



# AI 기반 마쿠쉬 화학 청구항의 신규성·진보성 평가 고도화 연구

CAIPEX 학회지 · 2026 년 9 월  
E(on). Jun  
inwardis@hotmail.com

## 초록

본 연구는 화학구조식과 치환기 정의가 결합된 마쿠쉬(Markush) 청구항의 선행기술 검색 및 신규성·진보성 평가를 지원하는 인공지능 기반 절차를 제안한다. 국제조사 지침과 유럽·미국 심사기준, 화학구조 인식 및 마쿠쉬 검색 연구를 검토하고, 구조 그래프·치환기 범위·변수 간 제약·원문 출처를 함께 기록하는 중간표현(PMIR)을 설계하였다. 제안 절차는 구조식 인식, 청구범위 제약 추출, 특정 화합물 및 일반식 병렬 검색, 문헌별 요소 대응, 실시 가능성 검토, 진보성 관련 증거 구성으로 이루어진다. 신규성에는 단일 문헌의 개시와 실시 가능성을, 진보성에는 차이점 및 기술적 동기의 증거를 각각 적용한다. 성능 검증을 위해 특허 패밀리와 출원 시점을 분리한 한국어 평가자료, 전문가 검색 대조군 및 심사관 참여 관찰평가를 제시한다. 본 연구의 결과는 운영 시스템의 실측 성능이 아니라 설계와 평가계획이며, 법적 결론은 심사관의 검증을 전제로 한다.

주제어: 마쿠쉬 청구항, 신규성, 진보성, 화학구조 인식, 선행기술 검색, 증거 추적

## 1. 서론 및 법적 평가 기준

마쿠쉬 청구항은 고정 기본 구조와 다수의 치환기, 치환 위치 및 변수 간 조건을 한 문장 또는 여러 도면·정의문에 분산하여 기술한다. 개별 화합물의 구조 인식 결과만으로는 청구범위가 보존되지 않는다. 기존 지식재산 심사품질 연구에서 제안된 사건 단위 자료구조, 사실-증거 연결, 규칙 기반 점검과 사람의 최종 확인 원칙은 이 과제에 적용 가능한 설계 기반을 제공한다. [36]

특히 추출 사실의 출처, 해석 확신도 및 규칙 버전을 기록하고 불확실한 사항을 별도로 표시하는 체계가 필요하다. 본 연구는 이를 화학 구조의 도면 위치와 청구항 문언까지 확장하여, 동일한 검색 결과에서 어떠한 구조적·법적 판단이 도출됐는지 재검토할 수 있도록 한다. 이러한 구성은 시스템 설계안이며 아직 실제 심사·행정 현장에서의 검증 결과를 뜻하지 않는다. [36]

신규성 판단은 관련 법역의 적용 규칙을 전제로 단일 선행문헌에서 청구항의 필수 요소가 직접적·명백하게 개시되는지 검토해야 한다. 필연적인 내재적 개시는 추측이나 단순한 개연성과 구별된다. 특히 화합물질의 명칭 또는 화학식이 언급됐다는 사실만으로 개별 화합물의 실시 가능한 개시가 확정되지는 않는다. 따라서 구조검색의 일치 결과와 신규성 판단 결과는 별도로 기록해야 한다. [5]

진보성 판단에서는 선행기술과의 차이점, 그 차이로부터 주장되는 기술효과, 통상의 기술자가 해당 구성을 채택할 이유를 검토한다. 복수 문헌의 결합은 적용 법역의 판단 구조에 따라 정당화되어야 한다. 구조 유사도나 임베딩 점수는 관련 문헌의 탐색과 정렬에 사용할 수 있지만, 자명성의 법적 결론을 직접 대체하지 않는다. [6][7]

마쿠쉬에 관한 법적 문제 역시 구분해야 한다. 유럽특허청의 마쿠쉬 대안에 관한 지침은 발명의 단일성 판단에서 공통 구조·성질을 다루며, 그 자체가 신규성의 판정 규칙은 아니다. 선택발명에서 일반식의 선행 개시가 특정 선택항목의 신규성을 언제 배제하는지는 별도 기준에 따른다. 미국 심사지침은 선행문헌에 명시된 특정 종(species)이 상위 속(genus) 청구항의 신규성에 미치는 효과를 설명한다. 따라서 본 연구는 종의 속 포함관계, 일반식에 의한 종의 개시, 특정 선택군 및 실시 가능성을 서로 다른 관계로 저장한다. 법역과 청구항 문언에 따라 결론은 달라질 수 있다. [8][9][37]

## 2. STN 기반 마쿠쉬 검색과 표현상의 한계

CAS의 Markush 검색은 단순히 수백만 개 화합물을 미리 전개(enumerate)해 저장하는 시스템이 아니다. MARPAT의 역사적 설계 자체가 generic chemical structure를 직접 저장·검색하기 위해 개발됐으며, CAS의 초기 논문들은 generic structure 표현, screening descriptor, reduced graph, atom-bond 수준 refined matching 같은 계층적 검색 문제를 다뤘다. [10] 현대 CAS 인터페이스에서도 Markush search는 “generic structures matching a query”를 별도로 검색하며 As Drawn과 As a Substructure 같은 검색모드를 제공한다. [11]

CAS의 공개 자료에 따르면 MARPAT은 130만 건 이상의 검색 가능한 마쿠쉬 구조와 54만 건 이상의 관련 특허 인용을 수록하며, 한국 특허는 2008년 이후의 범위를 명시한다. 다른 CAS 데이터베이스 안내문은 더 큰 수치를 제시하므로 수차·집계 시점·수록 범위를 구분해야 한다. MARPAT은 일부 무기화합물·고분자 등의 영역을 포함하지 않으며, 전체 한국 화학특허를 완전하게 색인한 자료로 전제할 수 없다. [3][12]

검색에서는 특정 화합물과 일반식의 두 경로를 병행한다(표 1).

