Budapesti Műszaki és Gazdaságtudományi Egyetem
Villamosmérnöki és Informatikai Kar
Méréstechnika és Információs Rendszerek Tanszék

Moduláris RoBERTa alapú architektúra heterogén orvosbiológiai adatfúzióra

TDK dolgozat

Készítette:

Marosi Márk
Antal Mátyás

Konzulens:

dr. Antal Péter
dr. Juhász Gabriella

2023

Tartalomjegyzék

Kivonat

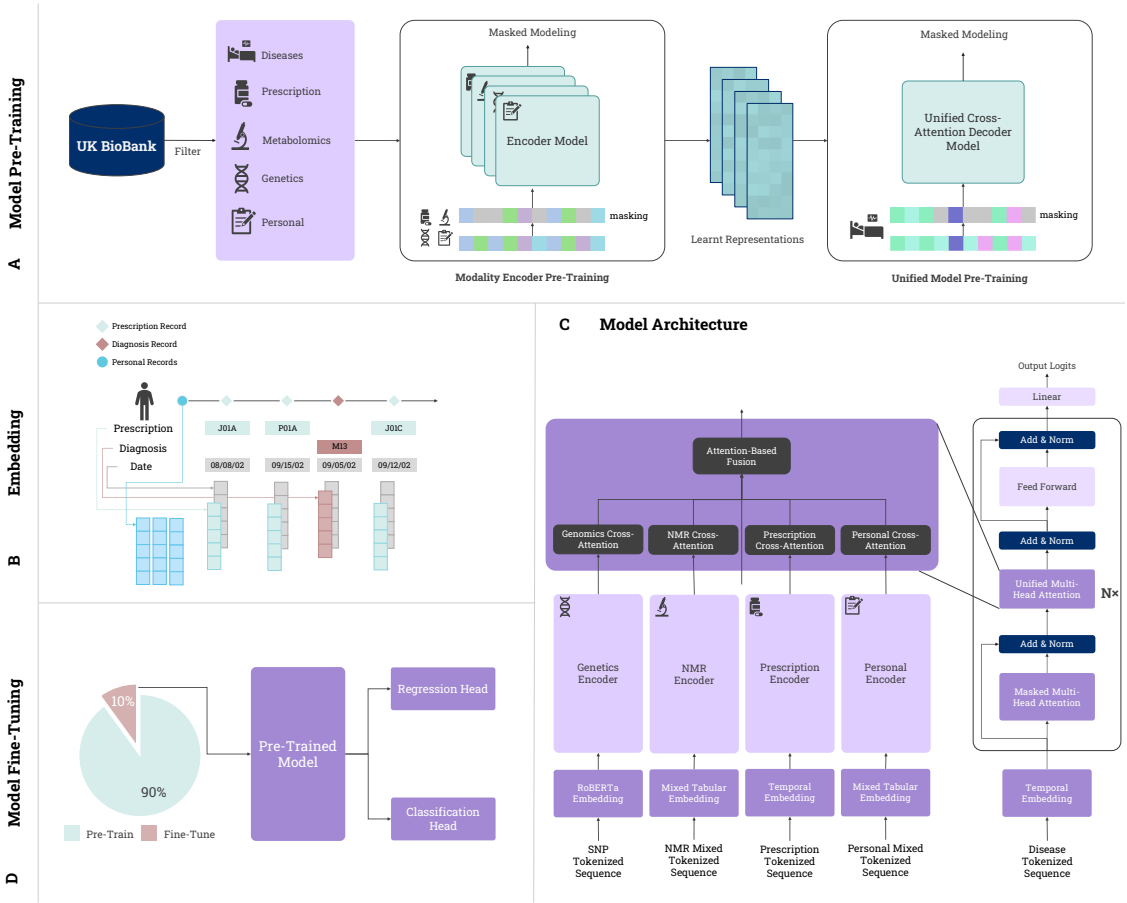
A transzformerek új megközelítést kínálnak a nagyméretű, különböző modalitásokban és omikai területeken elterjedt orvosbiológiai adatok egységesítésére. A multimodális adatokhoz egy új moduláris architektúrát javaslunk, amely robusztus mechanizmust kínál a hiányzó információk kezelésére. A modellt genetikai, demográfiai, laboratóriumi, diagnosztikai és gyógyszerfelírási adatok felhasználásával értékeljük a depresszióval kapcsolatos multimorbiditásra és polifarmáciára összpontosítva. Megvizsgáljuk a harmonizált és moduláris reprezentációkat, különös tekintettel a betegek klaszterezésére. Továbbá kiértékeljük a megtanult reprezentációk felhasználását jövőbeni megbetegségek és gyógyszerfogyasztás előrejelzésére, összehasonlítva hagyományos gépi tanulási módszerekkel.

Abstract

Transformers offer a new approach to consolidating biomedical data spread across large sizes, various modalities, and omics domains. We propose a new modular architecture for multimodal data, providing a robust mechanism to handle missing information. We evaluate the model using genetic, demographic, laboratory, diagnostic, and drug prescription data, focusing on depression-related multimorbidity and polypharmacy. We examine the harmonized and modular representations, with a special emphasis on clustering patients. Furthermore, we assess the utilization of the learned representations for predicting future illnesses and drug consumption, comparing it to traditional machine learning methods.

1. Bevezetés

Az élettudományokban alapvető fontosságú az információk integrációja, különösen a különböző tárgyterületeket lefedő adatbázisokban, melyek célja a multimorbiditások mögött rejlő közös molekuláris tényezők feltérképezése. Az elektronikus betegnyilvántartások és a biobankok egyre növekvő elérhetősége példátlan mennyiségű heterogén adatot biztosít a látens betegreprezentációk tanulásához, a betegek sztratifikálásához és a betegségek altípusainak meghatározásához. Ugyanakkor az idősori adatok, mint például a rendszertelenül mintavételezett gyógyszerfelírási ('recept') adatok integrációja, valamint a szisztematikusan hiányzó modalitásokkal rendelkező multimodális adatfúzió még nyitott kérdés.



1.1. ábra. Az adatok, módszerek és architektúra áttekintése. (A) Tanítás: A modell modularitása nem csak architekturális szempontból, hanem a tanítási szakaszban is megnyilvánul. A különböző modalitások enkóderjeit a modalitáson szűrt adatkészleteken tanítjuk elő. Ezután az előtanított enkódereket használjuk az egységes modell betegségszekvenciákon tanítására. **(B) Beágyazás:** Két új réteget vezetünk be a dátum és a táblázatos adat beágyazásának megkönnyítésére. **(C) Architektúra:** A RoBERTa dekódert új, Egységes Kereszt-Figyelmi Rétegekkel egészítjük ki, amely a modalitások input szekvenciáját figyelem alapú fúzióval ötvözi. **(D) Finomhangolás:** Az integrált multimodális egységes reprezentáció hatékonyságának értékelésére adataink 10%-át használjuk fel, hogy összehasonlítsuk a modell teljesítményét a hagyományos gépi tanulási technikákkal szemben.

A transzformer architektúrák hatékony jelölteknek bizonyultak a multimodális adatfúzió és az idősori adatelemzés területén [Xu et al., 2023, Nerella et al., 2023, Wen et al., 2022, Chen, 2020]. A transzformer modell család az elmúlt években forradalmasította a reprezentáció tanulást, különösen mély tanulási környezetekben. Az architektúra hatékonysága és sokoldalúsága miatt számos feladatban sikerrel alkalmazták, nem csak a természetes nyelvfeldolgozásban, hanem az élettudományok területén is.

1.1. Biológiai adattárak

A biobankokban tárolt adatok mennyisége az elmúlt években rohamosan nőtt, amint az élettudományok és az orvostudomány egyre inkább felismerték a nagy mennyiségű adatok jelentőségét (lásd például UK Biobank [Bycroft et al., 2018]). Ezek az adattárak azonban igen heterogének és hiányosak, ami bár rugalmasságot ad az adatok gyűjtésében, egyben kihívások elé is állítja a kutatókat az adatok egységes elemzése szempontjából. A biobankokban tárolt adatok sokfélesége az adatok közötti összefüggések sokrétűségét követik és jelzik azoknak a módszereknek a fontosságát, amelyek képesek kezelni ezt a strukturálatlan és heterogén adatbázist. A multimodális, összekapcsolt információk általában különböző forrásból származnak és különböző formátumokban állnak rendelkezésre. Ezek a különböző modalitások - például genomikai adatok, proteomikai profilok, klinikai mérések vagy képalkotó adatok - hatalmas potenciált hordoznak magukban az integratív elemzés szempontjából [Acosta et al., 2022, Steyaert et al., 2023]. A megfelelő módszerek kiválasztása és alkalmazása, melyek képesek összekapcsolni és interpretálni ezeket az összetett adatrétegeket, alapvető jelentőségű a biobankokban tárolt információ teljes mértékű kiaknázása érdekében.

1.2. Kapcsolódó kutatások

A nagyméretű biobankokban elérhető betegségek együttelőfordulási mintázatainak, multimorbiditásának az elemzése lehetővé tette a résztvevők újfajta csoportosítását ('sztratifikációját') és a betegek klasztereinek közös molekuláris hátterének azonosítását a betegek reprezentációinak látens terében [Jensen et al., 2014, Prasad et al., 2022]. A UK Biobank (UKB) metabolomikai adatain alapuló látens reprezentációk kapcsán is sikerült megmutatni, hogy növelik a teljesítményt betegségek bekövetkezésére vonatkozó prediktív feladatokban [Buergel et al., 2022]. A nem-longitudinális adatokon kívül a gyógyszerfelírási idősori adatok további jelentős erőforrást képeznek, de alkalmazásukat gyakran akadályozza a heterogén kódolás, a korlátozott elérhetőség és a longitudinális jellege [Wu et al., 2019, Schwarz et al., 2022, Kiiskinen et al., 2023, Stroganov et al., 2022, Darke et al., 2022].

A transzformerek gyorsan elterjedtek a multimodális és idősori biomedikai adatok körében [Nerella et al., 2023]. A BEHRT modell 301 állapotot tudott előre jelezni a jövőbeli orvoslátogatásokhoz $1,6$ millió ember elektronikus egészségügyi adataira alapozva [Li et al., 2020]. Meng et al. [2021] a BERT modellt javasolja Multimodális EHR esetében (Bidirectional Representation Learning model with a Transformer architecture on Multimodal, BRLTM), amelyet a súlyos depresszió (Major Depressive Disorder, MDD) előrejelzésére alkalmaz, diagnózisok, kezelések, gyógyszerek, demográfiai információk és klinikai jegyzetek együttes kezelésével. A Med-BERT modell az $ICD-9$ és $ICD-10$ kódokkal strukturált bemenetek, a transzfer tanulás és a finomhangolás előnyeit $100-50000$ terjedő mintaszámok esetében mutatta be [Rasmy et al., 2021]. A Hi-BEHRT modell egy hierarchikus kiterjesztést tartalmaz a hosszú beteg trajektóriák kezelésére és a transzfer tanulási hatások növelésére [Li et al., 2022]. Yin et al. [2023] strukturált multimodális

tanulást javasolt a beteg-sztratifikáció elősegítésére. Az ExBEHRT modell tovább bővíti a felhasznált adatok halmazát laboratóriumi tesztekkel, integrál idősori mellékinformációkat is, és bemutatja a modell reprezentációk alkalmazhatóságát a betegség altípusainak és a kockázati csoportok azonosítására [Rupp et al., 2023]. A Hybrid Value-Aware Transformer (HVAT) modell lehetővé teszi a nem-longitudinális és kvantitatív longitudinális adatok közös tanulását [Shao et al., 2023]. Placido et al. [2023] összehasonlította a mély tanulási módszerek teljesítményét 6 millió beteggel a Dán Nemzeti Betegnyilvántartásból (Danish National Patient Registry, DNPR) a hasnyálmirigyrák kockázatának előrejelzésére betegség trajektóriákból, melyben a transzformerek a legjobb AUROC értékeket érték el. Zhou et al. [2023] kibővítette a multimodális tanulás hatókörét klinikai képekhez, szövegekhez, kvantitatív laboratóriumi tesztekhez és strukturált információkhoz.

A transzformerek idősori előrejelzésekben való alkalmazhatósága azonban több kérdést is felvet, lásd például [Zeng et al., 2023], illetve az MDD előrejelzése kapcsán [Zou et al., 2023]. A kiértékelések szabványosítása miatt megjelentek szabványos EHR referencia adathalmazok is, mint például a MIMIC adatbázis [Johnson et al., 2023]; illetve a leggyakrabban használt egészségügyi adathalmazon a UK Biobank-on is megjelentek már nagy nyelvi modellek (LLM) kiértékelései, lásd például [Belyaeva et al., 2023]. Végül az Attention-based cRoss-MOdal fUsion with contRast (ARMOUR) architektúrát említenénk, amely a rendszeresen hiányzó modalitások kezelésére lett tervezve [Liu et al., 2023] és több ami hasonlóságot mutat a javasolt modellünkkel.

1.3. Kutatás célja

A kutatás fő célkitűzése egy új architektúra kifejlesztése és bemutatása, amely képes hatékonyan kezelni a multimodális adatok sajátosságait és kihívásait. Ahogy az adattudomány és az élettudományok területe egyre inkább elmozdul a komplex és heterogén adatforrások felé, szükségessé válik olyan új eszközök és módszertanok kialakítása, amelyek képesek integrálni és értelmezni ezeket az információkat. Az általunk javasolt architektúra a multimodális adatokban rejlő potenciált és azt célozza, hogy a különböző adatformátumokat és -forrásokat egy egységes és koherens modellbe és koherens, de az értelmezhetőséget támogató modalitásonként dekomponált reprezentációba integrálja.

Ehhez több új módszert vezetünk be:

- **Moduláris, hierarchikus architektúra:** A modell több különböző részből épül fel. A modulok szemantikailag értelmezhetőek, és egymástól függetlenül vannak tanítva. A modulok együttes tanításához egy új figyelem alapú fúziós réteget vezetünk be.
- **Vegyes Táblázatos Beágyazás:** Az új beagyazási (Embedding) réteg lehetővé teszi, hogy a transzformerek keverve használjanak tokenizált és folytonos bemeneteket.

Ezen kívül, a javasolt modellt az UKB adathalmazon fogjuk előtanítani (pre-train), ami lehetővé teszi számunkra, hogy a kialakított architektúrát egy specifikus és jól definiált adatkészleten értékeljük.

