Probengröße von lediglich 100 Simulationen zwar qualitative Trends, reicht aber nicht für belastbare statistische Aussagen zur Chip-Fertigung. Schließlich stellt die CPU-basierte Referenzimplementierung mit Rechenzeiten von mehreren

Stunden pro Realisation von realisierbaren Strukturen einen wesentlichen Engpass für umfangreichere Parameterstudien dar.

Unsere Pilotstudie, welche zugleich unseren Halbzeitmeilenstein für das QUASAR Projekt darstellte, zeigte damit, dass der neue Tight-Binding-Workflow funktionsfähig ist und physikalisch plausible Trends liefert: Keilprofile erhöhen zwar das Valley-Splitting, tun dies aber auf Kosten größerer Streuungen, während ein externes elektrisches Feld das Splitting sowohl anhebt als auch statistisch stabilisiert. Damit ist eine solide Basis für das in den folgenden Abschnitten beschriebene GPU-beschleunigte Hochskalierungsprogramm geschaffen.

GPU-beschleunigte Großsimulationen und Analyse des Valley-Splitting

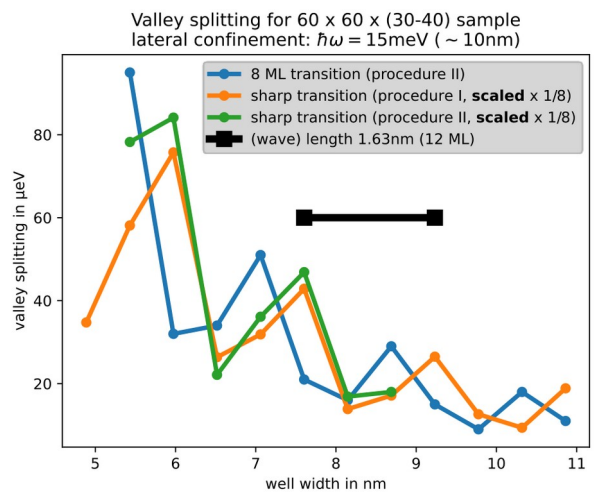
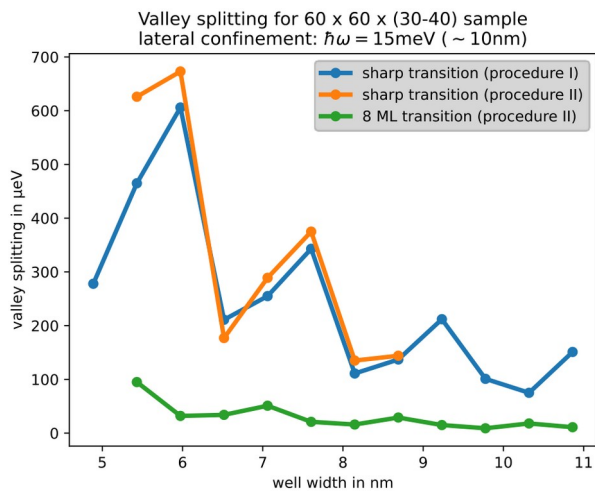
Der im dritten Projektjahr vollzogene Port des Tight-Binding-Workflows auf Grafikprozessoren (GPUs) war ein entscheidender Meilenstein: Erstmals konnten wir Si/Ge-Heterostrukturen in den für Spin-Qubits relevanten Abmessungen von bis zu $80 \times 80 \times 20 \text{ nm}^3$ – das entspricht rund 10 Millionen Atomen – mit atomarer Auflösung simulieren. Die Berechnungszeiten reduzierten sich von mehreren Tagen auf wenige Stunden pro Struktur, womit die Simulation realer Quantenpunktgrößen in einen produktiv nutzbaren Rahmen stattfinden konnte.

Um die physikalische Korrektheit dieser großskaligen Modelle abzusichern, wurden sämtliche Ergebnisse systematisch mit (semi-)analytischen Effektive-Masse-Berechnungen unsere Konsortialpartner von der Universität Konstanz sowie mit Literaturwerten verglichen. Schlüsseltendenzen – etwa die Abhängigkeit des Valley-Splittings von Quantentopfbreite, Punktdurchmesser oder externen Feldern – stimmten im Wesentlichen mit Literatur-bekannten Ergebnissen überein.