| 단계              | STN 중심 현행·권장 작업                                                                  | AI가 보조할 부분                                                          |
|-----------------|----------------------------------------------------------------------------------|---------------------------------------------------------------------|
| 청구범위 파악         | scaffold, 필수 atom/bond, R-group, 위치변이, stereo, salt/solvate 범위를 사람이 파악           | 구조식·청구문·표·정의문을 하나의 PMIR로 자동 추출                                      |
| Specific search | exemplified compound 및 claim의 대표 species를 REGISTRY 등에서 exact/substructure search | 명칭·SMILES·구조식 간 cross-check, tautomer/salt/stereo query variants 생성 |
| Generic search  | MARPAT Markush search, 필요시 variable atom/generic node/attachment 조건을 조정          | 보수적·중간·확장 query family 자동 제안                                        |
| 보완 검색           | text, IPC/CPC, assignee, inventor, citation/family, reaction/compound names      | dense retrieval과 citation/family graph로 reranking                   |
| 결과 검증           | hit patent의 원문·청구항·실시에·우선일 검토                                                    | claim-to-prior-art element map과 evidence span 자동 생성                 |
| 법률 판단           | 신규성/진보성/기재요건을 심사관이 분리 판단                                                         | 규칙 위반·누락 탐지, UNKNOWN routing                                        |
| 기록              | 검색식, DB, 날짜, 주요 hit 기록                                                           | 자동 provenance와 reproducible search package 생성                       |

이 방식은 PCT의 현행 검색기록 원칙과도 잘 맞는다. 2026 Guidelines는 가능한 경우 심사관이 사용한 chemical structure query, database 명칭 및 실제 query를 포함한 search history를 기록하도록 권장한다. [13] 따라서 AI 시스템이 “관련특허 10개”만 보여주는 것보다 어떤 구조에서 어떤 일반화가 이뤄졌고, 어떤 STN query가 실행됐으며, 어떤 query expansion이 hit를 추가했는지를 재현 가능하게 보존하는 것이 심사품질 관점에서 더 중요하다.

## 2.1. 마쿠쉬 청구항의 계층적 표현

특정 화합물의 표준식 하나로 마쿠쉬 청구범위를 대표하면 변형 가능한 치환기와 관계 제약이 손실될 수 있다. 이를 방지하기 위해 네 층의 표현을 사용한다(표 2).

| 층                        | 저장내용                                                                                                     | 권장 식별·표현                                                          |
|--------------------------|----------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------|
| Discrete Molecular Graph | 명확한 단일 화합물, stereo, charge, isotope                                                                      | canonical SMILES +<br>Standard InChI/InChIKey<br>+ 원본 graph       |
| Generic Scaffold         | 고정 scaffold와 variable attachment sites                                                                   | Markush graph /<br>CXSMILES 계열<br>serialization + 내부 graph<br>AST |
| Constraint Layer         | R1 ∈ ..., 치환가능성, 범위, 제외, positional variation, repeat range, R1/R2 연동                                    | PMIR constraint<br>AST/graph                                      |
| Patent Semantics         | claim number, definition source, embodiment, example, priority/family, cited paragraph/image coordinates | Canonical IP Case Package<br>+ provenance                         |

이 구분이 필요한 결정적 이유는 Standard InChI가 현재 Markush structures를 표현하지 못하기 때문이다. InChI Trust의 기술 FAQ는 Markush, polymers, mixtures 등을 현재 표현할 수 없는 대상으로 명시한다. [4] InChI 1.07은 IUPAC와 InChI Trust가 승인한 현재 세대의 identifier이지만, generic claim의 canonical identifier 역할을 맡길 수는 없다. [14]

Discrete compound의 표준화에는 PubChem 방식처럼 validation, normalization, valence 처리와 unique compound aggregation을 참고할 수 있다. PubChem standardization은 contributor Substance를 구조표준화해 Compound로 통합하고, tautomer/resonance 등 표현차이를 다룬다. [15] RDKit MolStandardize도 normalization·reionization·fragment/metal 처리 및 canonical tautomer 기능을 제공하지만, canonical tautomer는 “가장 안정한 tautomer”를 물리화학적으로 주장하는 것이 아니라 일관된 canonical representation을 고르는 것임을 명확히 한다. [16] 이 차이는 특허의 tautomer claim scope에서 특히 중요하다.

## 2.2. 구조검색 결과의 해석상 한계

첫째, 구조 색인은 자연어로 기술된 모든 예외와 의존조건을 보존하지 못할 수 있다. 치환기 정의가 조건문이나 다른 청구항의 정의를 참조하는 경우, 검색결과를 원 청구항 문맥과 다시 대조해야 한다. [1]

둘째, 전체 화합물의 완전 열거는 조합 폭발과 변수 간 제약 때문에 적절한 기본 검색법이 아니다. 예컨대 각각 20개 대안을 갖는 변수 10개만으로도 단순 조합이  $20^{10}$ 개이다. 공동고리 형성, 위치 선택, 반복 단위 등의 조건을 반영하면 독립적인 조합 모델 자체가 부정확해진다.

셋째, 부분구조 일치 문헌이 곧바로 해당 청구항의 신규성을 배제하는 것은 아니다. 개별 조합의 개시, 해당 실시예의 제조 가능성, 관련 낱자와 문헌의 법적 지위를 원문에서 확인해야 한다. [5]

넷째, 공개 이미지 인식 자료와 전문적으로 색인된 상용 일반식 자료는 범위와 품질이 다르다. 초기 구현에서는 허용된 검색 서비스와 기관 자료를 병행하고, 별도 구축 자료의 오류율을 측정하는 편이 타당하다. [1]

### 3. 화학구조 인식 기술과 청구범위 해석

일반 화학구조식 인식의 성능은 공개 연구에서 개선되고 있다. MolScribe, MolNexTR 및 OCSRGlyph은 각기 다른 자료와 평가기준에 따른 인식 결과를 보고한다. 예를 들어 OCSRGlyph의 특정 USPTO 이미지 평가에서는 입체화학을 포함한 완전 일치율 93.8%가 보고되었다. 이 수치는 마쿠쉬 청구항의 전체 의미 또는 법적 판단의 정확도를 측정한 값이 아니다. [17][18][19]