- **A reprezentáció struktúrájának vizsgálata:** Célunk a kialakított reprezentációs modell mélyebb megértésének segítése és struktúrájának elemzése, hogy jobban feltárhatak legyenek az adatokban rejlő összefüggéseket.
- **Prediktív teljesítmény:** A reprezentációt különböző, akár több éves időtávú betegség és gyógyszerfelírás előrejelzésére fogjuk használni, majd az eredményeket összehasonlítjuk a hagyományos gépi tanulási módszerekkel. Ezen összehasonlítás révén célunk annak felmérése, hogy milyen mértékben nyújt többletet az új architektúra a jelenleg alkalmazott módszerekhez képest.

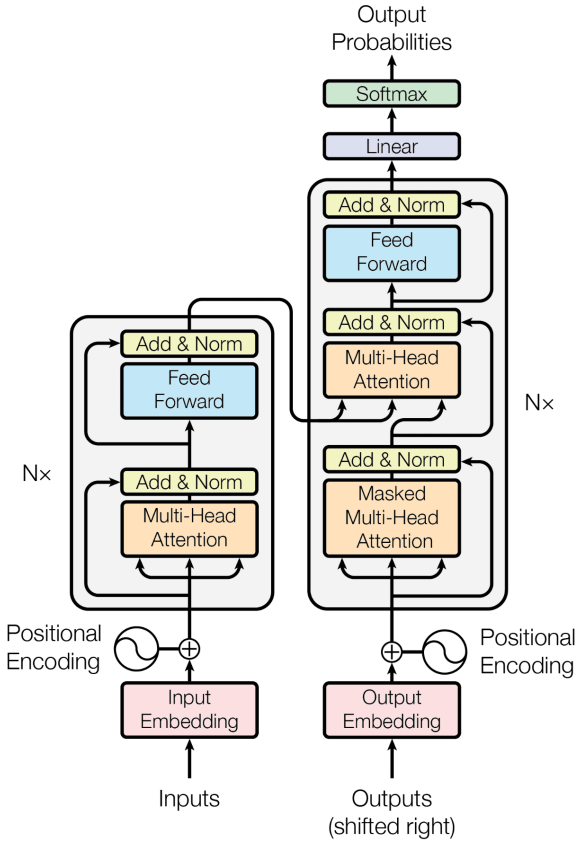
2. Elméleti háttér

A mély tanulás területén végzett kutatások az elmúlt évtizedben rohamos fejlődést mutattak [Si et al., 2021, Xie et al., 2022]. Az új modellek és architektúrák megjelenésével a gépi tanulás alkalmazási területei is kibővültek. Ebben a fejezetben az elméleti alapokat vizsgáljuk meg, amelyek a legújabb fejlesztések alapjául szolgálnak, különös tekintettel a transzformer architektúrára és annak alapvető komponenseire.

Áttekintést nyújtunk a figyelmi mechanizmusokról és a tokenizáló algoritmusokról, amelyek nélkülözhetetlenek a transzformer modellek hatékony működéséhez. Végül, de nem utolsósorban, beszélünk az előtanítás jelentőségéről és annak szerepéről a mély tanulási modellek teljesítményének javításában.

2.1. Transzformer architektúra

A transzformer (transformer) architektúra az "Attention is All You Need" [Vaswani et al., 2017] című cikkben került bemutatásra, és azóta az egyik legnépszerűbb és legjobban teljesítő architektúra az összetett természetes nyelvi szekvenciális feladatok, például a gépi fordítás, szöveg elemzés és szöveg generálás területén.



2.1. ábra. Az eredeti transzformer architektúra az "Attention is All You Need" cikkből. A kép a Vaswani et al. [2017] cikkből származik.

A transzformer architektúra két fő részből áll: egy enkóderből és egy dekóderből. Az enkóder és a dekóder különböző rétegekből állnak, amelyek speciálisan tervezett összetevők, például a többfejű figyelmi-mechanizmusok és a pozíció-specifikus előrecsatolt (feed-forward) hálózatok révén képesek a szekvencia-adatok kezelésére és feldolgozására.

Az enkóder rétegei közül mindegyik tartalmaz egy többfejű figyelmi-mechanizmust, amely a skálázott skaláris szorzaton alapul. Ezen mechanizmus lehetővé teszi a modell számára, hogy a bemeneti szekvencia különböző részeire összpontosítson, és megtanulja a különböző részek közötti összefüggéseket. A figyelmi-mechanizmus után egy pozíció-specifikus előrecsatolt hálózat következik, amely további feldolgozást végez a bemeneti adatokon, és segít azoknak a jellemzőknek a kinyerésében, amelyek segítenek a későbbi dekódolási folyamatban.

A dekóder, hasonlóan az enkóderhez, több rétegből áll, de egy további többfejű figyelmi-réteggel is rendelkezik. Ez a réteg kifejezetten az enkóder kimeneti jellemzőire összpontosít, és segít fúzionálni az enkóder és a dekóder reprezentációit a szekvencia különböző részei közötti kapcsolatok jobb megértése érdekében. A dekóder rétegei az enkóderben lévő rétegekhez hasonlóan tartalmaznak egy pozíció-specifikus előrecsatolt hálózatot is, amely segít a kimeneti szekvencia generálásában és finomításában.

2.1.1. Figyelmi-mechanizmusok

A figyelmi-mechanizmus (attention) lényegében súlyokat rendel az adatok bemeneti sorozatának minden eleméhez. Ezek a súlyok azt határozzák meg, hogy a modell mely részeire "koncentrál" a leginkább egy adott kimeneti lépés során.

A figyelmi-mechanizmus skálázott skaláris szorzat formájában kerül definiálásra:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

ahol:

- Q a lekérdezés mátrixa (query)
- K a kulcs mátrixa (key)
- V az érték mátrixa (value)
- d_k a kulcs dimenziója

Ezen egyenlet alábbi része adja a figyelmi-mátrixot:

$$\text{a}(Q, K) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

A multi-head attention további figyelmi-fejek kombinációjaként jelenik meg:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O$$

ahol minden egyes figyelmi-fej a következő:

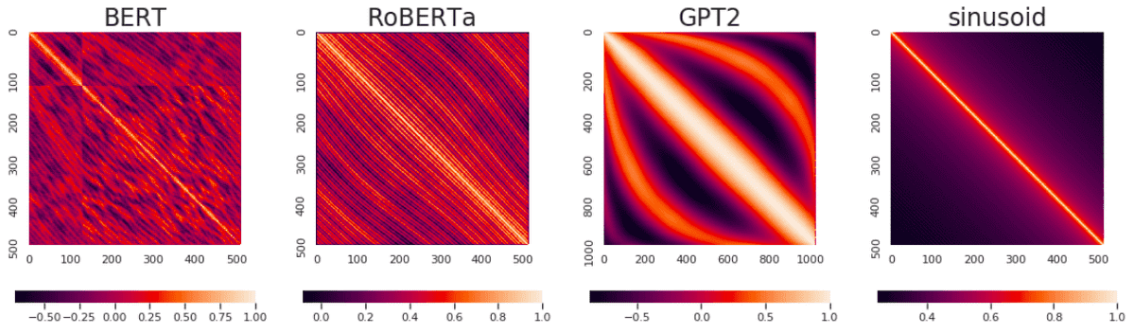
$$\text{head}_i = \text{Attention}(QW_{Qi}, KW_{Ki}, VW_{Vi})$$

2.1.2. Pozíciós kódolás

A transzformer modellek egyik központi eleme pozíciós információk hozzáadása a bemenethez. Az eredeti Transformer architektúrában szinuszos pozíciós kódolásokat használtak, amelyeket a modell bemenetéhez adtak hozzá, biztosítva ezzel, hogy a modell képes legyen a sorrendi információk észlelésére. Azonban a pozíciós beágyazások egy új megközelítést kínálnak, amely eltér a szinuszos kódolástól.

A pozíciós beágyazások nagyon hasonlóak a szóbeágyazásokhoz, de itt sorrendi információt ágyazzuk be. Minden egyes pozíció egy tanulható vektorhoz van rendelve, így a modell idővel képes finomítani ezen beágyazásokat a tanulási folyamat során.

Ezenkívül, míg a pozíciós kódolásokat a bemeneti szekvenciához adják hozzá, a pozíciós beágyazások közvetlenül a többfejű figyelmi-rétegeken belül is elhelyezhetők. Ez azt jelenti, hogy ezeket a beágyazásokat a modell több rétegén keresztül is elérhetjük, nem csak az elején. Ez különösen fontos a képfeldolgozási feladatoknál, ahol a strukturált adatok sorrendje kritikus jelentőségű.



2.2. ábra. A különböző pozíciós beágyazások pozíció szerinti koszinusz hasonlóságának vizualizációja. Az ábrán a világosabb területek nagyobb hasonlóságot jelölnek. A kép a Wang and Chen [2020] cikkből származik.

2.1.3. Tokenizáló algoritmusok

A mély tanulási modellek, különösen a transzformerek esetében, a bemeneti szekvenciák előfeldolgozása kritikus fontosságú. A transzformerek, melyeket gyakran használnak összetett szekvenciális feladatokhoz, speciális követelményeket támasztanak a bemeneti adatok formátumával és szerkezetével szemben. A tokenizálás az egyik kulcslépés, melynek során a szekvenciák kisebb, kezelhető egységekre, ún. tokenekre bontódnak. A tokenek egységes hosszúságú vektorokká konvertálódnak, amelyeket a modell könnyedén képes feldolgozni.

2.2. Előtanítás

Az előtanítás (Pre-training) mint fogalom az elmúlt években vált kulcsfontosságúvá a mély tanulás terén. Az előtanítás alapvetően egy mély neurális háló előzetes tanítását jelenti egy általános, nagy mennyiségű adathalmazon, mielőtt azt specifikus, sokszor kevesebb adattal rendelkező feladatokon finomítanánk. Ez a módszertan segít a modelleknek abban, hogy kialakítsanak és rögzítsenek bizonyos általános jellemzőket és struktúrákat, amelyek később hasznosak lehetnek a célirányos feladatokban.

Az előtanítási folyamatban részt vevő mély tanulásos modellek, különösen a transzformerek, millió, sőt milliárd paraméterrel rendelkezhetnek. Ezen paraméterek előzetes beállítása segít a modelleknek abban, hogy hatékonyabban tanuljanak kisebb adathalmazokon és javítsák teljesítményüket. Az előtanítás során a modellek általában nem felügyelt módon tanulnak, például önellenőrző feladatokon keresztül, ahol a cél a bemeneti adatok hiányzó részeinek előrejelzése.

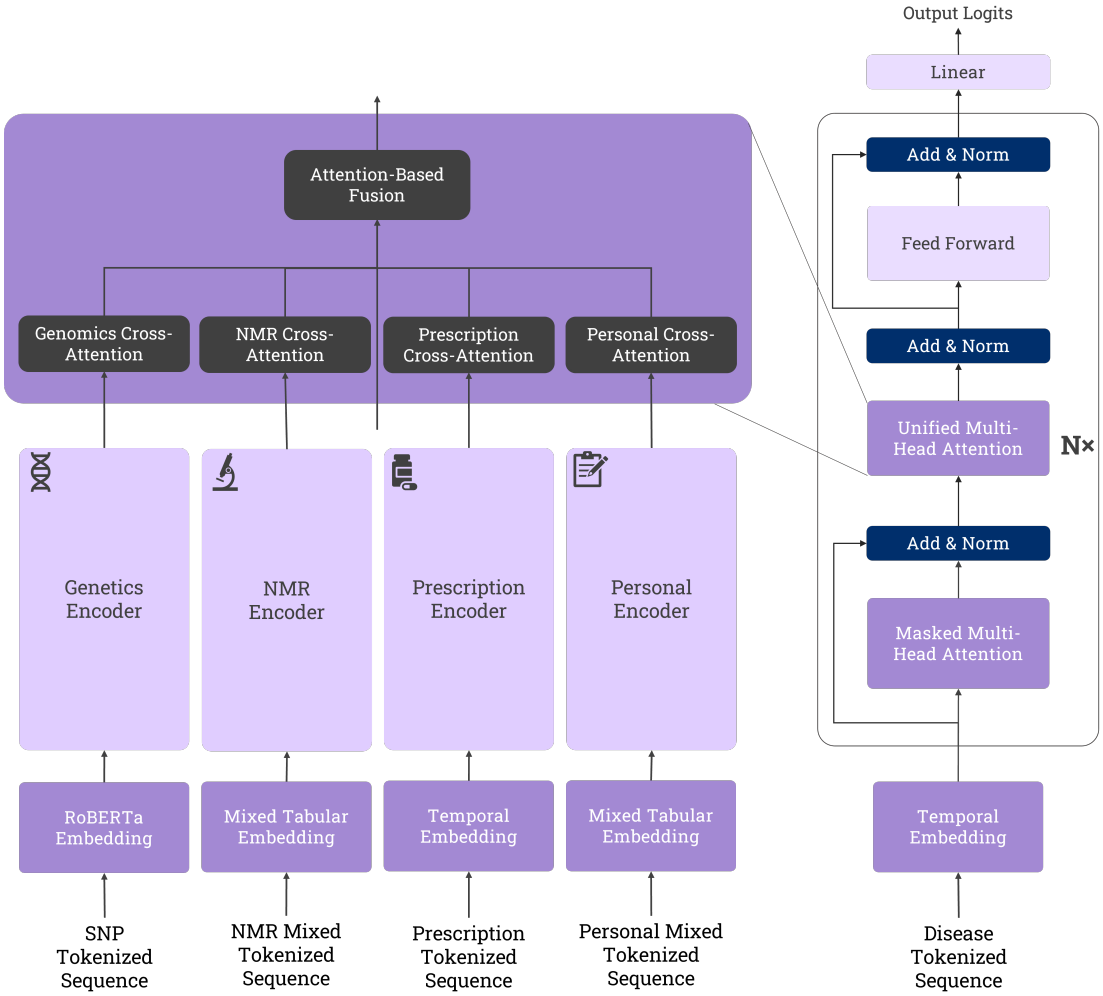
Az előtanítás előnyei, hogy a modellek, amelyek előzetesen nagy adathalmazokon lettek tanítva, gyakran jobb teljesítményt mutatnak kisebb adathalmazokon történő finomhangolás során. Az előtanítás által létrehozott általános jellemzők és tudás átültethető specifikus feladatokba, növelve ezzel az effektivitást és pontosságot, amelynek speciális területei a *transzfer tanulás* és *a kis mintaszámos (Few-shot)* tanulás.