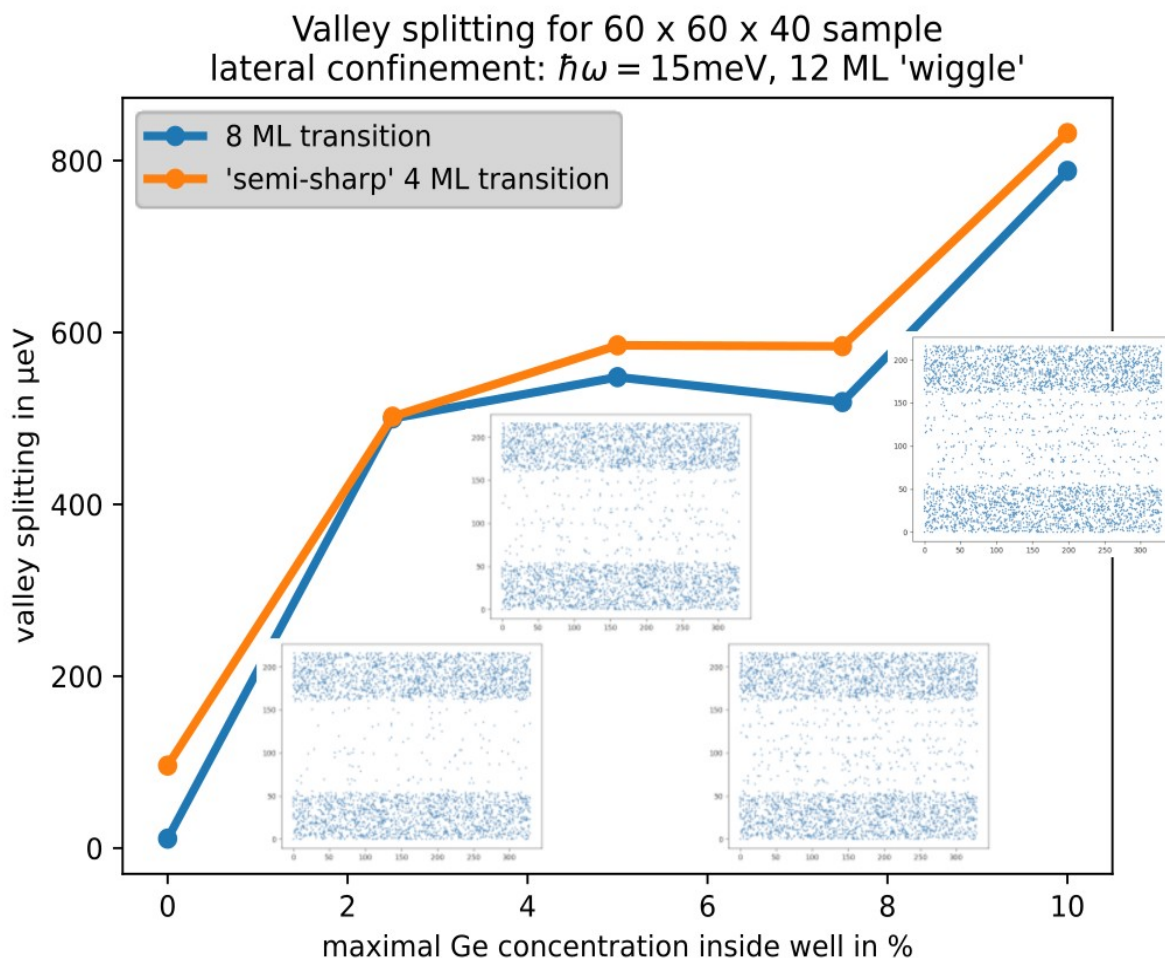
Vollständig kontrollierte Random-Alloy-Konfigurationen. Ein wesentlicher Vorteil des GPU-Workflows, gegenüber der industriellen Fabrikation, liegt in der vollständigen Kontrolle der Zufallslegierungskonfigurationen: Für jede Simulation kann die Quantentopfbreite variiert werden, während beide Grenzflächen Atom für Atom identisch bleiben. Dadurch isolieren wir geometrische Effekte (z. B. die Quantentopfbreite) von stochastischen Legierungsfluktuationen – ein in Experimenten nicht erreichbarer Idealzustand. Die begleitende Abbildung zeigt exemplarisch das Valley-Splitting als Funktion der Quantentopfbreite für zwei Extremfälle:

- „**sharp transition**“ – abruptes Stufenprofil,
- „**8 ML transition**“ – linear auslaufendes „trailing edge“ über 8 atomare Lagen.

Oszillatorisches Verhalten des Valley-Splittings. Für beide Oberflächentypen beobachten wir die in der Literatur erwarteten Oszillationen des Valley-Splittings mit einer Wellenlänge von etwa 1,63 nm. Allerdings verschiebt sich die Phase der Oszillation je nach zufälliger Legierungskonfiguration. Diese unkontrollierte Phasenlage stellt eine zentrale Herausforderung für deterministisches Qubit-Design dar. Gleichzeitig bestätigt die Studie, dass sehr scharfe Grenzflächen zu Valley-Splittings oberhalb von 200 μeV führen, allerdings ist es experimentell bislang wohl eher nicht möglich solch klar definierten Grenzflächen zuverlässig zu synthetisieren. Die folgenden Abbildungen zeigen die Abhängigkeit des Valley-Splittings von der Quantentopfbreite *bei festbleibenden Grenzflächen*. In der rechten Abbildung wurde das Valley-Splitting der „sharp transition“ durch die Anzahl der atomaren Lagen der „8 ML transition“ zwecks einer besseren Vergleichbarkeit geteilt.



„Wiggle-well“ als potenzieller Design-Hebel. In einer Einzelsimulation wurde das Germanium-Profil innerhalb des Quantentopfes sinusförmig moduliert („wiggle well“). Eine Erhöhung der Modulationsamplitude von 0 % auf 10 % Ge-Gehalt führte – in dieser Fallstudie – zu einer anekdotischen (Einzelfallstudie) Steigerung des Valley-Splittings um knapp 40 %. Obwohl diese Beobachtung noch keiner statistischen Absicherung unterzogen wurde, zeigt sie das Potenzial nicht-trivialer Wachstumsprofile als zusätzlicher Optimierungsparameter. Die folgende Abbildung zeigt die Abhängigkeit des Valley-Splittings von der Amplitude der Modulation im Quantentopf für zwei unterschiedliche Grenzflächen.