마쿠쉬 구조에는 도면과 주변 텍스트를 연결하는 추가 문제가 있다. MarkushGrapher 및 뒤이은 연구는 구조 이미지·문자·배치를 함께 처리하며 공개 평가자료를 제시한다. 저자들은 실제 자료에서의 완전 자동 처리 성능에 한계가 있음을 인정한다. [20][21]

MarkushGlyph은 도면과 정의문의 결합 인식을 개선했지만, 동일한 공개 평가기에 따른 정확도는 IP5-M, M2S, USPTO-Markush 순으로 60.6%, 62.1%, 63.5%였다. 데이터 규모와 평가 지표가 다르므로 일반 구조식 인식 또는 MarkitS의 수치와 직접 비교해서는 안 된다. [19]

#### 3.1. 공개 마쿠쉬 영상 인식의 평가결과

MarkushGrapher-2의 각 데이터셋 점수는 53.2%, 56.0%, 55.0%, MarkushGlyph MV@8의 점수는 60.6%, 62.1%, 63.5%로 보고되었다. 이는 동일 평가를 적용한 이미지 인식 비교이며, 청구항의 전체 치환기 정의나 법률판단의 성능을 뜻하지 않는다. [19]

그림과 치환기 주석의 일치도만으로 법적 의미까지 검증할 수는 없다. 기존 평가기는 잘못 추가된 주석에 대한 감점이나 치환기 정의문의 의미 평가가 제한될 수 있다. 실제 심사 보조용 평가자료에는 구조, 치환기 텍스트, 위치·범위·예외 조건의 동치성 검사가 포함되어야 한다. [1]

MarkitS가 별도의 마쿠쉬 이미지에서 보고한 85–90% image-to-SMILES 정확도는 주로 영상에서 구조 연결관계를 복원하는 과제에 해당한다. 이전 전체 인식 평가와 측정 대상이 달라 수치의 차이를 성능 우열로 해석할 수 없다. 기술별 기능과 제한은 표 3에 정리하였다. [2]

| 기술/도구            | 주 용도                               | 장점                                         | 검토 단계에서의 활용 범위 | 핵심 제한                                                            |
|------------------|------------------------------------|--------------------------------------------|----------------|------------------------------------------------------------------|
| CAS MARPAT / STN | Generic structure prior-art search | 전문 curated Markush index, 실무 검색체계          | 검색 단계에 사용 가능   | proprietary, coverage 범위 존재, hit≠법률판단 [3]                        |
| MarkushGlyph     | 전체 Markush 이미지 → CXSMILES 계열       | 전체 이미지를 직접 처리, 2026 공개 benchmark 보고된 최고 수준 | 중간             | 공개 benchmark exact 성능이 아직 60%대, substituent semantics 평가 미완 [19] |
| MarkushGrap      | multimodal                         | OCR+vision+te                              | 중간             | 저자도 자동 대규모 처리를                                                   |

| 기술/도구                   | 주 용도                               | 장점                                                     | 검토 단계에서의 활용 범위            | 핵심 제한                                                                                         |
|-------------------------|------------------------------------|--------------------------------------------------------|---------------------------|-----------------------------------------------------------------------------------------------|
| her-2                   | Markush recognition                | xt+layout fusion, IP5-M 공개                             |                           | 아직 불충분하다고 평가 [21]                                                                             |
| MarkitS                 | Markush image→SMILES               | 특화 correction, 최근 85~90% 보고                            | 중상 — topology 추출          | claim prose·법률범위 전체 모델은 아님 [2]                                                                |
| MolScribe               | 일반 구조식 image→graph                 | atom/bond alignment 와 confidence                       | 일반식 높음 / Markush 낮음       | R-table·condition semantics 가 별도 필요 [17]                                                      |
| MolNexTR                | 일반 구조식 OCSR                        | local/global vision, symbolic chemistry                | 일반식 높음 / Markush 낮음~중간    | generic claim logic 까지 직접 모델링하지 않음 [18]                                                       |
| OPSIN                   | systematic chemical name→structure | grammar 기반, 원 논문에서 precision 99.8%+, recall 98.7~99.2% | 이름 해석 높음                  | 복합 Markush prose·비체계명·법률조건과는 별개 [22]                                                          |
| RDKit MolStandardize    | discrete structure 정규화             | open, tautomer/charge/fragment 처리                      | 매우 높음 — discrete layer    | generic Markush canonicalizer 로는 부족 [23]                                                      |
| Standard InChI          | discrete chemical identifier       | DB interoperability 에 매우 유용                            | discrete 높음 / Markush 부적합 | 현재 Markush 미지원 [4]                                                                            |
| GNN/chemical embeddings | prior-art reranking·similarity     | 비정형 유사성 탐지 가능                                          | 중간 — retrieval only       | 높은 similarity 가 anticipation/obviousness 를 뜻하지 않음<br>구조 정확성·stereo·R constraint 에서 잘못된 생성 위험; |
| General-purpose VLM     | fallback OCR·context understanding | text+image 범용성                                         | 낮음~중간                     | chemistry-specialized model 필요 [21]                                                           |

Chemical name parsing 에는 OPSIN 류 deterministic parser 를 우선 적용하고, 실패한 이름·abbreviation·Markush prose 만 LLM 이 해석하도록 해야 한다. OPSIN 은 systematic organic nomenclature 에 매우 높은 정확도를 보여주지만 원 연구의 평가대상과 특허 Markush 문언은 동일하지 않으므로, “OPSIN 성공 = claim scope 확정”으로 취급해서는 안 된다. [22]

실제 multimodal 시스템이 오류가 발생하기 쉬운 부분은 단순 atom recognition보다 다음의 cross-modal binding이다. 예컨대 R<sup>1</sup>, R<sup>1</sup>, R1, R1의 OCR 혼동, 구조 그림의 R1과 몇 문단 뒤의 definition 연결, “R1 and R2 together with the atoms to which they are attached form ...” 같은 공동고리 조건, “independently selected”, “provided that”, “except”, “optionally substituted”, 위치가변결합, 반복단위, salt/solvate/stereoisomer/tautomer 범위가 그것이다. MarkushGlyph 연구도 variable definitions가 prose-조건문·상호참조로 표현되고 patent office와 시대별 drawing convention이 다르다고 지적한다. [1]

한국 특허 데이터에서도 기반자료는 상당 부분 마련할 수 있다. KIPRIS Plus는 현재 산업재산권 정보를 Bulk/Open API 방식으로 제공하며, 공개·등록공보에서 전문과 대표도면을 제공한다. [24] 청구항 변동이력은 1999년 이후를 제공하고 2026년 8월 기준 약 344만 건 규모이며, 표준화된 특허·실용 인용문헌도 별도 API/BULK로 제공된다. [25] 이는 한국어 청구항 이력, 선행기술 인용, 이미지 데이터를 결합한 국내 마쿠쉬 벤치마크를 구축하기에 기초자료로 활용할 수 있다.