Transzfer tanulás az előtanításban: A transzfer tanulás lényege, hogy a korábban tanult ismereteket átvigyünk és alkalmazzuk új feladatokon. Az előtanított modellek képesek általános tudást felhalmozni, amelyet később specifikusabb feladatokhoz használnak.

Kis mintaszámos tanulás: A korlátozott mennyiségű címkézett adatokkal rendelkező feladatokban az előtanított modellek különösen hasznosak. Ebben a kontextusban a modell az előzetes tudására támaszkodik és csak néhány példányon keresztül finomít, hogy megfeleljen az adott feladatnak.

3. A javasolt multimodális fúziós transzformer modell

Az általunk javasolt modell a heterogén orvosbiológiai tárgyterületek és adatmodalitások integrálására összpontosít, amelyet a továbbiakban egységesen modalitásként hivatkozunk. Kiemelten foglalkozunk egy új, általunk kifejlesztett beágyazással, a "Mixed Tabular Embedding" (vegyes táblázatos beágyazás), amely innovatív módon kombinálja a különböző adattípusokat.Az architektúra a genomikai, gyógyszerfelírási, személyes fiziológiai és életvitellel kapcsolatos, valamint a metabolomikai adatok összefüggéseinek mélyebb megértését célozza meg.



3.1. ábra. A javasolt multimodális architektúra sémája. A rendszer négy különálló enkódert használ: Genetics (genetikai leírók), NMR (NMR metabolomikai profilok), Prescription (gyógyszeres kezelések és Personal (alapvető fiziológiai egyéni leírók). Kezdeti adatbeágyazásokat RoBERTa, Vegyes Táblázatos és Időbeli technikákkal végezzük. Ezeket a kódolt modalitásokat egyedi kereszt-figyelmi mechanizmusokon keresztül feldolgozzák, ami lehetővé teszi a modalitások közötti figyelmet, mielőtt egy figyelem-alapú fúziós mechanizmussal aggregáljuk.

Megvizsgáljuk a különböző adattípusok közötti inter-modális kapcsolatok azonosításának és integrálásának specifikus módszereit, valamint a kereszt-figyelmi mechanizmusokat, amelyek lehetővé teszik az adatok hatékony kombinálását és megmagyarázhatóságát. Összefoglalva, a kutatás során kifejlesztettünk egy új figyelmen alapuló fúziós módszert, melynek segítségével sikerült az egyes adatsorozatokból származó információkat egységes reprezentációvá aggregálni. Ezen kívül a dolgozatban részletesen ismertetjük a tokenizációs folyamatokat, amelyeket a különböző adatmodalitások konzisztens reprezentációjának biztosítására alkalmaztunk.

3.1. Tokenizálási módszertan

A tokenizálási folyamatunkat több adatmodalitáson keresztül alkalmaztuk annak érdekében, hogy konzisztens reprezentációt biztosítsunk. A betegségeket és a gyógyszerfelírásokat specifikus kódolási rendszerek segítségével ábrázoltuk. A betegségeket a 3 karakter hosszú $ICD10$ kódokkal reprezentáltuk, amely $1,127$ különböző tokent eredményezett. A gyógyszerfelírás esetében 4 karakter hosszú ATC szint 3 kódokat használtunk, amely 176 tokent eredményezett. A genomikai adatok, amelyek egy nukleotidos polimorfizmusok (SNP-k), úgy lettek kódolva, hogy csak a minor és a major allélt jelenítsék meg a szekvenciában, elhagyva a többi jellemzőt. Ez a módszer SNP-énként 2 tokent eredményezett, összesen $1,400$ tokennel.

A metabolomikai és környezeti adatokat lebegőpontos tokeneket tartalmazó szekvenciákra tokenizáltuk, amelyek alkalmasak voltak a vegyes táblázatos beágyazási rétegek számára. A lebegőpontos tokenek nem tudják jelezni a mértékegységüket, ezért egy további tokent vezettünk be minden folytonos mértéket megelőzően annak típusának meghatározására. Összességében a modalitás kategorikus méréseinek tokenjeivel, a betegségi tokenekkel és a gyógyszerfelírás tokenekkel $3,155$ tokent tartalmaz a szótár.

Az időbeli komponens kezelésére az események dátumát évekre, hónapokra és napokra osztottuk, és ezeket külön tokenizáltuk. Míg a hónapokat és napokat lehetett csak sorszámuk szerint tokenezni, az évek tokenjei csak 1934-től 2020-ig terjedő tartományt fedtek le, ezért ott más évekhez tartalmazó tokent nem használtunk. Mindegyik kategóriában szerepel egy speciális "nincs dátum" token a különleges tokenek jelzésére.

3.2. Beágyazási technikák

A beágyazási technikák (embedding) képesek magas dimenziós adatokat sűrített, alacsony dimenziós térben ábrázolni. Ezek a beágyazások kiváló eszközök az információ tömörítésére és a komplex adatstruktúrák közötti összefüggések feltárására, lehetővé téve a modellek számára, hogy mélyebb "intuíciót" és "megértést" nyerjenek az adatokból. Ebben a fejezetben megvizsgáljuk, hogyan alkalmazhatók ezek a technikák az orvosi adatok kontextusában, és bemutatjuk, hogyan segíthetnek a betegség lefolyásának és kezelési stratégiáinak jobb megértésében.

3.2.1. Időbeli beágyazások

Az orvosi adatok terén a longitudinális tanulmányok egyedülálló kihívásokat és lehetőségeket jelentenek a beteg egészségi állapotának időbeli jellegzetességei miatt. A beteg egészségi mutatóinak folyamatos monitorozása releváns mintázatokat fedhet fel a betegség lefolyásában, a terápiás válaszban és a külső események által vezérelt hatásokban. Ebben a fejezetben bemutatjuk, hogyan építjük be az alapvető időbeli információkat modellünk beágyazási rétegeibe, lehetővé téve a komplex időbeli jellemzők észlelését.

Időbeli beágyazások. Egy architekturális komponens, amely az időbeli információt folytonos vektoros reprezentációkká alakítja át. Ebben a modulban egy hagyományos, szótáralapú beágyazási megközelítést alkalmazunk, amely az időbeli egységeket beágyazásokká transzformálja.

- **Év beágyazások.** Az orvosi adatokon belüli longitudinális mintázatok és trajektóriák reprezentálásáért felelnek. Tekintettel a több év alatt kialakuló vagy átalakuló állapotok és betegségek fontosságára, ezek a beágyazások nélkülözhetetlenek prognosztikai feladatokhoz, például betegség trajektóriáinak előrejelzéséhez vagy hosszú távú kezelési hatások megfejtéséhez.
- **Hónap és nap beágyazások.** Ezek lehetővé teszik a modell számára a rövid távú időbeli változások, mint például a ciklikus orvosi jelenségek vagy szezonális állapotok érzékelését. Biztosítják, hogy a modell érzékeny maradjon mind a nagy, mind a kis időbeli változásokra.

Az összetett időbeli reprezentációt az egyes időbeli egységek (év, hónap, nap) beágyazásainak összegzésével állítjuk elő. Ez az összegző mechanizmus garantálja, hogy minden egyes időbeli granularitás hozzájáruljon saját, egyedi időbeli jellegzetességével az összetett időbeli beágyazáshoz.

3.2.2. Vegyes Táblázatos Beágyazások

Ez a megközelítés beágyazások létrehozását célozza meg a token bemenetek, numerikus értékek, token kategóriák, pozíciós és látogatási adatok kombinálásával. Az adatok tokenizált része egy beágyazási réteget használ, míg a folytonos adatok egy lineáris réteget használnak a transzformációhoz. Maszkoló tenzort alkalmazunk a különböző típusok megkülönböztetésére.

1. **Szó beágyazások:**

   Tekintettel a token ID-k bemeneti tenzorára, I, és egy maszkoló tenzorra M a folytonos mérésekhez:

   $$T_e = E_w((I \odot \sim M)_{int})$$

   Ahol E_w a szó beágyazások mátrixát jelenti, és T_e az eredményül kapott token beágyazások. $\odot$ az elemenkénti szorzást jelenti, és $\sim M$ az M logikai negációját.

2. **Mérés beágyazások:**

   A bemeneti tenzor I folytonos mérései beágyazásokká alakulnak:

   $$M_e = E_m(I \odot M)$$

   Ahol E_m a lineáris réteg, amely a folytonos méréseket azonos dimenziójú beágyazásokká alakítja, és M_e az eredményül kapott mérési beágyazások.

3. **Token és mérés beágyazások kombinálása:**

   A token és mérési forrásokból származó végső beágyazások a következőképpen nyerhetők el:

$$I_e = (\sim M \odot T_e) + (M \odot M_e)$$

   Az egyes tokenek teljes beágyazása annak megfelelő token vagy mérési beágyazásának, pozíciós beágyazásának, token típus beágyazásának és látogatási típus beágyazásának összege, a token típus- TT_e és a látogatások beágyazása V_e az alapvető architektúrákhoz hasonlóan beágyazási rétegekkel érhetők el:

$$E = I_e + P_e + TT_e + V_e$$

   Ezt követően normalizálási és dropout műveleteket alkalmaz E-n a regulációhoz.

A Vegyes Táblázatos Beágyazások rugalmas keretet kínálnak a táblázatos adathalmazokban gyakran előforduló heterogén adattípusok beágyazására. Figyelembe véve a tokeneket, folytonos méréseket, pozíciós információt, token típusokat és látogatási típusokat, ez a modul széles körű információt rögzít beágyazásaiban.

3.3. A javasolt architektúra

A modell a transzformer architektúrának egy olyan továbbfejlesztett változata, amely minden egyes modalitáshoz különálló enkóder egységet rendel, az adott modalitás specifikus beágyazásának megfelelően. Az általunk kijelölt primer modalitás, melynek predikciós feladatait vizsgáljuk, Specifikált Kereszt-Figyelmi Modulok segítségével integrálja az összes többi modalitásból származó információt.

Paraméter	
Enkóder rétegek száma	2
Rejtett rétegek száma	4
Rejtett réteg mérete	312
Előrecsatolt réteg mérete	1200
Tanítható paraméterek száma	13,177,787
Paraméterek száma	69,007,835

3.1. táblázat. Az általunk készített modell konfigurációja a TinyBERT[Jiao et al., 2020] architektúráját követi.

Az általunk javasolt architektúra négy modalitás enkódert tartalmaz, amelyek gyógyszerfelírás, NMR, személyes és genetikai adatok feldolgozására szolgálnak. Az adatbeágyazások RoBERTa, vegyes táblázatos és időbeli technikákkal készülnek, lehetővé téve a bonyolult információk pontos ábrázolását. Ezeket a kódolt adatokat kereszt-figyelemi mechanizmusokkal dolgozzák fel az enkóderek, elősegítve az intermodális információcsere dinamikáját. A központi figyelem alapú fúzió koordinálja és egyesíti ezeket a modalitásokat egy koherens reprezentációvá, optimálisan támogatva a diagnosztizált betegségszekvenciák elemzését. Az adatokról részletes leírást a 4. fejezet tartalmaz. A modell konfigurációját a 3.1. táblázat mutatja.

3.4. Az Egyesített Kereszt-Figyelmi Dekóder

Az eltérő orvosi adatmodalitásoknak, esetünkben a beteg általános állapotleírói kontextusának, a gyógyszerfelírási profilnak, a genomi profiloknak és a laboratóriumi eredményeknek az együttes kezelése egy kifinomult integrációs mechanizmust igényel. Az általunk javasolt Egyesített Kereszt-Figyelmi Dekóder (Unified Cross-Attention Decoder, UCAD) olyan szerkezetet kínál, amely kifejezetten erre a problémára lett szabva, és hatékonyan, illetve pontosan integrál több adatmodalitást.

- **Specializált Kereszt-Figyelmi Egységek.** Az architektúra több specializált kereszt-figyelmi mechanizmust tartalmaz, mindegyiket egy adott adatmodalitás számára használják. Ezek az egységek az inter-modális korrelációk felismerésére és összehangolására szolgálnak.
- **Figyelmen Alapuló Fúzió.** Ez a módszer figyelmi súlyokat használ a modalitások különböző fontosságának felismeréséhez, dinamikusan integrálva azokat a kontextusbeli relevancia alapján.

Figyelmen Alapuló Fúzió. Egy adott bemeneti tenzor X esetén, amelynek alakja $[num\_sequences, batch\_size, seq\_len, hidden\_dim]$, és egy figyelmi súlyvektor w dimenzióval $[hidden\_dim, 1]$:

1. **Figyelmi Pontszámok.** A pontszámokat az X és a súlyvektor w szorzásával számolják:
$$s = X \cdot w$$
amely egy s tenzort eredményez átalakítva $[num\_sequences, batch\_size, seq\_len]$ formára.
2. **Figyelmi Valószínűségek.** A valószínűségek p a softmax művelet alkalmazásával nyerhetők ki a $num\_sequences$ dimenzió felett:
$$p = \text{softmax}(s)$$
3. **Súlyozott Összeg.** A végső reprezentáció O a bemeneti sorozatok súlyozott összegzése azok figyelmi valószínűségei alapján:
$$O = \sum_{i=1}^{num\_sequences} p_i \odot X_i$$
ahol $\odot$ az elemenkénti szorzást jelöli, összegzve a $num\_sequences$ dimenzió mentén.

Lényegében a Figyelmen Alapuló Fúzió (Attention-Based Fusion, ABF) módszer súlyokat rendel minden X-ben lévő sorozathoz annak tartalma és a figyelmi súlyvektor alapján. Ezek a súlyok lehetővé teszik a sorozatok egy egységes reprezentációba történő összegzését, megtartva a sorozatok összességéből származó kiemelkedő jelentőségű információkat.

3.5. A tanítás folyamata

Minden enkódert előtanítottunk a saját szűrt adathalmazukon, 10 epochon keresztül, 2 NVIDIA V100 GPU használatával. Egy epoch hozzávetőleg 1 óra GPU időt vett igénybe minden egyes enkóder számára. Ezután felhasználtuk az összes adatot és 10 epochon tanítottuk a fúziós modellt. A ráépülő (downstream) feladatokat a fúziós modellen további 10 epochon keresztül tanítottuk, ahol minden epoch hozzávetőlegesen 20 percet vett igénybe a megfelelő ráépülő feladathoz tartozó adathalmazon.