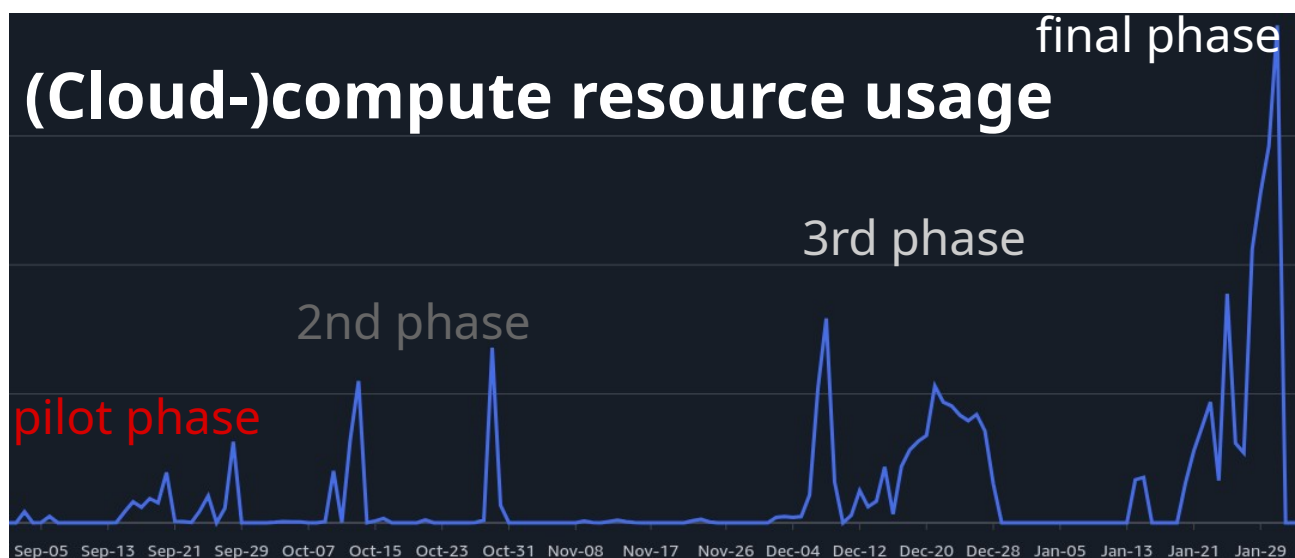


Zusammenfassung. Mit dem GPU-basierten Workflow haben wir somit eine validierte Plattform bereit, die realistische, vollständig kontrollierte Atom-Modelle von Spin-Qubit-Strukturen in industriell relevanten Zeiträumen berechnen kann – ein entscheidender Schritt vom Proof-of-Concept hin zur quantitativen Design-Optimierung zukünftiger Spin-Qubit-Chips. Nachdem die Modellpräzision erreicht war, verlagerte sich der Engpass auf den Durchsatz: Aussagekräftige Statistiken erfordern Zehntausende Zufallskonfigurationen, deren Berechnung selbst auf GPUs wochenlange Laufzeiten belegt. Diese Erkenntnis war Auslöser für die im Projekt bewilligte Mittelumwidmung hin zu skalierbaren Cloud-GPU-Ressourcen, mit denen wir die Zahl gleichzeitiger Simulationen um zwei Größenordnungen steigern konnten.

Abschließende Großsimulationen in der Cloud

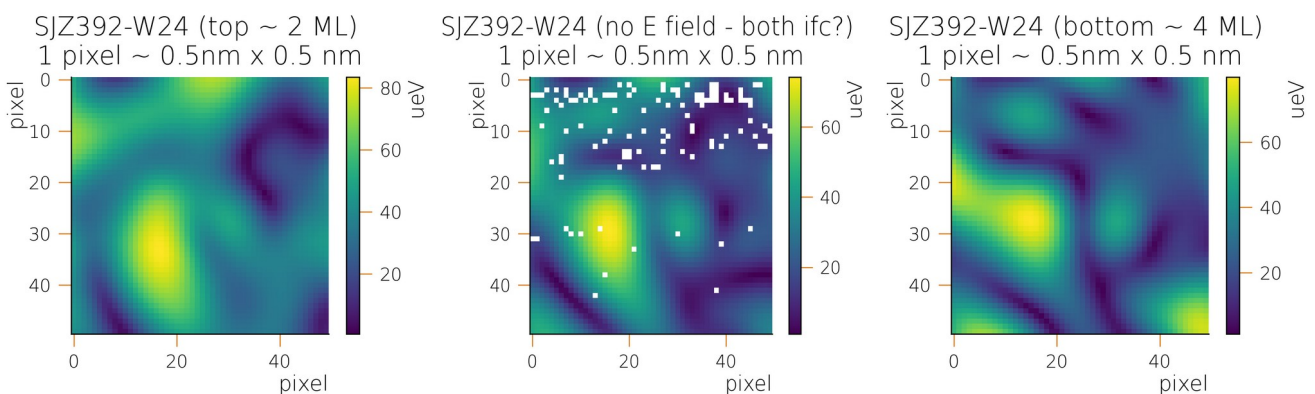
Nachdem wir gemeinsam mit den Konsortialpartnern erkannt hatten, dass Einzelsimulationen nicht ausreichen, um verlässliche Designregeln für die Shuttling-Architektur abzuleiten, genehmigte der VDI eine Mittelumwidmung für zusätzliche Cloud-Ressourcen. Damit stand uns in der AWS-Umgebung ein Pool von bis zu 512 NVIDIA-A10-GPUs gleichzeitig zur Verfügung. Wir konnten als externe Expertise das von uns entwickelte Workflowsystem HQS Tasks einbringen um damit den bisher ausgearbeiteten Workflow nahtlos in einen „Containerisierte“-Simulationsablauf umzuwandeln, welcher sowohl auf dem lokalen HPC-Cluster als auch in der Cloud ausführbar ist.

Simulationsablauf. Jeder Simulationszyklus dauerte mehrere Tage reiner Rechenzeit, die von einer ein- bis zweiwöchigen Analyse- und Optimierungsphase abgelöst wurde. In vier solcher Zyklen, die zwischen Oktober 2024 und März 2025 stattfanden, steigerten wir schrittweise die Datenausbeuten und passten die Parameterauswahl für die Simulation, nach Rücksprache mit unseren Konsortialpartnern, an. Der Pilotzyklus setzte noch auf einen verkleinerten Quantenpunkt, um eine zweidimensionale Valley-Splitting Scans möglichst schnell dem Konsortium zur Verfügung zu stellen und diskutieren zu können. Das in dieser Phase gewonnene Feedback floss unmittelbar in Code-Verbesserungen wie das Caching großer Kopplungsmatrizen und ein Cloud-paralleliertes Diagonalisierungsverfahren ein; dadurch vervierfachte sich die Produktivität schon im zweiten Zyklus. Spätestens im dritten Durchlauf, als die laterale Korrelationslänge von der Test- auf die reale Struktur (≈ 80 nm) umgestellt wurde, war klar, dass sich die Zeit pro Datenpunkt gegenüber dem Stand von 2023 um den Faktor 100 verkürzt hatte. Die folgende Abbildung stellt den Verlauf der Simulationen anhand der täglich benutzten (Cloud-)compute Ressourcen dar.