또한 대표적인 국내 특허문헌 KR102829082B1은 구조식 이미지에서 atom region과 bonding relation을 인식하고 adjacency matrix를 만든 뒤 문자열 포맷을 생성하는 AI 구조를 청구한다. 이는 image→atom/bond graph→serialization이라는 OCSR 파이프라인이 국내 특허기술로도 구체화됐음을 보여주는 사례다. [26]

## 4. 마쿠쉬 청구항의 인공지능 보조 평가체계

시스템은 언어모델의 단일 응답을 법적 결론으로 취급하지 않는다. 도면과 문언의 추출, 제약관계의 정규화, 선행기술 검색, 원문 대조, 법률요건별 증거표 작성, 심사관 검증을 분리한다. 각 단계는 수정 전 원본과 변경 근거를 보존한다. [36]

처리 흐름은 원문 수집 → 화학구조 및 치환기 추출 → PMIR 변환 → 검색식 구성 → 문헌별 증거 대조 → 신규성·진보성 검토 → 심사관 확인으로 구성한다.

### 4.1. 특허 마쿠쉬 중간표현

PMIR(Patent Markush Intermediate Representation)은 고정 기본 구조, 변수별 허용집합·제외집합, 변수 간 논리적 제약, 변형 범위, 실시예, 청구항·도면 위치와 추출 이력을 함께 저장하는 중간표현으로 정의한다. 아래의 자료구조는 구현 예시이다.

```
ClaimMarkush
├─ fixedScaffoldGraph
│  └─ atoms / bonds / aromaticity
│  └─ stereochemistry
│  └─ attachmentSites
├─ variables
│  └─ R1: allowed classes / explicit members / exclusions
│  └─ R2: ...
│  └─ positional attachment alternatives
├─ relationalConstraints
│  └─ independentlySelected
│  └─ togetherFormRing(R1,R2)
│  └─ atLeastOne(...)
│  └─ ifThen(...)
│  └─ excludedCombination(...)
├─ rangeConstraints
│  └─ carbon-count
│  └─ repeat-unit
│  └─ substitution count
├─ scopeModifiers
│  └─ salt / solvate / hydrate
│  └─ stereoisomer / tautomer
│  └─ optional substitution
└─ provenance
   └─ claim / paragraph
```

└ image bounding box  
└ OCR/name-parser/model version  
└ confidence / human verification

원본과 canonical representation 을 동시에 보존해야 한다. 구조정규화 단계에서 aromaticity, charge, tautomer, salt stripping 등을 변경하면 검색 recall 은 좋아질 수 있으나 법률적으로 원 disclosure 와 달라질 수 있기 때문이다. PubChem 의 Substance→standardized Compound 구조가 이 차이를 잘 보여준다. [27]

## 4.2. 제약 기반 선행기술 검색

검색은 일반식의 제약 대조, 선택적 개별 화합물 생성, 선행 특정 화합물의 청구범위 포함 검사라는 세 방법을 병행한다.

Symbolic generic matching 에서는 scaffold 와 variable constraints 를 직접 subgraph/constraint problem 으로 비교한다. 빠른 fingerprint/screening 단계 후 exact graph matching 을 하며,  $R1 \in \text{class}$ , positional variation, repeat range, inter-variable relationship 을 별도 constraint solver 가 확인한다. 전통적 generic structure retrieval 도 screening 과 정밀 atom-bond matching 을 단계화해 발전해 왔다는 점에서 구조적으로 타당하다. [28]

Lazy enumeration 은 전체 family 를 만들지 않고, 검색에 필요한 대표 species 만 생성한다. exemplified members, prior-art hit 와 가장 가까운 boundary members, claim constraint 를 차별화하는 counterexample 을 우선 생성한다.

Discrete reverse containment 에서는 선행기술의 특정 compound 가 claim 의 Markush constraints 를 만족하는지 solver 로 검사한다. 이는 USPTO 의 “species may anticipate genus” 논리와 직접 맞닿아 있다. [9]

## 4.3. 신규성의 문헌별 증거표

각 선행문헌을 별도로 평가하고 청구항의 필수 구성요소마다 개시 여부와 원문 위치를 표시한다. 표 4는 가상의 예시이며, 수치는 실측 결과가 아니다.

| 청구항 요소                  | D1의 개시      | 대응 유형               | 근거 위치        | 예시 신뢰도 |
|-------------------------|-------------|---------------------|--------------|--------|
| core scaffold           | Yes         | exact graph         | Fig./example | 0.99   |
| R1 = substituted alkyl  | Yes         | generic containment | claim 4      | 0.94   |
| R2 positional variation | uncertain   | inferred            | paragraph x  | 0.61   |
| stereochemistry         | No          | —                   | —            | 0.99   |
| enabling preparation    | Yes/Unknown | reaction/example    | Example 12   | 0.82   |

구조 유사성이 높더라도 동일 문헌의 개시와 실시 가능성이 확인되지 않으면 신규성 상실을 단정하지 않는다. 부족한 증거는 미확인 상태로 보존하고, 해당 부분을 심사관의 재검토 대상으로 표시한다. [5]

## 4.4. 진보성의 증거 구성

진보성 검토를 위해 가장 가까운 선행기술 후보, 청구항과의 차이, 관련 기술효과, 다른 문헌을 결합할 이유 및 성공 기대에 관한 증거를 연결한다. 검색과 증거 경렬 과정은 자동화할 수 있으나 법적 평가는 해당 법역의 기준과 심사관의 검증에 따른다. [6]