4. A felhasznált adatbázis

Az elemzés során a UK Biobank (UKB) adatait vettük alapul [Bycroft et al., 2018]. Az UKB adatai egyedülálló részletességű betekintést nyújtanak a résztvevők egészségügyi állapotába, amelyet különféle szempontok szerint lehet elemezni (lásd 4.1).

Adattípus	Változók száma	Minták száma
Életmódbeli	178	502 504
Betegségek (ICD-10)	1127	502 504
Genetikai (SNPs)	700	488 377
Laboratóriumi (NMR)	249	121 724
Gyógyszerfelírások (ATC)	176	158 154

4.1. táblázat. Az általunk használt UKB adatok tulajdonságai.

4.1. Életmódbeli tényezők

Az életmódbeli tényezők, amelyeket az UKB "személyes" kategóriában azonosított, fontosak az egyének egészségi állapotának és a betegségek kockázatának mélyreható megértéséhez. Az ilyen tényezők magukban foglalják az egyének alapvető leíró jellemzőit, mint például a táplálkozási szokásokat, a fizikai aktivitást, az alkohol- és dohányzásra vonatkozó információkat, a munkahelyi és környezeti expozíciót, valamint más, a mindennapi életben meghozott döntéseket. Az életmódbeli tényezők kritikus információt szolgáltatnak arra vonatkozóan, hogy milyen körülmények között élnek az emberek, és milyen kockázati tényezőkkel járnak együtt. Az életmódbeli adatok elemzése lehetővé teszi az egészséges és kockázatos életmódbeli mintázatok azonosítását, és segíti az intervenciók és megelőző stratégiák hatékony tervezését az egészség javítása és a betegségek kockázatának csökkentése érdekében.

4.2. Betegség kategóriák

A betegség kategóriák értékelése elengedhetetlen az egészségügyi tendenciák és kockázatok megértéséhez. Az 1127 három karakteres $ICD-10$ betegség kategóriát használtuk az UKB adataiból, amelyek részletes információt nyújtanak a betegségek eloszlásáról és gyakoriságáról. Az $ICD-10$ kódok standardizáltak és nemzetközileg elfogadottak, így lehetővé teszik az adatok összehasonlítását más nemzetközi adatbázisokkal is. Ezek az adatok segítenek azonosítani azokat a betegségeket és állapotokat, amelyek a legnagyobb terhet jelentik az egészségügyi rendszer számára, és lehetővé teszik az intervenciók és a megelőző stratégiák célzott kialakítását.

4.3. Genetikai adatok

A genetikai információk kulcsszerepet játszanak a betegségek kialakulásának megértésében és az egyének kockázatának azonosításában. A DisGeNET platform [Piñero et al., 2020] felhasználásával beazonosítottuk azokat a géneket a melyek közvetlen kapcsolatban állnak az MDD-vel, majd meghatároztuk az ezekhez tartozó 700 SNP-t (Single Nucleotide

Polymorphism). Az SNP-k egyetlen nukleotidot érintő gyakori genetikai változatok, amelyek befolyásolhatják az egyén betegségre való hajlamát, így elemzésük lehetőséget kínál a genetikai faktorok és az MDD közötti kapcsolat jobb megértésére. A genetikai adatok alapján kifejlesztett modellek előre jelezhetik a betegség kockázatát és elősegíthetik a megelőzés és a korai beavatkozás stratégiáinak kialakítását.

4.4. Laboratóriumi mérések

A laboratóriumi mérések, különösen a metabolomikai adatok, híd szerepet töltenek be az alapvető biológia és a klinikai gyakorlat között. Az UKB adatai között az NMR (Nuclear Magnetic Resonance) technikával mért metabolikus markereket használtuk, amelyek kritikus információval szolgálnak az egyén metabolikus állapotáról. A metabolomika átfogó képet nyújt az egyén biokémiai állapotáról, és segíthet a betegség korai stádiumának azonosításában, még a klinikai tünetek megjelenése előtt. Ezen adatok elemzése lehetőséget nyújt a metabolikus utak és a betegségek közötti összefüggések azonosítására, és elősegíti az új diagnosztikai és terápiás célkitűzések felfedezését.

4.5. Gyógyszerfelírások

A gyógyszeres kezelésekre vonatkozó információk fontosak az egészségügyi trendek és az orvosi ellátás minőségének értékeléséhez. Az UKB-ból származó adatok közel $165,000$ résztvevő elsődleges ellátási adatait tartalmazzák. Ezek az adatok a SystmOne gyakorlati kezelési rendszerből származnak, mely a Biobank egyik legmegbízhatóbb adatszolgáltatója. Az elérhető gyógyszerfelírási adatok az 1990 és 2016 közötti időszakot ölelik fel, amelyek jelentős információt nyújtanak a gyógyszerek használatának trendjeiről ebből az időszakból. Az adatok kódolása a British National Formulary (BNF) rendszer szerint történt, de elemzéseink során az Anatómiai Terápiás Kémiai Osztályozási Rendszerre (ATC) történő konverziót alkalmaztunk, amelyet az SE-BME együttműködésben dolgoztak ki, lehetővé téve számunkra a gyógyszerek szisztematikus csoportosítását és elemzését.

4.6. Hiányzó adatok kezelése

A transzformer architektúra adat-reprezentációjának köszönhetően képesek implicit módon kezelni az egyes modalitások részleges hiányzását. Amikor egy adott modalitásból az összes adat hiányzik, a hierarchikus-moduláris architektúra lehetővé teszi az adott enkóder kimenetének egyszerű mellőzését, biztosítva ezzel a rendszer rugalmasságát és az információ torzításának minimalizálását.

5. Eredmények

A vizsgálatunk elsődleges célja a javasolt architektúra tesztelése és valós adaton történő kiértékelése volt. Az alkalmazási terület az MDD potenciális altípusainak ("sztratifikációjának") feltárása volt, különös tekintettel a párhuzamosan jelen lévő krónikus betegségek (multimorbiditás) és a egyszerre szedett gyógyszerek (polifarmácia) számára. A multimorbiditási pontszám (MM) és a polifarmácia pontszám (PP) a betegség és a gyógyszer terheket méri. Ennek megfelelően ábrázoltuk a következő jellemzőket a transzformátor modellekből tanult reprezentációkban: (1) az elsőként előforduló betegség osztályok, (2) a leggyakoribb betegség osztályok, (3) a multimorbiditás és (4) a polifarmácia pontszámok.

A nagy nyelvi modellek területén a modellek előzetes általános, feladatfüggetlen, nem felügyelt előtanítása nagy mennyiségű általános adatokkal lehetővé tesz egy azt követő célzott alkalmazást. Az előtanított modellt, annak részletét, részbeni funkcionalitását felhasználva, amely egy új lehetőséget jelent az adatok teljes értékű hasznosításában. Ebben a szakaszban bemutatjuk azokat az eredményeket, melyeket egy Maszkolt Nyelvi Modell (MLM) módszerrel elő-tanított modellünkkel értünk el egy ráépülő (*downstream*) predikciós feladatban betegségeket előrejelezve. Ezen kívül szisztematikusan elemeztük a látens tér modalitások szerinti alstruktúráját, különös tekintettel azon információkra, amelyek a különböző betegségek gyógyszeres kezelésére vagy genetikai markerek közötti kapcsolatokra utalnak. A látens tér kvalitatív elemzését UMAP dimenziócsökkentési technikával végeztük el [McInnes et al., 2018]. Ez az elemzés lehetővé teszi a látens térben felfedezett összefüggések és mintázatok ábrázolását, így áttekintést és mélyebb megértést biztosít az általunk kifejlesztett modell működésével és hatékonyságával kapcsolatban.

5.1. Vizualizáció a moduláris látens reprezentációban

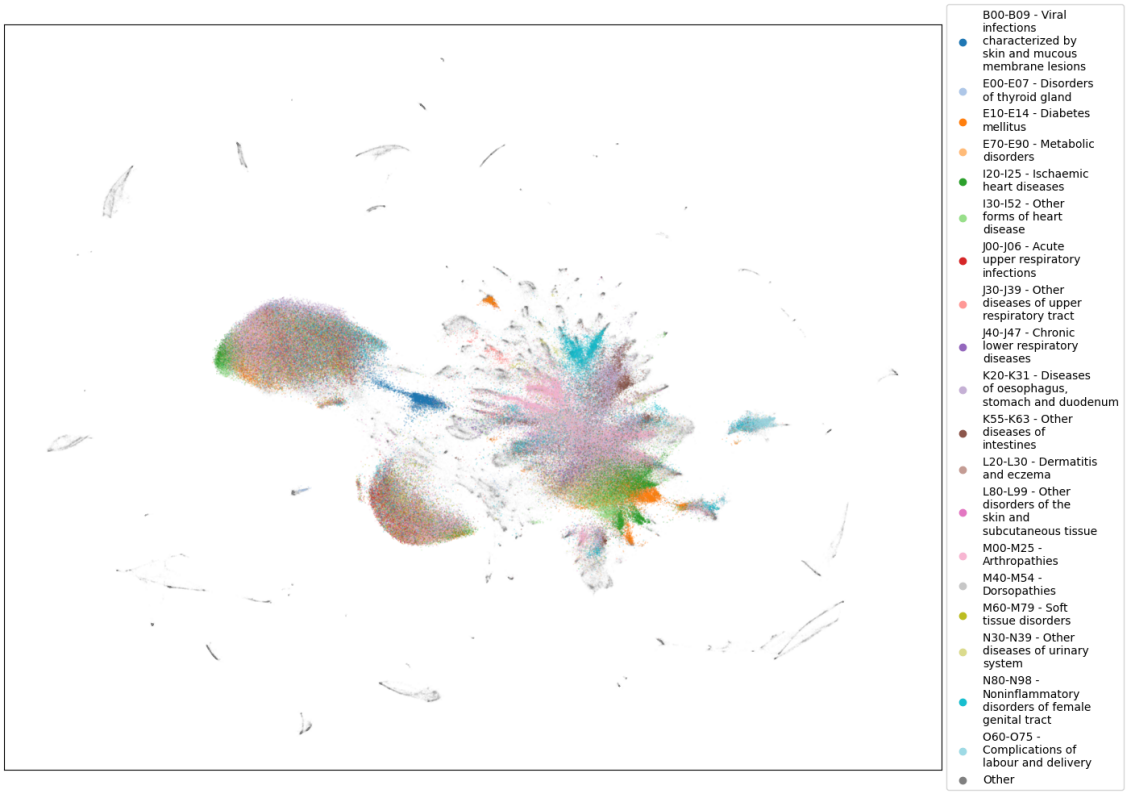
A modellünk előtanítás fázisa a Maszkolt Nyelvi Modellezés (MLM) költségfüggvényét használja. Ez az önfelügyelt tanítási stratégia magában foglalja az input tokenek egy részének elrejtését és a modell számára a kontextusuk alapján történő előrejelzésüket. Ezen folyamat során a modell olyan beágyazásokat tanul meg létrehozni látens terében, amelyek gazdagok kontextuális információban és szemantikai társításokban, így ideális alapot szolgáltatnak ráépülő feladatokban.

A modellünk többenkóderes architektúrája egyedülálló abban a tekintetben, hogy nem egyetlen látens térrel rendelkezik, hanem egymással összekapcsolt látens terek összességével, amelyben az egyes autonóm látens terek rendre az egyes modalitásokhoz és tárgyterületekhez tartoznak. Ez az összefüggő struktúra lehetővé teszi az információk szabad áramlását a reprezentációk között, erősítve azok szemantikai gazdagságát. A dolgozatbeli alkalmazásban ezek a látens alterek a genetikai leírók (**Genetics**), az NMR metabolomikai profilok (**NMR**), a gyógyszeres kezelések (**Prescription**) és az alapvető fiziológiai egyéni leírók (**Personal**) enkóderjeihez tartalmaznak; a kombinált teret a **Joint** megnevezés jelöli.

A longitudinális gyógyszerfelírási adatokat és az alapvető egyéni információkat figyelembe véve, amelyeket a többenkóderes architektúra kombinál, a modellünk képes a longitudinális multimorbiditás és polifarmácia előrejelzésére, így a klinikai döntéstámogatásban is felhasználható. A modellünkkel kapott beágyazások jól használhatók olyan klinikai eszközök létrehozásához, amelyek segítenek az orvosoknak a betegség progressziójának előrejelzésében, a kezelési stratégiák optimalizálásában és a betegséggel kapcsolatos következmények minimalizálásában.

A betegségek megjelenése a fuzionált látens reprezentációban

A krónikus betegségek megjelenése jelentősen befolyásolja az egyének életminőségét és hosszú távú egészségi kilátásait. Az UKB résztvevők közötti elemzés során az egyik legfontosabb információ, hogy mely krónikus betegségek vannak jelen egy adott személynél, hiszen ezek a betegségek határozzák meg az egészségben eltöltött életévek számát (krónikus megbetegedések nélkül). A következőkben bemutatjuk, hogy mely betegségek és betegségcsoportok jelennek meg leggyakrabban az elsőként kialakuló krónikus megbetegedések között az UKB résztvevői körében.



5.1. ábra. A látens tér ábrázolása a leggyakoribb betegségcsoport (ICD blokkok) szerint. A térben jól kivehetők a domináns "szigetek", melyek a leggyakoribb betegségcsoportokat képviselik, valamint a kisebb szigetek és elszigetelt pontok, amelyek a ritkábban előforduló betegségeket jelzik.

Betegségek eloszlása. A látens térben megfigyelhető, hogy néhány nagyobb "sziget" dominál, melyek körül a legtöbb adatpont csoportosul a népegészségügyben nagy jelentőségű gyakori betegségcsoportokkal. Emellett számos kisebb "sziget" és elszigetelt pont is megtalálható, amelyek ritkább betegségeket vagy kevésbé gyakori állapotjellemzőket képviselnek.