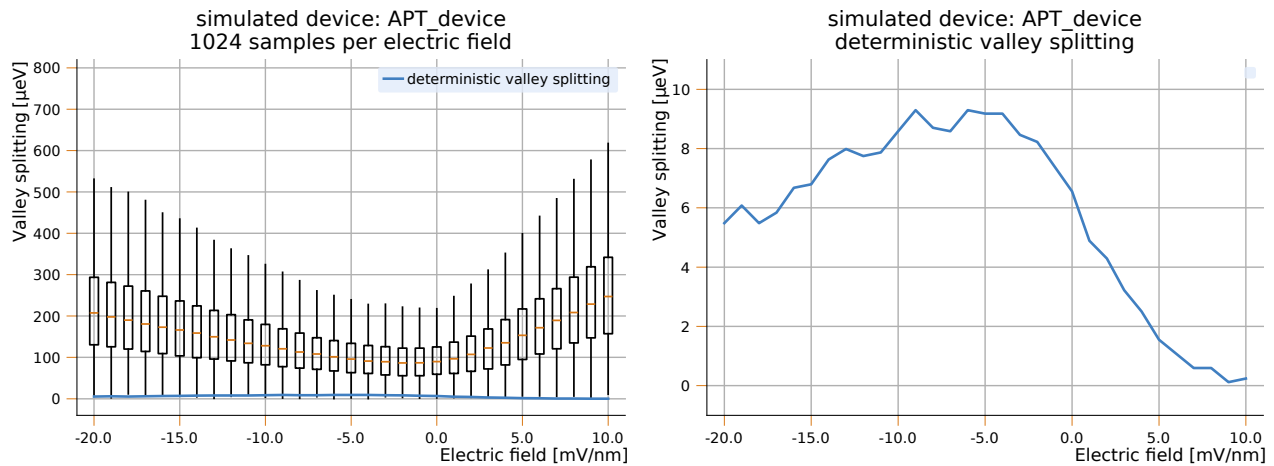
Setup. Für die abschließende Simulation wählten wir gemeinsam mit den Partnern drei charakteristische Germanium-Profile aus: einen kleinen Quantentopf (6.9 nm, SJZ392-W24), einen großen Quantentopf (13.7 nm, SJZ367-W14) und ein durch [Atom-Probe-Tomography](#) [4] (APT) bestimmtes Profil mit ausgeprägtem „trailing edge“. Für diese drei Legierungsprofile wurden 2D Scans durchgeführt wobei ein elektrisches Feld entlang und entgegen der Wachstumsrichtung angelegt wurde, sodass der Quantenpunkt-Zustand entweder in die obere oder untere Grenzfläche gedrückt wurde.

Pilotphase. In der Pilotphase wurden noch künstlich kleine Quantenpunkte mit einem Durchmesser von etwa 10nm (Abstand der klassischen Wendepunkten des Oszillators) berechnet und der 2D Scan mit der höchsten verfügbaren Auflösung durchgeführt, zur Erinnerung die „Einheitszelle“ entspricht etwa einem $0.5 \times 0.5 \times 0.5 \text{ nm}^3$ Würfel. Außerdem wurde lediglich zwei elektrische Felder von 1.34 mV / nm entlang und entgegen der Wachstumsrichtung angelegt. Die folgenden Abbildungen zeigen die Scans für den kleinen Quantentopf und den APT Chip. Die Pixelfehler im mittleren Scan entsprechen abgebrochenen Rechnungen im Zusammenhang des Setups der benutzen Cloud-Infrastruktur und verdeutlicht die Wichtigkeit die Simulation behutsam zu skalieren um Erfahrungen und Feedback kontinuierlich in die Simulation einfließen lassen zu können. Ein wichtiges Erkenntnis aus der Pilotstudie war, dass die Auflösung des 2D Scans, selbst bei den verwendeten äußerst kleinen Quantenpunkten deutlich heruntergefahren werden kann und somit ein bessere (unabhängigere) Statistik zu sammeln, oder äquivalent, einen größeren Bereich des (virtuellen) Chips zu charakterisieren.