GNN과 molecular embedding은 이 과정에서 D2 후보 reranking에는 가치가 있다. 그러나 embedding distance 자체를 “통상의 기술자가 그 치환기를 선택했을 것”의 증거로 사용해서는 안 된다. PCT는 진보성 평가를 발명 전체와 통상의 기술자 관점에서 수행하며, 복수문헌 결합 여부도 그 관점에서 평가한다. [6]

#### 4.5. 검색 기록과 심사관 검증

각 검색식의 목적과 확대 조건, 사용 데이터베이스, 실행시점, 결과의 출처를 검색 패키지로 기록한다. 다음 JSON은 실제 서비스 호출이 아닌 기록 양식의 예시이다.

```
{
  "claim_id": "KR-...-C1",
  "pmir_version": "1.3",
  "query_family": [
    {
      "purpose": "strict scaffold",
      "database": "MARPAT",
      "mode": "substructure",
      "assumptions": []
    },
    {
      "purpose": "R1 generalized",
      "database": "MARPAT",
      "mode": "substructure",
      "assumptions": ["R1 class relaxed one level"]
    },
    {
      "purpose": "exemplified compounds",
      "database": "REGISTRY",
      "mode": "exact/substructure"
    }
  ]
}
```

실제 STN 자동연동 범위는 기관의 라이선스·제공 API/batch interface에 맞춰야 하며, 직접 API가 허용되지 않는 경우에는 구조파일/검색식 package export + 심사관 실행 + result import 만으로도 가치가 있다. 중요한 것은 검색식, query intent, relaxation, DB, 실행시점 및 hit provenance가 모두 남는 것이다. PCT 역시 chemical structure query와 database를 포함한 search history 기록을 권장한다. [13]

심사관 화면은 청구항·원본 도면, 청구범위 구조화 결과, 선행기술 및 법적 증거를 각각 대조할 수 있도록 구성한다(표 5).

| 화면                         | 주요 기능                                                                                                              |
|----------------------------|--------------------------------------------------------------------------------------------------------------------|
| Claim / Source             | 원 청구항·구조이미지, OCR bounding box, R-group definition highlight                                                        |
| Markush Scope              | scaffold graph, variable tree, constraint graph, canonicalization variants, confidence                             |
| Prior Art / Legal Evidence | STN hits, atom mapping, R-group containment, priority/family, evidence spans, novelty matrix, inventive-step graph |

심사관이 도면의 원자나 치환기 정의를 수정하면 그 수정과 이유를 사실 자료에 반영하고 관련 검색식을 갱신한다. 수정 전후의 범위 차이와 문헌별 판단 변화도 기록한다. [36]

## 5. 자료 구축 및 평가 설계

학습·평가에는 이용 조건을 충족하는 공개자료와 기관 보유자료를 사용하고, 상용 MARPAT 결과는 허용된 범위에서 대조 검색자료로 이용한다. 검색 서비스의 라이선스가 허용하지 않는 자료를 학습에 재사용하는 것은 본 설계의 전제가 아니다. 데이터원별 오류와 수록 범위를 분리해 평가한다. [12]

| 데이터원             | 데이터 유형                                               | 권장 용도                                              | 주의점                                                                                                            |
|------------------|------------------------------------------------------|----------------------------------------------------|----------------------------------------------------------------------------------------------------------------|
| KIPRIS Plus      | 한국 공개/등록공보, 전문·도면, claim history, citations, IPC/CPC | 한국어 Markush 평가자료, claim amendment study            | license/수수료·공개범위 확인 필요 [25]                                                                                    |
| CAS MARPAT       | curated generic Markush                              | production prior-art 검색, gold retrieval comparison | proprietary, 일부 화학영역 제외, KR 2008+ 명시 [3]                                                                       |
| CAS REGISTRY     | discrete substances                                  | specific compound 검색·cross-check                   | proprietary [12]                                                                                               |
| SureChEMBL 2.0   | patent-extracted compounds/images                    | OCSR training, open chemical prior-art pre-search  | automated extraction error 를 검증된 정답으로 간주하면 안 됨. 최신 EBI 자료는 USPTO/EPO/WIPO 등 대규모 compound coverage 를 보고한다. [29] |
| PubChem          | discrete standardized compounds                      | canonicalization tests, augmentation, identifiers  | generic Markush 자체는 모델링하지 않음 [30]                                                                              |
| USPTO ODP        | patent bulk/public data                              | US patent text·metadata corpus                     | 2025 ODP 로 데이터 접근체계가 현대화됨. [31]                                                                                |
| WIPO PATENTSCOPE | PCT/global patent retrieval                          | multilingual text/family 보완검색                      | 2026 AI assistant 는 현재 keyword/date/country 중심 query construction 이며 화학 Markush 엔진은 아님. [32]                   |
| IP5-M            | 실제 IP5 Markush images                                | cross-jurisdiction OCSR benchmark                  | 공개분 878 cases 기준 비교 필요 [19]                                                                                    |
| M2S              | image+variable definition text                       | multimodal Markush benchmark                       | 103 cases 로 작아 별도 국내 gold 가 필요 [1]                                                                             |
| USPTO-Markush    | US Markush benchmark                                 | domain test                                        | 74 cases 로 규모 제한 [19]                                                                                          |

SureChEMBL의 초기 연구 자체도 자동 text/image mining 기반 chemical-patent DB는 유용하지만 정확성의 한계를 관리해야 한다는 점을 강조해 왔다. [33] 따라서 SureChEMBL에서 얻은 구조를 “선행문헌에 법률적으로 disclosure된 compound”라는 label로 바로 학습시키지 말고, 원 patent evidence span과 연결해 검증해야 한다.