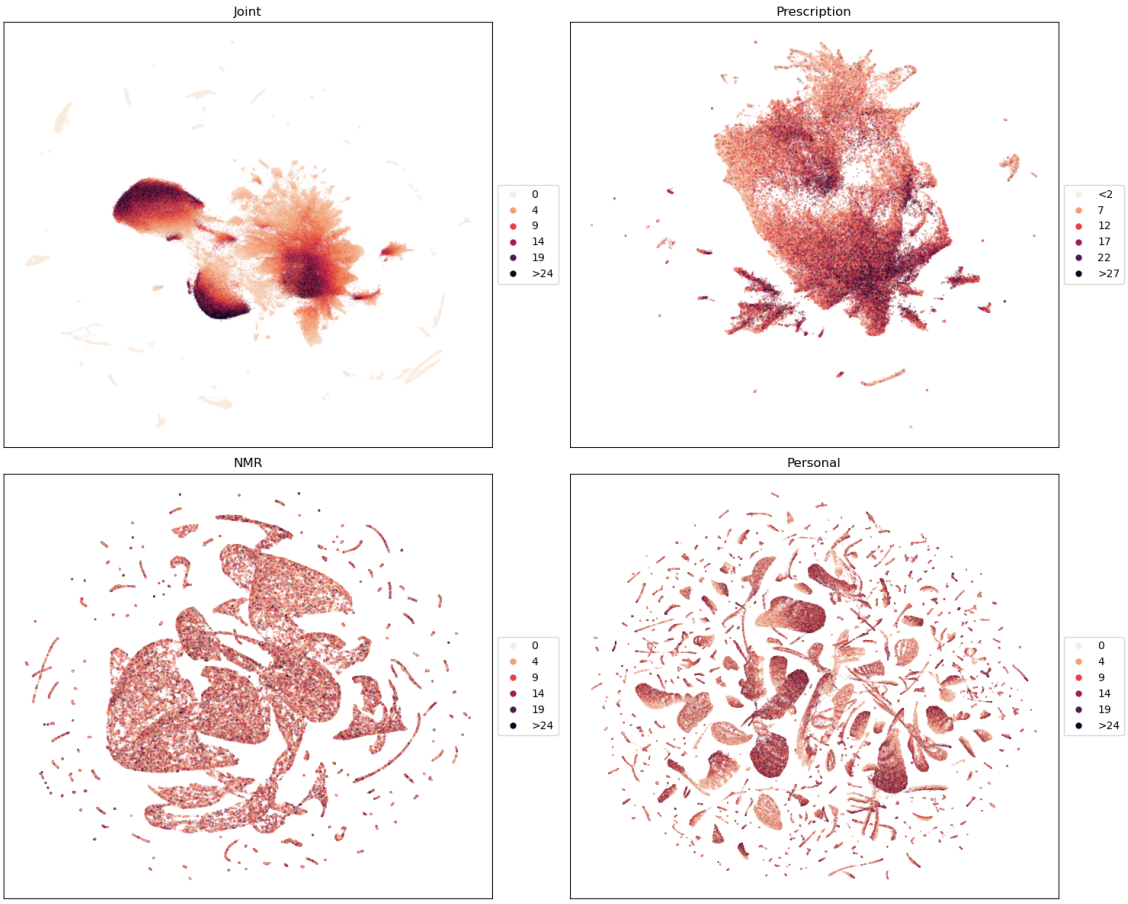
Átfedő régiók. Néhány betegség, mint például Diabetes melitus jól tükrözi az orvosi szakirodalomban jól ismert komorbiditási és multimorbiditási mintázatokat, nevezetesen a metabolikus megbetegedések, illetve a szív- és érrendszeri megbetegedések közeli szomszédságát. Hasonlóan a "Nyelőcső, gyomor és nyombél betegségei" és az "Egyéb bélbetegségek" átfedésben vannak a látens térben. Ez azt sugallja, hogy a betegek diagnosztizált állapotai között megosztott vagy hasonló alapvető jellemzők vannak, vagy jelezheti a társbetegségeket.

Kategóriák közötti elkülönülés. Bizonyos betegségcsoportok között egyértelmű az elkülönülés. Például a "Bőr- és nyálkahártya léziókkal jellemezhető vírusfertőzések" és a "Diabetes Mellitus" különböznek a többitől. Ez arra utalhat, hogy ezeknek az állapotoknak az alapvető jellemzői jelentősen különböznek egymástól.

Ritka területek. Vannak területek a látens térben, ahol kevés vagy nincs adatpont. Ezek a régiók egészséges alpopulációkat, ritkább betegségeket, vagy outlier mintákat jelenthetnek, amelyek nagyban különböznek az adathalmaz többi elemétől.

Összehasonlító elemzés az enkóderek látens reprezentációiról a diagnosztizált betegségek száma alapján

A multimorbiditás, azaz több krónikus betegség egyidejű jelenléte az egyénben, növekvő tendenciát mutat, különösen az idősebb korosztályok esetében. Az egyszerre több betegség megjelenése bonyolíthatja a terápiás stratégiákat, ami fokozza az nemkívánatos gyógyszerválaszok, terápiás ellentmondások és a terápiás adherencia csökkenésének kockázatát [Koné Pefoyo et al., 2015, Guthrie et al., 2015, Langenberg et al., 2023]. Ezen összefüggésben kiemelt jelentőséggel bír olyan számítási modellek kialakítása, amelyek adekvátan képesek leképezni a multimorbiditás komplex jellegét és feltárni közös molekuláris háttereit Rosenthal et al. [2023].



5.2. ábra. UMAP vizualizációk a négy látens reprezentációról ('Joint', 'Prescription', 'NMR' és 'Personal'). A szín intenzitása a betegek diagnosztizált betegségeinek számát jelzi.

Joint Enkóder. A 'Joint' vizualizáción egyértelműen három fő csoport látható. A nagyobb szigeteken erős gradiens jelenik meg, ami jól mutatja a multimorbiditás szerinti csoportusulást a látens térben. A kisebb de fő "szigetekhez" közeli csoportokon belül is erős gradiens jelenik meg a multimorbiditás szerint. Az elszigetelt "külső" régiókat pedig a betegséggel nem rendelkező egyének alkotják.

Prescription Enkóder. A 'Prescription' vizualizáció bár kevésbé ragadja meg a multimorbiditást, itt is jól kivehető a gradiens. A kisebb elkülönülő "szigetek" is jól megragadják a betegségek számát tehát a gyógyszerfelírások látens tere korrelál a betegségek számával.

NMR Enkóder. A 'NMR' ábrázolása látszólag a multimorbiditást független, tehát laboratóriumi eredmények alapján a modell nem fedett fel összefüggést a multimorbiditással, amely jelezheti, hogy az NMR adatok UKB-beli keresztmetszeti vizsgálatból származó jellege erősen behatárolja a longitudinális jellegű betegség és gyógyszerfelírási információkkal való összevetését.

Personal Enkóder. A 'Personal' ábrázolás jól tükrözi a multimorbiditást. Ez arra utalhat, hogy a kontextusos jellemzők információt tartalmaznak a gyógyszerek számáról és más szélesebb körű tényezőkkel, mint például környezeti és életvitelbeli jellemzőkkel is kapcsolatban lehetnek.

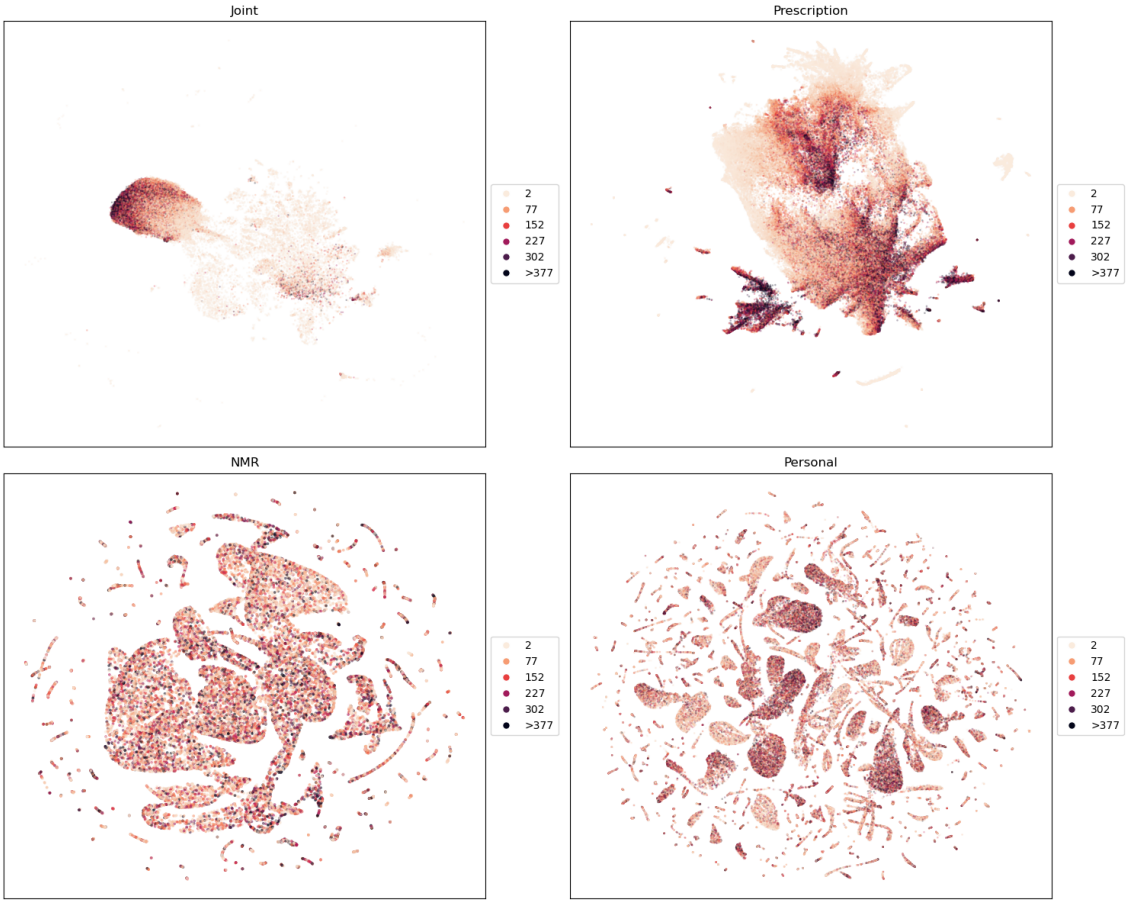
Összehasonlító elemzés az enkóderek látens reprezentációiról a felírt gyógyszerek száma alapján

A felírt gyógyszerek mennyisége és sokfélesége sokrétű képet nyújt a résztvevők egészségügyi profiljáról. A gyógyszerek nagyobb száma összetett társbetegségeket vagy multifaktoriális egészségi állapotot jelenthet, amely több gyógyszeres kezelést igényel. Ezzel szemben az alacsonyabb szám kevésbé komplex egészségi állapotot vagy krónikus betegség hatékony kezelését jelezheti minimális gyógyszeres kezeléssel. Ezeknek a reprezentációknak a megértése a gyógyszerfelírási mintákkal együtt döntő betekintést nyújt az egészségügyi állapotok, a kezelési stratégiák és az általános egészségügyi menedzsment közötti lehetséges kapcsolatokba. A gyógyszerfelírások gyakorisága és változatossága jelentős szerepet játszik az egyén egészségügyi trajektóriájának feltérképezésében [Koné Pefoyo et al., 2015, Guthrie et al., 2015, Schwarz et al., 2022].

Joint Enkóder. A 'Joint' vizualizáción egyértelműen két fő csoport látható. Egy erős gradienssel rendelkező sziget a bal oldalon, míg a tér többi részében ahol azok a résztvevők helyezkednek el akiknek nagyon kevés (2) gyógyszerfelírása van.

Prescription Enkóder. A 'Prescription' vizualizáció természetéből adódóan erős összefüggést mutat, ahol az erősebb színezésű területek (a magasabb gyógyszer számokat jelzők) jelentős mértékben elkülönülnek az alacsonyabb számú területektől. Itt a kisebb bal oldali szigeten láthatjuk a legtöbb gyógyszerrel rendelkező résztvevőket.

NMR Enkóder. A 'NMR' ábrázolása látszólag a gyógyszer számoktól független, tehát laboratóriumi eredmények alapján a modell nem fedett fel összefüggést a gyógyszerek számával.



5.3. ábra. A felírt gyógyszerek száma az integrált és modalitások szerinti látens reprezentációkban ('Joint', 'Prescription', 'NMR' és 'Personal' esetében, a Genetika nem szerepel). Az összes felírt gyógyszer számát a szín intenzitása jelzi.

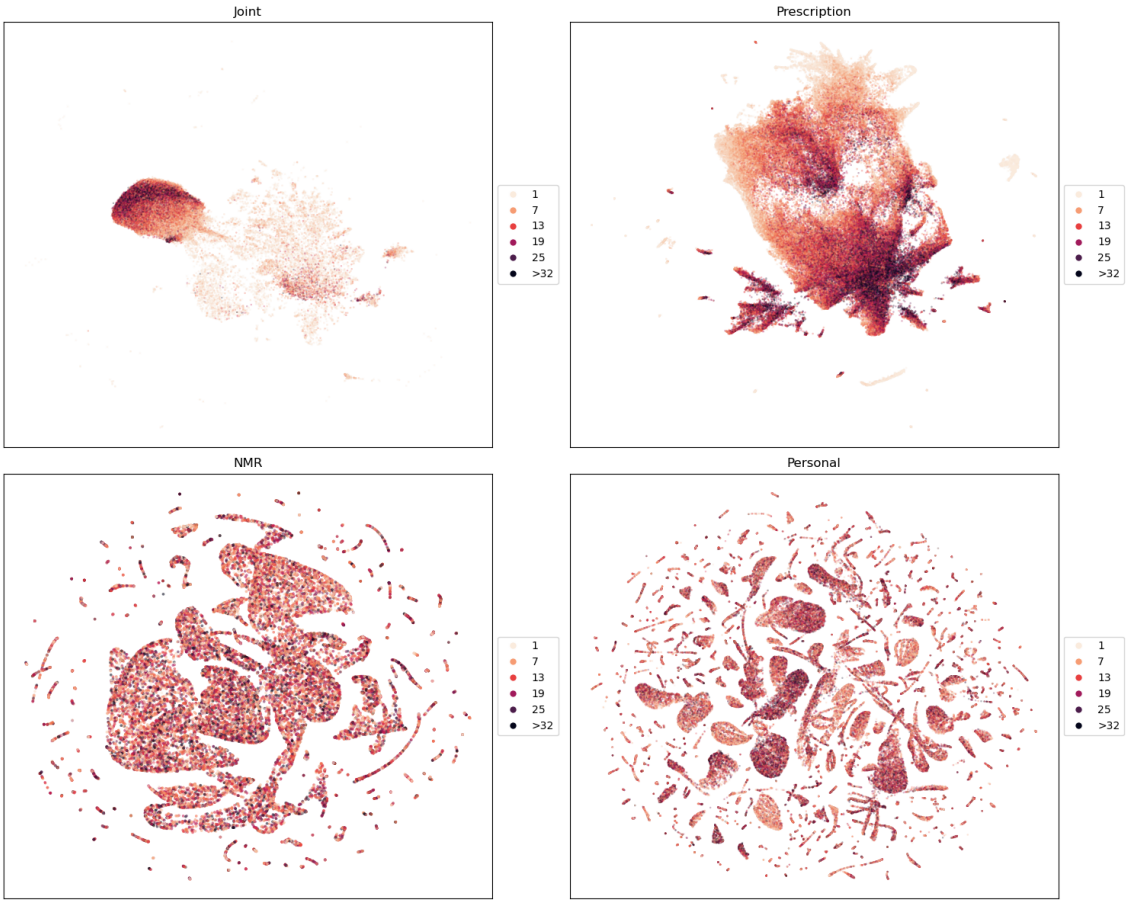
Personal Enkóder. A 'Personal' ábrázolás összefüggést mutat gyógyszer számokkal, bár az eloszlása szétszórtabb mint a 'Prescription' enkóderé, de néhány jelen jól látható csoportosulások vannak. Ez arra utalhat, hogy míg a kontextusos jellemzők kevesebb információt tartalmaznak a gyógyszerek számáról, de fő kapcsolatokat megragadhatnak.