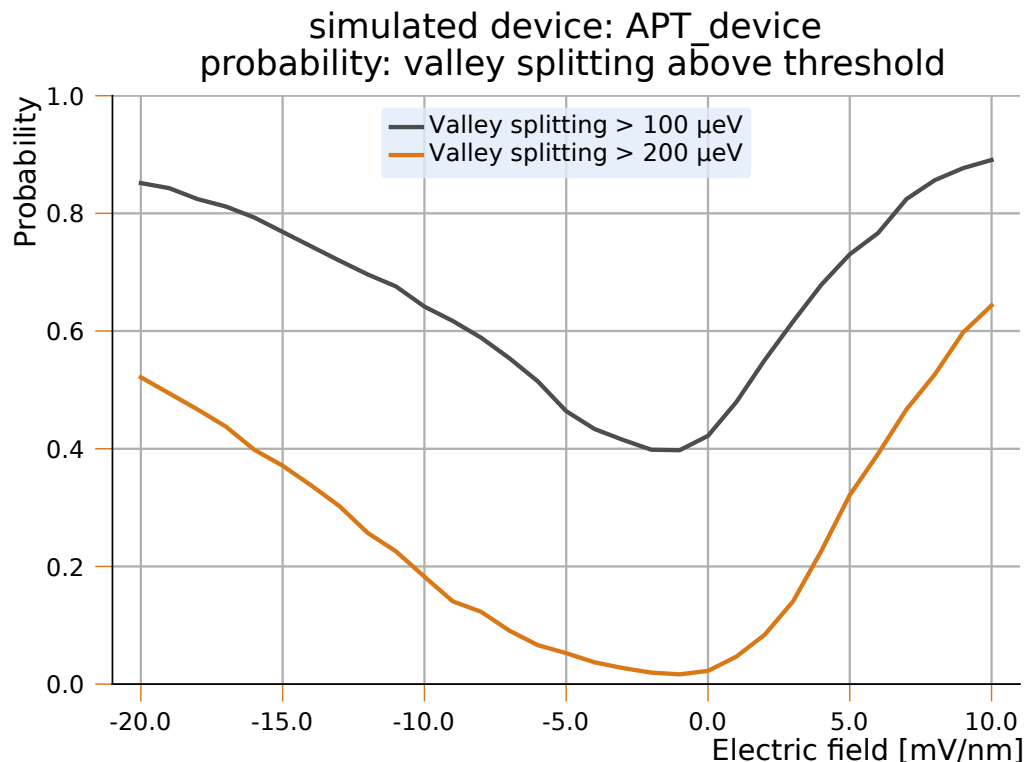


Simulation. Für die finalen Simulation wurden dann Quantenpunkte mit einem Durchmesser von etwa 25nm (Abstand der klassischen Wendepunkten des Oszillators), oder äquivalent mit einem induzierten „level splitting“ von 2.5 meV, berechnet. Für die drei ausgewählten Legierungsprofile wurden zunächst das deterministische Valley-Splitting (aus dem gemittelten VCA Hamiltonoperator) für 31 elektrische Felder von -20 mV / nm bis $+10 \text{ mV / nm}$ in Schritten von 1 mV / nm berechnet. Als Vorzeichenkonvention für das elektrische Feld verwenden wir, dass *negative* elektrische Felder den Quantenpunkt in die obere Grenzfläche verschieben. Anschließend wurden 1024 explizite Legierungsrealisationen generiert und deren Valley-Splitting bestimmt. Die zentralen Ergebnisse sind im Folgenden dargestellt. Für jedes elektrische Feld wurde die Verteilung der Valley-Splittings durch einen Boxplot zusammen mit dem Wert des entsprechenden deterministischen Valley-Splittings dargestellt. Die Valley-Splittings der drei untersuchten Legierungsprofile sind alle durch „Stochastik“ dominiert, d.h., das gemittelte Valley-Splitting ist *nicht* ähnlich dem deterministischen Valley-Splitting, weshalb wir das deterministische Valley-Splitting nochmals gesondert darstellen.

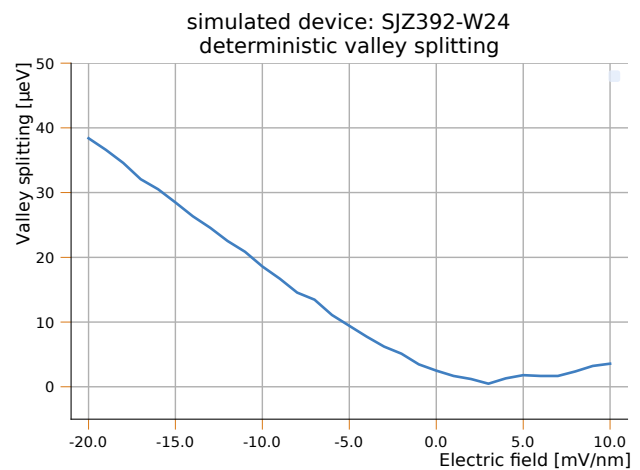
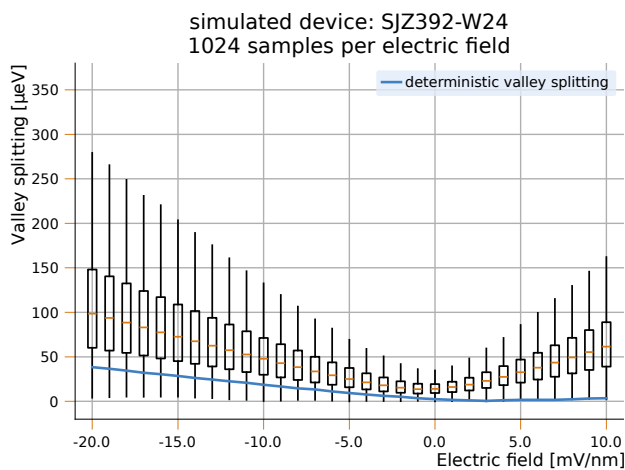
APT. Das APT Legierungsprofil ist aus [Klos et al. \(2024\)](#) [4] entnommen und wurde ausgewählt, da es mit einer sehr genauen, aber destruktiven, Methode vermessen wurde. Dabei wurde festgestellt, dass in Abhängigkeit des Fabrikationsvorgangs sich ein sogenanntes „trailing edge“ an der untere Grenzfläche des Quantentopfs ausbilden kann, welches unter anderem bedeutet, dass mehr Germanium in den Quantentopf diffundieren. Das deterministische Valley-Splitting ist hart an der Auflösungsgrenze der Energie in unserem Simulationsworkflows von etwa 2-5 μeV .



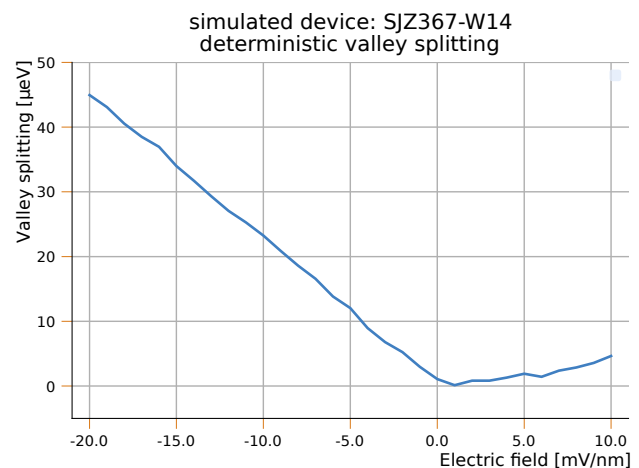
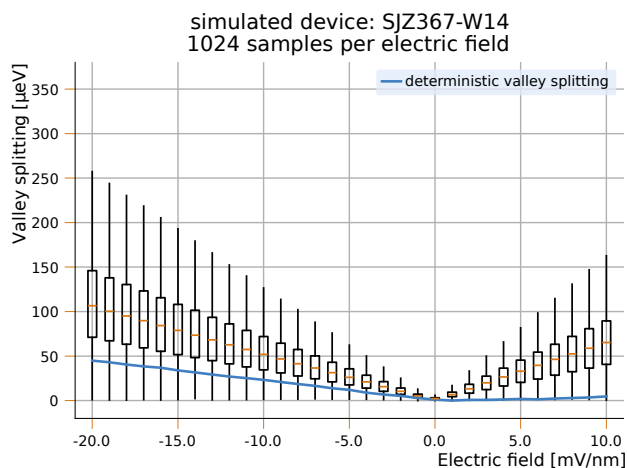
Eine weitere Fragestellung die im Konsortium diskutiert wurde, war mit welcher Wahrscheinlichkeit Valley-Splittings, für Chips ähnlich des mittels APT vermessenen Chips, gemessen werden sollten. In der folgenden Abbildung wird dies Frage basierend auf unseren Ergebnissen in Abhängigkeit des angelegten elektrischen Feldes für zwei Schranken, oberhalb von 100 μeV und 200 μeV , beantwortet.