### 5.1. 한국어 마쿠쉬 평가자료

한국어 청구항과 도면을 연결한 평가자료(K-MARKUSH-EXAM)를 구축한다. 초기 목표는 1,000~2,000개 청구항 패밀리이며 실제 확보량을 뜻하지 않는다. 동일 패밀리의 출원 간 중복을 제거하고 우선일을 기준으로 시점을 분리하여 평가자료의 누출을 줄인다. [36]

각 gold case에는 단순 SMILES 정답이 아니라 다음 annotation이 필요하다.

| Gold annotation                                 | 목적                                |
|-------------------------------------------------|-----------------------------------|
| 원 image crop + bounding boxes                   | OCR/OCSR 오류 위치 추적                 |
| exact scaffold graph                            | 기본 connectivity                   |
| stereo annotations                              | stereochemical novelty            |
| R-group labels + textual definitions            | image-text binding                |
| normalized chemical classes                     | alkyl/aryl/heterocycle 등 ontology |
| inter-variable constraints                      | 독립/공동고리/조건부/제외                    |
| exemplified species                             | enumeration/retrieval 검증          |
| claim amendment history                         | scope 변화 인식                       |
| examiner-cited prior art                        | retrieval baseline                |
| independent expert prior art                    | citation bias 보정                  |
| novelty element map                             | legal evaluation                  |
| inventive-step distinguishing features/evidence | reasoning evaluation              |
| ambiguity/UNKNOWN label                         | 강제판단 방지                           |

### 5.2. 평가 지표 및 사전 정의한 목표

영상 인식의 구조 일치뿐 아니라 치환기 정의와 변수 간 제약의 의미 일치를 검증한다. 표 8의 수치 조건은 실험으로 확인해야 할 제안 목표치이며 검증 완료된 성능이 아니다. [1]

| 평가층                  | 주요 지표                                                       | 사전 설정할 검증 목표의 예                                  |
|----------------------|-------------------------------------------------------------|--------------------------------------------------|
| Figure detection/OCR | bbox F1, label CER, R-label accuracy                        | R-label accuracy $\geq 99\%$ 목표                  |
| Discrete OCSR        | full stereo exact graph equality, validity                  | human-verified production path 기준 $\geq 98\%$ 목표 |
| Markush scaffold     | strict parsed-graph equality                                | 독립 공개 benchmark + KR benchmark                   |
| Variable semantics   | R-definition exact/F1, class normalization accuracy         | 핵심 R-definition $\geq 97-98\%$                   |
| Constraint semantics | positional/repeat/if-then/ring constraints precision-recall | critical constraint miss 최소화                     |

| 평가층                     | 주요 지표                                                                                        | 사전 설정할 검증 목표의 예                                |
|-------------------------|----------------------------------------------------------------------------------------------|------------------------------------------------|
| Membership test         | compound $\in$ claim family<br>precision/recall                                              | precision/recall $\geq 99\%$<br>gold subset 목표 |
| Prior-art retrieval     | Recall@50/100/200,<br>nDCG@20, MRR                                                           | 특히 known critical art<br>Recall 중심             |
| Novelty evidence        | claim-element coverage<br>F1, enablement evidence<br>precision                               | false anticipation 최소화                         |
| Inventive-step evidence | closest-art recall,<br>distinguishing-feature F1,<br>motivation/effect evidence<br>precision | 법률결론 정확도보다<br>evidence completeness<br>우선      |
| Calibration             | ECE, Brier score, risk-<br>coverage curve                                                    | 저신뢰가 실제 오류와<br>상관해야 함                          |
| Human workflow          | review time, correction<br>count, critical miss,<br>override rate                            | baseline 대비 통계적 개선                             |
| Auditability            | evidence provenance,<br>query reproducibility                                                | 공식 판단 후보는 100% 근거<br>연결(evidence-linked)       |

출처 연결 비율, 존재하지 않는 인용의 발생률, 핵심 선행기술의 재현율, 심사관 검토 시간 등을 품질 지표로 둔다. 기준값은 실제 기관 자료의 난이도와 오류 비용을 확인한 후 설정해야 하며, 별도 연구에서 제안된 목표치를 본 연구의 달성 결과로 인용하지 않는다. [36]

### 5.3. 비교실험과 심사관 관찰평가

검색 비교실험에서는 전문가의 기존 STN 검색, 공개자료 기반 검색, AI 검색식 보조와 STN 결합, 결과 재정렬 단계를 동일한 사건에 적용한다. 중요한 선행기술의 재현율과 검토 시간, 검색 누락을 비교한다.

Representation experiment에서는 canonical SMILES-only, CXSMILES 계열, PMIR constraint graph 세 방식을 비교한다. 동일한 1만 개의 query compound에 대해 claim membership 판정 precision/recall과 표현 round-trip loss를 측정하면 PMIR의 필요성을 직접 검증할 수 있다.

OCSR ensemble experiment에서는 MarkushGlyph, MarkitS, MolScribe/MolNexTR 계열의 단독 결과와 ensemble disagreement를 비교한다. 모델들이 동일한 graph를 산출한 경우와 불일치한 경우의 실측 오류율을 구하면, agreement confidence를 examiner routing에 사용할 수 있다. 관련 연구에서도 여러 OCSR 결과를 Tanimoto/MCS로 비교하고 사람이 선택하는 human-in-the-loop 방식이 활용되고 있다. [34]

Claim semantics experiment가 특히 중요하다. 이미지가 완벽히 인식된 상태에서 R-definition text만 주고 PMIR constraints를 생성하게 하여, OCSR와 language interpretation 오류를 분리해야 한다. 그렇지 않으면 전체 end-to-end 정확도가 낮아도 원인을 알 수 없다.

법률 판단의 경계 사례에는 일반식의 선행 개시와 선택된 종, 특정 종과 속 청구항, 실시 불가능한 개시, 입체이성질체 및 염·호변이성질체, 치환기 간 종속관계가 누락된 경우를 포함한다. 사건별로 적용 법역과 관련 원문을 기록하여 평가자의 판단을 대조한다. [5][9]

심사관 참여 관찰평가는 30~60 명, 8~12 주를 잠정 설계치로 하며 실제 모집 규모와 기간은 기관 협의에 따라 정한다. 기존 검색과 보조 시스템을 비교하되, 자료 난이도·심사관 경험을 통제하고 누락률, 검토 시간, 수정 건수와 신뢰도 판단을 측정한다. 자동 생성 결과는 공식 처분에 직접 반영하지 않는다. [36]

USPTO 가 실제 AI-assisted prior-art search 를 도입하는 방식도 같은 방향이다. SimSearch 는 application text 를 이용해 ranked prior art 를 제공하지만 examiner 의 재량을 유지하고 기존 search tool 을 대체하지 않는다고 명시한다. [35] WIPO 의 2026 PATENTSCOPE AI-Assisted Search 도 자연어를 structured query 로 바꾸는 검색 보조에 초점을 두며, 모델을 WIPO premises 에서 운영해 query confidentiality 를 유지한다고 설명한다. [32] 이는 “AI 가 심사판정을 대신하기보다 query construction 과 evidence retrieval 을 고도화한다”는 본 설계와 일치한다.