Összehasonlító elemzés az enkóderek látens reprezentációiról a különböző felírt gyógyszerek alapján

A betegek által egyidejűleg alkalmazott több gyógyszer, vagyis a polifarmácia, különösen az idősek vagy a több betegségben szenvedők körében növekvő jelenség [Koné Pefoyo et al., 2015, Guthrie et al., 2015]. Bár a polifarmácia sok esetben terápiás előnyöket is hordoz, veszélyekkel is járhat, mint az adverse gyógyszerreakciók és gyógyszerinterakciók. Emiatt kulcsfontosságú olyan modellek kialakítása, amellyel feltárhatjuk a polifarmácia komplex jellegét [Schwarz et al., 2022]. Ez a megközelítés új nézőpontot nyújthat a különböző gyógyszerkombinációk jellegzetességeire. Az enkóderek polifarmáciával kapcsolatos értelmezése és ábrázolási módja új információt nyújthat a gyógyszerterápia menedzseléséről, esetleges terápiás ellentmondásokról és a gyógyszeres kezelés finomításáról.

Joint Enkóder. A 'Joint' vizualizáción egyértelműen két fő csoport látható. Egy erős gradienssel rendelkező sziget a bal oldalon, míg a tér többi részében ahol azok a résztvevők helyezkednek el akiknek nagyon kevés (2) gyógyszerfelírása van.

Prescription Enkóder. A 'Prescription' vizualizáció természetéből adódóan erős összefüggést mutat, ahol az erősebb színezésű területek (a magasabb egyedi gyógyszer számokat jelzők) jelentős mértékben elkülönülnek az alacsonyabb számú területektől. Itt a kisebb bal oldali szigeten láthatjuk a legtöbb egyedi gyógyszerrel rendelkező résztvevőt.



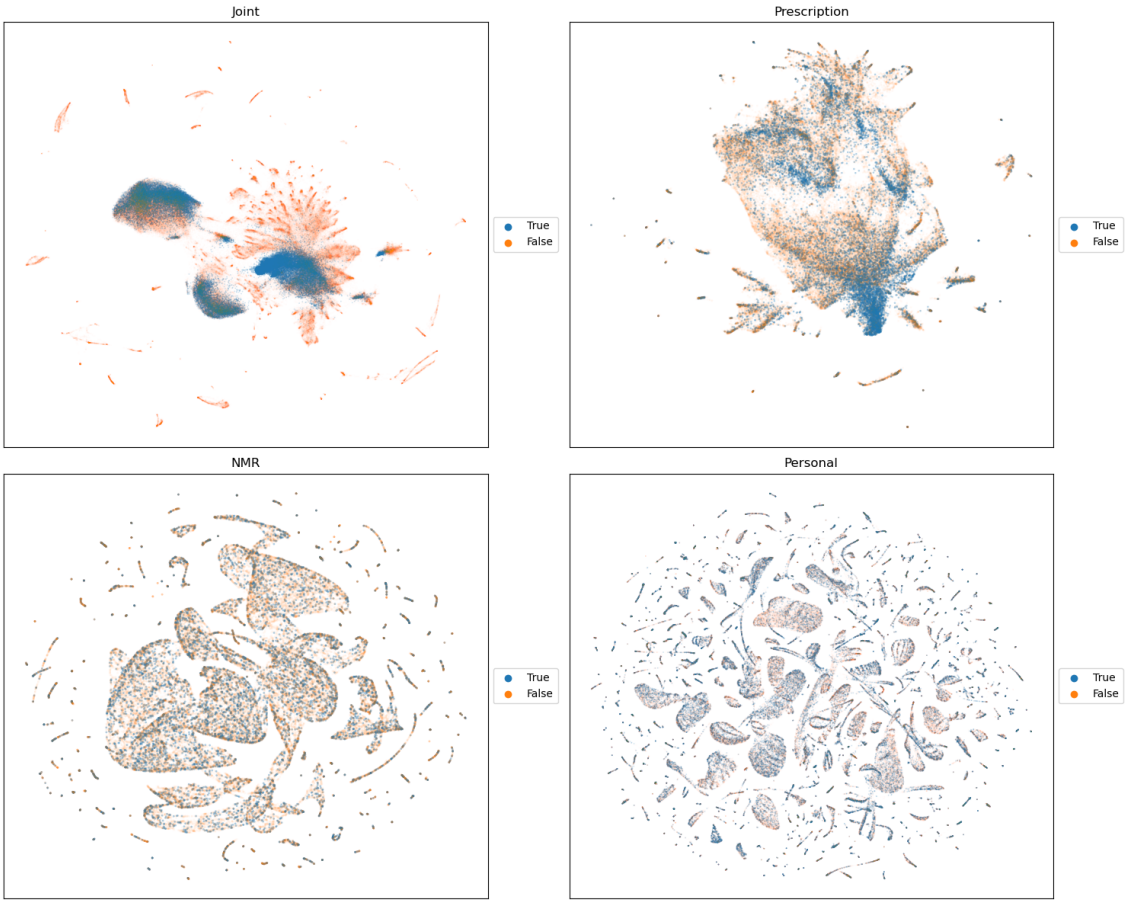
5.4. ábra. A különböző felírt gyógyszerek száma a 'Joint', 'Prescription', 'NMR' és 'Personal' látens reprezentációkban UMAP vizualizációval. A szín intenzitása a különböző felírt gyógyszerek számát tükrözi, tehát egy gyógyszer többszöri felírása nem befolyásolja.

NMR Enkóder. A 'NMR' ábrázolása látszólag az egyedi gyógyszer számoktól független, tehát laboratóriumi eredmények alapján a modell nem fedett fel összefüggést az egyedi gyógyszerek számával.

Personal Enkóder. A 'Personal' ábrázolás, összefüggést mutat gyógyszer számokkal, bár az eloszlása szétszórtabb mint a 'Prescription' enkóderé, de néhány jelen jól látható csoportosulások vannak. Ez arra utalhat, hogy míg a kontextusos jellemzők kevesebb információt tartalmaznak az egyedi gyógyszerek számáról, de szélesebb körű tényezőkkel kapcsolatban lehetnek.

Összehasonlító elemzés az enkóderek látens reprezentációiról a hangulatzavarokkal kapcsolatos betegségek jelenléte alapján

Az UKB résztvevőinek egészségügyi adatainak kiterjedt feltárása során a különböző enkóderekből származó látens reprezentációk megértése kiemelkedővé válik, különösen akkor, ha ezek a reprezentációk hangulatzavarokhoz, például a major depresszív zavarhoz (MDD) kapcsolódnak. Az MDD, amelyet tartós szomorúság, érdeklődésvesztés, valamint számos fiziológiai és kognitív zavar jellemez, az egyik legelterjedtebb hangulatzavar világszerte. Sokrétű jellege megköveteli, hogy minden analitikai eszköz, beleértve az enkóderek is, nagy pontossággal rögzítse bonyolultságát. Annak értékelése, hogy a különböző enkóderek hogyan különböztetik meg az MDD-vel kapcsolatos finom és gyakran összefonódó jellemzőket, kritikus betekintést nyújthat a rendellenesség mögöttes szerkezetébe és potenciális biomarkereibe.



5.5. ábra. Hangulatzavarral kapcsolatos betegségek a 'Joint', 'Prescription', 'NMR' és 'Personal'látens reprezentációkban UMAP vizualizációval. A kék színnel színezett betegeknél jelen van hangulatzavar, a naracssárgával színezett betegeknél nincs.

Joint Enkóder. A 'Joint' vizualizációján határozott elkülönülést látunk a hangulatzavar betegségcsoportra vonatkozóan, amely azonban kifejezett tagolódást mutat. A kék régiók hangulatzavarral rendelkező egyéneket képviselik (igaz), jól láthatóan több különálló klaszterben csoportosulnak.

Prescription Enkóder. A 'Prescription' vizualizáción jól elkülönül a hangulatzavar jelenléte a gyógyszerbevitel alapján. A kék régiók (igaz) bizonyos területeken csoportosulnak, ami arra utal, hogy a rendellenességgel rendelkező egyének gyógyszerfelírási profilja elkülönül a többi embertől.

NMR Enkóder. A 'NMR' alapú ábrázolás bizonyos kék (igaz) klasztereket mutat be, amelyek megfelelhetnek a hangulatzavarra jellemző specifikus laboratóriumi eredményeknek. Vannak azonban átfedő területek, amelyek arra utalnak, hogy egyes laboratóriumi eredmények nem meggyőzőek vagy mindkét csoportban közösek lehetnek.

Personal Enkóder. A 'Personal' alapú csoportosításban kék (igaz) és narancssárga (hamis) régiók keveréke található. A nehány jelentősen elkülönülő csoport azt sugallja, hogy a kontextuális tényezők, mint például az életmód vagy a környezet, szerepet játszhatnak hangulatzavar jelenlétében.

Összehasonlító elemzés az enkóderek látens reprezentációiról a cukorbetegség megléte alapján

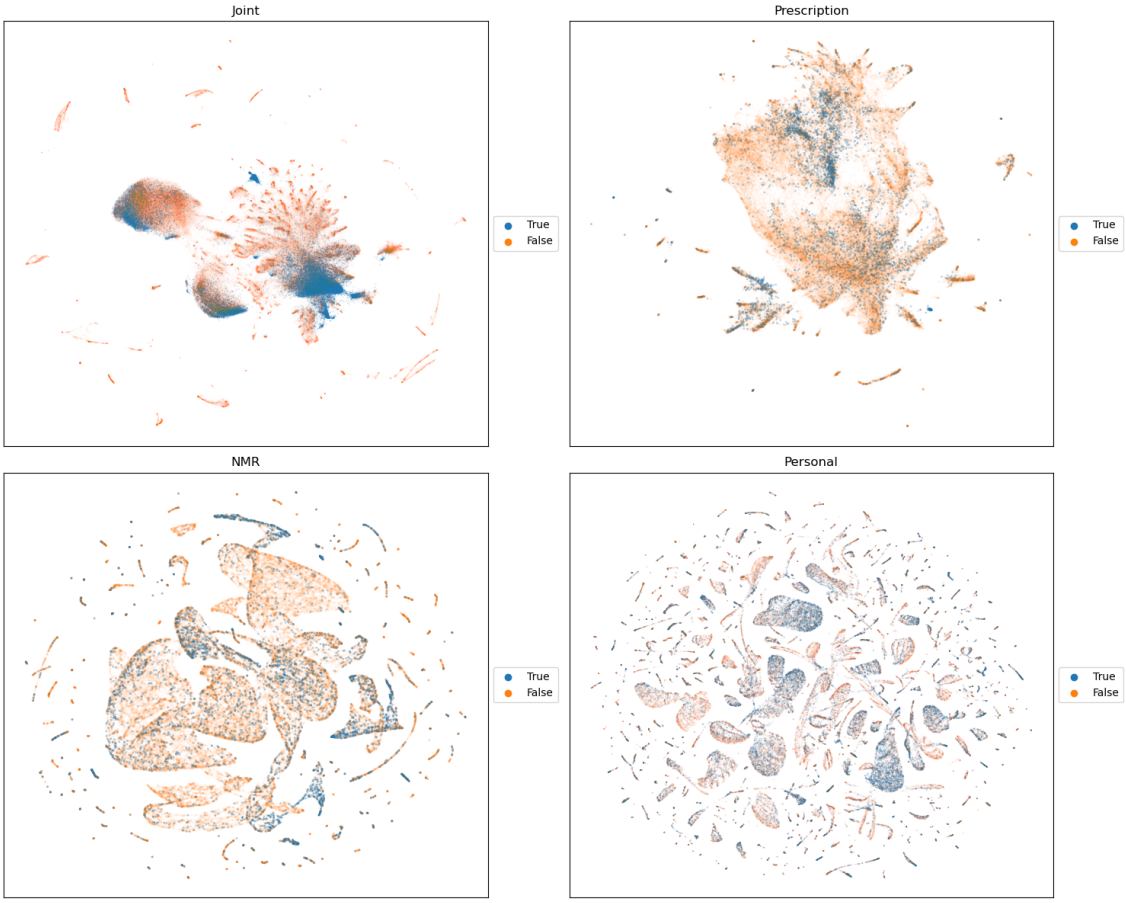
Az UKB résztvevőinek átfogó elemzésében a különböző enkóderek által előállított látens reprezentációk kifejezett jelentőséggel bírnak, amikor konkrét egészségügyi állapotokkal, például cukorbetegséggel összefüggésében értékeljük ki őket. A cukorbetegség, mélyreható következményekkel jár a különböző fiziológiai rendszerekre, és jelentősen befolyásolja a beteg általános egészségügyi profilját. A metabolikus szindróma az MDD esetében is egy fontos komorbiditás, amely az obezitással és cukorbetegséggel, mind az MDD, de a szív- és érrendszeri megbetegedésekkel is egy nagy jelentőségű multimorbiditási csoportot alkot [Marx et al., 2017]. Az a mód, ahogyan a enkóderek rögzítik a cukorbetegség jelenlétével vagy hiányával kapcsolatos árnyalatokat, segíthet megérteni a diabetes mellitus és más egészségügyi változók közötti kapcsolatokat. A cukorbetegséggel kapcsolatos látens minták felismerése megnyitja az utat a lehetséges prediktív modellek előtt.