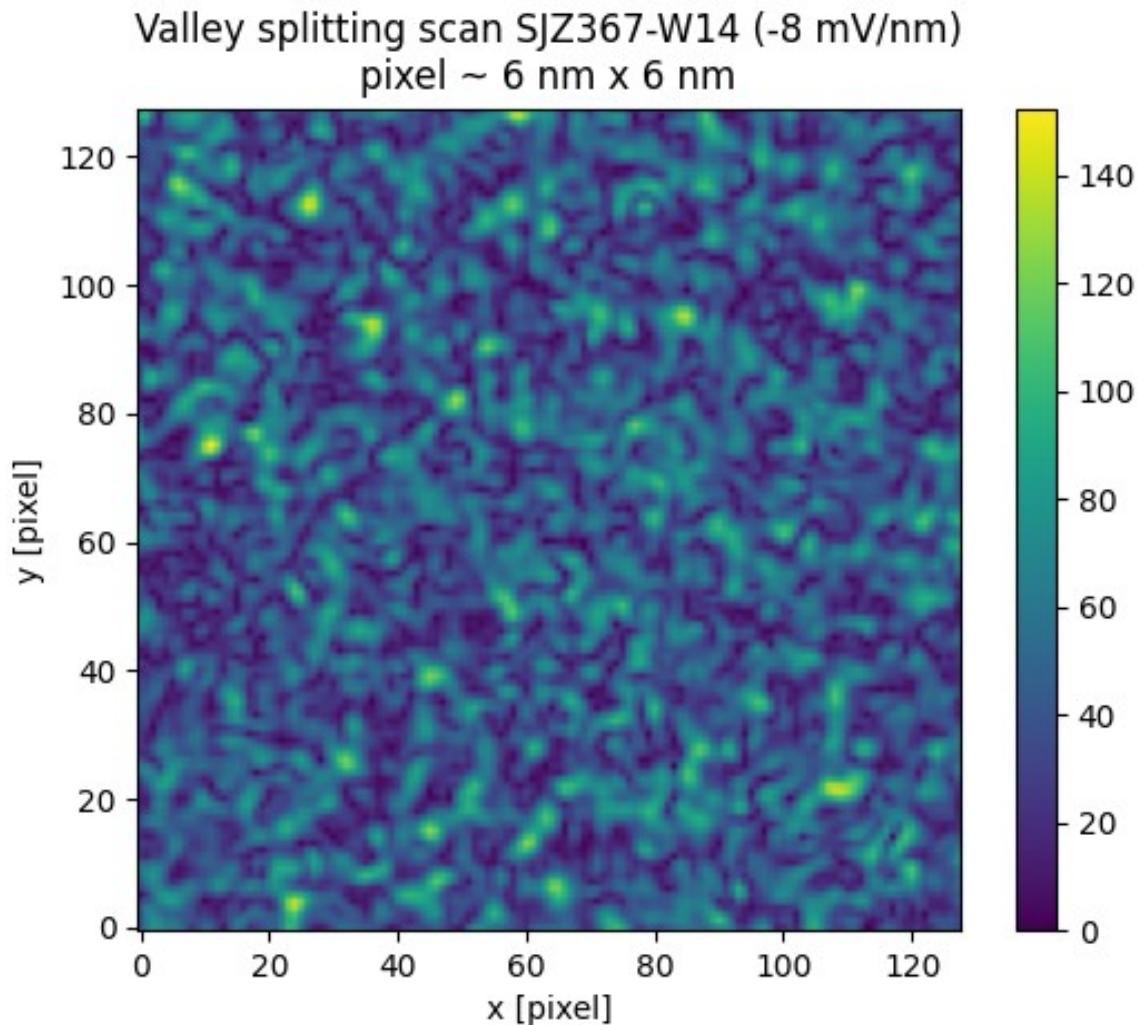
Kleiner Quantenpunkt. Der kleinere Quantenpotf hat eine Breite von 6.9 nm und Grenzflächen welche durch eine Standardabweichung von 0.15 nm (obere Grenzfläche) und 0.3 nm (untere Grenzfläche) charakterisiert sind. Der Germaniumanteil ist (asymptotisch) 34% in den Grenzgebieten des Quantenpotfs. Für das deterministische Valley-Splitting ist eine klare Asymmetrie des Einflusses des elektrischen Feldes sichtbar: das Valley-Splitting erhöht sich stärker, wenn der Quantenpunkt in die „schärfere“ (obere) Grenzfläche gedrückt wird. Dieser Trend ist umgekehrt (zumindest für starke elektrische Felder) das „statistische“ Valley-Splitting, d.h., für die „weichere“ (untere) Grenzfläche ist das mittlere Valley-Splitting größer.