## 6. 단계별 구현 및 검증

구현은 청구범위의 손실 없는 표현과 평가자료 구축을 우선한 뒤, 구조식 인식과 제약 검증, 검색식 생성, 증거 연결의 순서로 진행한다. 표 9 는 제안 일정과 단계별 검증 조건을 정리한 것이다.

| 단계                       | 기간       | 최우선 산출물                                                                                         | Go/No-Go 기준                                                         |
|--------------------------|----------|-------------------------------------------------------------------------------------------------|---------------------------------------------------------------------|
| Foundation               | 0-3 개월   | PMIR schema,<br>claim/image<br>provenance,<br>discrete<br>normalization, STN<br>search protocol | 100 개 gold claims<br>round-trip review                              |
| Parsing MVP              | 3-6 개월   | Markush OCSR<br>ensemble,<br>chemical OCR,<br>OPSIN, R-definition<br>parser                     | 500 claims 에서<br>error taxonomy 와<br>calibrated<br>confidence 확보    |
| Search MVP               | 6-10 개월  | REGISTRY/<br>MARPAT query<br>planner, lazy<br>enumeration,<br>KIPRIS/SureChEMB<br>L pre-search  | STN expert<br>baseline 대비<br>critical-art recall<br>non-inferiority |
| Legal Evidence<br>MVP    | 9-14 개월  | novelty matrix,<br>enablement<br>evidence,<br>inventive-step<br>evidence graph,<br>PEQ rules    | 모든 공식 판단<br>후보에 provenance<br>100% 기록                               |
| Shadow Pilot             | 14-20 개월 | examiner UI,<br>correction loop,<br>shadow mode,<br>regression suite                            | critical misses 악화<br>없음, review-time<br>유의미 감소                     |
| Controlled<br>Production | 20-26 개월 | approved case<br>classes 에서<br>assistive<br>deployment                                          | error/abstention<br>threshold 충족                                    |
| Scale & Rule             | 26-30 개월 | active learning,                                                                                | 연속 audit 와                                                          |

| 단계        | 기간 | 최우선 산출물                                                         | Go/No-Go 기준     |
|-----------|----|-----------------------------------------------------------------|-----------------|
| Evolution |    | Korean gold expansion, rule evolution, institutional governance | release gate 통과 |

각 개발 단계에서 유지해야 할 원칙은 다음과 같다.

첫째, 단일 화합물의 구조 표준화와 마쿠쉬 청구항의 제약표현을 분리한다. InChI/InChIKey 는 특정 화합물의 식별에 유용하지만, 마쿠쉬 일반식의 전체 범위를 나타내는 식별자로 사용할 수 없다. [4]

둘째, 신규성은 단일 문헌의 개시와 실시 가능성을, 진보성은 차이점·기술효과·변형 동기에 관한 복수의 증거를 구분해 다룬다. 관할권에 따른 차이는 규칙의 적용 범위와 함께 기록한다. [5][6]

셋째, 허용되는 범위에서 MARPAT 등 전문 일반식 색인을 검색 채널로 활용한다. AI 는 청구항을 검색 가능한 구조와 조건으로 변환하고, 실행 이력 및 검토 근거를 심사관에게 제공한다. [3]

넷째, 보고된 구조식 인식 성능을 실제 심사 성능으로 확대 해석하지 않는다. 오류 가능성이 높은 도면·치환기 조건은 검토 대상으로 지정하고, 신뢰도가 낮거나 근거가 부족한 법률 결론은 유보한다. [19][2]

다섯째, 사실의 출처·버전·수정 내역을 저장하고 규칙 평가와 사람의 판단을 연결한다. 데이터와 법적 기준이 달라지면 이전 사건을 다시 평가할 수 있도록 모델·규칙·검색식의 버전을 함께 보존한다. [36]

## 7. 결론

본 연구는 마쿠쉬 청구항을 도면, 정의문과 제약관계를 함께 표현한 뒤, 특정 화합물 및 일반식 검색결과를 청구항 요소별 증거에 연결하는 심사 보조체계를 제안했다. 신규성과 진보성의 요건을 별도로 다루고 원문 추적과 판단 유보 기능을 두면, 검색결과를 법적 결론으로 잘못 치환하는 위험을 줄일 수 있다. 다만 제안한 PMIR 및 평가 절차는 아직 실제 심사자료에서 검증되지 않았다. 향후 한국어 청구항 평가자료를 구축하고 전문가 검색 대조군 및 심사관 관찰평가를 통해 선행기술 누락률, 범위 해석 오류, 검토 시간의 변화를 측정해야 한다.