Joint Enkóder. A 'Joint' vizualizációján határozott elkülönülést látunk a diabetes mellitusra vonatkozó "igaz" és "hamis" címkék között. A kék régiók, amelyek a diabetes mellitusban szenvedő egyéneket képviselik (igaz), jól láthatóan csoportosulnak.

Prescription Enkóder. A 'Prescription' vizualizáción jól elkülönül a diabetes mellitus állapot megoszlása a gyógyszerbevitel alapján. A kék régiók (igaz) bizonyos területeken csoportosulnak, ami arra utal, hogy bizonyos gyógyszereket gyakran írnak fel cukorbetegeknek.

NMR Enkóder. A 'NMR' alapú ábrázolás bizonyos kék (igaz) klasztereket mutat be, amelyek megfelelhetnek a cukorbetegekre jellemző specifikus laboratóriumi eredményeknek. Vannak azonban átfedő területek, amelyek arra utalnak, hogy egyes laboratóriumi eredmények nem meggyőzőek vagy mindkét csoportban közösek lehetnek.

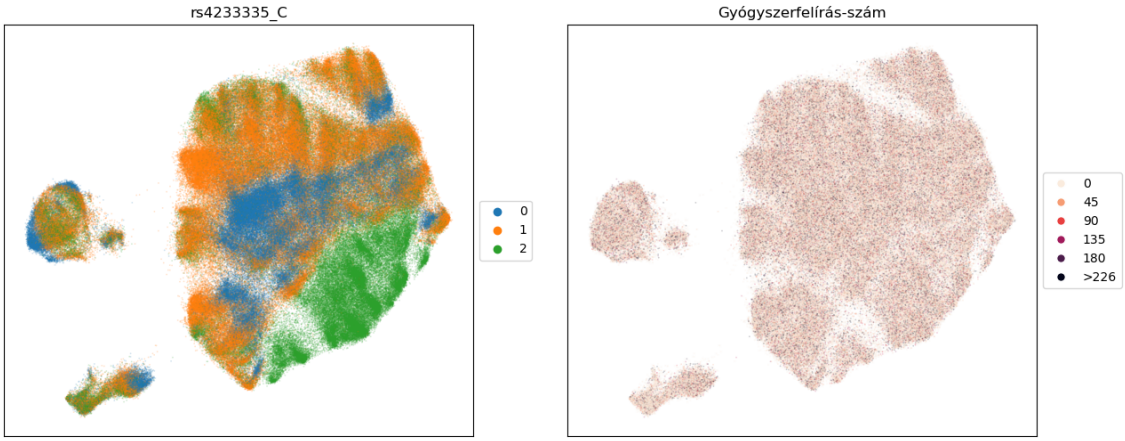
Personal Enkóder. A 'Personal' alapú csoportosításban kék (igaz) és narancssárga (hamis) régiók keveréke található. A nehány jelentősen elkülönülő csoport azt sugallja, hogy a kontextuális tényezők, például az életmód vagy a környezet, szerepet játszhatnak a betegségben, de nem feltétlenül tesznek egyértelmű különbséget a két csoport között.



5.6. ábra. A cukorbetegség a 'Joint', 'Prescription', 'NMR' és 'Personal' látens reprezentációkban UMAP vizualizációval. A színek cukorbetegség meglétét jelzik.

A genetikai enkóder látens tere

A genetikai enkóder látens tere magas fokú rendezettséget mutat a tanító adatban szereplő SNP-k szerint. Mivel az egyes SNP-k összefüggése a beteg későbbi állapotával gyenge összefüggést mutat, az általunk használt 700 SNP-s látens tér sem mutat a többi adatunk szerint struktúrát.



5.7. ábra. A Genetikai enkóder látens tere az 'rs4233335_C' SNP variánsok (bal), valamint a gyógyszerfelírások száma szerint (jobb).

5.2. Predikció a látens reprezentáció felhasználásával

A tanult reprezentáció hasznosságát predikciós szempontból is megvizsáltuk, amely a reprezentációknak a szakértői értelmezéstől eltérő másik felhasználását jelentik. Ennek során egy finomhangolt *(fine-tuned)* modellt hasonlítottunk össze hagyományos gépi tanulási módszerekkel. A finomhangoláshoz, a tanult reprezentációkon tanítunk egy Multi Layer Perceptron-t (MLP), és az előtanított modell súlyait rögzítettük.

Adatok előkészítése. A kiértékelésre félretett adatokat *1:9* arányban osztottuk fel finomhangoláshoz és a finomhangolt modellek kiértékeléséhez. A teljesítményt összehasonlítottuk a nyers adaton futó hagyományos gépi tanulási módszerekkel, hogy megvizsgáljuk a megtanult reprezentáció alkalmazhatóságát.

Minden ráépülő feladat a beteg jövőbeli állapotának előrejelzésére irányul az eddigi adatai alapján. Minden 2009. december 31. előtti adatot, bemeneti adatként használtuk, és az ezután történ eseményeket használtuk a célváltozók meghatározásához. A szigorú elválasztás biztosítja, hogy semmilyen előzetes információ nem befolyásolja a prediktált eseményeket. A teszteléshez két típusú tulajdonság előrejelzését céloztuk meg:

- **Regressziós feladatok:** specifikus események előfordulásának száma. Különösen fókuszálva a polifarmáciára, valamint a multimorbiditásra.
- **Bináris osztályozás:** egy betegségcsoportból származó betegség előfordulása, egy bizonyos időintervallumon belül.

A regressziós feladatok, a 2010-ben bekövetkezett eseményeket használták.

- **Betegségek száma:** Ez lehetővé teszi a betegek állapotának és egészségügyi profiljának jövőbeli monitorozását.
- **Gyógyszerfelírások száma:** Ennek segítségével felmérhetjük az egyéni gyógyszerterhelést és a gyógyszerellátás komplexitását.
- **Különböző típusú gyógyszerfelírások száma:** Ezzel a beteg gyógyszeres kezelésének sokféleségét és változatosságát tudjuk értékelni.

A bináris osztályozás során a leggyakoribb betegségcsoportokra összpontosítottunk. Az alacsony előfordulás miatt, itt 5 éves periódust (2010-2015) vizsgáltunk:

- **Hangulatzavar** *(F30-F39 ICD kódok)*: Ilyenek például a depresszió, a bipoláris zavar.
- **Diabetes Mellitus** *(E10-E14 ICD kódok)*: A cukorbetegség fajtái, melyek közül a leggyakoribbak az 1-es és 2-es típusú cukorbetegség.
- **Hipertóniával kapcsolatos betegségek** *(I10-I15 ICD kódok)*: Például a természetesen kialakuló magas vérnyomás, a más betegségekből adódó magas vérnyomás és a veseproblémák által okozott magas vérnyomás.

Ráépülő feladatok eredményei. A regresszió és a bináris osztályozás esetében a hagyományos gépi tanulási algoritmusok hiperparamétereit rácskereséssel optimalizáltuk. A finomhangoláshoz használt hiperparamétereket nem optimalizáltuk, iparági szabványos paramétereket használtunk. A bináris osztályozás esetén, mivel az osztályok jelentősen kiegyensúlyozatlanok voltak, az AUPRC *(area under precision-recall curve)* metrikára optimalizáltuk a modelleket, amely jól tükrözi a modell stabilitását. A regressziós feladatok esetében az MSE-re optimalizáltuk.

Regression Datasets			Classification Datasets		
Metric	**Mean**	**Std**	**Dataset**	**Pos Samples**	**Pos Percentage**
Disease Burden	0.429	1.05	F30-F39	787	1.62
Unique Drug Burden	0.668	1.79	E10-E14	1237	2.55
Drug Burden	2.44	7.55	I10-I15	4244	8.76

5.1. táblázat. A ráépülő adatok statisztikái: Mindegyik adathalmaz 48452 mintát tartalmaz.

A **regressziós** feladatokhoz a tanult reprezentáció effektíven kódolja ezt az információt, mivel modellünk jelentősen felülmúlja a többi modell teljesítményét. Fontos megjegyezni, hogy a moduláris megközelítés nélkül a teljesítmény csak enyhén javul. Ez jól látható a 5.2 alapján, ahol a standard RoBERTA modell, csak kis (3% MAPE) javulást ért el.

Model	**MSE**	**RMSE**	**MAE**	**MAPE**
MLP	1.08	1.04	0.624	89.4
rf_reg	1.08	1.04	0.649	95.8
XGB	1.08	1.04	0.646	95.6
RoBERTa Encoder	1.02	0.918	0.603	92.2
Saját Modell	**1.01**	**0.91**	**0.561**	**80.7**

5.2. táblázat. Betegség-szám predikció: A különböző modellek metrikái a betegség-szám predikció során. A MAPE metrika alapján a modellek nagy átlagos eltéréssel prediktálnak (80-90%), de a legjobb teljesítményt a finomhangolt modell éri el.

Model	**MSE**	**RMSE**	**MAE**	**MAPE**
MLP	52.4	7.24	3.98	105
rf_reg	52.1	7.22	3.85	99
XGB	52	7.21	3.87	100
RoBERTa	46.8	6.33	3.25	94.4
Saját Modell	**23.8**	**4.28**	**1.8**	**51.8**

5.3. táblázat. Gyógyszerfelírás-szám predikció: A különböző modellek metrikái a Gyógyszerfelírás-szám predikció során. Az új módszer kimagaslóan alacsony hibával rendelkezik az eddigi algoritmusokhoz képest.

Model	**MSE**	**RMSE**	**MAE**	**MAPE**
MLP	3.22	1.8	1.04	96.7
rf_reg	3.2	1.79	1.05	98.5
XGB	3.2	1.79	1.04	98.5
RoBERTa	2.21	1.42	0.811	81.9
Saját Modell	**1.15**	**0.989**	**0.467**	**44.6**

5.4. táblázat. Egyedi gyógyszerfelírás-szám predikció: Az új módszer kimagaslóan alacsony hibával rendelkezik az eddigi algoritmusokhoz képest.

A **bináris osztályozási** eredmények értékelése nehézkes volt, mivel a metrikák minden modellnél alacsonyak voltak a probléma bonyolult természete miatt. Figyelemre méltó, hogy az AUPRC-t használva meghatározó metrikaként, a javasolt modell mind a három feladatban felülmúlta az eddigi módszereket, és a moduláris architektúra a háromból kettő feladaton javította a teljesítményt.

Model	Accuracy	Precision	Recall	F1 Score	AUROC	AUPRC
DT	0.963	0.019	0.025	0.022	0.653	0.027
LR	0.910	0.027	0.127	0.044	0.658	0.028
MLP	**0.968**	**0.085**	0.101	**0.092**	0.666	0.044
XGB	0.047	0.016	**0.962**	0.032	0.655	0.038
RoBERTa	0.798	0.033	0.172	0.054	0.711	**0.272**
Saját Modell	0.875	0.038	0.136	0.057	**0.742**	0.269

5.5. táblázat. $F30$-$F39$ **Hangulatzavar:** AUPRC alapján a transzformer alapú modellek stabilabbak a kiegyensúlyozatlan adathalmazon, mint a eddigi módszerek. A két modell között AUROC metrikában van jelentős eltérés, ahol az új modell jobban teljesít.

Model	Accuracy	Precision	Recall	F1 Score	AUROC	AUPRC
DT	0.738	0.071	**0.766**	0.130	0.798	0.127
LR	0.924	0.140	0.379	0.204	0.801	0.147
MLP	0.502	0.042	0.839	0.079	0.782	0.114
XGB	**0.948**	**0.183**	0.298	**0.227**	**0.859**	0.162
RoBERTA	0.882	0.090	0.220	0.114	0.729	0.313
Saját Modell	0.916	0.142	0.291	0.176	0.807	**0.472**

5.6. táblázat. $E10$-$E14$ **Diabetes Mellitus:** AUPRC alapján a transzformer alapú modellek stabilabbak a kiegyensúlyozatlan adathalmazon, mint a eddigi módszerek. Az új modell kiemelkedően jó AUPRC értékkel rendelkezik.

Model	Accuracy	Precision	Recall	F1 Score	AUROC	AUPRC
DT	0.567	0.131	0.701	0.221	0.667	0.140
LR	0.666	0.137	0.529	0.217	0.657	0.138
MLP	0.408	0.116	0.871	0.205	0.681	0.145
XGB	0.261	0.103	**0.960**	0.186	0.687	0.152
RoBERTa	0.684	0.145	0.521	0.214	0.652	0.296
Saját Modell	**0.747**	**0.183**	0.546	**0.260**	**0.720**	**0.369**

5.7. táblázat. $I10$-$I15$ **Hipertóniával kapcsolatos betegségek:** Az új modell, a 'Recall'-tól eltekintve minden metrikában kiemelkedő. 'Recall' esetében az XG-Boost a pontosságát áldozza fel a kimagasló metrikáért.

5.3. Hiperparaméter optimalizálás

Megvizsgáltuk a tanult reprezentációkat egy prediktív beállításban is, nem longitudinális és longitudinális ('gyógyszeres') adatokat használva 2010 előtt a képzésben és a betegség osztályok és multimorbiditás/polifarmácia pontszámok előrejelzésében az azt követő 1, 3 és 5 évben. A kis prevalencia és így a kiegyensúlyozatlan osztályozási feladat miatt az általunk preferált értékelési mérték a PRC alatti terület (AUPRC) volt. A regressziós feladatban a standard MSE metrikát használtuk.

Rendszeresen megvizsgáltunk adatszűréseket az adat homogenitásának növelése érdekében, például etnikai hovatartozás és születési év alapján, de az adatveszteség meghaladta a potenciális előnyét (ezek eredményét nem közöljük).

Az előtanítás fázis előtt az UKB adathalmazt különbontottuk, a pre-traininghez (90%) és a finomításhoz (10%) elkülönítve az adathalmazokat. A modell későbbi értékelése korábban nem látott adatokon történt, biztosítva az általános teljesítményének a becslését.