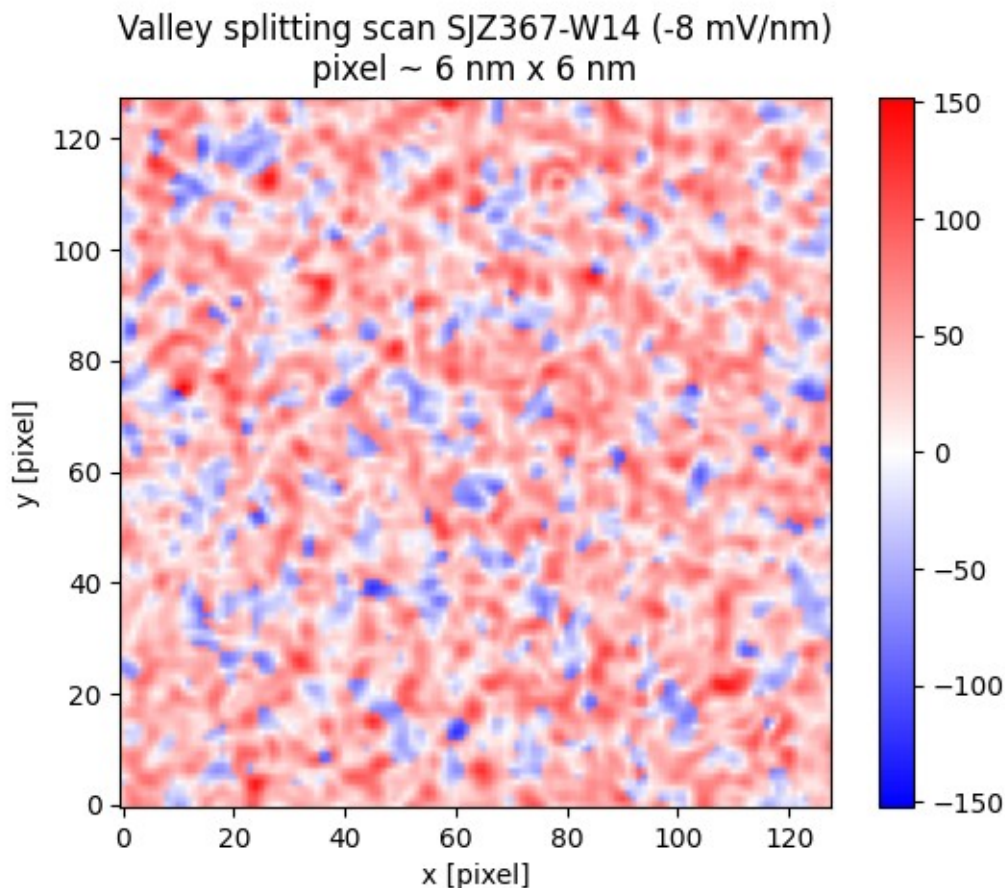
Großer Quantenpunkt. Der große Quantenpotf hat eine Breite von 13.7 nm und ansonsten die gleichen Charakteristika wie der kleine Quantenpunkt. Die Trends für das deterministische und das statistische Valley-Splitting sind sehr ähnlich zu den Ergebnissen für den kleinen Quantenpunkt. Ein wichtiger Unterschied ist das Verhalten des statistischen Valley-Splittings für kleinen elektrische Felder (kleiner 1 mV / nm). Für den großen Quantenpotf ergeben unsere Simulationen, dass das statistische Valley-Splitting quasi kollabiert, d.h., unterhalb der Energieauflösung der Simulation liegt, während es für den kleinen Quantenpotf auf minimal etwa 15 μeV abfällt.



2D Scan großer Quantenpunkt. Für einige externen elektrischen Felder haben wir schließlich 2D Scans des Valley-Splitting, wie zum Beispiel in [Volmer et al. \(2024\)](#) [5] experimentell gemessen, durchgeführt. Die folgende Abbildung veranschaulicht einen solchen Scan: der 128 x 128 pixel große Scan entspricht einer etwa 16 Tausend Datenpunkten, welche eine etwa eine 0.5 μm x 0.5 μm Ausschnitt des Spin-Qubit Chips repräsentieren.



„Signed“ Valley-Splitting. Die folgende Abbildung zeigt die gleichen Daten mit folgender wichtigen Modifikation: Statt den Absolutwert der Energieaufspaltung der untersten beiden Quantenpunktzustände darzustellen, vergleichen wir den Quantenpunktzustände mit den Quantenpunktzuständen welche über den VCA Hamiltonoperator, der Hamiltonoperator welcher das deterministische Valley-Splitting hervorbringt, berechnet wurden. Damit können wir den Charakter der Quantenzustände vergleichen und feststellen ob die beiden untersten Quantenpunktzustände für eine explizite Realisierung des Germaniumprofils ihr Plätze, im Vergleich zur VCA Lösung, getauscht haben. Falls ja, wird dem Valley-Splitting ein negatives Vorzeichen „angehängt“. Die Idee hinter dieser Darstellung ist es die Anordnung der relevanten Zustände für das Spin-Qubit räumlich zu verfolgen um mögliche Orte für einen Kollaps des Valley-Splittings zu identifizieren, welches für das im Chip angestrebte Shuttling ein Hindernis sein könnte.



Zusammenfassung und Verwertungsaussichten

Die Teilnahme von HQS Quantum Simulations am Arbeitspaket „Simulation und Modellierung“ im Rahmen des QUASAR-Projekts war in mehrfacher Hinsicht eine außergewöhnlich bereichernde Erfahrung. In der Regel konzentriert sich unsere Arbeit auf die Entwicklung Quanten-basierter Algorithmen und Softwarelösungen, die direkt auf aufstrebenden Quantencomputern angewendet werden. Innerhalb von QUASAR jedoch lag unser Fokus darauf, mit klassischen numerischen Simulationen auf herkömmlicher IT-Infrastruktur direkt die Entwicklung von Hardware für Quantencomputer, konkret die Bauelemente für Silizium-Germanium Spin-Qubits, zu unterstützen. Auf den ersten Blick mag dies wie ein Perspektivwechsel erscheinen, doch im weiteren Verlauf des Projekts wurde deutlich, dass sich hier ein zentraler Kreislauf von Forschung und Innovation abzeichnet: Die gleichen quantenmechanischen Simulationen, die heute mit klassischen Mitteln neuartige Hardware ermöglichen, werden künftig zu den wichtigsten Anwendungen von Quantencomputern selbst zählen – insbesondere im Bereich der Materialentwicklung und Quantentechnologien.

Der Einblick in die Welt der konkreten Quantenhardware – speziell der Spin-Qubit-Realisierung in Halbleitersystemen – war für HQS ebenso lehrreich. Durch die enge Zusammenarbeit mit akademischen und industriellen Partnern, die teils seit Jahrzehnten Halbleiterbauelemente mit immer präziseren Methoden fertigen, eröffnete sich für HQS ein tiefgreifendes Verständnis für die technologische Kontinuität und Evolution in diesem Forschungsfeld. Es wurde klar, dass die Nutzung bestehender industrieller Standards und Fertigungsverfahren – also das Fortschreiben der “Moore’s Law”-Erfolgsgeschichte – eine vielversprechende Strategien zur Skalierung von

Quantencomputern darstellen dürfte. Diese grundlegende Erkenntnis stärkte unsere Überzeugung, dass unsere Kompetenzen in numerischer Physik und quantenmechanischer Simulation künftig nicht nur für die Softwareentwicklung an Quantencomputern, sondern ebenso für die modellbasierte Unterstützung der Chipentwicklung wichtig sind.