## 참고문헌

- [1] MarkushGlyph and OCSRGLyph: Improved Chemical Structure Recognition <https://arxiv.org/abs/2607.28532>
- [2] MarkitS: an image-to-SMILES parsing workflow for Markush structures | Journal of Cheminformatics | Springer Nature Link <https://link.springer.com/article/10.1186/s13321-026-01301-7>
- [3] Markush - MARPAT | CAS <https://www.cas.org/training/documentation/markush>
- [4] InChI Technical FAQ - InChI Trust <https://www.inchi-trust.org/technical-faq/>
- [5] PCT International Search and Preliminary Examination Guidelines [https://www.wipo.int/en/web/pct-system/texts/ispe/12\\_01\\_02](https://www.wipo.int/en/web/pct-system/texts/ispe/12_01_02)
- [6] PCT International Search and Preliminary Examination Guidelines [https://www.wipo.int/en/web/pct-system/texts/ispe/13\\_01\\_02](https://www.wipo.int/en/web/pct-system/texts/ispe/13_01_02)
- [7] 3. Obvious selection? [https://www.epo.org/en/legal/guidelines-epc/2026/g\\_viia\\_3.html](https://www.epo.org/en/legal/guidelines-epc/2026/g_viia_3.html)
- [8] 3.2.5 Markush-Gruppe (Alternativlösungen in einem einzigen Anspruch) [https://www.epo.org/de/legal/guidelines-epc/2026/f\\_v\\_3\\_2\\_5.html](https://www.epo.org/de/legal/guidelines-epc/2026/f_v_3_2_5.html)
- [9] 2131-Anticipation — Application of 35 U.S.C. 102 <https://www.uspto.gov/web/offices/pac/mpep/s2131.html>
- [10] The Chemical Abstract's Service generic chemical (Markush) structure storage and retrieval capability. 1. Basic concepts | Journal of Chemical Information and Modeling | ACS Publications <https://pubs.acs.org/doi/10.1021/ci00066a008>
- [11] Markush Structure Patent Search [https://help.patentexplorer.cas.org/cur/Searching\\_in\\_Patent\\_Explorer/Patent\\_Search/Markush\\_Structure\\_Patent\\_Search.htm](https://help.patentexplorer.cas.org/cur/Searching_in_Patent_Explorer/Patent_Search/Markush_Structure_Patent_Search.htm)
- [12] CAS Databases Documentation | Chemical Information Resources <https://www.cas.org/training/documentation/cas-databases>

- [13] PCT International Search and Preliminary Examination Guidelines  
[https://www.wipo.int/en/web/pct-system/texts/ispe/15\\_52\\_62](https://www.wipo.int/en/web/pct-system/texts/ispe/15_52_62)
- [14] InChI 1.07 available on GitHub - IUPAC | International Union of Pure and Applied Chemistry <https://iupac.org/inchi-1-07-available-on-github/>
- [15] PubChem chemical structure standardization - PMC <https://pmc.ncbi.nlm.nih.gov/articles/PMC6086778/>
- [16] RDKit: RDKit::MolStandardize::TautomerEnumerator Class Reference  
[https://rdkit.org/new\\_docs/cppapi/classRDKit\\_1\\_1MolStandardize\\_1\\_1TautomerEnumerator.html](https://rdkit.org/new_docs/cppapi/classRDKit_1_1MolStandardize_1_1TautomerEnumerator.html)
- [17] MolScribe: Robust Molecular Structure Recognition with Image-to-Graph Generation | Journal of Chemical Information and Modeling | ACS Publications <https://doi.org/10.1021%2Facs.jcim.2c01480>
- [18] MolNexTR: A Generalized Deep Learning Model for Molecular Image Recognition <https://arxiv.org/abs/2403.03691>
- [19] GitHub - EdisonScientific/glyph: Glyph: optical chemical structure recognition (OCSRGlyph) and Markush structure recognition (MarkushGlyph) · GitHub <https://github.com/EdisonScientific/glyph>
- [20] MarkushGrapher: Joint Visual and Textual Recognition of Markush Structures <https://arxiv.org/abs/2503.16096>
- [21] MarkushGrapher-2: End-to-end Multimodal Recognition of Chemical Structures <https://arxiv.org/abs/2603.28550>
- [22] Chemical Name to Structure: OPSIN, an Open Source Solution | Journal of Chemical Information and Modeling | ACS Publications <https://pubs.acs.org/doi/abs/10.1021/ci100384d>
- [23] rdkit.Chem.MolStandardize.rdmolstandardize module — The RDKit 2026.03.6 documentation  
<https://www.rdkit.org/docs/source/rdkit.Chem.MolStandardize.rdmolstandardize.html>
- [24] 지식재산정보 활용 서비스 <https://plus.kipris.or.kr/portal/main.do>
- [25] KIPRIS Plus [https://plus.kipris.or.kr/portal/data/clasList/DBII\\_000000000000327/view.do](https://plus.kipris.or.kr/portal/data/clasList/DBII_000000000000327/view.do)
- [26] KR102829082B1 - An artificial intelligence apparatus for recognizing structural formula image and method for the same - Google Patents <https://patents.google.com/patent/KR102829082B1/en>
- [27] Data Organization - PubChem <https://pubchem.ncbi.nlm.nih.gov/docs/data-organization>
- [28] Computer storage and retrieval of generic chemical structures in patents. 5. Algorithmic generation of fragment descriptors for generic structure screening | Journal of Chemical Information and Modeling  
<https://pubs.acs.org/doi/abs/10.1021/ci00042a003>
- [29] SureChEMBL 2.0 is here | EMBL-EBI <https://www.ebi.ac.uk/about/news/updates-from-data-resources/surechembl-2-0-is-here/>
- [30] Structures - PubChem <https://pubchem.ncbi.nlm.nih.gov/docs/structures>
- [31] USPTO launches new Open Data Portal for easy, quick access to data | USPTO <https://www.uspto.gov/about-us/news-updates/uspto-launches-new-open-data-portal-easy-quick-access-data>
- [32] PATENTSCOPE AI-Assisted Search Now Available <https://www.wipo.int/en/web/patentscope/w/news/2026/patentscope-ai-assisted-search-now-available>
- [33] Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents | Journal of Cheminformatics | Springer Nature Link <https://link.springer.com/article/10.1186/s13321-015-0097-z>
- [34] MARCUS: molecular annotation and recognition for curating unravelled structures | Digital Discovery | The Royal Society of Chemistry <https://doi.org/10.1039/D5DD00031J>
- [35] Another USPTO AI-assisted examination tool ready for prime time | USPTO <https://www.uspto.gov/subscription-center/2025/another-uspto-ai-assisted-examination-tool-ready-prime-time>
- [36] Jun, E(on). 생성형 AI 기반 지식재산 심사품질 진단 고도화 연구·구축안. CAIPEX, 2026.9. (관련 설계 연구)
- [37] European Patent Office. Guidelines for Examination in the European Patent Office, G-VI, 7, “Selection inventions” (2026). [https://www.epo.org/en/legal/guidelines-epc/2026/g\\_vi\\_7.html](https://www.epo.org/en/legal/guidelines-epc/2026/g_vi_7.html)