A finomításhoz kijelölt adathalmazból egy részt a finomítási fázisban az utó/rátanításhoz, egy kisebb részletet pedig a modell értékeléséhez különítettünk el. A felosztás során megőriztük különböző attribútumok reprezentatív eloszlását, így biztosítva a modell finomított teljesítményének robusztus értékelését.

6. Konklúzió

Az elektronikus egészségügyi nyilvántartások (EHR) kezelése komplex kihívásokkal jár, különösen amikor több modalitású adatokról van szó. Kutatásunk során bemutattunk egy új modellarchitektúrát, amely specifikusan a multimodális EHR adatok kezelésére lett tervezve.

Megvizsgáltuk a modell által tanult reprezentációkat, és megmutattuk, hogy a résztvevő egyének jellemzői alapján jól elkülönülő struktúrákat mutatnak. Ezek a reprezentációk nem csupán a betegek aktuális állapotát tükrözik, hanem mélyebb betekintést nyújtanak az egyes orvosbiológiai profilok közötti összefüggésekbe is.

Ahogy a 6.1-es és a 6.2-es ábra mutatja, több különböző feladatban is sikerült jobb teljesítményt elérnünk ezen reprezentációk felhasználásával, ami megfelel az előtanított és finomhangolt transzformer modellekkel kapcsolatos eddigi tapasztalatoknak, és alátámasztja a modell hatékonyságát és alkalmazhatóságát az egészségügyi területen. A transzformer képes longitudinális információkat kihasználni, amit a kutatási élvonalba tartozó UK Biobank adatainak felhasználásával mutattunk be. Fontos megjegyezni, hogy ezek az eredmények nemcsak az új architektúra erényeit mutatják be, hanem az alapvető fontosságúnak ítélt multimodális adatok felhasználására is egy új megközelítést mutat be az egészségügyi kutatások terén a kapcsolt látens terek révén.

	Disease Burden		Drug Burden		Unique Drug Burden	
Model	**MSE**	**MAPE**	**MSE**	**MAPE**	**MSE**	**MAPE**
MLP	1.08	89.4	52.4	105	3.22	96.7
RF	1.08	95.8	52.1	99	3.2	98.5
XGB	1.08	95.6	52	100	3.2	98.5
RoBERTa	1.02	92.2	46.8	94.4	2.21	81.9
Saját Modell	**1.01**	**80.7**	**23.8**	**51.8**	**1.15**	**44.6**

6.1. táblázat. Aggregált regressziós teljesítmények. Az új modell jelentős javulást ért el.

	F30-F39		**E10-E14**		**I10-I15**	
Model	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
DT	0.653	0.027	0.798	0.127	0.667	0.140
LR	0.658	0.028	0.801	0.147	0.657	0.138
MLP	0.666	0.044	0.782	0.114	0.681	0.145
XGB	0.655	0.038	**0.859**	0.162	0.687	0.152
RoBERTa	0.711	**0.272**	0.729	0.313	0.652	0.296
Saját Modell	**0.742**	0.269	0.807	**0.472**	**0.720**	**0.369**

6.2. táblázat. Aggregált bináris klasszifikációs teljesítmények. A transzformerek jelentős javulást értek el a kiegyensúlyozatlan adathalmazokon. A 3-ból 2 feladaton az új modell jelentős javulást ért el.

Összességében a bemutatott architektúra előrelépés a multimodális EHR adatok hatékonyabb kezelése felé, és további kutatási lehetőségeket kínál a biomedikai információk optimalizált reprezentációjának és elemzésének terén.

6.1. Potenciális alkalmazási területek

A transzformerek és az általunk bemutatott moduláris architektúra széles körű alkalmazásra nyit lehetőséget a orvosbiológiai kutatásokon túl is. A klinikai döntéstámogató rendszerek, ahol a betegek egészségügyi előzményei és a hosszútávú kimenetek előrejelzése kritikus, különösen előnyös lehet. Ezen túlmenően a gyógyszerkutatásban és a célzott terápiák kifejlesztésében is nagy potenciállal bír, ahol a genetikai adatok és más omikai információk egységesítése és elemzése létfontosságú a hatékony és személyre szabott kezelések kifejlesztéséhez. A transzformerek alkalmazása a különböző orvosbiológiai adatforrások közötti információk egységesítésével és kinyerésével új utakat nyithat az egészségügyi kutatásban és gyakorlatban egyaránt.

6.2. Továbbfejlesztési lehetőségek

Az újabb figyelmi mechanizmusok lehetővé tehetik, hogy még jobban skálázhatóvá váljunk a genetikai adatokra, ami a biomedikai területen rendkívül releváns. A konvolúciós technikák alkalmazása, mint például az 1D konvolúciós hálózatok, hosszabb inputok, például teljes genom szekvenciák kezelését is lehetővé tehetik. A maszkolt nyelvi modellek (MLM) variánsainak további vizsgálata szintén ígéretes lehet a modell teljesítményének optimalizálásában és a releváns információk jobb extrahálásában az adatokból.

Köszönetnyilvánítás

Elsősorban szeretnénk köszönetet mondani Gézsi Andrásnak, amiért felhasználhattuk a DisGeNET platformon az MDD-vel közvetlen kapcsolatban álló beazonosított géneket, valamint a UKB genetikai adatainak meghatározásához, ami az ezekhez tartozó 700 SNP-t (Single Nucleotide Polymorphism) tartalmazza. Köszönetet szeretnénk mondani mind Gézsi Andrásnak, mind Nagy Tamásnak az adatok a British National Formulary (BNF) rendszer kódolásából az Anatómiai Terápiás Kémiai (ATC) Osztályozási Rendszerre történő konverziójáért, amelyet az SE-BME együttműködésben dolgoztak ki.
A kutatásban felhasználásra került a UK Biobank Resource a 1602. számú pályázat keretében. Linked health data Copyright © 2019, NHS England. Újrafelhasználás a UK Biobank engedélyével. Minden jog fenntartva.
A kutatás az Európai Unió támogatásával valósult meg, az RRF-2.3.1-21-2022-00004 azonosítójú, Mesterséges Intelligencia Nemzeti Laboratórium projekt keretében, illetve a TKP2021-EGA-02 számú projekt a Kulturális és Innovációs Minisztérium Nemzeti Kutatási Fejlesztési és Innovációs Alapból nyújtott támogatásával, a TKP2021-EGA pályázati program finanszírozásában. Továbbá, a kutatást a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal (K 143391); a Magyar Agykutatási Program 3.0 (NAP2022-I-4/2022); valamint a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból a Nemzeti Innovációs és Technológiai Minisztérium a TKP2021-EGA támogatási rendszer keretében (TKP2021-EGA-25) támogatta.

Irodalomjegyzék

Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784, 2022.

Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Cory Y McLean, and Nicholas A Furlotte. Multimodal llms for health grounded in individual-specific data. *arXiv preprint arXiv:2307.09018*, 2023.

Thore Buergel, Jakob Steinfeldt, Greg Ruyoga, Maik Pietzner, Daniele Bizzarri, Dina Vojinovic, Julius Upmeier zu Belzen, Lukas Loock, Paul Kittner, Lara Christmann, et al. Metabolomic profiles predict individual multidisease outcomes. *Nature Medicine*, 28(11):2309–2320, 2022.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

Peiying Chen. Representation learning for electronic health records: A survey. In *Journal of Physics: Conference Series*, volume 1487, page 012015. IOP Publishing, 2020.

Philip Darke, Sophie Cassidy, Michael Catt, Roy Taylor, Paolo Missier, and Jaume Bacardit. Curating a longitudinal research resource using linked primary care ehr data—a uk biobank case study. *Journal of the American Medical Informatics Association*, 29 (3):546–552, 2022.

Bruce Guthrie, Boikanyo Makubate, Virginia Hernandez-Santiago, and Tobias Dreischulte. The rising tide of polypharmacy and drug-drug interactions: population database analysis 1995–2010. *BMC medicine*, 13(1):1–10, 2015.

Anders Boeck Jensen, Pope L Moseley, Tudor I Oprea, Sabrina Gade Ellesøe, Robert Eriksson, Henriette Schmock, Peter Bjødstrup Jensen, Lars Juhl Jensen, and Søren Brunak. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature communications*, 5(1):4022, 2014.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding, 2020.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Tuomo Kiiskinen, Pyry Helkkula, Kristi Krebs, Juha Karjalainen, Elmo Saarentaus, Nina Mars, Arto Lehisto, Wei Zhou, Mattia Cordioli, Sakari Jukarainen, et al. Genetic predictors of lifelong medication-use patterns in cardiometabolic diseases. *Nature Medicine*, 29(1):209–218, 2023.

Anna J Koné Pefoyo, Susan E Bronskill, Andrea Gruneir, Andrew Calzavara, Kednapa Thavorn, Yelena Petrosyan, Colleen J Maxwell, YuQing Bai, and Walter P Wodchis. The increasing burden and complexity of multimorbidity. *BMC public health*, 15(1): 1–11, 2015.

Claudia Langenberg, Aroon D Hingorani, and Christopher JM Whitty. Biological and functional multimorbidity—from mechanisms to management. *Nature Medicine*, 29(7): 1649–1657, 2023.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):7155, 2020.

Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. Hi-behrt: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE journal of biomedical and health informatics*, 27(2):1106–1117, 2022.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities. *Journal of Biomedical Informatics*, 145:104466, 2023.

Peter Marx, Peter Antal, Bence Bolgar, Gyorgy Bagdy, Bill Deakin, and Gabriella Juhasz. Comorbidities in the diseasome are more apparent than real: what bayesian filtering reveals about the comorbidities of depression. *PLoS computational biology*, 13 (6):e1005487, 2017.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Yiwen Meng, William Speier, Michael K Ong, and Corey W Arnold. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3121–3129, 2021.

Subhash Nerella, Sabyasachi Bandyopadhyay, Jiaqing Zhang, Miguel Contreras, Scott Siegel, Aysegul Bumin, Brandon Silva, Jessica Sena, Benjamin Shickel, Azra Bihorac, et al. Transformers in healthcare: A survey. *arXiv preprint arXiv:2307.00067*, 2023.

Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, Francesco Ronzano, Emilio Centeno, Ferran Sanz, and Laura I Furlong. The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic acids research*, 48(D1):D845–D855, 2020.

Davide Placido, Bo Yuan, Jessica X Hjaltelin, Chunlei Zheng, Amalie D Haue, Piotr J Chmura, Chen Yuan, Jihye Kim, Renato Umeton, Gregory Antell, et al. A deep learning algorithm to predict risk of pancreatic cancer from disease trajectories. *Nature Medicine*, pages 1–10, 2023.

Bodhayan Prasad, Anthony J Bjourson, and Priyank Shukla. Data-driven patient stratification of uk biobank cohort suggests five endotypes of multimorbidity. *Briefings in Bioinformatics*, 23(6):bbac410, 2022.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

Sara Brin Rosenthal, Sarah N Wright, Sophie Liu, Christopher Churas, Daisy Chilin-Fuentes, Chi-Hua Chen, Kathleen M Fisch, Dexter Pratt, Jason F Kreisberg, and Trey Ideker. Mapping the common gene networks that underlie related diseases. *Nature protocols*, pages 1–15, 2023.

Maurice Rupp, Oriane Peter, and Thirupathi Pattipaka. Exbehrt: Extended transformer for electronic health records. In *ICLR 2023 Workshop on Trustworthy Machine Learning for Healthcare*, 2023.

Kyriakos Schwarz, Daniel Trejo Banos, Giulia Rathmes, and Michael Krauthammer. Drug administration clusters in the uk biobank: An assessment of drug-drug interactions and patient outcomes in a large patient cohort. *arXiv preprint arXiv:2207.08665*, 2022.

Yijun Shao, Yan Cheng, Stuart J Nelson, Peter Kokkinos, Edward Y Zamrini, Ali Ahmed, and Qing Zeng-Treitler. Hybrid value-aware transformer architecture for joint learning from longitudinal and non-longitudinal clinical data. *medRxiv*, pages 2023–03, 2023.

Yuqi Si, Jingcheng Du, Zhao Li, Xiaoqian Jiang, Timothy Miller, Fei Wang, W Jim Zheng, and Kirk Roberts. Deep representation learning of patient data from electronic health records (ehr): A systematic review. *Journal of biomedical informatics*, 115:103671, 2021.

Sandra Steyaert, Marija Pizurica, Divya Nagaraj, Priya Khandelwal, Tina Hernandez-Boussard, Andrew J Gentles, and Olivier Gevaert. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nature Machine Intelligence*, 5(4):351–362, 2023.

Oleg Stroganov, Alena Fedarovich, Emily Wong, Yulia Skovpen, Elena Pakhomova, Ivan Grishagin, Dzmitry Fedarovich, Tania Khasanova, David Merberg, Sándor Szalma, et al. Mapping of uk biobank clinical codes: Challenges and possible solutions. *Plos one*, 17 (12):e0275816, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

Yu-An Wang and Yun-Nung Chen. What do position embeddings learn? an empirical study of pre-trained language model positional encoding. *CoRR*, abs/2010.04903, 2020. URL `https://arxiv.org/abs/2010.04903`.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*, 2022.

Yeda Wu, Enda M Byrne, Zhili Zheng, Kathryn E Kemper, Loic Yengo, Andrew J Mallett, Jian Yang, Peter M Visscher, and Naomi R Wray. Genome-wide association study of medication-use and associated disease in the uk biobank. *Nature communications*, 10 (1):1891, 2019.

Feng Xie, Han Yuan, Yilin Ning, Marcus Eng Hock Ong, Mengling Feng, Wynne Hsu, Bibhas Chakraborty, and Nan Liu. Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. *Journal of biomedical informatics*, 126:103980, 2022.

Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Qing Yin, Linda Zhong, Yunya Song, Liang Bai, Zhihua Wang, Chen Li, Yida Xu, and Xian Yang. A decision support system in precision medicine: contrastive multimodal learning for patient stratification. *Annals of Operations Research*, pages 1–29, 2023.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.

Hong-Yu Zhou, Yizhou Yu, Chengdi Wang, Shu Zhang, Yuanxu Gao, Jia Pan, Jun Shao, Guangming Lu, Kang Zhang, and Weimin Li. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nature Biomedical Engineering*, pages 1–13, 2023.

Bochao Zou, Xiaolong Zhang, Le Xiao, Ran Bai, Xin Li, Hui Liang, Huimin Ma, and Gang Wang. Sequence modeling of passive sensing data for treatment response prediction in major depressive disorder. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:1786–1795, 2023.