Ein zentrales Lernerlebnis des Projekts war die konsequente Erweiterung unseres technischen Horizonts bis in Größenordnungen (von Nanometer bis hin zu Mikrometern) hinein, die in normalen quantenmechanischen Elektronenstruktursimulationen selten erreicht werden. Zu Projektbeginn arbeiteten wir im Wesentlichen mit der vergleichsweise überschaubaren Atomzahl rein quantenmechanischer Softwaresimulationen. Im Projektverlauf haben wir, getrieben durch die Anforderungen der Realbauteile und im engen Austausch mit Partnern aus Industrie und Forschung, unsere Algorithmen sowie unsere numerische Infrastruktur derart skaliert, dass am Ende Strukturen mit mehreren Millionen Atomen effizient simuliert werden konnten. Möglich wurde dies einerseits durch methodische Innovationen, insbesondere bei der Optimierung und GPU-Beschleunigung unseres Tight-Binding-Simulations-Codes, andererseits durch die kreative Lösung administrativer Hürden: Mit Unterstützung des VDI konnten zur rechten Zeit Mittel umgewidmet und die notwendigen Ressourcen in kommerziellen Cloud-Rechenzentren bereitgestellt werden. Dies war, auch in Bezug auf Datenorganisation, Austausch mit mehreren Partnern aus der Industrie und Forschung und den Aufbau effizienter Workflows in der Cloud, eine wertvolle Übung mit starken Lerneffekten über das eigentliche Forschungsthema hinaus.

Aus technischer Sicht konnten im Projekt wesentliche Meilensteine erreicht werden: Der von HQS konzipierte Simulations-Workflow deckte die vollständige Wertschöpfungskette ab – von der Grundlagenuntersuchung mittels Dichtefunktionaltheorie (DFT) über multiskalige Modell-Implementierungen bis hin zur großskaligen Tight-Binding-Simulation in industrierelevanten Größenordnungen. Die Workflow-Architektur wurde in enger Rückkopplung mit experimentellen Daten iterativ verbessert. Der Austausch mit den Partnern ermöglichte es uns, unsere Ergebnisse an reale Chipstrukturen zu spiegeln, verschiedene Materialprofile und Herstellungsvarianten gezielt zu vergleichen und so Simulationen direkt in die experimentelle Entwicklung von Prototypen einfließen zu lassen.

Von besonderer Bedeutung war auch der intensive Austausch zwischen den verschiedenen Fachcommunities. Die Schnittstellen zwischen numerischer Materialsimulation, Halbleitertechnologie und Quantenpunktphysik förderten bei HQS ein neues Verständnis für die Anforderungen und Herausforderungen der jeweiligen Disziplinen. Die Zusammenarbeit mit weltweit führenden Experten auf dem Gebiet der Spin-Qubit-Technologie eröffnete unser Team nicht zuletzt auch für zukünftige Partnerschaften und neue Märkte im Bereich der Hardware-nahen Modellierung und Co-Design von Quantenhardware.

Der direkte Nutzen für die Firmenentwicklung von HQS lässt sich zusammenfassend in mehreren Dimensionen festhalten:

- **Technologische Kompetenz:** Der Aufbau und die “Industrialisierung” unseres Tight-Binding-Workflows sowie der Sprung zu GPU-optimierten Großsimulationen stärken unser Portfolio weit über das Projekt hinaus. Damit erschließen sich neue Geschäftsfelder z.B. in der Beratung für Materialentwicklung, Bauteildesign und Chipoptimierung.
- **Netzwerk und Reputation:** Die erfolgreiche Zusammenarbeit und der Beitrag zum Gesamtprojekt steigerten unsere Sichtbarkeit bei Forschungs- und Industriepartnern. Gerade

im hochdynamischen europäischen Quantenökosystem ist dies ein wesentlicher Faktor für nachhaltiges Wachstum.

- **Zukunftsperspektive:** Die Übertragung numerischer Methoden auf industrielle Fragestellungen – hier im Chipdesign für Spin-Qubits – bietet den idealen Anknüpfungspunkt für zukünftige hybride Ansätze, bei denen Simulationaufgaben sukzessive an Quantencomputer ausgelagert werden. Damit steht HQS an der Schnittstelle von klassischem und quantenbasiertem Computing und kann entscheidende Brücken zwischen Theorie, Simulation und Industrieanwendung schlagen.

Abschließend lässt sich feststellen, dass das QUASAR-Projekt für HQS vielfältige positive Impulse gegeben hat. Das Spektrum reicht von konkreten technologischen Fortschritten und methodischen Innovationen bis hin zur Neubewertung und Erweiterung unserer Rolle innerhalb der Wertschöpfungskette der Quantenindustrie. Die gemeinsam erarbeiteten Simulationsergebnisse dokumentieren und stärken das Know-how und die internationale Wettbewerbsfähigkeit von HQS Quantum Simulations im zunehmend integrierten Feld von Quantenhardware und -software.

Referenzen:

- [1] *GPAW: An open Python package for electronic structure calculations*, Mortensen et al., *J. Chem. Phys.* **160**, 092503 (2024)
- [2] *Maximally localized Wannier functions: Theory and applications*, Marzari et al., *Rev. Mod. Phys.* **84**, 1419 (2012)
- [3] *Onsite matrix elements of the tight-binding Hamiltonian of a strained crystal: Application to silicon, germanium, and their alloys*, Niquet et al., *Phys. Rev. B* **79**, 245201 (2009)
- [4] *Atomistic Compositional Details and Their Importance for Spin Qubits in Isotope-Purified Silicon Quantum Wells*, Klos et al., *Advanced Science* **11**, 42, 2407442 (2024)
- [5] *Mapping of valley splitting by conveyor-mode spin-coherent electron shuttling*, Volmer et al., *npj Quantum Information* **10**, 61 (2024)